LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Silke Janitza & Gerhard Tutz

Prediction Models for Time Discrete Competing Risks

Technical Report Number 177, 2015
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Prediction Models for Time Discrete Competing Risks

Silke Janitza & Gerhard Tutz

Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München

January 2, 2015

Abstract

The classical approach to the modeling of discrete time competing risks consists of fitting multinomial logit models where parameters are estimated using maximum likelihood theory. Since the effects of covariates are specific to the target events, the resulting models contain a large number of parameters, even if there are only few predictor variables. Due to the large number of parameters classical maximum likelihood estimates tend to deteriorate or do even not exist. Regularization techniques might be used to overcome these problems. This article explores the use of two different regularization techniques, namely penalized likelihood estimation methods and random forests, for modeling time discrete competing risks using both, extensive simulation studies and studies on real data. The simulation results as well as the application on three real world data sets show that the novel approaches perform very well and distinctly outperform the classical (unpenalized) maximum likelihood approach.

Keywords: Competing Risks, Event History Models, Discrete Survival, Prediction, Penalized Likelihood, Random Forests, Survival Forests.

1 Introduction

In survival analysis often the prediction of a specific event is of interest. The predicted event may, for example, be death, diagnosis of a certain disease, the recurrence of cancer, birth of a child, time of failure for an electric device or getting a job after a time of unemployment. In many applications, however, there is more than one possible event type that can occur. In a clinical or epidemiological study the cause of death or the occurrence of different diseases may be of interest. When predicting the time of birth one might want to differentiate between live births and stillbirths, and in labor market studies it is often of interest whether employment is permanent or temporary. The modeling of event times in the presence of multiple events is usually referred to as competing risks modeling. The literature on competing risks mostly deals with the case where time is measured as a continuous variable (see, Beyersmann et al.; 2012; Putter et al.; 2007; Kalbfleisch and Prentice; 2011; Kleinbaum and Klein; 2005). In some cases, however, time is measured on a discrete scale, for example, in weeks, months or years. As an example, assume that a yearly screening is performed in order to detect a disease at an early stage. If a disease is detected at a screening visit, it is most likely not possible to specify the exact onset of the disease; instead a time interval enclosing disease onset is specified, which may be the time period between two consecutive screening visits. In other cases the time to the occurrence of a specific

event is intrinsically discrete, meaning that the discrete time variable is not a coarsened version of an underlying continuous time variable. This is the case when considering time to graduation or dropping out of university with time measured in semesters. When time is discrete, classical survival and competing risks models for continuous time become inappropriate (Andersen et al.; 1993), and one has to apply special approaches for discrete time. The standard approach to the modeling of time discrete competing risks consists of fitting a multinomial logit model which links an individual's covariates to the risk for observing a specific event. However, the use of the multinomial model for the prediction of time discrete competing risks is restricted to applications with few predictor variables since maximum likelihood estimates deteriorate quickly resulting in inaccurate predictions, or estimates may even not exist. Recently, penalized versions of the multinomial logit model have been used in the context of time discrete competing risks. They can be applied in settings with a large number of predictors possibly including interaction terms (Möst et al.; 2014). Penalization-based approaches like that of Möst et al. (2014) might be promising for the development of accurate prediction models. However, parametric specifications may be too restrictive in the presence of complex data settings with non-linear effects and interactions of higher order. Alternative procedures, such as random forests (Breiman; 2001), have been shown to yield high prediction performance in various applications. Recently, a tree-based approach was proposed by Schmid et al. (2013) for modeling discrete survival times. This approach makes use of the fact that the likelihood of a time discrete survival model is equivalent to that of a regression model for binary outcome data. An extension of this approach to the competing risks case with multiple event types is also considered as a second promising alternative to classical maximum likelihood models, which are currently in use.

The paper is structured as follows: Section 2 gives an overview of the classical and the novel methods for modeling time discrete competing risks. In the first part of this section we outline the standard approach which makes use of the fact that likelihood estimation for time discrete competing risks can be embedded into the framework of classical multivariate generalized linear models (GLMs). Subsequently, we introduce two alternative promising modeling strategies which have the advantage that they can be applied even if the number of parameters to be estimated exceeds the number of observations. The first approach is based on penalized maximum likelihood estimation, while the second approach makes use of random forest methodology. In Section 3 we show the results of extensive simulation studies, in which prediction ability of the two modeling approaches is assessed and compared to the classical approach. In Section 4 we compare the methods by using three real world datasets. A summary and discussion of our results are given in Section 5.

2 Competing Risks Models for Discrete Time

Let time be divided into intervals $[a_0, a_1), \ldots, [a_{k-1}, a_k), [a_k, \infty)$ and let $t \in \{1, \ldots, q\}$ with $q = k+1$ denote the failure within interval $[a_{t-1}, a_t)$. In a competing risks analysis one usually models the so-called *cause-specific hazard functions*. The *discrete cause-specific hazard function* for event type $r \in \{1, \ldots, m\}$ of an observation with covariates $\boldsymbol{x}_i$ is defined as

$$\lambda_r(t|\boldsymbol{x}_i) = P(T_i = t, R_i = r|T_i \geq t, \boldsymbol{x}_i).$$

The discrete cause-specific hazards describe the probability for failure at t from a specific event type, provided that the observation is still under risk prior to t. The *overall hazard function* describes the probability of failing from any of the m event types at time point t, again, provided that the observation is still under risk prior to t. It is computed from the sum of cause-specific hazards as

$$\lambda(t|\boldsymbol{x}_i) = P(T_i = t|T_i \geq t, \boldsymbol{x}_i) = \sum_{r=1}^{m} \lambda_r(t|\boldsymbol{x}_i). \tag{1}$$

The *discrete survival function* is the probability of surviving the first t time intervals:

$$S(t|\boldsymbol{x}_i) = P(T_i > t|\boldsymbol{x}_i) = \prod_{s=1}^{t} (1 - \lambda(s|\boldsymbol{x}_i)). \tag{2}$$

For the unconditional probability for failure from event type r at time interval t, in presence of all other event types, one obtains

$$P(T_i = t, R_i = r|\boldsymbol{x}_i) = \lambda_r(t|\boldsymbol{x}_i) \prod_{s=1}^{t-1} (1 - \lambda(s|\boldsymbol{x}_i)) \tag{3}$$

$$= \lambda_r(t|\boldsymbol{x}_i) S(t-1|\boldsymbol{x}_i). \tag{4}$$

Estimation

Let C_i be a random variable for the time interval at which observation i is censored. In practical applications only $\min(T_i, C_i)$, the minimum of failure and censoring time, is observed. Often it is helpful to define an indicator variable δ_i which indicates whether an event was observed for i (then $\delta_i = 1$) or not ($\delta_i = 0$). Here we assume that censoring occurs at the end of the time interval. Then δ_i is defined as

$$\delta_i = \begin{cases} 1, & T_i \leq C_i \\ 0, & T_i > C_i. \end{cases}$$

Under the assumption of random censoring, that is T_i and C_i are assumed to be conditionally independent, the likelihood contribution of observation i is given by

$$L_i = P(T_i = t_i, R_i = r_i|\boldsymbol{x}_i)^{\delta_i} P(T_i > t_i|\boldsymbol{x}_i)^{1-\delta_i} P(C_i \geq t_i|\boldsymbol{x}_i)^{\delta_i} P(C_i = t_i|\boldsymbol{x}_i)^{1-\delta_i}. \tag{5}$$

In the case of non-informative censoring (non-informative in the sense of Kalbfleisch and Prentice; 2011), the latter two factors in (5) that describe the censoring process, can be ignored, and the likelihood reduces to

$$L_i = \lambda_{r_i}(t_i|\boldsymbol{x}_i)^{\delta_i} (1 - \lambda(t_i|\boldsymbol{x}_i))^{1-\delta_i} \prod_{s=1}^{t_i-1} (1 - \lambda(s|\boldsymbol{x}_i)). \tag{6}$$

In order to show that this likelihood corresponds to the likelihood of a multinomial response model, we introduce indicator variables for the transition to the next time interval. These are defined as

$$y_{itr} = \begin{cases} 1, & \text{if event type } r \text{ occurs at time interval } t, \text{ given } t \text{ is reached} \\ 0, & \text{otherwise,} \end{cases}$$

and

$$y_{it0} = \begin{cases} 1, & \text{if no event occurs at time interval } t, \text{ given } t \text{ is reached} \\ 0, & \text{otherwise.} \end{cases}$$

