Gene expression

Boosting for high-dimensional time-to-event data with competing risks

Harald Binder\textsuperscript{1,2,}\textsuperscript{*}, Arthur Allignol\textsuperscript{1,2}, Martin Schumacher\textsuperscript{2} and Jan Beyersmann\textsuperscript{1,2}

\textsuperscript{1}Freiburg Center for Data Analysis and Modeling, University of Freiburg, Eckerstr. 1 and \textsuperscript{2}Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany

Received on October 16, 2008; revised on January 19, 2009; accepted on February 12, 2009

Advance Access publication February 25, 2009

Associate Editor: Joaquín Dopazo

ABSTRACT

\textbf{Motivation:} For analyzing high-dimensional time-to-event data with competing risks, tailored modeling techniques are required that consider the event of interest and the competing events at the same time, while also dealing with censoring. For low-dimensional settings, proportional hazards models for the subdistribution hazard have been proposed, but an adaptation for high-dimensional settings is missing. In addition, tools for judging the prediction performance of fitted models have to be provided.

\textbf{Results:} We propose a boosting approach for fitting proportional subdistribution hazards models for high-dimensional data, that can e.g. incorporate a large number of microarray features, while also taking clinical covariates into account. Prediction performance is evaluated using bootstrap \textit{632+} estimates of prediction error curves, adapted for the competing risks setting. This is illustrated with bladder cancer microarray data, where simultaneous consideration of both, the event of interest and competing events, allows for judging the additional predictive power gained from incorporating microarray measurements.

\textbf{Availability:} The proposed boosting approach is implemented in the \textsf{R} package \textsf{CoxBoost} and prediction error estimation in the package \textsf{peperr}, both available from CRAN.

\textbf{Contact:} binderh@fdm.uni-freiburg.de

1 INTRODUCTION

For building prognostic models for cancer patients, often techniques are needed that were developed specifically for time-to-event responses. For example, when a gene expression signature for prediction of an event such as progression or recurrence is wanted, there will typically be censored observations, i.e. binary or continuous response techniques are not adequate. As an alternative, for example the Cox proportional hazards model allows for adequately dealing with censoring, and several approaches for fitting such models to high-dimensional data have been developed (see Binder and Schumacher, 2008b; Gui and Li, 2005; Park and Hastie, 2007, for example).

When there are competing event types, such as ‘progression’ versus ‘death from non-cancer cause’, often either a Cox proportional hazards model for a composite endpoint, such as ‘progression free survival’, is fitted, or a proportional hazards model is fitted for the event of interest only, where occurrence of other events is treated as censoring. While the former approach does no longer allow to specifically investigate effects on occurrence of the event of interest, the latter is problematic, because censoring by a competing event is informative (Andersen \textit{et al.}, 1993). Informative censoring is, in particular a concern, if there is influence of the covariates on the competing cause-specific hazard. Furthermore, one cause-specific hazard alone is not sufficient to describe the cumulative incidence function, i.e. the proportion of patients that suffer the event of interest over the course of time, and therefore does not allow for interpretation of covariate effects in terms of effects on event probabilities (Pepe and Mori, 1993). In the following, we will investigate high-dimensional modeling approaches, that address these shortcomings.

One approach would be to fit two Cox proportional hazards models, one for the event of interest, and a second for the other events. These would then have to be interpreted together to get insight into the structure underlying the data. For example, one would need to investigate whether a gene that is protective with respect to cancer events is potentially harmful with respect to death from other causes. A more formal synthesis interpretation in terms of the cumulative incidence functions can already be challenging in a low-dimensional setting. In a high-dimensional setting, combined interpretation of two cause-specific model fits is even more difficult, as signatures, i.e. lists of genes deemed informative, obtained from high-dimensional models typically are rather unstable. For example, if one gene is seen to have influence on the hazard for the event of interest, but seemingly not for the other events, it is unclear whether the latter is the case because there really is no effect or whether the effect was just overlooked due to unstability of the signature.

Fine and Gray (1999) suggested a methodological framework for a synthesis interpretation. They proposed a formal direct synthesis model, adapting the Cox proportional hazards model for the subdistribution hazard, which is the hazard attached to the cumulative incidence function. As the subdistribution hazard relates directly to the cumulative incidence function, which is not the case for the cause-specific hazards (Beyersmann \textit{et al.}, 2007), only one
model has to be fitted for describing the cumulative incidence function of the event of interest. Also, modeling of effects on the cumulative incidence function is distinctly more specific than employing a composite endpoint. For example, an increasing effect of a covariate on the cumulative incidence function of interest may be due to an increasing effect on the cause-specific hazard of interest. It may also be due to a decreasing effect on the competing cause-specific hazard. These effect constellations may lead to both increasing and decreasing effects on the composite hazard. Consequently, the effect on the cumulative incidence function (CIF) may be obscured in the analysis of the composite endpoint.

