Biostatistics (2016), **17**, 4, *pp.* 708–721
doi:10.1093/biostatistics/kxw017
Advance Access publication on April 26, 2016

Penalized estimation for competing risks regression with applications to high-dimensional covariates

FEDERICO AMBROGI*

Department of Clinical Sciences and Community Health, University of Milan, Via Vanzetti 5, 20133 Milano, Italy

federico.ambrogi@unimi.it

THOMAS H. SCHEIKE

Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5 entr. B, PO Box 2099, DK-1014 Copenhagen, Denmark

SUMMARY

High-dimensional regression has become an increasingly important topic for many research fields. For example, biomedical research generates an increasing amount of data to characterize patients' bio-profiles (e.g. from a genomic high-throughput assay). The increasing complexity in the characterization of patients' bio-profiles is added to the complexity related to the prolonged follow-up of patients with the registration of the occurrence of possible adverse events. This information may offer useful insight into disease dynamics and in identifying subset of patients with worse prognosis and better response to the therapy. Although in the last years the number of contributions for coping with high and ultra-high-dimensional data in standard survival analysis have increased (Witten and Tibshirani, 2010. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research* **19**(1), 29–51), the research regarding competing risks is less developed (Binder and others, 2009. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* **25**(7), 890–896). The aim of this work is to consider how to do penalized regression in the presence of competing events. The direct binomial regression model of Scheike *and others* (2008. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**(1), 205–220) is reformulated in a penalized framework to possibly fit a sparse regression model. The developed approach is easily implementable using existing high-performance software to do penalized regression. Results from simulation studies are presented together with an application to genomic data when the endpoint is progression-free survival. An R function is provided to perform regularized competing risks regression according to the binomial model in the package `timereg` (Scheike and Martinussen, 2006. *Dynamic Regression models for survival data*. New York: Springer), available through CRAN.

1. INTRODUCTION

The impact of genomics on biomedical research has been enormous over the past 15 years producing an impressive amount of biomedical data. Microarray gene expression and recently NGS technologies opened

*To whom correspondence should be addressed.

unprecedented possibilities for the discovery and understanding of molecular mechanisms, biomarker discovery, and personalized medicine. The data produced by such platforms are characterized by a number of covariates, the transcripts, that greatly exceeds the number of samples under study (high-throughput assay), $p \gg n$. In clinical studies, researchers consider high-dimensional genomic data together with a possibly censored survival time for each patient in study. One important research question is to find the transcripts that most associate with the survival outcome. The popular BRB-ArrayTools (Simon *and others*, 2007) provides a survival analysis tool able to handle correctly the information from censored subjects applying the Cox proportional hazard (PH) model on a gene per gene basis (using Wald or Score tests). The ordered gene list using the p-values can then be used and interpreted after correction and controlling the proportion of false discoveries. Such a strategy can be replaced by fitting a multiple Cox regression in high-dimensional data settings.

The definition of the primary endpoint of the study must be carefully considered according to the study goals: progression-free survival, cause-specific crude cumulative incidence, or transition probabilities are endpoints receiving more and more attention in the methodological and applied/clinical settings. When dealing with competing risks, the simple use of the Cox PH model makes it possible only to estimate the cause-specific hazard ratio, a measure not directly linked to the cumulative incidence. While the cause-specific hazard is useful for investigating the disease dynamics to get insights in disease mechanisms and biological processes, it is less appropriate for clinical decision support for which it is preferable to consider the cumulative incidence probability, the marginal probability of failure for a specific cause.

In this case, the estimation of measures of association directly connected with the cumulative incidences is preferable to evaluate the specific risk of interest. As pointed out by a Referee, a proper analysis would require modeling the competing event(s) as well. In fact, as has been noted by many authors, the relation between factors and the event of interest can only be understood through the relation between these factors and the competing events, i.e. competing risks analyses must be interpreted in the presence of competing causes of failure; see Latouche *and others* (2013) for a complete discussion.

The aim of this work is to propose a direct estimation method for cumulative incidences in the presence of many covariates through penalization.