From this definition it directly follows that $y_{it0} = 1 - y_{it1} - \ldots - y_{itm}$. Let $\boldsymbol{y}_{it}^\top = (y_{it0}, y_{it1}, \ldots, y_{itm}), t = 1, \ldots, t_i$ be the response vector of observation i. Using the indicator variables, the likelihood contribution (6) for observation i can be rewritten as

$$L_i = \prod_{s=1}^{t_i} \{ \prod_{k=1}^{m} \lambda_k(s|\boldsymbol{x}_i)^{y_{isk}} \} \{1 - \lambda(s|\boldsymbol{x}_i)\}^{y_{is0}}, \tag{7}$$

which corresponds to the likelihood of a multinomial response model with observations $\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{it_i}$. Accordingly, the response $\boldsymbol{y}_{it}$ is multinomially distributed with $\boldsymbol{y}_{it} \sim \mathcal{M}(1, \lambda_0(t|\boldsymbol{x}_i), \ldots, \lambda_m(t|\boldsymbol{x}_i))$, where $\lambda_0(t|\boldsymbol{x}_i) = 1 - \sum_{k=1}^{m} \lambda_k(t|\boldsymbol{x}_i))$ denotes the probability of survival of the t-th interval. The corresponding log-likelihood contribution for observation i is thus given by

$$l_i = \sum_{s=1}^{t_i} \left\{ \sum_{k=1}^{m} \{ y_{isk} \log \lambda_k(s|\boldsymbol{x}_i) \} + y_{is0} \log \left(1 - \sum_{k=1}^{m} \lambda_k(s|\boldsymbol{x}_i)\right) \right\},$$

and the total log-likelihood is obtained by the sum of likelihood contributions for all observations $i = 1, \ldots, n$:

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \sum_{s=1}^{t_i} \left\{ \sum_{k=1}^{m} \{ y_{isk} \log \lambda_k(s|\boldsymbol{x}_i) \} + y_{is0} \log \left(1 - \sum_{k=1}^{m} \lambda_k(s|\boldsymbol{x}_i)\right) \right\}.$$

Usually the cause-specific hazards are modeled via the multinomial model given by

$$\lambda_r(t|\boldsymbol{x}_i) = \frac{\exp(\eta_{itr})}{1 + \sum_{k=1}^{m} \exp(\eta_{itk})}, \tag{8}$$

with $\eta_{itr} = \gamma_{0tr} + \boldsymbol{x}_i^\top \boldsymbol{\gamma}_r$. Maximum likelihood estimates can then be obtained by using statistical software for multinomial models with an appropriate design matrix. The design matrix is composed of $\sum_{i=1}^{n} t_i$ observations with corresponding design variables, yielding a blown-up design. The t_i observations and design variables for person i are given by

$$\begin{bmatrix} \boldsymbol{y}_{i1} \\ \vdots \\ \boldsymbol{y}_{it_i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_{i1} \\ \vdots \\ \boldsymbol{Z}_{it_i} \end{bmatrix} \tag{9}$$

with

$$\boldsymbol{Z}_{it} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & \boldsymbol{x}_i^T & 0 & \cdots & 0 \\ \vdots & & \vdots & 0 & \ddots & & \vdots & \vdots & & \vdots & 0 & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & & \ddots & 0 & \vdots & & \vdots & \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & \boldsymbol{x}_i^T \end{bmatrix}$$

and corresponding parameter vector $\boldsymbol{\beta}^\top = (\gamma_{011}, \ldots, \gamma_{01m}, \gamma_{021}, \ldots, \gamma_{0qm}, \boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_m^\top)$.

2.1 Penalized Likelihood Models

Penalized maximum likelihood estimation uses a penalized version of the likelihood by including a penalty term. Let $\boldsymbol{\beta}^\top = (\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}^\top)$ be the parameter vector with $\boldsymbol{\gamma}_0^\top = (\gamma_{011}, \ldots, \gamma_{01m}, \gamma_{021}, \ldots, \gamma_{0qm})$ containing the baseline parameters and $\boldsymbol{\gamma}^\top = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_m^\top)$ containing the covariate effects. A penalized version of the log-likelihood derived from (7) is defined by

$$l_{\zeta_1,\zeta_2}(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}) = l(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}) - J_{\zeta_1,\zeta_2}(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}).$$

The first term, $l(\boldsymbol{\gamma}_0, \boldsymbol{\gamma})$, denotes the ordinary log-likelihood and the second term is a penalty term that includes the tuning parameters ζ_1 and ζ_2. The penalty that is used,

$$J_{\zeta_1,\zeta_2}(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}) = \zeta_1 J_1(\boldsymbol{\gamma}_0) + \zeta_2 J_2(\boldsymbol{\gamma}), \tag{10}$$

is split into two parts, $\zeta_1 J_1(\boldsymbol{\gamma}_0)$ and $\zeta_2 J_2(\boldsymbol{\gamma})$. The first part, $\zeta_1 J_1(\boldsymbol{\gamma}_0)$, represents a penalization of the baseline parameters $\boldsymbol{\gamma}_0$ and is chosen as

$$\zeta_1 J_1(\boldsymbol{\gamma}_0) = \zeta_1 \sum_{k=1}^{m} \sum_{t=2}^{q} (\gamma_{0tk} - \gamma_{0,t-1,k})^2. \tag{11}$$

It smoothes the baseline hazards over time by penalizing the differences between coefficients of adjacent time intervals, with the tuning parameter ζ_1 controlling the amount of penalization. The second part of the penalty term, $\zeta_2 J_2(\boldsymbol{\gamma})$, regularizes the estimates of the explanatory variables.

The simplest choice is a ridge type penalty given by

$$\zeta_2 J_2(\boldsymbol{\gamma}) = \zeta_2 \sum_{r=1}^{m} \sum_{j=1}^{p} \gamma_{rj}^2. \tag{12}$$

The penalty stabilizes estimates but no regression coefficients are set exactly to zero. Thus no variables are selected. Ridge type estimators for generalized linear models were investigated by Nyquist (1991) and Segerstedt (1992), the extension to multinomial responses was considered by Zahid and Tutz (2013).

More promising candidates that enforce variable selection are lasso type penalty terms (Tibshirani; 1996). However, simple lasso penalties, which consist in replacing γ_{rj}^2 in (12) by $|\gamma_{rj}|$, select parameters but not variables. Better penalty terms that enforce true variable selection penalize all the parameters that are linked to one variable simultaneously. Therefore, we consider the penalty proposed by Tutz et al. (2015), which was recently extended to the modeling of time

discrete competing risks (Möst et al.; 2014). It has the form

$$\zeta_2 J_2(\boldsymbol{\gamma}) = \zeta_2 \sum_{j=1}^{p} \phi_j \, ||\boldsymbol{\gamma}_{\bullet j}||_2, \tag{13}$$

where $\boldsymbol{\gamma}_{\bullet j}^\top = (\gamma_{1j}, \ldots, \gamma_{mj})$ comprises all parameters related to the j-th variable, ϕ_j is an (adaptive) weight which adjusts the penalty levels on parameter vectors $\boldsymbol{\gamma}_{\bullet j}^\top$ for their dimension, and $||\boldsymbol{u}||_2 = \sqrt{\boldsymbol{u}^\top \boldsymbol{u}}$ denotes the L_2-norm. The penalty (13) enforces true variable selection, meaning that all parameters which are related to the same variable are simultaneously shrunk toward zero. Its use yields sparse models which are easy to interpret. The penalty is closely related to the grouped lasso penalty (Yuan and Lin; 2006; Meier et al.; 2008), but in contrast to the original group lasso, the grouping of parameters arises from the multivariate response model, not from the predictors.

2.2 Prediction with Random Forests

The random forest method, introduced by Breiman (2001), is an ensemble of classification or regression trees. By aggregating several unstable trees a stable classification rule is built which has the advantage that the association between the predictors and the response is modeled in a highly flexible way. It has been shown that random forests have a much better prediction accuracy than single trees. Random forests incorporate complex interaction patterns between predictors and can also be applied to high-dimensional data where the predictor space is higher than the number of observations. This makes random forests especially suitable for complex genetic data that include hundreds or thousands of variables measured on a comparably small number of individuals. For detailed information on random forest methodology, we refer the reader to the existing literature (see, e.g., Boulesteix et al.; 2012; Strobl et al.; 2009, for an overview).

Several tree-based approaches have been developed to model survival times and their advantages over (semi-)parametric methods have been extensively discussed (see Bou-Hamad et al.; 2011, for an overview). Here we use the concept of Schmid et al. (2013) who make use of the fact that the likelihood of a time discrete survival model is equivalent to the likelihood of a regression model for binary outcome data (this follows directly from Eq. (7) in the special case of $m = 2$). This equivalence allows one to apply tree construction methods for binary outcomes. Analogously, the likelihood equivalence of a time discrete competing risks model and of a regression model for multinomially distributed outcomes, which was shown earlier in this section, allows to apply tree-based approaches for multicategory outcomes. The input data for observation i is given by

$$\begin{bmatrix} 0 & 1 & \boldsymbol{x}_i^\top \\ 0 & 2 & \boldsymbol{x}_i^\top \\ \vdots & \vdots & \vdots \\ 0 & t_i - 1 & \boldsymbol{x}_i^\top \\ r_i & t_i & \boldsymbol{x}_i^\top \end{bmatrix}. \tag{14}$$

Note that the time variable t has to be supplied to the software either as an ordered factor variable or as a metric variable, to only allow for splits at value c that yield partitions $\{t \leq c\}$ and $\{t > c\}$. After having fit a tree using input data of form (14), tree predictions can be obtained by computing the class proportions in the terminal nodes.