As the number of observations for microarray data is usually small, only a small number of influential genes can reliably be identified. Therefore, an approach is wanted that fits sparse models, i.e. where most of the estimated parameters are equal to zero. The few parameter estimates that are not equal to zero then provide for a short list of important genes. In the following, a boosting approach, that has been developed for standard time-to-event settings (Binder and Schumacher, 2008b), will be adapted for fitting proportional subdistribution hazards models to high-dimensional data. We chose that specific approach because it results in sparse model fits and can incorporate clinical covariates into microarray-based models in a way that allows for convenient comparison to purely clinical models. It builds only on the weighted likelihood, proposed by Fine and Gray (1999), and does not rely on variance estimates which might be more problematic, especially in a high-dimensional setting. The aim of such an analysis is, e.g. to identify a gene expression signature that is of relevance for an individual to experience a cancer event.

A further aim of our analysis is to investigate whether there is predictive information in the data or whether microarray features can improve prediction performance over a purely clinical model. Therefore, an appropriate measure of prediction performance is needed. Prediction error curves, i.e. the mean squared difference attached to the likelihood (see Gui and Li, 2005; Park and Hastie, 2007, for example). Boosting approaches, on the other hand, attempt minimization of a loss function in the course of a large number of boosting steps by gradient descent in function space (Friedman, 2001). With microarray data, often sparse models are needed. Prediction error curves, i.e. the mean squared difference

\[ F_1(t) = P(T^* \leq t, \epsilon = 1), \]

i.e. the expected proportion of patients suffering event 1 over the course of time. For investigating the influence of covariates on this quantity, Fine and Gray (1999) suggest to build models for the subdistribution hazard

\[ h_1(t) = \frac{dF_1(t)}{dt} \]

which coincides with the standard definition of a hazard used in time-to-event analysis, i.e. the instantaneous risk of having an event, in the absence of competing events. Specifically, they propose to fit a Cox proportional hazards model for the subdistribution hazard, i.e. employing a model

\[ h_1(t|x_i) = h_{1,0}(t) \exp(x_i'\beta), \]

where \( h_{1,0}(t) \) is an unspecified baseline hazard, and \( \beta = (\beta_1, \ldots, \beta_p)' \) is a parameter vector that can be estimated by maximizing a partial likelihood with modified risk set and inverse probability weights. However, for \( p > n \) direct maximization is no longer possible, requiring techniques developed specifically for high-dimensional settings.

### 2.2 Boosting for subdistribution hazards models

There are two major approaches for maximizing a likelihood in cases where the number of covariates is larger than the number of observations: in penalized estimation approaches, a penalty term, that enforces a restriction on the parameter vector, is attached to the likelihood (see Gui and Li, 2005; Park and Hastie, 2007, for example). Boosting approaches, on the other hand, attempt minimization of a loss function in the course of a large number of boosting steps by gradient descent in function space (Friedman, 2001). With microarray data, often sparse models are wanted, i.e. only few elements of the estimated parameter vector should have non-zero values, resulting in parsimonious signatures. To achieve this, either penalized approaches with a Lasso-like penalty (Tibshirani, 1996) or componentwise boosting approaches (Bühlmann and Yu, 2003) can be employed. The latter update only one element of the parameter vector in each boosting step, resulting in model fits similar to those obtained from Lasso-like approaches.

Besides wanting sparse model fits, another objective is direct comparability to purely clinical models. The latter is only possible, if clinical covariates can be incorporated into a predictive microarray model in an unpenalized way. Binder and Schumacher (2008b) propose a componentwise, likelihood-based boosting approach for the Cox proportional hazards model that allows for this, while still
resulting in sparse model fits. We will therefore adapt this approach for fitting proportional subdistribution hazards models.