Alternatively to the penalization directly on the cumulative incidence function (CIF), which is the aim of our contribution, some works have suggested to do penalization for each of the cause-specific hazards, thus basing it, for example, on Cox models for each cause and then penalizing each cause-specific hazard. Binder *and others* (2009) discussed the difficulties of this approach, namely the need to combine all the cause-specific hazard regressions to get information on the effect on the cause-specific cumulative incidence of the different features. In particular, selection of covariates for cause-specific hazards is the not same as selecting covariates for the cumulative incidence. Assuming that cause 1 is the primary cause of interest, the CIF for cause 1, that is the probability of experiencing cause 1 before time t, given covariate sets X and Z, is given by

$$P_1(t; X, Z) = P(T \leqslant t, \epsilon = 1 \mid X, Z),$$

where ϵ indicates the type of failure cause. The first covariate set X may include covariates that have to be considered in the regression model, such as clinical important covariates (Age, disease stage, etc.); the second set, Z, may include the covariates for which selection is requested, i.e., the high-dimensional set of covariates. Several methods have been developed to directly model the cumulative incidence probability of a specific cause of failure (Fine and Gray, 1999; Fine, 2001; Scheike *and others*, 2008; Scheike and Zhang, 2008). The key issue here is how to deal with incomplete data due to the fact that observations are observed subject to right-censoring; that is some subjects are not observed dying from either of the causes, because of a limited observation period, and thus only seen to survive T units. We

consider models of the form

$$g\{P_1(t; X, Z)\} = \eta(t) + X^T\gamma_x + Z^T\gamma_z \tag{1.1}$$

extended to competing risks by Fine (2001). To estimate the parameters of this model, we will use inverse probability of censoring weighting (IPCW) techniques or pseudo-value techniques as in Klein and Andersen (2005) and Scheike *and others* (2008).

We here consider our preferred $logit$ link function. For more on this, see Ambrogi *and others* (2008), Zhang and Fine (2008), Gerds *and others* (2012), and Eriksson *and others* (2015). We find that the regression coefficients are more easily understood on the probability scale, which is indeed one of the aims of a competing risks regression method. The aim of this work is to develop a method for penalized regression when considering the CIF, $P_1(t; X, Z)$, and Z is potentially a high-dimensional covariate. The actual implementation relies on state of the art software which will keep being developed and extended according to advances in penalized regression. Previous works have considered the use of boosting in connection with Fine & Gray's model leading to a partial subdistribution criterion (Binder *and others*, 2009). There is no extensions and practical implementations, at present, for the proportional subdistribution hazard models in penalized settings. In particular, the choice of Binder *and others* (2009) was explicitly to develop boosting instead of penalization, probably because of the possibility to use the highly efficient available software by using a boosting algorithm, instead of directly modifying the score equations.

A key point of our suggestion for covariate selection for the competing risks model is that an IPCW weighted approach can rely on standard software (for the unweighted case) provided a `weights` option is available. We here show how it can be done using IPCW for a binomial regression problem. The advantages of the binomial regression method is that it easily generalizes to other settings, such as transition probabilities in multistate models (see, for example, Scheike and Zhang, 2007), and the models may be extended in terms of additional flexibility. For example, the choice of the link function is completely open. Also the penalized fitting for these standard models is highly developed and, as a consequence, it is possible to do ridge, lasso (with all its variants) and elastic net penalization. The same could be done for the approach of Fine & Gray that can be programmed using standard software for the Cox regression model (Geskus, 2011). In this case, the software should handle left truncation which is not readily available with functions for censored penalized regression. Although the "standard" regression approach, in many practical applications, is the subdistribution hazard model by Fine and Gray, we feel that it is often important to use a link function that more readily gives a simple interpretation of the covariate effects. Whether one model is to be preferred or not can be investigated by some of the available goodness of fit tools or computing prediction error estimates, see Gerds and Schumacher (2007).

The proposed approach is described in Section 2, while Section 3 reports some simulation results. A practical application is also presented using publicly available data. The description of the `R` code for the analysis is reported in the supplementary material together with a mathematical appendix (available at *Biostatistics* online).