Instead of using single trees, in this article we make use of the random forest method, in which

a prediction is obtained by averaging over tree predictions. The proposed random forest approach to the modeling of time discrete competing risks has the advantage that it does not require a modification of the standard random forest algorithm, so any available random forest software can be used. The results presented in this paper were obtained using the random forest version of Hothorn et al. (2006), in which an unbiased split selection is implemented.

3 Simulation Studies

The data for the time discrete competing risks analysis were simulated by use of the multinomial logit model. The number of competing risks was always set to $m = 3$. Simulation studies were conducted with $n = 100$ and three different choices for the number of time intervals: $q \in \{5, 10, 20\}$. The simulations were performed for two settings. *Case 1* ("low-dimensional") denotes the setting where more observations are available than parameters in a multinomial logit model, and maximum likelihood estimates exist. In our studies we compare the performance of unpenalized multinomial logit models to that of penalized multinomial logit models and random forests. The second setting is denoted by *Case 2* ("high-dimensional"), and describes the setting where the number of parameters is larger than the number of observations. For this setting, maximum likelihood estimates do not exist and unpenalized multinomial logit models cannot be fit. We use these studies to investigate if the considered regularization techniques give reasonable predictions in settings where traditional approaches cannot be applied anymore. For this purpose, we compare the considered regularization techniques to a null model which does not include any covariates.

For *Case 1* we generated 100 datasets and 50 for *Case 2*. For both settings, six different scenarios that differ in the complexity of the data structure were simulated. These are described in the following.

3.1 Simulation Scenarios

Data was simulated for different scenarios that differ in

- the presence/absence of correlations between predictor variables,
- the inclusion of time-varying predictor effects in the linear predictor η_{itr} of the multinomial logit model given in Eq. (8),
- the inclusion of non-linear predictor effects in η_{itr},
- the inclusion of interaction terms in η_{itr}.

Table 1 gives an overview of the complexity of data in the six scenarios. In *Scenario 1* all predictor variables were uncorrelated and had time-constant and linear effects and no interaction terms were included. For *Scenarios 2-5* the data structure was more complex: exactly one of the "complexity components" (i.e., correlated predictor variables, time-varying predictor effects, non-linear predictor effects or interactions) was present in each scenario (see Table 1). Data structure was most complex in *Scenario 6* where all the "complexity components" are present.

3.2 Data Generation

For *Case 1* (low-dimensional setting), the number of predictor variables was set to $p = 8$. Variables X_1, X_3, X_5, X_7 were drawn from $Bin(1, 0.4)$, and X_2, X_4, X_6, X_8 were drawn from a multivariate

Scenario	Correlated predictors	Time-varying predictor effects	Non-linear predictor effects	Interacting predictors
1	–	–	–	–
2	✓	–	–	–
3	–	✓	–	–
4	–	–	✓	–
5	–	–	–	✓
6	✓	✓	✓	✓

Table 1: Overview of the complexity of data (in terms of predictor variable correlations or the inclusion of time-varying effects, non-linear effects or interaction terms in the linear predictor) in the different scenarios.

normal distribution with mean $\boldsymbol{\mu}_1^\top = (0, 0, 0, 0)$ and covariance matrix

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix},$$

with ρ specifying the correlation between metric predictor variables. The parameter ρ was set to 0 for scenarios without any correlations, and to 0.8 for scenarios with correlations.

For *Case 2* (high-dimensional setting), the number of predictor variables was set to $p = 500$. All predictor variables were drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}_2^\top = (0, 0, \ldots, 0) \in \mathbb{R}^{500}$ and block diagonal covariance matrix

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} \boldsymbol{A}_1 & 0 & \ldots & 0 \\ 0 & \boldsymbol{A}_2 & \ldots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & \boldsymbol{A}_{100} \end{bmatrix},$$

with

$$\boldsymbol{A}_j = \begin{bmatrix} 1 & a_j & a_j & a_j & a_j \\ a_j & 1 & a_j & a_j & a_j \\ a_j & a_j & 1 & a_j & a_j \\ a_j & a_j & a_j & 1 & a_j \\ a_j & a_j & a_j & a_j & 1 \end{bmatrix}$$

for $j = 1, \ldots, 100$. For scenarios without any correlations, the off-diagonal elements a_j were set to 0 for $j = 1, \ldots, 100$. For scenarios which include correlated predictors, a_j were independently drawn from the set $\{0.1, 0.2, 0.4, 0.6, 0.8\}$ for $j = 1, \ldots, 100$, accounting for a positive correlation within each block of five predictor variables and no correlation between the blocks. Strong correlations among a set of variables is typical, for example, in microarray data where genes highly correlate due to their spatial proximity in the genome.

Cause-specific hazards

The cause-specific hazards $\lambda_r(t|\cdot), r \in \{1, 2, 3\}$ were modeled via the multinomial logit model given in Eq. (8). However, more generally we will allow the parameter vector to depend on time. Thus,

$\boldsymbol{\gamma}_{tr}$ was specified depending on the considered scenario.

In *Case 1* the predictors $X_1, \ldots, X_4$ had an effect, while the other predictors $X_5, \ldots, X_8$ had no effect. In *Scenarios 1* and *2* including no interaction terms, time-varying or non-linear terms, the parameter vector for covariates $\boldsymbol{x}^\top = (x_1, x_2, \ldots, x_8)$ is simply given by $\boldsymbol{\gamma}_{tr} = \boldsymbol{\gamma}_r = (\gamma_{r1}, \ldots, \gamma_{r8})^\top$, for $r \in \{1, 2, 3\}$. In all other scenarios the parameter vector depends on time, or includes some additional terms which account for non-linear associations or interactions.

In all scenarios the coefficients $\gamma_{tr5}, \ldots, \gamma_{tr8}$ of the non-influential predictor variables $X_5, \ldots, X_8$ were always set to zero. In scenarios with time-constant predictor effects, γ_{r2} and γ_{r4} related to the metric predictor variables X_2 and X_4, were randomly drawn from the set $M_{\text{met}} = \{0.2, 0.4, 0.6, 0.8, 1\}$, and coefficients γ_{r1}, γ_{r3} for the binary predictor variables X_1 and X_3, were randomly drawn from $M_{\text{bin}} = \{-1, -0.5, 0.5, 1\}$ (cf. Table A1). In *Scenarios 3* and *6* (scenarios with time-varying effects), the effect γ_{tr1} of the binary predictor variable X_1 was specified as a functional in t. The effects of all other predictors did not depend on time (i.e., $\gamma_{trj} = \gamma_{rj}, j \in \{2, 3, \ldots, 8\}$). For influential predictors X_1, X_2, X_3, X_4, these were drawn from the set M_{met} and M_{bin}, respectively, and for non-influential predictors X_5, X_6, X_7, X_8, they were set to zero. The linear predictor η_{itr} was extended by an interaction term between predictor variables X_3 and X_4 in *Scenarios 5* and *6*, and by a quadratic term for X_2 in *Scenarios 4* and *6*. Detailed information are given in the appendix.

In *Case 2* only the first 20 of 500 predictor variables had an effect. In addition, predictor variables X_{21}, X_{22}, X_{23} had an effect in *Scenarios 5* and *6*, in which interaction terms related to these predictor variables were included. Coefficients $\gamma_{tr,21}, \ldots, \gamma_{tr,500}$ related to $X_{21}, \ldots, X_{500}$, were set to zero in all scenarios. The coefficients $\gamma_{tr1}, \ldots, \gamma_{tr,20}$ related to $X_1, \ldots, X_{20}$, always took a different value than zero for at least one $r \in \{1, 2, 3\}$. As with *Case 1*, coefficient values were randomly drawn from a set of appropriate values if the effect of a variable was constant over time (see Table A1 for details). For scenarios including variables with time-varying effects, the effects of $X_1, \ldots, X_{10}$ were modeled as a functional in t for some $r \in \{1, 2, 3\}$ (see Table A2). In *Scenarios 5* and *6* several interaction terms of different forms were included in the linear predictor, and in *Scenario 4* and *6* quadratic terms for variables $X_3, \ldots, X_{20}$ were integrated. Details are given in the appendix.

For both, *Case 1* and *Case 2*, the cause-specific baseline hazard functions γ_{0tr} for $r = 1, 2, 3$ were defined as follows:

$$
\begin{aligned}
\gamma_{0t1} &= a_1 t + b_1, \\
\gamma_{0t2} &= a_2 \frac{1}{\sqrt{t}} + b_2, \\
\gamma_{0t3} &= \begin{cases} a_3, & t \in \{1, 5, 9, 13, \ldots\} \\ a_3 + 1.5, & t \in \{2, 4, 6, 8, \ldots\} \\ a_3 + 3, & t \in \{3, 7, 11, 15, \ldots\}. \end{cases}
\end{aligned}
$$

For the cause-specific hazard function of the first event type, a linear function in t was imposed which gives a constant difference between baseline values of adjacent time intervals. For the second event type the difference between baseline values of adjacent time intervals was smaller for later time intervals, and for the third event type the baseline hazard function was periodic and repeats over four time intervals mimicking a seasonal effect. Values for a_1, b_1, a_2, b_2 and a_3 were chosen such that at each time interval a reasonable number of individuals failed from any event type

(Table A3).