The proposed approach is based on two main ideas: first, there are $M$ boosting steps, where in each step some elements of the estimated parameter vector are updated. The updates are determined by penalized maximum partial likelihood estimation, where the previous boosting steps are incorporated as an offset. Second, there is a distinction between a set of indices of mandatory covariates $\mathcal{J}_{\text{mand}} \subset \{1, \ldots, p\}$ and the set of indices of optional covariates $\mathcal{J}_{\text{opt}} = \{j \in \{1, \ldots, p\}, j \notin \mathcal{J}_{\text{mand}}\}$. Typically, clinical covariates will be mandatory, and covariates such as microarray features will be optional. In each boosting step, only one element of the estimated parameter vector, corresponding to one optional covariate, is updated. The elements corresponding to the mandatory covariates are updated simultaneously before each boosting step.

The details of the algorithm are as follows:

1. Initialize the offset $\hat{\eta}_{0,i} = 0, i = 1, \ldots, n$ and the estimated parameter vector $\hat{\beta}_0 = (0, \ldots, 0)'$.
2. For each boosting step $k = 1, \ldots, M$,
   a. Update the elements $j \in \mathcal{J}_{\text{mand}}$ of $\hat{\beta}_{k-1}$ by one maximum partial likelihood Newton–Raphson step and update the offset via $\hat{\eta}_{k-1,i} = \hat{\gamma}_{k-1,i} + \hat{x}_i \hat{\beta}_{k-1}$.
   b. Estimate the parameters $\hat{\gamma}_{k,j}$ in candidate models $h_1(t|x_i) = h_1(t|x_i|\hat{\gamma}_{k-1,i} + \hat{\gamma}_{k,j}x_{ij}), j \in \mathcal{J}_{\text{opt}}$.
   c. Determine the best candidate model $j^*$ with parameter estimate $\hat{\gamma}_{k,j^*}$ and perform the update $\hat{\beta}_{k,j} = \left\{ \begin{array}{ll} \hat{\beta}_{k-1,j} + \hat{\gamma}_{k,j^*} & \text{if } j = j^* \\ \hat{\beta}_{k-1,j} & \text{otherwise} \end{array} \right.$
   d. Update the offset via $\hat{\eta}_{k,i} = \hat{\gamma}_{k,j} + \hat{x}_i \hat{\beta}_k$.

To avoid overfitting, the number of boosting steps should not be too large. It can, for example, be determined by cross-validation with respect to the partial log-likelihood (Verweij and van Houwelingen, 1993).

To avoid boosting steps that are too large, the parameters $\hat{\gamma}_{k,j}$ are determined by penalized estimation. The partial-log-likelihood provided by Fine and Gray (1999) is augmented by a penalty term, resulting in

$$l_{\text{pen}}(\hat{\gamma}_{k,j}) = \sum_{i=1}^n \left( \delta_i \frac{\exp(h_{1,i|-\hat{\beta}_0})}{\sum_{j'=1}^p \exp(h_{1,i|-\hat{\beta}_0})} \right)$$

$$- \log \left( \frac{w_i(t_i) \exp(h_{1,i|-\hat{\beta}_0})}{\sum_{j'=1}^p w_i(t_i) \exp(h_{1,i|-\hat{\beta}_0})} \right) + \frac{\lambda}{2} \hat{\gamma}_{k,j}^2,$$

where $R_i = \{ t_j \geq t_i \text{ or } h_i t_j > 1 \}$ is the risk set that arises naturally when the competing risks process for an individual is stopped just before the time of a competing event (Beyersmann and Schumacher, 2008), $w_i(t) = \tilde{G}(t)/\tilde{G}(t_i) | h_i < \tilde{G}(t)$ are weights to account for censoring (with $\tilde{G}(t)$ being a Kaplan–Meier estimate of $P(C > t)$) and $\lambda$ is a penalty parameter that determines the size of the boosting steps. The latter is of minor importance and is typically chosen such that the number of boosting steps, selected e.g. by cross-validation, is larger than 50, as this number limits the maximal number of non-zero coefficients of the fitted model.

The estimates are determined by one Newton–Raphson step, i.e.

$$\hat{\gamma}_{k,j} = \frac{\lambda}{2} \gamma_{k-1,j}, \quad \hat{\eta}_{k,i} = 0,$$

where $U_{\text{pen}}(\gamma) = \frac{\partial l_{\text{pen}}(\gamma)}{\partial \gamma}$ is the score function and $I_{\text{pen}}(\gamma) = \frac{\partial^2 l_{\text{pen}}(\gamma)}{\partial \gamma^2}$ is the information matrix. Correspondingly, the best candidate model is taken to be the one that maximizes the score statistic $U_{\text{pen}}(\hat{\gamma}_{k-1}) l_{\text{pen}}(0)/U_{\text{pen}}(0)$, which performs evaluation at parameter value $\gamma = 0$, still incorporating the offset $\hat{\eta}_{k-1,i}$.