2. METHODS

2.1 *Binomial regression modeling*

Let T and C be the event time and right-censoring time of the ith individual and let $\epsilon_i \in \{1, 2\}$ denote the failure type. The basic assumption is that, given covariates Z and X, we have that $P_1(t; X, Z)$ is on the form (1.1) for some link function $g(\bullet)$, where $\eta(t)$ is a non-parametric baseline, and γ_z and γ_x are regression effects related to Z and X, respectively. The entire set of regression coefficients is denoted by $\gamma = (\gamma_x, \gamma_z)$. In addition to these basic assumptions, we also have right-censoring present for our competing risks data,

such that we observe $\tilde{T} = \min(T, C)$, $\Delta = I(T \leqslant C)$ and $\tilde{\epsilon} = \epsilon \cdot \Delta$. We define also $\Delta_i(t) = I(\min(T_i, t) \leqslant C_i)$, the indicator of being uncensored at time t. We here assume that the censoring times may depend on X, such that the conditional distribution of C given X is given by $S_C(t; X) = P(C > t|X)$. The censoring distribution is generally unknown and can be estimated using a regression model, such as Cox or Aalen regression; see, for example, Sun and Zhang (2009). Define also $N(t) = I(\tilde{T} \leqslant t, \tilde{\epsilon} = 1)$.

To estimate the competing risks regression model, we first define the basic responses as $Y_i(t) = N_i(t)$ and organize them into vectors $n \times 1$ vectors $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))$ and the model mean $\mathbf{P_1}(t; X, Z)$ with elements $P_1(t; X_i, Z_i)$; further the regression design is written as XZ with rows (X_i, Z_i). We further need the derivatives of $\mathbf{P_1}(t; X, Z)$ with respect to the parameters, and let $\mathbf{D}_\eta(t)$ and $\mathbf{D}_\gamma(t)$ be matrices with the ith rows equal to $D_{\eta,i}(t) = \partial P_1(t; x_i, z_i)/\partial\eta(t)$ and $D_{\gamma,i}(t) = \partial P_1(t; x_i, z_i)/\partial\gamma$, respectively. We also define IPCW weight matrices $\mathbf{W}(t) = diag(W_i(t))$ with $W_i(t) = \Delta_i(t)/S_C(\min(t, \tilde{T}_i); X_i)$, and the similarly estimated quantity as $\hat{\mathbf{W}}(t)$ with $\hat{W}_i(t) = \Delta_i(t)/\hat{S}_C(\min(t, \tilde{T}_i); X_i)$. The regression function $\eta(t)$ and regression parameter γ can be estimated based on the following estimating equations:

$$U_\eta(t) = U_\eta(t, \eta, \gamma) = \mathbf{D}_\eta{}^{\mathsf{T}}(t)\hat{\mathbf{W}}(t)\{\mathbf{Y}(t) - \mathbf{P_1}(t; X, Z)\} = 0, \tag{2.1}$$

$$U_\gamma(\tau) = U_\gamma(\tau, \eta, \gamma) = \int_0^\tau \mathbf{D}_\gamma{}^{\mathsf{T}}(t)\hat{\mathbf{W}}(t)\{\mathbf{Y}(t) - \mathbf{P_1}(t; X, Z)\}\,\mathrm{d}t = 0, \tag{2.2}$$

where τ is the last time-point considered. Note that the estimates of $\eta(t)$ will be piecewise constant functions that change their value only after events of type 1, so we only need to consider the score equations for $\eta(t)$ in the jump times. Alternatively, we may consider the score equations for only one specific time-point, or any specific set, to get time-specific analysis. In the case of high-dimensional covariates later, we consider a specific set of time-points only to reduce the computations, and the baseline is thus reduced to a finite-dimensional parameter. We return to this point later. In this case, the integral in the above equation is replaced with a sum of the score contributions for the different time-points considered.

These equations are a modification of the direct binomial regression approach of Scheike *and others* (2008) and seem to have slightly improved finite sample properties. Estimators solving these estimating equations have properties that can be derived along the lines of the arguments of, for example, Sun and Zhang (2009). Basically, it follows that the solutions to $U_\eta(t)$ and $U_\gamma(\tau)$ will be consistent and asymptotically normal with variances that can be estimated. We have shown that the distributions of $\sqrt{n}\{\hat{\eta}(t) - \eta(t)\}$ and $\sqrt{n}(\hat{\gamma} - \gamma)$ are asymptotically equivalent to the following (approximate) i.i.d. decompositions:

$$\sqrt{n}\{\underline{I}_\eta(t)\}^{-1}\sum_i \mu_{i1}(t) \quad \text{and} \quad \sqrt{n}\{\underline{I}_\gamma\}^{-1}\sum_i \mu_{i2},$$