Failure times

Failure times and the type of failure or censoring were generated using a sequential approach: a multinomial experiment was performed successively for each time interval until either failure from an event or censoring occurred. The algorithm which describes the generation of failure times is outlined in the following.

For an observation i, we start with $t = 1$ and repeat the following steps until $t > q$, or until the algorithm is stopped at an earlier stage.

1. A multinomial experiment is performed with $\boldsymbol{y}_{it} = (y_{it0}, y_{it1}, y_{it2}, y_{it3}) \sim \mathcal{M}(1, 1-\sum_{k=1}^{3} \lambda_k(t|\boldsymbol{x}_i), \lambda_1(t|\boldsymbol{x}_i),\ \lambda_2(t|\boldsymbol{x}_i),\ \lambda_3(t|\boldsymbol{x}_i))$. Independent of this process, a realization c_{it} is drawn from $Bin(1, \pi)$, where the parameter π controls the amount of censoring.

2. For $y_{it0} = 1 \wedge c_{it} = 0$, neither an event nor censoring occurs at time interval t. In this case the individual reaches the next time interval $t + 1$, and one proceeds with the first step. Otherwise the observation time for individual i ends at t. Then event type $r \in \{1, 2, 3\}$ is observed at t if $y_{itr} = 1$, and censoring occurs at t if $y_{it0} = 1$.

This simulation process was repeated for $i = 1 \ldots, n$. It is noteworthy that in our simulations the parameter π did not depend on covariates $\boldsymbol{x}_i$. We chose a value for π which yielded a moderate amount of censoring (20 – 30%).

3.3 Application to Simulated Data

When building prediction models we made use of all covariates $X_1, \ldots, X_p$, with $p = 8$ for *Case 1* and $p = 500$ for *Case 2*. Penalized and unpenalized multinomial logit models were modeled using a linear predictor of form $\eta_{itr} = \gamma_{0tr} + \boldsymbol{x}_i^\top \boldsymbol{\gamma}_r$, with $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $\boldsymbol{\gamma}_r^\top = (\gamma_{r1}, \ldots, \gamma_{rp})$. By using a linear predictor of this form, one does not account for non-linear relationships, time-dependent predictor effects and interactions. However, non-linear effects, time-varying effects and interactions are present in the data generated according to *Scenarios 3 - 6*. For these scenarios the penalized and unpenalized multinomial logit models are thus misspecified, while for *Scenarios 1* and *2* the models are correctly specified.

In contrast to these models, random forests are non-parametric and do not require any specification of the underlying structure. Random forests employ several parameters that have to be specified, such as the number of predictors randomly drawn for a split (`mtry`) or the size of a tree. Since the random forest approach for modeling discrete time competing risks, which was described in Section 2.2, has not been tested, we used different tree sizes in our studies to investigate if prediction performance is affected by the size of trees. Tree size was controlled by making restrictions on the minimum number of observations in a node. Other random forest parameters were specified as proposed by Strobl et al. (2007) to guarantee an unbiased tree construction (Table A4).

Considered prediction models

The models that were considered in our studies are outlined in the following.

- *Traditional approach*: Two standard multinomial logit models were considered which differ in the baseline hazard. For the first model ("GLM") a coefficient for each time interval was

estimated, while for the second model ("GLM sm.") the baseline hazard was smoothed by use of the penalty (11).

- *Penalized maximum likelihood*: Two penalized multinomial logit models with smoothed baseline hazard were fit. One model ("Lasso") uses penalty (13) which enforces true variable selection, while the other ("Ridge") makes use of penalty (12).

- *Random forests*: Five random forest models were tested that differ in the size of the trees. These models are referred to as "RF a, b", with a denoting the value for the minimum number of observations that is required in a node, and b denoting the value for the minimum number of observations that is required in a node in order to split the node into two daughter nodes. Values (a, b) were chosen as $(0, 0), (5, 20), (10, 20), (10, 40), (20, 60)$.

Parameter tuning

The parameters ζ_1 and ζ_2 in (10), which determine the strength of the penalization for the penalized likelihood models, and the parameter `mtry` which denotes the number of randomly drawn variables for a split in random forests, were chosen by 5-fold cross-validation. Due to the data blow up, a modification of the classical cross-validation procedure was used, in which all t_i data entries from the same observation i were attributed to the same cross-validation fold. For penalized likelihood models, the combination of ζ_1 and ζ_2 values from a two-dimensional grid of possible values was chosen that yielded the smallest cross-validated predictive deviance. Similarly, the parameter `mtry` in the random forest model was chosen from a grid of appropriate values.

3.4 Prediction Accuracy

The models described in the previous section were fit on a training set and evaluated on an independent test set of size $n_{\mathcal{T}} = 1000$, that follows the same distribution as the training set. Prediction accuracy was evaluated by using the predictive deviance which measures the discrepancy between independent data and the model fit. Models with smaller predictive deviances have higher accuracy in predicting future data than models with larger predictive deviances. The predictive deviance evaluated for the test set $\mathcal{T} := \{(t_i^{\mathcal{T}}, \delta_i^{\mathcal{T}}, r_i^{\mathcal{T}}, \boldsymbol{x}_i^{\mathcal{T}}), i = 1, \ldots, n_{\mathcal{T}}\}$, is given by

$$D(\mathcal{T}) = -2 \sum_{i=1}^{n_{\mathcal{T}}} \sum_{s=1}^{t_i^{\mathcal{T}}} \left\{ \sum_{k=1}^{m} y_{isk}^{\mathcal{T}} \log \hat{\lambda}_k(s|\boldsymbol{x}_i^{\mathcal{T}}) + y_{is0}^{\mathcal{T}} \log \left(1 - \sum_{k=1}^{m} \hat{\lambda}_k(s|\boldsymbol{x}_i^{\mathcal{T}})\right) \right\},$$

where y_{is0} and y_{isk} are indicator variables for the transition to the next time interval (cf. Section 2), and $\hat{\lambda}_r(t|\cdot)$ are the hazards for event types $r = 1, \ldots, m$, which are estimated from the training data.

The predictive deviance, however, is an unbounded measure. Thus we also considered the corresponding R^2 coefficient which is defined by

$$R^2 = \frac{1 - \exp((\sum_{i=1}^{n_{\mathcal{T}}} t_i^{\mathcal{T}})^{-1}(D - D_0))}{1 - \exp(-(\sum_{i=1}^{n_{\mathcal{T}}} t_i^{\mathcal{T}})^{-1} D_0)}, \tag{15}$$

where D_0 corresponds to the predictive deviance obtained from the null model that does not use any covariates. In our studies we considered a null model with smoothed baseline hazard via

applying penalty (11). The R^2 coefficient takes value 1 for a perfect prediction accuracy, and is 0 for a model which does not give better predictions than the null model (Nagelkerke; 1991).

3.5 Results

Case 1

Figure 1 shows the performance of all considered methods. Prediction performance was measured in terms of R^2 for the different scenarios and for a different number of time intervals ($q \in \{5, 10, 20\}$). The penalized multinomial logit models, Lasso and Ridge, always showed the best performance. They significantly outperformed the penalized multinomial logit models, GLM and GLM sm., in all considered scenarios. In *Scenario 4* with $q = 20$, the traditional multinomial logit models did not have any predictive ability at all. The penalized models, in contrast, still performed well. Lasso was often slightly better than Ridge, however, the difference was marginal. Both, GLM and GLM sm., had almost the same performance; obviously penalizing the baseline hazard did not result in better predictions.

There were large differences in performance between prediction models by random forests that differ in the size of trees: RF 0, 0 always had worst performance, and was very often not better than the null model. While RF 0, 0 had consistently the worst performance, there was no clear winner among the other four random forest prediction models. Which model performed best, was not only specific to the considered scenario, but also to the number of time intervals. Smaller tree sizes were especially of advantage for larger numbers of time intervals: for $q = 5$, prediction performance of RF 5, 20, RF 10, 20 and RF 10, 40 was clearly better than that of RF 20, 60, but for $q = 20$, RF 20, 60 outperformed the other two random forests.Compared to GLM and GLM sm. the random forest models (except for RF 0, 0) often had comparable performance. In the most complex scenario (*Scenario 6*) the best random forest models significantly outperformed the traditional models.

Note that in *Scenario 6* all parametric models (GLM, GLM sm., Lasso, Ridge) were poorly specified, as the linear predictor did not reflect true relationships. Although true relationships were not well captured, Lasso and Ridge did not have worse prediction performance than the non-parametric models by random forests. More precisely, the best random forest models had very similar performance. In all other scenarios the penalized multinomial logit models showed better prediction performance than the models based on random forests. It was somewhat surprising that Lasso and Ridge consistently outperformed random forest models, in particular because the linear predictors in Lasso and Ridge models were not specified correctly in most of the scenarios (*Scenarios 3 - 6*).