## 3 METHODS

### 3.1 Bootstrap .632+ prediction error curve estimates

Having estimated the parameter vector $\hat{\beta}$, a prediction for the cumulative incidence function is obtained by

$$\hat{\pi}(t|\hat{x}) = 1 - \exp(-\hat{H}_{1,0}(t) \exp(x' \hat{\beta})).$$

where $\hat{H}_{1,0}(t)$ is the Breslow estimator of the cumulative baseline subdistribution hazard $H_{1,0}(t) = \int_0^t h_{1,0}(s)ds$.

The true prediction error curve then is

$$Err(t; \hat{\pi}) = E[(N(t) - \hat{\pi}(t|x))^2],$$

i.e. the Brier score tracked over time, where $N(t)$ is the true status of an individual at time $t$ (Gerrits and Schumacher, 2006; Graf et al., 1999). In a competing risks setting, the true status is defined as $N(t) = I(T^* < t \text{ and } \delta = 1)$, corresponding to modified event times for stopped processes, where the event time for an individual with competing event is set to infinity. This is the natural definition for judging prediction of the cumulative incidence function, as the average true status at each time will correspond to the latter.

When the prediction error curve is to be estimated from data, weights have to be introduced to account for censoring. For example, the apparent error, evaluated in the training data, is

$$\overline{err}(t; \hat{\pi}) = \frac{1}{n} \sum_{i=1}^n (N(t) - \hat{\pi}(t|x_i))^2 W_i(t; \hat{G}),$$

with inverse probability of censoring weights $W_i(t; \hat{G}) = \frac{W_i(t; \hat{G})}{\frac{W_i(t; \hat{G})}{\frac{W_i(t; \hat{G})}{W_i(t; \hat{G})}}}$.

If $\hat{G}(t|x_i)$ is a consistent estimate of $P(C > t|x_i)$, the empirical prediction error curve will also be a consistent estimate (Gerds and Schumacher, 2006; Schoop, 2008).

Another useful quantity for judging competing risks data is the conditional cumulative incidence function or conditional probability function

$$CP_1(t) = \frac{F_1(t)}{1 - F_{\text{censor}}(t)},$$

where $F_{\text{censor}}(t)$ is the cumulative incidence function for all events $\epsilon > 1$ (Pepe and Mori, 1993). It is the proportion of the events of interest after removing observations that have had a competing event up to time $t$. As this quantity has a larger range compared with the cumulative incidence function, there is also more room for improvement over a null model, in terms of prediction performance, given that the variability of the predictions for $CP_1(t)$ is not (much) larger than that of predictions for $F_1(t)$.

Predictions for the conditional probability function are obtained by

$$\hat{\pi}_{\text{cond}}(t|x) = \frac{\hat{\pi}(t|x)}{1 - \hat{F}_{\text{censor}}(t)}.$$
estimate for \( F_{\text{obs}}(t) \), such as the Aalen–Johansen estimator, is sufficient. Otherwise, approaches for direct modeling of the conditional probability function would be needed, which are not readily available.

For calculation of the empirical prediction error curves for predictions of the conditional probability function, the weights \( W_i(t; \hat{G}) \) have to be modified

\[ W_{\text{cond}}(t; \hat{G}) = I(t_j > t \text{ or } b_i \epsilon = 1)W_i(t; \hat{G}), \]

i.e. observations are removed from the calculation at the time a competing event occurs. The number by which the sum of weighted squared differences is divided then also has to be decreased accordingly.

The apparent error (1) will typically underestimate the true prediction error, as the same data is used for fitting a model and evaluating its performance. We therefore draw bootstrap samples \( b = 1, \ldots, B \), each containing observations with indices \( \pi_b \), fit a model in each of these samples, including all model building steps, such as selection of the number of boosting steps, resulting in \( \hat{R} \) predictors for the cumulative incidence function \( \hat{R}_b(t)(x) \). The bootstrap cross-validation estimate of the prediction error curve then is

\[ \overline{Err}_{bg0}(t; \hat{R}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i \in I} (N_i(t) - \hat{R}_b(t(xi)))^2 W_i(t; \hat{G}), \]

where \( b_o \) is the number of original observations not contained in bootstrap sample \( b \). To avoid a complexity selection bias, bootstrap samples of size 0.632n are drawn without replacement (Binder and Schumacher, 2008a), therefore \( b_o = (1 - 0.632)n \).