respectively, where $\underline{I}_\eta(t) = \mathbf{D}_\eta(t)^{\mathsf{T}}\mathbf{W}(t)\mathbf{D}_\eta(t)$, $\underline{I}_{\eta,\gamma}(t) = \mathbf{D}_\eta(t)^{\mathsf{T}}\mathbf{W}(t)\mathbf{D}_\gamma(t)$, $\underline{I}_\gamma = \int_0^\tau \mathbf{D}_\gamma(t)^{\mathsf{T}}W(t)H(t)\mathbf{D}_\gamma(t)\,\mathrm{d}t$ and $H(t) = I - \mathbf{D}_\eta(t)\{\mathcal{I}_\eta(t)\}^{-1}\mathbf{D}_\eta(t)^{\mathsf{T}}\mathbf{W}(t)$, and

$$\begin{aligned}
\mu_{1i}(t) &= (\eta_{1i}(t)W_i(t) + \psi_{1i}(t)) - \mathcal{I}_{\eta,\gamma}(t)\mu_{2i},\\
\eta_{1i}(t) &= D_{\eta,i}(t)(Y_i(t) - P_1(\gamma, \eta, Z_i, t)),\\
\mu_{2i} &= \int_0^\tau \eta_{2i}(s)W_i(s) + \psi_{2i}(s)\,\mathrm{d}s,\\
\eta_{2i}(t) &= (D_{\gamma,i}(t) - \mathcal{I}_{\gamma,\eta}(t)\{\underline{I}_\eta(t)\}^{-1}D_{\eta,i}(t))(Y_i(t) - P_1(\gamma, \eta, Z_i, t)).
\end{aligned}$$

The ψ terms are due to the uncertainty of the estimation of the censoring distributions, and follow from the derivations of, for example, Fine and Gray (1999) or Sun and Zhang (2009) and are given in the appendix in the supplementary material (available at *Biostatistics* online). We require that the functions

$\mathcal{I}_\eta(t)$, $\mathcal{I}_{\eta,\gamma}(t)$, and $\mathcal{I}_\gamma$ converge uniformly in probability, in addition to the standard assumptions of Sun and Zhang (2009).

These score equations are equivalent to those of a standard binomial regression model with IPCW weights on a stacked version of the data. The binomial regression model can thus be easily estimated using standard regression algorithms for generalized estimating equations with binomial error and cluster robust variance computation. Clusters are identified by the replications of each subject through the event times or through a grid of times of convenience (for example all jump times of the cause of interest).

When turning to the penalized estimation for covariate selection or prediction in the next section, we rely on these score equations. A couple of comments are therefore important. We first note that a profile score for all $\gamma = (\gamma_x, \gamma_z)$ parameters can be achieved after profiling out the baseline. When we consider the score equation (2.1) for only a fixed set of time-points $t_1, \ldots, t_m$, as we do later for computational simplicity, we can write the profiled score for γ as $U_p(\gamma) = U(\hat{\eta}_\gamma(t_1), \ldots, \hat{\eta}_\gamma(t_m), \gamma)$. As already said, in this case the score equation for γ becomes a sum over the time-points, but we keep the notation with the integral anyway to avoid too much notation. In this situation the total set of parameters for the full score are (γ, α) with $\alpha = (\eta(t_1), \ldots, \eta(t_m))$ representing the baseline parameters. We have this situation in mind when we start regularizing the score equations in the next section.

The score equation is asymptotically linear in these parameters and satisfies the condition that there exist an M and a non-singular matrix A such that $n^{-1/2} \sup_{|\gamma-\gamma_0| \leqslant M/\sqrt{n}} |U_p(\gamma_0) - U_p(\gamma) - n^{1/2} A (\gamma - \gamma_0)| = o_p(1)$, where γ_0 is the true parameter. This is exactly the conditions that we need for the penalized estimating equations in the next section, thus relying on the results of Johnson *and others* (2008). In the case of a known censoring distribution, A is given by $A_1 = E(D_{\gamma_0,i} U^{\otimes 2} W_i(t)) - E(D_{\gamma_0,i} U D_{\alpha_0,i} U W_i(t)) E((D_{\alpha_0,i} U)^{\otimes 2} W_i(t))^{-1} E(D_{\alpha_0,i} U D_{\gamma_0,i} U W_i(t))$. This follows after a Taylor expansion and taking limits in probability (that are evaluated at the true values of the parameters). When the censoring distribution is also estimated, A is given by $A_1 + \Psi$, where Ψ is given in the appendix in the supplementary material (available at *Biostatistics* online). Similarly, when γ_x is also profiled out, we can still establish this condition for the score solely in γ_z due to smoothness of the model. When profiling out both the baseline and γ_x, we denote the score as $U_{px}(\gamma_z)$ or $U_{px}(\gamma)$ in the next section.