Case 2

The results for *Case 2* are shown in Figure 2. Note that – as previously stated – results for GLM and GLM sm. are not shown because maximum likelihood estimates do not exist. For *Case 2* the performance of penalized multinomial logit models depends highly on the penalty that was used. Ridge models did not have any predictive ability, which is seen from the corresponding distributions of R^2 lying around or below zero. The poor performance of Ridge has a specific reason. In these high-dimensional settings almost all of the predictors were completely unassociated with the occurrence of event types. Since Ridge shrinks the parameter coefficients of noise predictors toward zero but without explicitly removing noise predictors from the model, the performance

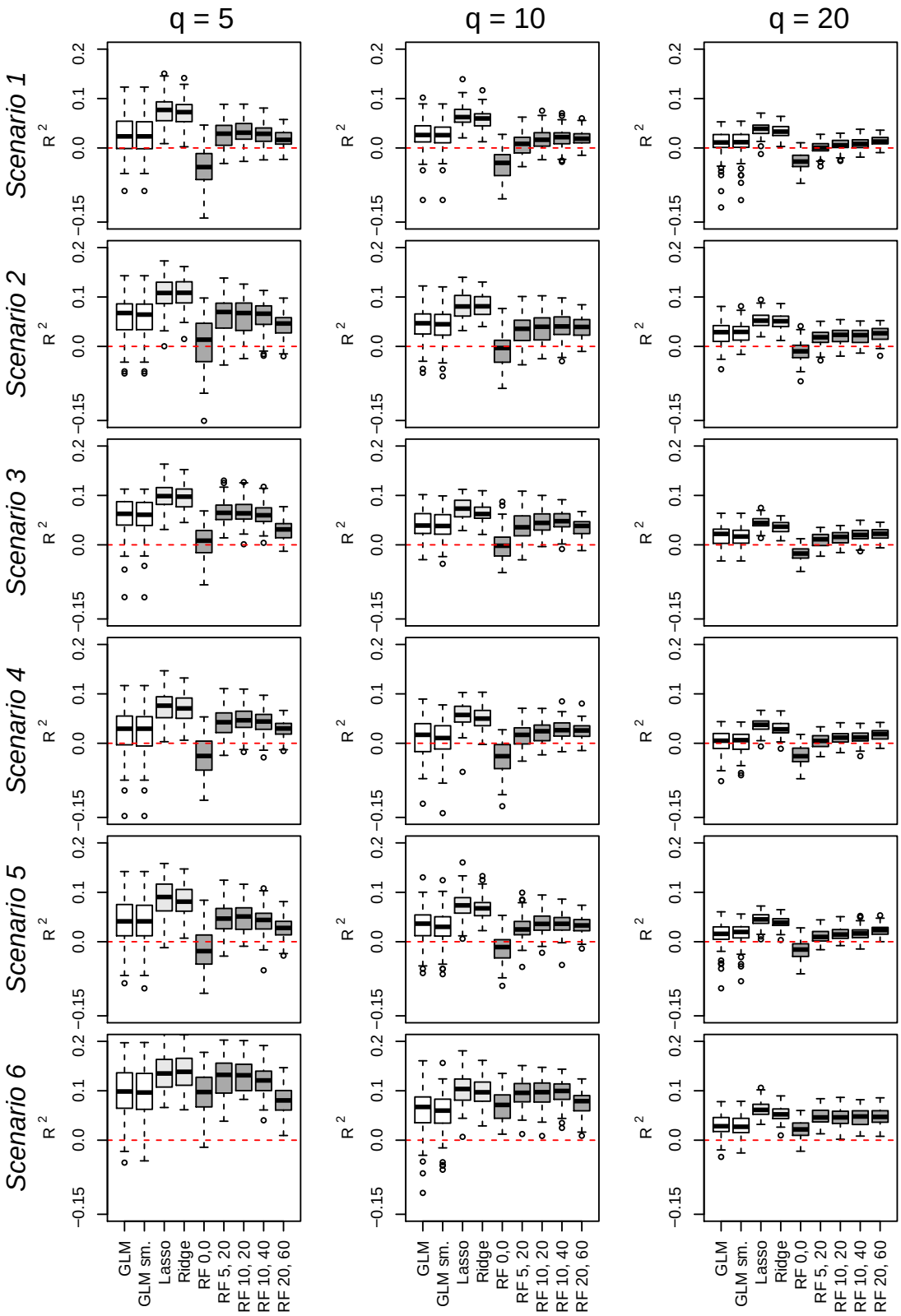

Figure 1: Comparison of methods in terms of prediction accuracy measured by R^2 coefficient for *Case 1*, for 100 simulated datasets with $n = 100$. Values close or below zero (red line) indicate that the respective model does not perform better than the null model, i.e., has no predictive ability. A smoothed baseline hazard was considered for the null model.

Data	No. event types m	No. individuals n	No. covariates p	No. time intervals q	Censoring (in %)
Bone Marrow Transplantation Data	2	137	11	8	39
Hodgkin's Disease Data	2	865	6	8	51
Bladder Cancer Data	2	304	1386	10	62

Table 2: Overview of real datasets (as used for the analysis).

suffers. This is a severe disadvantage if there are many predictors without any effect. In contrast, Lasso enforces variable selection and removes noise predictors from the model with the effect that prediction is much better.

The results obtained with the random forest method are in line with those for *Case 1.* RF 0, 0 had poor performance and should not be used for prediction purposes. The other four random forest models showed better performance. However, in some of the scenarios they did not have any predictive ability, either. Moreover, Lasso always clearly outperformed the considered random forest models. Even though the linear predictor did not capture the true association in *Scenarios 3 - 6*, Lasso performed much better than the non-parametric models by random forests.

4 Real Data Studies

The proposed methods were tested and compared to the traditional maximum likelihood approach based on three publicly available real world datasets from the medical field. Table 2 gives a rough overview of the datasets. The first two datasets reflect low-dimensional settings, in which unpenalized multinomial logit models can be applied, while the last dataset includes so many predictors that maximum likelihood estimates do not exist and is used to investigate the practical utility of the proposed regularization techniques.

In the following we give a brief description of the datasets. A description of the covariates can be found in the appendix.

4.1 Data

Bone Marrow Transplantation Data

The Bone Marrow Transplantation Data includes $n = 137$ acute leukemia patients who have received a bone marrow transplant. Bone marrow transplantation was considered to have failed if either leukemia returns (relapse) or if the patient dies while being in remission (treatment-related death). These are the two competing events that were considered for the analysis. Patient-related as well as donor-related factors are expected to play a role in the patients' recovery process. A total of 11 patient- and donor-related variables were documented which may help in predicting the two events. Neither relapse nor death was observed for 39% of the patients. Time was originally given in months from transplantation. For performing a competing risks analysis with discrete time, due to stability reasons time was coarsened into the 8 time intervals $[0, 0.25], (0.25, 0.5], (0.5, 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 7.5]$, with the numbers corresponding to years from transplantation. The dataset is provided in Appendix C of the textbook of Klein and Moeschberger (2005) and is also part of the R package KMsurv. For a detailed description of the