Bootstrap cross-validation tends to overestimate prediction error. Therefore, Efron and Tibshirani (1997) proposed the bootstrap .632+ estimate, which is a linear combination of the apparent error and the bootstrap cross-validation estimate. Gerds and Schumacher (2007) adapted this idea for prediction error curve estimation, employing the bootstrap .632+ prediction error curve estimate

\[ \overline{Err}_{bg0}(t; \hat{R}) = (1 - \hat{\omega}(t))\overline{Err}_{bg0}(t; \hat{R}) + \hat{\omega}(t)\overline{Err}_{bg0}(t; \hat{R}), \]

where the weights \( \hat{\omega}(t) \) adapt to the potential of the evaluated fitting procedure for overfitting and are given by \( \hat{\omega}(t) = 0.632 / (1 - 0.369\overline{R}(t)) \), where \( R(t) \) is the relative overfitting rate

\[ \overline{R}(t) = \frac{\overline{Err}_{bg0}(t; \hat{R}) - \overline{Err}_{bg0}(t; \hat{R})}{\text{NoInf}(t; \hat{R}) - \overline{Err}_{bg0}(t; \hat{R})} \]

with the no-information error

\[ \text{NoInf}(t; \hat{R}) = \frac{1}{n} \sum_{i=1}^{n} (N_i(t) - \hat{R}_b(t(xi)))^2 W_i(t; \hat{G}), \]

representing a worst case scenario with no association between covariates and response. For further details see Gerds and Schumacher (2007), and for illustration in a high-dimensional setting see Schumacher et al. (2007). Note that for evaluation of predictions of the conditional probability function the denominators \( n, b_o \), and \( n^2 \) in (1), (2) and (3), respectively, have to be decreased at times where observations with competing events are removed from evaluation.

### 3.2 Simulation study

Combined interpretation of fitted cause-specific hazards models is expected to be complicated in high-dimensional settings, as only a small number of covariates will receive non-zero parameter estimates, i.e. if a true coefficient has a small non-zero value, the estimate may nevertheless be equal to zero. This makes a summary proportional subdistribution hazards model even more desirable. In the following, we therefore investigate whether the latter can indeed deliver a reasonable summary analysis in a high-dimensional setting and whether it also provides good prediction performance.

If the proportionality assumption does hold for the cause-specific hazards, it will typically not hold for the subdistribution hazard, i.e. proportional hazards models for the latter will be misspecified (Latouche et al., 2007). We base the simulation study on this setting, for investigating whether even misspecified high-dimensional subdistribution hazards models provide a reasonable summary analysis. If this would be found to be the case, good performance can also be expected when the proportionality assumption for the subdistribution hazards holds.

#### 3.2.1 Design

We simulate competing risks data with two possible events, the event of type 1 being the one of interest. Mimicking microarray data structure, \( p = 5000 \) correlated covariates are generated for \( n = 400 \) observations, following the design employed by Binder and Schumacher (2008a).

We consider 16 informative covariates with an effect on the cause-specific hazards for events of type 1 and/or 2. This corresponds to a sparse true model, i.e. this setting favors approaches that fit sparse models. In real data many more genes might have an effect, i.e. the true model might not be sparse. The performance judgement of sparse model fits in the present simulation setting might therefore be somewhat optimistic. However, the main objective is the comparison of proportional subdistribution hazards models to models fitted for the cause-specific hazards. As both are fitted to be sparse, the comparison should still be valid. Each informative covariate is located in one of three blocks of correlated covariates, where the correlation in each block is 0.5, 0.35 and 0.05, respectively. In the first block, four covariates have an increasing effect on both hazards. The second block has four informative covariates with an increasing effect on the cause-specific hazard for type 1 events, and a decreasing effect on the competing cause-specific hazard. In the third block, there are four covariates that have a decreasing effect on the event type 1 hazard only, and four other covariates that have an increasing effect on the event type 2 hazard. For an increasing effect, the true coefficient \( \alpha_{ej}, e \in \{1, 2\} \) (for the event types), takes value 0.5, for a decreasing effect value \(-0.5\). The remaining covariates have no direct effect on the hazards, i.e. \( \alpha_{ej} = 0 \).