2.2 *Penalized binomial regression modeling*

The extension of estimating equations (2.2) to penalized estimating equations follows Johnson *and others* (2008) who developed a general theory for penalized semiparametric estimating equations. The penalized estimating functions for the γ parameters can be written as

$$U_{px}(\gamma) - n\mathbf{q}_\lambda(|\gamma|)sgn(\gamma), \tag{2.3}$$

where $\mathbf{q}_\lambda(|\gamma|) = (q_{\lambda,1}(|\gamma_1|), q_{\lambda,p}(|\gamma_p|))$, and $q_{\lambda,j}(\cdot)$ for $j = 1, \ldots, p$ are functions depending on the coefficients. The interest here is in cases where $q_{\lambda,j}(\cdot)$ is the derivative of some penalty function $p_{\lambda,j}(\cdot)$. In some cases the $q_{\lambda,j}(\cdot)$ functions do not vary with j. The parameter γ can refer only to covariates to be selected γ_z in contrast to parameters for which selection is not to be done γ_x. Therefore the score in equation (2.3) is the score after profiling out the fixed covariates in addition to the baseline.

Considering the software availability, the penalized function generally considered are:

- the LASSO penalty: $p_{\lambda,j}(|\gamma|) = \lambda|\gamma|$;
- the adaptive LASSO: $p_{\lambda,j}(|\gamma|) = \lambda|\gamma|\omega_j$;
- the Elastic Net penalty: $p_{\lambda,j}(|\gamma|) = \lambda_1|\gamma| + \lambda_2(\gamma)^2$;
- the SCAD penalty: $q_{\lambda,j}(|\gamma|) = \lambda\{I(|\gamma| < \lambda) + ((a\lambda - |\gamma|)_+/(a-1)\lambda) I(|\gamma| \geqslant \lambda)\}$.

In order to derive asymptotic results for the penalized estimating equations, Johnson *and others* (2008) outline two conditions: Condition C1 is explained in the present setting in paragraph (2.1) and is the linear approximation of the score, while Condition C2 pertains to the oracle property and is satisfied in the present case by adaptive LASSO and SCAD penalty functions. We also apply in simulations LASSO, Elastic Net, and Relaxed LASSO for the sake of comparison. For LASSO and Elastic Net condition C2 does not hold (see Meinshausen and Bühlmann, 2006), while for Relaxed LASSO there is a conjecture about the oracle property, Meinshausen (2007). Relaxed LASSO corresponds to a sort of two-stage LASSO, and was proposed, as for adaptive LASSO, to overcome the shrinkage problem and the low converging rate of standard LASSO. Under C1 and C2 we have that asymptotically: (a) there exist a root-n solution $\hat{\gamma}$; (b) this solution is sparse; and (c) This solution is asymptotically normal. Here we focus on sparse solutions that are generally advisable with high-dimensional data. Sparse solutions are in fact preferred considering the generally small sample sizes and according to the consideration that a few genes, with respect to the thousand genes measured, are expressed. There are efficient algorithms for the computation of the entire solution path for the coefficients γ with elastic net penalty. In particular, here we use the `glmnet` software (Friedman *and others*, 2010), which is available in `R` and `MATLAB`. Very importantly, the algorithm allows to specify a set of variables whose coefficients must not be penalized. In particular, the coefficients used for baseline estimation, i.e. score equation (2.1), must not be penalized together with the coefficients of the X variables. The function uses an efficient coordinate descent algorithm. We also use the implementation of SCAD regression from `R` package `SIS` version 0.6, Fan *and others* (2010), as it allows the use of weights contrary to more recent implementations. For relaxed LASSO we used the package `relaxnet`, which uses the `glmnet` function. We modified the package to properly handle the `weights` we need for estimation.