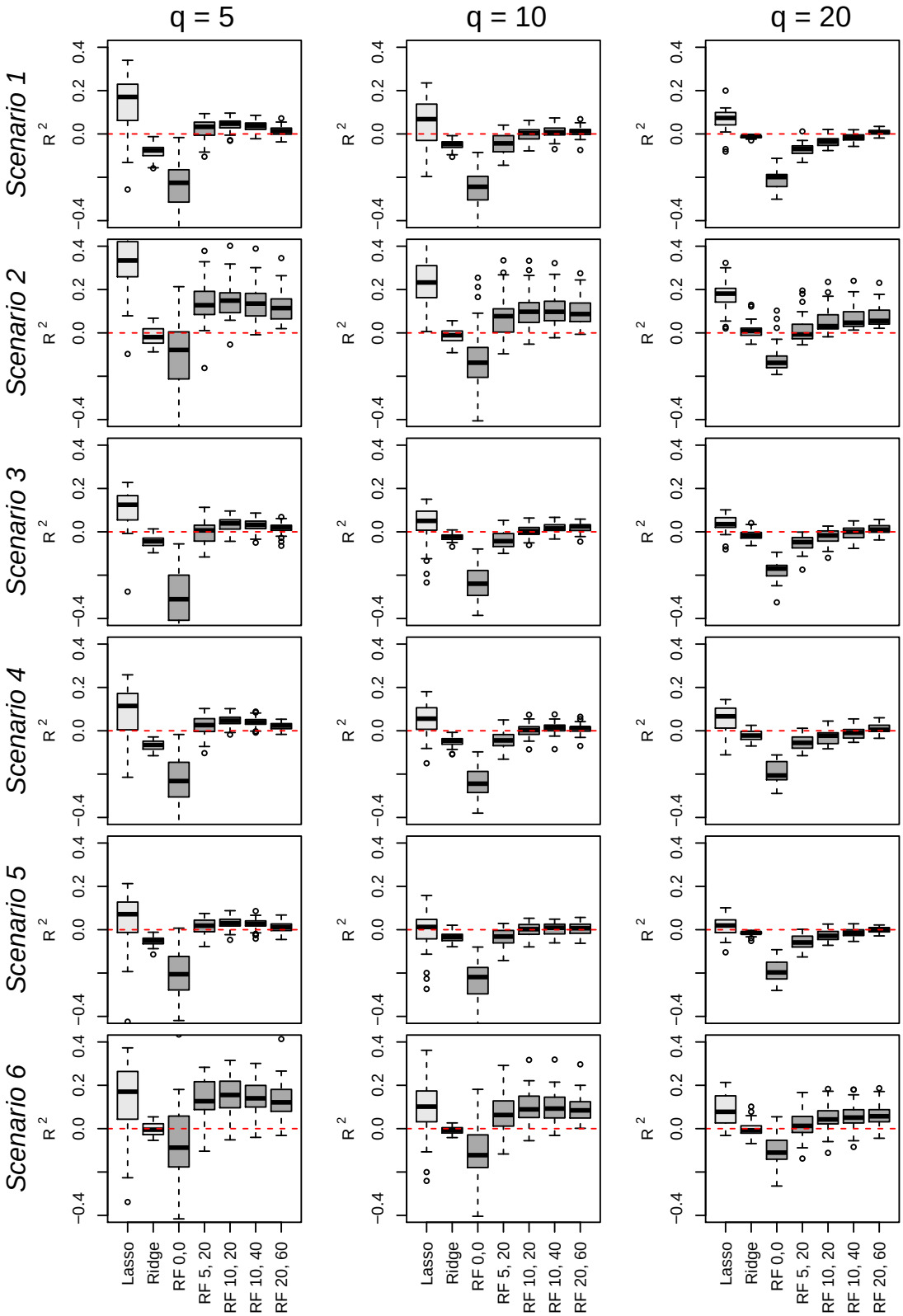

Figure 2: Comparison of methods in terms of prediction accuracy measured by R^2 coefficient for *Case 2*, for 50 simulated datasets with $n = 100$. Values close or below zero (red line) indicate that the respective model does not perform better than the null model, i.e., has no predictive ability. A smoothed baseline hazard was considered for the null model.

study, we refer the reader to Klein and Moeschberger (2005) and Copelan et al. (1991).

Hodgkin's Disease Data

The Hodgkin's Disease Data comprises information on $n = 865$ patients with stage I or II Hodgkin's disease recorded in the years 1968 to 1986 at the Princess Margaret Hospital (Petersen et al.; 2004). The two event types which we considered, were relapse and death without preceding relapse. For half of the patients neither type of event was observed giving a high amount of censored observations. Six covariates were included in the analysis. For stability reasons time was coarsened to 4 year intervals. Due to the small number of observed events or censoring for later follow-up times, the last interval was extended and comprised all individuals whose failure time is beyond 28 years. This gave us a total of 8 time intervals. The dataset is publicly available under `http://www.uhnresearch.ca/labs/hill/People_Pintilie.htm` and is also integrated in the R package randomForestSRC.

Bladder Cancer Data

In a retrospective multicenter study Dyrskjøt et al. (2007) validated previously reported gene signatures for predicting progression in bladder cancer patients. Biological material was taken from bladder cancer patients who were operated in the years 1987 to 2000 in hospitals in Denmark, Sweden, Spain, France and England. Information on 1381 preprocessed microarray features was extracted. In addition, there was information on five clinical covariates. The analysis presented here was restricted to $n = 304$ patients with non muscle-invasive tumors (stage pT_a and pT_1 tumors), for whom clinical important covariates (age, sex, tumor stage, grade) and genetic information were available. Here the two competing events "death from bladder cancer" and "death from another or unknown reason" were considered. The amount of censoring was very high in this dataset (62% in the considered patient population). Follow-up time was originally given in months from sampling visit. We coarsened it to one year intervals. Information from years 10 to the maximal follow-up time of 15 after sampling visit were aggregated to one time interval due to the sparse number of observed events or censored individuals for later follow-up times. The Bladder Cancer Data is publicly available from the Gene Expression Omnibus (GEO) database (series accession no. GSE5479).

4.2 Application to Real Data

We applied the prediction models outlined in Section 3.3 to the three real datasets. Penalized and unpenalized multinomial logit models were modeled using a linear predictor of form $\eta_{itr} = \gamma_{0tr} + \boldsymbol{x}_i^\top \boldsymbol{\gamma}_r$, with covariate vector $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and coefficient vector $\boldsymbol{\gamma}_r^\top = (\gamma_{r1}, \ldots, \gamma_{rp})$. Thus effects are assumed to be constant over time, and no interaction terms or non-linear terms are assumed.

We fit random forest models in addition to those described in 3.3, if there were indications that the parameters a and b controlling tree size were not appropriate. For the Hodgkin's Disease Data, for example, we also considered models with smaller tree sizes (RF 40, 200; RF 40, 300 and RF 40, 400), since there was a tendency that these might perform better. It is noteworthy that prediction error will be biased downwards when fitting several random forest models with different tree sizes and choosing the best one. Thus in practical applications one should tune

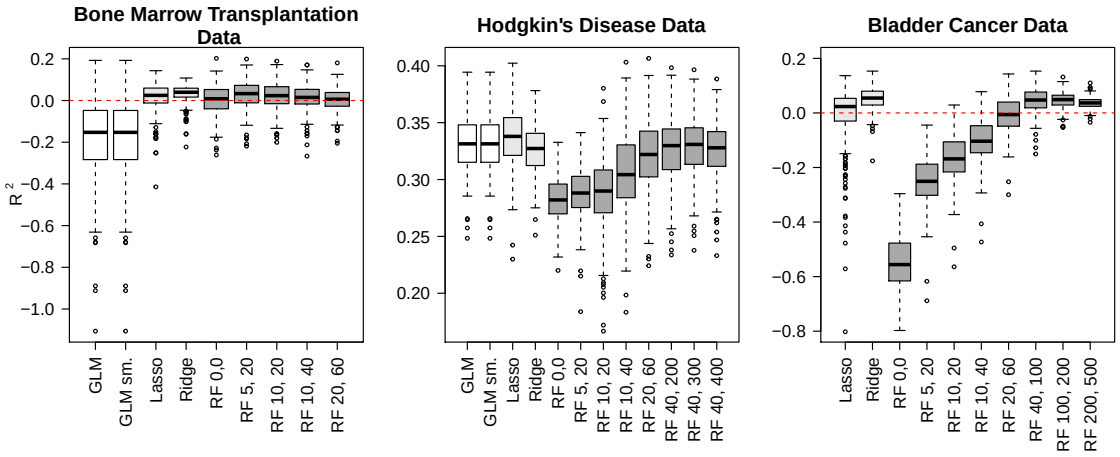

Figure 3: Comparison of methods for the Bone Marrow Transplantation Data (left panel), the Hodgkin's Disease Data (middle panel) and the Bladder Cancer Data (right panel). Prediction accuracy measured in terms of the R^2 coefficient in combination with 50 repetitions of 5-fold cross-validation. Values close or below zero (red line) indicate that the respective model does not perform better than the null model, i.e., has no predictive ability. A smoothed baseline hazard was considered for the null model.

the parameters that control the size of trees. Here we did not tune these parameters because we wanted to investigate the influence of tree size on prediction performance.

To study prediction performance of the methods on real datasets, we computed the R^2 as defined in (15), in combination with cross-validation. This was done as follows: a real dataset was split into 5 parts, while ensuring that data entries from the same observation were attributed to the same cross-validation fold. Each part of the data was used once as test set, while the other parts were used as training data. Prediction models were fit using the training observations and the models' prediction performance was evaluated using the respective test observations.

The parameters ζ_1 and ζ_2 in Eq. (10) for penalized multinomial logit models and `mtry` for random forests were chosen through nested cross-validation. More precisely, within each training step of the cross-validation procedure we chose values for the parameters which minimized the cross-validated predictive deviance, as has already been described in Section 3.3. Since the models' performance depends highly on the partition of the data into the five folds, we used 50 random partitions. This resulted in 250 values for the R^2.

4.3 Results

Figure 3 shows the predictive performance of the considered prediction models for the Bone Marrow Transplantation Data (left panel), for the Hodgkin's Disease Data (middle panel) and for the Bladder Cancer Data (right panel).

The results are in line with those obtained for simulated data. Penalized multinomial logit models were among the best methods for the three considered datasets. For the Bone Marrow Transplantation Data, Lasso and Ridge clearly outperformed the unpenalized models. The random forest models, RF 5, 20 and RF 10, 20, however, had comparable performance.

For the Hodgkin's Disease Data, the traditional models performed well. This is possibly related to the small number of parameters to be estimated and the large sample size ($p = 2$; $n = 865$; $m = 2$). The penalized multinomial models had almost the same performance. The random forest

approach had competitive performance if very high values for the parameters controlling tree size were chosen.

Traditional models could not be estimated for the Bladder Cancer Data. For this dataset penalized multinomial logit models and random forests with small tree sizes (RF 40, 100, RF 100, 200, and RF 200, 500) performed better than the null model. Random forest models with larger tree size, in contrast, did not have any predictive ability. Ridge performed better than Lasso which might be attributable to a possibly large number of genes with small effects.