Event times are generated cause-specific hazard driven, where both cause-specific hazards follow a Cox-exponential model (Binder et al., 2005), with baseline hazards equal to 0.1. As the baseline hazard for the competing event has the same size as the baseline hazard for the event of interest, effects on the hazard for the competing event are likely to be reflected in the cumulative incidence function for the event of interest.

Censoring times are taken from a uniform distribution with values between 0 and 9, resulting in censoring for \( \approx 36\% \) of the observations. Events of type 1 are observed for \( \approx 38\% \) of the observations, and events of type 2 for 26%. The 100 datasets and corresponding test sets of size \( n_{\text{new}} = 1000 \), for determining the prediction performance, are generated.

#### 3.2.2 Results

As the absolute value of the estimated parameters will be highly biased, we investigate only whether an estimate is non-zero.

Table 1 illustrates the relation of the estimated parameters from the cause-specific hazards models (fitted by the boosting approach described in Binder and Schumacher, 2008b) and the proportional subdistribution hazards model, fitted by the proposed approach, the number of boosting steps being determined by 10-fold cross-validation.

The proportion of covariates that have no effect but nevertheless receive non-zero parameter estimates is similar for the subdistribution hazards model and the cause-specific models. While the covariates wrongly deemed informative by the subdistribution hazards model mostly are the same as those from the cause-specific model for event type 1, there is a disagreement with the model for event type 2. The subdistribution model identifies at least some covariates that have an effect on the event type 1 hazard only (\( \approx 40\% \)), none of the event type 2 only covariates are recovered. The performance with respect to the former seems to be better than that of the cause-specific model, at least in this setting. With respect to the missed covariates with effects on the type 2 event hazard only, one has to consider that the proportional cause-specific hazards model has also a low sensitivity. For the covariates that have an effect on both hazards, all models perform very well, with a slight advantage for the subdistribution hazards model. The reason for this
As these are only available for the proposed techniques are illustrated with publicly available preprocessed 1381 custom platform microarray features (GEO with series accession no. GSE5479) for fitting and validating a signature for predicting progression, Dyrskjøt et al. (2007) employed samples from patients with pT1 tumors with no previous or synchronous muscle-invasive tumors. 

A cross tabulated proportions of covariates included into the fitted models (i.e. with non-zero parameter estimates) for the proportional subdistribution hazard model (SH; with estimates \( \hat{\beta}_j \)) versus each of the cause-specific hazards models (CSH1 and CSH2; with parameter estimates \( \hat{\alpha}_{1,j} \) and \( \hat{\alpha}_{2,j} \), for the two cause-specific proportional hazards model), with empirical 95% CIs.

<table>
<thead>
<tr>
<th>True effect</th>
<th>SH</th>
<th>CSH1</th>
<th>CSH2</th>
<th>SH overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\alpha}_{1,j} = 0 )</td>
<td>( \hat{\alpha}_{1,j} \neq 0 )</td>
<td>( \hat{\alpha}_{2,j} = 0 )</td>
<td>( \hat{\alpha}_{2,j} \neq 0 )</td>
</tr>
<tr>
<td>No effect</td>
<td>( \hat{\beta}_j = 0 )</td>
<td>99.88 [99.76; 99.98]</td>
<td>0.01 [0; 0.04]</td>
<td>99.77 [99.61; 99.91]</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_j \neq 0 )</td>
<td>0.02 [0; 0.10]</td>
<td>0.10 [0.01; 0.22]</td>
<td>0.12 [0.02; 0.24]</td>
</tr>
<tr>
<td>Effect on CSH1 only</td>
<td>( \hat{\beta}_j = 0 )</td>
<td>95.25 [50.00; 100]</td>
<td>0.50 [0; 1]</td>
<td>95.75 [50.00; 100]</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_j \neq 0 )</td>
<td>2.75 [0; 38.13]</td>
<td>1.50 [0; 25.00]</td>
<td>4.25 [0; 50.00]</td>
</tr>
<tr>
<td>Effect on CSH2 only</td>
<td>( \hat{\beta}_j = 0 )</td>
<td>100 [100; 100]</td>
<td>0 [0; 0]</td>
<td>96.50 [61.88; 100]</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_j \neq 0 )</td>
<td>0 [0; 0]</td>
<td>0 [0; 0]</td>
<td>0 [0; 0]</td>
</tr>
<tr>
<td>Effect on CSH1 and CSH2</td>
<td>( \hat{\beta}_j = 0 )</td>
<td>18.63 [37.50]</td>
<td>1.63 [0; 19.06]</td>
<td>9.13 [0; 37.50]</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_j \neq 0 )</td>
<td>4.37 [25.00]</td>
<td>75.37 [50.00; 100]</td>
<td>33.12 [0; 69.06]</td>
</tr>
</tbody>
</table>