2.3 *Simulation strategy*

The proposed estimating equations (EEs) were investigated through different simulation scenarios:

1. unpenalized EE in low dimension (5 covariates);
2. penalized EE in standard variable selection setting (20 covariates);
3. penalized EE in the high-dimensional setting (1000 covariates).

Competing risks data were simulated from the proportional odds model $P_1(t; X) = H(t)\exp(X\beta)/(1 + H(t)\exp(X\beta))$, where $H(t) = 0.5t$. The CIF for the competing event was modeled as $(1 - p)1 - \exp(t\exp(X\beta))$, where $p = P_1(\infty; X)$. Simulations according to the proportional subdistribution hazard model using the strategy outlined in Fine and Gray (1999) were also performed to account for model mis-specification. Basically the CIF for cause 1 was generated according to $1 - [1 - p(1 - \exp(-t))]^{\exp(X\beta)}$. In the first scenario, we generated 5 i.i.d. covariates from a $N(0, 1)$, only two of which have an impact on $P_1(t; X)$ ($\gamma = 0.5$) to evaluate the efficiency of the new estimating equations. In the second simulation scenario 20 covariates were generated, both independent and correlated. Two sample sizes were considered, 400 and 500 times. Three cases were considered for the covariates with true impact on CIF (coefficient 0.5). Namely out of 20 covariates the ones with impact were: (1) predictors 1 and 10; (2) predictors 1, 5, 10 and 15; (3) predictors 1, 2, 3, 5, 8, 10, 13, 15, 18, and 20. Data were generated using both proportional odds and proportional subdistribution hazards. The model under comparison was penalized binomial regression using `glmnet` (the only link available is logit) and different penalizations and the boosted subdistribution hazard regression approach of Binder *and others* (2009) implemented in `R` package `CoxBoost`, Binder (2013).

The third simulation scenario was performed to investigate the performance of the penalized binomial regression in high-dimensional settings with sparsity. As the coefficient estimates are biased toward zero,

the focus of the simulations is on the probability of selection of the relevant covariates. In this case, 1000 covariates were generated with only 16 covariates with an impact on the CIF of interest. The coefficient for the relevant covariates was set at 0.5. Three high-dimensional covariates settings were simulated: in the first the covariates are i.i.d.; in the second the covariate vector X is generated from a p-variate normal distribution whose marginal distributions are standard normal and the correlation between x_i and x_j is $0.5^{|i-j|}$; in the third scenario the covariate vector Z was generated with block correlations according to the simulations in Binder and Schumacher (2008) to mimic high-throughput experimental data. Four covariate blocks are generated with different patterns of correlation. In the first block covariates have 0.5 correlation, in the second 0.35, in the third 0.05, and in the fourth 0.32. Four covariates were selected in each block.

Censoring times were generated from a $U[0, 6]$. This leads to ∼25% censorings, 55% of the cause of interest, and 20% of the competing cause. A second scenario with 50% censoring was also investigated using $U[0, 1.5]$. Four different models were compared: binomial regression with LASSO, adaptive LASSO and relaxed LASSO penalties and the boosting approach.

3. RESULTS

3.1 *Simulation study application*

Results for the first scenario are reported in Table 1. It seems that estimating equations (2.2) have slightly improved finite sample properties with respect to those of Scheike *and others* (2008) and implemented in the `comp.risk` function. The empirical SE are close to the average of the estimated within simulation SE, accounting for the unbiasedness of the estimates.

Results for the second scenario are reported in Table 2 for 20% censoring and the independence setting. Results with other settings are very similar. All the considered approaches have maximum sensitivity, that is the percentage of selection for the important covariates. Considering specificity, that is the percentage of non-selection for truly 0 coefficients, adaptive LASSO seems to have the best performance, while boosting is in line with LASSO. Results appear to be quite robust with respect to model mis-specification. In general, we observe a tendency toward selecting larger models than the true one using cross-validation. Regarding SCAD the number of parameters effectively set to 0 is less than for LASSO. In any case the average estimate of coefficients with true 0 coefficient is of the order of 10^{-3} (the same is true for Elastic Net). We did not study SCAD and the Elastic Net penalties further the high-dimensional setting.