5 Conclusion

The classical approach to the modeling of competing risks with discrete duration time is to fit a multinomial logit model. However, the amount of parameters increases rapidly when the number of predictor variables grows, as multinomial logit models employ several coefficients for each explanatory variable. Therefore, maximum likelihood estimates tend to deteriorate quickly which leads to a worsening in prediction performance. In some cases maximum likelihood estimates do even not exist. This motivates the development of alternative approaches to the modeling of time discrete competing risks.

This paper investigates the use of alternative approaches that are based on regularization techniques. As regularization techniques we considered penalized maximum likelihood models with ridge- and lasso type penalties and the non-parametric random forest method using the random forest version of Hothorn et al. (2006). While penalized maximum likelihood models have already been used for modeling competing risks, the random forest approach to the modeling of time discrete competing risks has not been described before. It has the advantage that it does not require a modification of the standard random forest algorithm, so any available random forest software can be used.

The prediction performance of the considered methods was investigated through simulation studies and three real data applications. Our studies show that regularization-based models give more accurate predictions than unpenalized multinomial logit models in many settings. Penalized multinomial logit models overall had the best performance. In all our studies the performance of these models was at least as good as the performance of unpenalized multinomial logit models.

Predictive abilities of random forest models highly depended on the size of the trees. If trees were grown without employing stopping criteria, the resulting random forest models had poor performance, and very often did not have any predictive ability. In practical applications one should thus tune parameters that control the size of trees. Compared to the penalized multinomial logit models, the random forest models had sometimes equal but never better performance.

We conclude from our results that regularization-based parametric approaches considered here, are promising tools for prediction purposes in a competing risks settings with discrete time. In particular, penalized multinomial logit models have shown the best performance, and our results suggest that they give accurate predictions even in cases of model misspecifications.

Acknowledgements

We thank Wolfgang Pößnecker for technical assistance and for providing us with an implementation of the Lasso approach for discrete time competing risks models.

References

Andersen, P. K., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer, New York.

Beyersmann, J., Allignol, A. and Schumacher, M. (2012). *Competing risks and multistate models with R*, Springer.

Bou-Hamad, I., Larocque, D., Ben-Ameur, H. et al. (2011). A review of survival trees, *Statistics Surveys* **5**: 44–71.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(6): 493–507.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Copelan, E. A., Biggs, J. C., Thompson, J. M., Crilley, P., Szer, J., Klein, J. P., Kapoor, N., Avalos, B. R., Cunningham, I. and Atkinson, K. (1991). Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with BuCy2, *Blood* **78**(3): 838–843.

Dyrskjøt, L., Zieger, K., Real, F. X., Malats, N., Carrato, A., Hurst, C., Kotwal, S., Knowles, M., Malmström, P.-U., de la Torre, M. et al. (2007). Gene expression signatures predict outcome in non–muscle-invasive bladder carcinoma: a multicenter validation study, *Clinical Cancer Research* **13**(12): 3545–3551.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics* **15**(3): 651–674.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, Vol. 360, John Wiley & Sons.

Klein, J. and Moeschberger, M. (2005). *Survival analysis: Techniques for censored and truncated data*, Springer.

Kleinbaum, D. G. and Klein, M. (2005). *Survival analysis*, Springer.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1): 53–71.

Möst, S., Pößnecker, W. and Tutz, G. (2014). Variable selection for discrete competing risks models, *Technical Report 161*, Department of Statistics, University of Munich.
URL: *http://epub.ub.uni-muenchen.de/20923/*

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination, *Biometrika* **78**(3): 691–692.

Nyquist, H. (1991). Restricted estimation of generalized linear models, *Applied Statistics* **40**: 133–141.

Petersen, P., Tsang, R., Gospodarowicz, M., Pintilie, M., Wells, W., Hodgson, D., Sun, A. and Crump, M. (2004). Stage I and II Hodgkin's disease: Long term outcome and second cancer risk, *Radiotherapy and Oncology* **72**(S23).

Putter, H., Fiocco, M. and Geskus, R. (2007). Tutorial in biostatistics: Competing risks and multi-state models, *Statistics in Medicine* **26**(11): 2389–2430.

Schmid, M., Küchenhoff, H. and Tutz, G. (2013). Survival trees for discrete failure times, *Proceedings of the Joint Statistical Meetings (JSM 2013), Section on Statistical Learning and Data Mining*, Montreal, Canada, pp. 4196–4207.

Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models, *Communications in Statistics – Theory and Methods* **21**: 2227–2246.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics* **8**: 25.

Strobl, C., Malley, J. and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests, *Psychological Methods* **14**(4): 323–348.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**(1): 267–288.

Tutz, G., Pößnecker, W. and Uhlmann, L. (2015). Variable selection in general multinomial logit models, *Computational Statistics & Data Analysis* **82**: 207–222.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1): 49–67.

Zahid, F. M. and Tutz, G. (2013). Ridge estimation for multinomial logit models with symmetric side constraints, *Computational Statistics* **28**(3): 1017–1034.

A Simulation Studies

A.1 *Case 1*

Time-varying effects

In the presence of time-varying effects the linear predictor η_{itr} in the multinomial model

$$\lambda_r(t|\boldsymbol{x}_i) = \frac{\exp(\eta_{itr})}{1 + \sum_{k=1}^{m} \exp(\eta_{itk})}$$

can be formulated as

$$\eta_{itr} = \gamma_{0tr} + \boldsymbol{x}^\top \boldsymbol{\gamma}_{tr},$$

with time-varying covariate weights $\boldsymbol{\gamma}_{tr}^\top = (\gamma_{tr1}, \ldots, \gamma_{tr8})$ for predictor variables $X_1, \ldots, X_8$. For *Scenarios 3* and *6* the effect of the binary predictor variable X_1 on the cause-specific hazards for event types $r = 1, 2, 3$ was modeled as

$$\begin{aligned}
\gamma_{t11} &= 2\left(\frac{t}{q}\right)^2, \\
\gamma_{t21} &= -2\frac{t}{q}, \\
\gamma_{t31} &= 4I(t > \left\lceil \frac{q}{2} \right\rceil).
\end{aligned}$$

The effects for all other predictor variables are time-constant and the corresponding coefficients were independently drawn from the sets $\{0.2, 0.4, 0.6, 0.8, 1\}$ (for metric variables) and $\{-1, -0.5, 0.5, 1\}$ (for binary variables).

Non-linear effects

For *Scenarios 4* and *6*, besides the linear term, a quadratic term for the metric predictor variable X_2 was included in the linear predictor η_{itr} for event types $r = 1$ and $r = 3$. The corresponding coefficients for the quadratic terms were 0.2 and 0.6, respectively.

Interactions

For *Scenarios 5* and *6* an additional (time-constant) interaction term between the binary predictor variable X_3 and the metric predictor variable X_4 was included in the linear predictor η_{itr}. The interaction term for $r = 1$ was 1.5, for $r = 2$ it was -1 and for $r = 3$ it was 0.5.

A.2 *Case 2*

Time-varying effects

For *Scenarios 3* and *6* time-varying parameter vectors $\boldsymbol{\gamma}_{tr}^\top = (\gamma_{tr1}, \ldots, \gamma_{tr,500})$ were defined. Table A2 gives an overview of the parameters.

Non-linear effects

For *Scenarios 4* and *6*, besides the linear term, quadratic terms were included in the linear predictor. For $r = 1$ and predictor variables $X_6, \ldots, X_{10}$, the coefficients for the quadratic terms

Case	Event Type	Coefficients	Effect / Effect set
1	$r \in \{1,2,3\}$	γ_{r1}, γ_{r3}	$\{-1, -0.5, 0.5, 1\}$
		γ_{r2}, γ_{r4}	$\{0.2, 0.4, 0.6, 0.8, 1\}$
		$\gamma_{r5}, \ldots, \gamma_{r8}$	0
2	$r = 1$	$\gamma_{11}, \ldots, \gamma_{1,10}$	$\{0.6, 0.8, 1, 2\}$
		$\gamma_{1,11}, \ldots, \gamma_{1,500}$	0
	$r = 2$	$\gamma_{21}, \ldots, \gamma_{2,20}$	$\{-2, -1, 1\}$
		$\gamma_{2,21}, \ldots, \gamma_{2,500}$	0
	$r = 3$	$\gamma_{31}, \ldots, \gamma_{35}$	$\{-2, 2\}$
		$\gamma_{36}, \ldots, \gamma_{3,500}$	0

Table A1: Parameters specified for *Case 2 Scenarios 1, 2, 4, 5* which include no time-varying predictor effects.

were $0.5, 0.5, -0.5, -0.5$ and -0.5. For $r = 2$ and predictor variables $X_{11}, \ldots, X_{20}$, the values were $-0.5, -0.5, -0.5, -0.5, -0.5, 0.5, 0.5, 0.5, 0.5, 0.5$, and for $r = 3$ and predictor variables X_3, X_4, X_5, values were $-0.5, 0.5$ and 0.5.