Finally, we investigate the prediction performance of the fitted proportional subdistribution hazards models, by inspecting the difference between the integrated prediction error curve for this model on new data and the prediction error curve for a null model, i.e. the Aalen–Johansen estimator. We integrate this difference up to the 90% quantile of the observed event times for the considered event type. For events of type 1, with a 90% quantile of 2.34, the mean integrated difference is 0.12 (95% CI: [0.08; 0.18]), in favor of the proportional subdistribution hazards model. It is seen that the fitted models have good prediction performance. Therefore, in addition to providing a reasonable summary analysis, we also expect good prediction performance on real data, even if the proportional hazards assumption holds for the cause-specific hazards and not for the subdistribution hazard.

4 APPLICATION EXAMPLE

The proposed techniques are illustrated with publicly available data from patients with bladder cancer (Dyrskjøt et al., 2007). For validating a signature for predicting progression, Dyrskjøt et al. (2007) employed samples from patients with pT\(_{\alpha}\) and pT\(_{1}\) tumors with no previous or synchronous muscle-invasive tumors.

In the following, we are going to employ the corresponding publicly available preprocessed 1381 custom platform microarray features (GEO with series accession no. GSE5479) for fitting and evaluating a new predictive model for progression. Potentially important clinical covariates are age, sex, stage (pT\(_{\alpha}\) versus pT\(_{1}\)), grade (PUNLMP/low versus high) and treatment (for adjustment).

As these are only available for \( n = 301 \) observations, analyses will be restricted to this subset.

The response of interest is the time to progression. For 84 patients progression or death from bladder cancer was observed, which is the event of interest. There is censoring for 194 patients. Therefore, there is also a competing event, death from other or unknown causes, that is observed for 33 patients. Therefore, not only model fitting techniques that can deal with censoring are needed, but techniques that can deal with competing events.

We fit a proportional subdistribution hazards model to the data by the proposed boosting technique, where the clinical covariates are included as mandatory and the microarray features are optional. For comparison, employing the boosting approach described in Binder and Schumacher (2008b), we fit separate, cause-specific Cox proportional hazards models for the events ‘progression or death from bladder cancer’ and ‘death from other or unknown cause’, where the respective other events are treated as censored.

The cause-specific model for the event ‘progression or death from bladder cancer’ comes closest to the analyses performed by Dyrskjøt et al. (2007), who consider only one cause-specific hazard. This does not allow for a straightforward probability interpretation, which is in contrast to the proportional subdistribution hazards approach. For the cause-specific model, 10-fold cross-validation selects 84 boosting steps, resulting in eight microarray features with non-zero estimates. This is much smaller compared with the 88 gene progression classifier proposed by Dyrskjøt et al. (2007) and is due to the sparseness enforced by componentwise boosting. For example, if there is a group of highly correlated microarray features that is related to the response, typically only one or two features will receive non-zero estimates. Therefore, it does not come as a surprise, that there is only a small overlap, with only one common microarray feature (SEQ820).

There is no overlap with the five microarray features selected in the second cause-specific model (at 99 boosting steps, chosen by 10-fold cross validation). This, however, should not be interpreted as a sign that there are no common effects for the cause-specific hazards models, as the coefficients corresponding to weak effects may have gotten parameter estimates equal to zero, due to the attempt to fit sparse models. The only conclusion that can be drawn is, that there are no strong common effects. This illustrates the difficulties of interpreting the fit of two cause-specific models together in a high-dimensional setting.

In contrast, the fitted proportional subdistribution hazards model provides a simpler summary analysis. At 117 boosting steps, selected by 10-fold cross-validation, 12 microarray features received non-zero estimates. There is an overlap of five microarray features (SEQ34, SEQ820, SEQ1384, SEQ162, SEQ164) with the cause-specific model for progression, indicating that the model for the subdistribution hazard incorporates the main effects of the latter. There is also an overlap of one microarray feature (SEQ250) with the other cause-specific model, further illustrating that the proportional subdistribution hazards model (even if potentially misspecified) can offer a summary analysis. This microarray feature has a positive effect on CSH1 and CSH2.
Therefore, given that the latter already is the quantity of interest, while already the results of Dyrskjøt et al. (2007) indicated that the null model (grey curves), resulting in an overlap of two genes, which is more than what would be expected by chance.