Results for the third simulation scenario are reported in Tables 3 and Figure 1. Also in this setting we report the sensitivity and the specificity of the variable selection method. The tables also report the probability of simultaneous selection of all 16 important covariates. Figure 1 shows the different balancing between false positive and false negative results for the different methods. For the proportional odds simulation, the adaptive LASSO shows the best sensitivity. This is at the cost of a greater number of selected variables; see the median number of variables selected in Tables 3. Adaptive LASSO has the best performance in terms of the proportion of selected models including all 16 important variables for proportional odds simulation. The best performance with the proportional subdistribution hazard simulation is that of boosting. This property is often referred to as the variable screening property (Meijer and Goeman, 2013; Bühlmann and van de Geer, 2011). Under the proportional subdistribution hazard simulation, of particular interest is the behavior of relaxed LASSO, which shows very good performances in terms of specificity, maintaining a high sensitivity. In particular, in the block correlated simulation (Binder) relaxed LASSO appears as a valuable alternative to the boosting approach. The estimate of the coefficients was, as expected, biased downward. Adaptive and relaxed LASSO perform very well in this respect with shrinkage less than that of the other methods. Relaxed LASSO (Zou, 2006), presents with an exceptionally high computational cost; only 100 simulations were then performed for each scenario.

Table 1. *Comparison of different estimating equations for the binomial regression model fit*

Correlation		Model	Bias	%Bias	Empirical SE	Average SE	MSE	Coverage	Length
$\rho = 0$	β_1	CR	0.009	0.019	0.104	0.011	0.011	0.950	0.406
		Bin	0.006	0.013	0.101	0.010	0.010	0.948	0.394
	β_1	CR	0.008	0.017	0.105	0.011	0.011	0.948	0.406
		Bin	0.005	0.011	0.103	0.010	0.011	0.950	0.394
	β_1	CR	0.001		0.100	0.010	0.010	0.950	0.387
		Bin	−0.000		0.098	0.009	0.010	0.946	0.378
	β_1	CR	−0.002		0.102	0.010	0.010	0.946	0.387
		Bin	−0.003		0.098	0.009	0.010	0.948	0.378
	β_1	CR	−0.001		0.101	0.010	0.010	0.947	0.387
		Bin	−0.001		0.098	0.009	0.010	0.947	0.378
$\rho = 0.5$	β_1	CR	0.015	0.030	0.128	0.016	0.017	0.946	0.489
		Bin	0.008	0.016	0.121	0.014	0.015	0.944	0.461
	β_1	CR	0.013	0.025	0.126	0.016	0.016	0.954	0.488
		Bin	0.009	0.019	0.119	0.014	0.014	0.947	0.461
	β_1	CR	0.000		0.121	0.014	0.015	0.947	0.462
		Bin	0.001		0.118	0.013	0.014	0.941	0.443
	β_1	CR	−0.000		0.122	0.014	0.015	0.947	0.463
		Bin	0.000		0.115	0.013	0.013	0.947	0.443
	β_1	CR	−0.003		0.122	0.014	0.015	0.945	0.462
		Bin	−0.001		0.115	0.013	0.013	0.945	0.443

The original estimating equations implemented in the `comp.risk` function are compared with that using a binomial error. It is possible to note a small efficiency gain for the binomial error function. Logit simulation, with 20% and 50% censoring, and true coefficient values for β_1 and β_2 are equal to 0.5. The other covariates have no impact on the CIF. CR: function `comp.risk` from `R` package `timereg`; Bin: estimating equations (2.2).

3.2 *Application to bladder carcinoma*

We use publicly available data considered in a validation study about a gene expression signature to predict outcome in non-Muscle-Invasive Bladder Carcinoma (Dyrskjøt *and others*, 2007). The same data were also used by Binder *and others* (2009) for the study of boosting and we are therefore reporting the same analysis here just for the comparison. In particular, 301 patients used for training and validating the progression classifier and with complete information on clinically important covariates (age, sex, BCG/MMC treatment, Reevaluated WHO grade and Reevaluated pathological disease stage) were considered. One hundred eighty-four patients were censored, 84 had progression or death from bladder cancer, while 33 died from other causes. Genomic and clinical data are available as supplementary material on the journal website (available at *Biostatistics* online). In particular, microarray data on 1381 gene expressions are available. Stacked data were generated by replicating the data of each patient for 10 time-points, corresponding to deciles of the unique event time distribution for the cause of interest. A 10-fold cross-validation is used to select the optimal λ value. Cross-validation is performed across patients rather than rows. The adaptive weights for LASSO were taken from a first-stage ridge regression. Adaptive Lasso selected a total of 90 probes. Ten probes are in common with those selected in the work of Dyrskjøt *and others* (2007). LASSO selected 60 probes, relaxed LASSO selected only 4 probes, while the boosted Fine & Gray's model selected 30 probes. In the top left corner of Figure 2, we report the overlap of the probes selected