Interactions

For *Scenarios 5* and *6* additional (time-constant) interaction terms between two or three variables were included in the linear predictor. Interaction effects were modeled for (i) variables, all with (main) effects, (ii) variables, both, with effect and without effect, and (iii) variables, all without any effect. Note that (main) effects were present for variables $X_1, \ldots, X_{20}$, and were defined as described in Table A2. For *Scenarios 5* and *6* the following interaction terms were included:

- For $r = 1$:
  $\gamma_{1,18:19}\ x_{18}\ x_{19} := 0.5\ x_{18}\ x_{19}$,
  $\gamma_{1,20:21}\ x_{20}\ x_{21} := -1\ x_{20}\ x_{21}$,
  $\gamma_{1,22:23}\ x_{22}\ x_{23} := -0.5\ x_{22}\ x_{23}$.
- For $r = 2$:
  $\gamma_{2,6:7:8}\ x_6\ x_7\ x_8 := 2\ x_6\ x_7\ x_8$,
  $\gamma_{2,9:10:11}\ x_9\ x_{10}\ x_{11} := -2\ x_9\ x_{10}\ x_{11}$,
  $\gamma_{2,12:13:14}\ x_{12}\ x_{13}\ x_{14} := -2\ x_{12}\ x_{13}\ x_{14}$.
- For $r = 3$:
  $\gamma_{3,3:4}\ x_3\ I(x_4 > 0) := 1\ x_3\ I(x_4 > 0)$,
  $\gamma_{3,5:6}\ x_5\ I(x_6 > 0) := -1\ x_5\ I(x_6 > 0)$,
  $\gamma_{3,7:8}\ x_7\ I(x_8 > 0) := -1\ x_7\ I(x_8 > 0)$.

Event Type	Coefficient	Effect / Effect set	
$r = 1$	γ_{t11}	$\frac{2}{q}t$	Time-varying effects
	γ_{t12}	$\frac{2}{q}(q-t+1)$	
	γ_{t13}	$-\frac{2}{q}t$	
	γ_{t14}	$-\frac{2}{q}(q-t+1)$	
	γ_{t15}	$-2+4I(t \geq \frac{q}{2})$	
	$\gamma_{t16} = \gamma_{16}$	$\{0.6, 0.8, 1, 2\}$	Time-constant effects
	$\vdots$	$\vdots$	
	$\gamma_{t1,10} = \gamma_{1,10}$	$\{0.6, 0.8, 1, 2\}$	
	$\gamma_{t1,11} = \gamma_{1,11}$	0	No effect
	$\vdots$	$\vdots$	
	$\gamma_{t1,500} = \gamma_{1,500}$	0	
$r = 2$	γ_{t21}	$\frac{2}{\sqrt{q}}\sqrt{t}$	Time-varying effects
	γ_{t22}	$\frac{2}{\sqrt{q}}\sqrt{q-t+1}$	
	γ_{t23}	$\frac{-2}{\sqrt{q}}\sqrt{t}$	
	γ_{t24}	$\frac{-2}{\sqrt{q}}\sqrt{q-t+1}$	
	γ_{t25}	$\frac{2}{q^2}t^2$	
	γ_{t26}	$\frac{2}{q^2}(q-t+1)^2$	
	γ_{t27}	$\frac{-2}{q^2}t^2$	
	γ_{t28}	$\frac{-2}{q^2}(q-t+1)^2$	
	γ_{t29}	$-2I(t \geq \frac{q}{2})$	
	$\gamma_{t2,10}$	$-2I(t < \frac{q}{2})$	
	$\gamma_{t2,11} = \gamma_{2,11}$	$\{-2, -1, 1\}$	Time-constant effects
	$\vdots$	$\vdots$	
	$\gamma_{t2,20} = \gamma_{2,20}$	$\{-2, -1, 1\}$	
	$\gamma_{t2,21} = \gamma_{2,21}$	0	No effect
	$\vdots$	$\vdots$	
	$\gamma_{t2,500} = \gamma_{2,500}$	0	
$r = 3$	γ_{t31}	$2I(t \geq \frac{q}{2})$	Time-varying effects
	γ_{t32}	$2I(t < \frac{q}{2})$	
	$\gamma_{t33} = \gamma_{33}$	$\{-2, 2\}$	Time-constant effects
	$\vdots$	$\vdots$	
	$\gamma_{t35} = \gamma_{35}$	$\{-2, 2\}$	
	$\gamma_{t36} = \gamma_{36}$	0	No effect
	$\vdots$	$\vdots$	
	$\gamma_{t3,500} = \gamma_{3,500}$	0	

Table A2: *Case 2*: Parameters specified for *Scenarios 3* and *6*, both including time-varying predictor effects.

A.3 Baseline Hazards

Cause-specific baseline hazard functions γ_{0tr} for $r = 1, 2, 3$ were defined as follows:

$$\gamma_{0t1} = a_1 t + b_1,$$
$$\gamma_{0t2} = a_2 \frac{1}{\sqrt{t}} + b_2,$$
$$\gamma_{0t3} = \begin{cases} a_3, & t \in \{1, 5, 9, 13, \ldots\} \\ a_3 + 1.5, & t \in \{2, 4, 6, 8, \ldots\} \\ a_3 + 3, & t \in \{3, 7, 11, 15, \ldots\}, \end{cases}$$

with a_1, b_1, a_2, b_2 and a_3 given in Table A3.

Setting	q	a_1	b_1	a_2	b_2	a_3
Case 1	5	$5.5\frac{1}{5-1}$	$-5.5 - b_2$	$-11\frac{\sqrt{5}}{\sqrt{5}-1}$	$-11 - a_1$	-5
(low-dimensional)	10	$5\frac{1}{10-1}$	$-5 - b_2$	$-9\frac{\sqrt{10}}{\sqrt{10}-1}$	$-10 - a_1$	-5.5
	20	$6\frac{1}{20-1}$	$-6 - b_2$	$-9\frac{\sqrt{20}}{\sqrt{20}-1}$	$-11 - a_1$	-6.5
Case 2	5	$7\frac{1}{5-1}$	$-7 - b_2$	$-14\frac{\sqrt{5}}{\sqrt{5}-1}$	$-14 - a_1$	-6
(high-dimensional)	10	$9\frac{1}{10-1}$	$-8 - b_2$	$-16\frac{\sqrt{10}}{\sqrt{10}-1}$	$-16 - a_1$	-7
	20	$12\frac{1}{20-1}$	$-8 - b_2$	$-20\frac{\sqrt{20}}{\sqrt{20}-1}$	$-20 - a_1$	-7.5

Table A3: Parameter values specified for the baseline hazard functions.

A.4 Random Forests Parameters

Details on the parameters passed to the `cforest` function from R package party are given in Table A4.

Parameter	Value	Default #
`ntree`	500	yes
`mtry`	determined via cross-validation	no
`replace`	`FALSE`	yes
`teststat`	"`quad`"	yes
`testtype`	"`Univ`"	yes
`mincriterion`	0	no
`minsplit`	{0, 20, 40, 60}	no
`minbucket`	{0, 5, 10, 20}	no

Table A4: Parameters passed to `cforest` or `cforest_control`. # Default settings in party version 1.0-10.

B Dataset Descriptions

Brief descriptions of the variables for the Bone Marrow Transplantation Data, the Hodgkin's Disease Data and the Bladder Cancer Data are given in Tables A5 – A7.

Variable	Description
Event type	1: treatment-related death 2: relapse
Time	time since transplantation
Disease group	ALL AML (low risk) AML (high risk)
Patient's age	age of patients (range: $[7, 52]$)
Donor's age	age of the patient's donor (range: $[2, 56]$)
Patient's sex	patient's gender
Donor's sex	gender of the patient's donor
Patient's CMV status	positive / negative
Donor's CMV status	positive / negative
Waiting time to transplant	waiting time in months from diagnosis (range: $[0.8, 87.2]$)
FAB classification	Grade 4 or 5 and AML / otherwise
Hospital	the hospital where transplantation took place: Ohio State University Hospitals (Columbus) Alferd Hospital (Melbourne) St. Vincent's Hospital (Sydney) Hahnemann University (Philadelphia)
MTX used as graft-versus-host-prophylactic	yes / no

Table A5: Description of variables for the Bone Marrow Transplantation Data.

Variable	Description
Event type	1: death 2: relapse
Time	time since diagnosis
Treatment	the treatment a person received, which is either radiation or radiation in combination with chemotherapy
Age	person's age (range: $[15.6, 90]$)
Sex	person's gender
Size of mediastinum involvement	either no involvement, of small size or of large size
Extranodal disease	has the disease spread? (yes / no)
Clinical stage	clinical stage of lymphoma, either stage I or stage II

Table A6: Description of variables for the Hodgkin's Disease Data.

Variable	Description
Event type	1: death from bladder cancer 2: death from other or unknown reason
Time	time after sampling visit
Treatment	received treatment, which is either one of instillations of Bacillus Calmette-Guerin (BCG) or mitomycin-C
Age	person's age (range: $[27, 95]$)
Sex	person's gender
Clinical stage	clinical stage of tumor which is either pT_a or pT_1
Grade	PUNLMP, low, high
Seq. 1	Microarray measurement 1
$\vdots$	$\vdots$
Seq. 1381	Microarray measurement 1381

Table A7: Description of variables for the Bladder Cancer Data.