For judging whether microarray features can improve prediction performance over a purely clinical model, bootstrap .632+ prediction error curve estimates were determined based on $B = 100$ bootstrap samples drawn without replacement. Figure 1 shows the estimates for both prediction of the cumulative incidence function (left panel) and for prediction of the conditional probability function (right panel). The performance of the Aalen–Johansen estimator, that does not employ any covariate information and therefore serves as a null model, is given as a reference. Both the prediction performance of a purely clinical model (dashed curves), fitted by standard maximum likelihood, and of the boosting estimates of models that in addition include microarray features (solid black curves) is seen to be better than that of the null model (grey curves), indicating that valuable information is contained in the data. Also, the combined model is seen to clearly improve over the purely clinical model.

While already the results of Dyrskjøt et al. (2007) indicated that the predictive information in the microarray data might be independent of the clinical covariates, fitting a proportional subdistribution hazards model together with prediction error curve estimates allows for a better quantification of the additional benefit.

The prediction error curve estimates for the conditional probability function are seen to be further apart, compared with the prediction error curves for the cumulative incidence function. Therefore, given that the latter already is the quantity of interest, it seems to have the additional advantage of allowing for better performance comparison of different models.

5 CONCLUSION

While several techniques are available for fitting high-dimensional models to time-to-event data when there is a single type of event, the competing risks setting is more challenging. If interest is focussed on one of the competing events, the corresponding cumulative incidence function and subdistribution hazard are a natural starting point for modeling.

We adapted a boosting approach, that was developed for fitting Cox proportional hazards models, for fitting proportional subdistribution hazards models. In this approach, important clinical predictors can be included as mandatory covariates, allowing for comparison of models that combine clinical and microarray information to purely clinical models.

For judging whether there is predictive information in the covariates, and for judging whether microarray features can improve over purely clinical models, an adequate measure of prediction performance is needed. We illustrated how prediction error curve techniques can be adapted for the competing risks setting. In particular, bootstrap .632+ estimates allowed for evaluation of prediction performance without having to set aside data for a test set. Prediction error curve estimates were given for evaluating prediction of the cumulative incidence function as well as for evaluation of conditional probability function predictions, where the latter, given that it is the quantity of interest, might have the additional benefit of allowing for better discrimination of model performance.

We investigated the performance of the proposed boosting approach in a simulation study, specifically addressing scenarios where a proportional subdistribution hazards model might be misspecified. The fitted models were nevertheless found to identify the important covariates, providing a summary analysis of the cause-specific hazards and to result in good prediction performance.

The proposed subdistribution approach was illustrated in an application example with microarray time-to-event data and competing risks. Prediction error curve techniques allowed to quantify the benefit in terms of prediction performance that can be gained from employing microarray features in addition to clinical covariates. The fitted models for the subdistribution hazard was seen to identify genes that are relevant for an individual to experience a cancer event.

The application illustrated that the subdistribution analysis is more specific than the usual analysis of a combined endpoint. The analysis is also adequate for prediction purposes. Complementing this analysis, the aim of a further investigation would be to study...
how a relevant impact of microarray features is mediated through the cause-specific hazards. As noted earlier, we should not conclude that a certain feature has no effect on the other cause-specific hazard, simply because it had not been included in the list of relevant features. Therefore, the cause-specific hazards analyses would need to be coupled. The subdistribution analysis could serve as a starting point for such an endeavor. For example, relevant features identified in the subdistribution analysis could be included as mandatory covariates in models for the cause-specific hazards.

In a preliminary analysis of this type, we found that a microarray feature that is beneficial in terms of the cause-specific cancer hazard may be beneficial or harmful in terms of the competing cause-specific hazard. We found the same picture for microarray features that are harmful in terms of the cause-specific cancer hazard. These effect constellations are challenging in the absence of CIs. They also highlight the need for research work on coupling transition hazards analyses, potentially also in models that are more complex than competing risks models.

In the presence of these difficulties, and given that the subdistribution approach worked well for simulated data and in the application example, the proposed techniques may readily be applied whenever a more specific competing risks analysis is desired in a setting with high-dimensional covariates.

ACKNOWLEDGEMENT

We gratefully acknowledge support from Deutsche Forschungsgemeinschaft (DFG Forschergruppe FOR 534).

Conflict of Interest: none declared.

REFERENCES