Table 2. *Performance of different penalties in a non-high-dimensional setting for model selection*

		400		Logit		500		400		cll		500	
Model		2	4	10	2	4	10	2	4	10	2	4	10
L	sp	0.76	0.63	0.38	0.77	0.62	0.36	0.76	0.62	0.35	0.76	0.63	0.34
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	6	10	16	6	10	17	6	10	17	6	10	17
	ME	0.36	0.38	0.42	0.37	0.39	0.43	0.45	0.50	0.56	0.47	0.51	0.58
AL1	sp	0.83	0.82	0.73	0.86	0.83	0.77	0.84	0.83	0.77	0.84	0.82	0.78
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	4	7	12	4	6	12	4	6	12	4	6	12
	ME	0.45	0.46	0.45	0.46	0.47	0.47	0.53	0.55	0.57	0.54	0.55	0.58
AL2	sp	0.84	0.82	0.76	0.85	0.83	0.77	0.85	0.81	0.76	0.85	0.81	0.77
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	4	6	12	4	6	12	4	7	12	4	7	12
	ME	0.46	0.46	0.45	0.46	0.46	0.46	0.53	0.55	0.58	0.54	0.56	0.58
EN	sp	0.36	0.21	0.10	0.34	0.19	0.08	0.28	0.16	0.06	0.27	0.15	0.05
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	14	17	19	14	17	19	15	18	20	15	18	20
	ME	0.31	0.36	0.41	0.33	0.38	0.42	0.41	0.48	0.55	0.43	0.50	0.58
SCAD	sp	0.34	0.32	0.42	0.33	0.33	0.42	0.45	0.42	0.49	0.46	0.42	0.50
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	15	16	16	15	15	16	13	14	15	13	13	15
	ME	0.52	0.52	0.52	0.51	0.51	0.51	0.60	0.62	0.65	0.60	0.62	0.65
Boost	sp	0.73	0.63	0.49	0.72	0.62	0.49	0.72	0.61	0.42	0.72	0.62	0.41
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	6	10	15	6	10	15	7	10	16	7	10	16
	ME	0.28	0.29	0.30	0.29	0.29	0.30	0.40	0.42	0.45	0.42	0.43	0.45
RL	sp	0.76	0.62	0.36	0.78	0.63	0.36	0.85	0.85	0.76	0.85	0.84	0.79
	sn	1	1	1	1	1	1	1	1	1	1	1	1
	Median	6	10	16	5	10	16	4	5	11	3	5	10
	ME	0.37	0.40	0.43	0.38	0.40	0.44	0.45	0.51	0.56	0.47	0.52	0.56

For each penalty the sensitivity (sn), i.e. the average percentage of selected correct non-0 coefficients, and specificity (sp), i.e. the correct percentage of non-selected 0 coefficients, are reported together with the median number (Median) of variables included in the model and the average value (ME) of the true non-0 coefficients. One thousand simulations were performed with 20 independent covariates. The numbers 2, 4, and 10 refer to the number of parameters with true impact on CIF (with coefficient 0.5). LASSO, Lasso penalty; ADALASSO1, adaptive LASSO with weights from ridge regression; ADALASSO2, adaptive LASSO with weights from standard binomial GLM; EN, elastic net penalty; SCAD, SCAD penalty; Boost, boosted Fine & Gray's model; RL, relaxed LASSO. Data were simulated according to the proportional odds model (logit) and to the proportional subdistribution hazard model (cll).

by the different methods. The comparison highlights once more that the selection of probes is model-dependent. A few probes were chosen in common between the boosting and the penalized approaches and with respect to the 88 probes selected in the clinical work. This is probably due to the block correlated structures and redundancy in the gene expression data. We compared the estimated prediction error of the different models. The strategy outlined in Binder *and others* (2009) was used. Basically bootstrap samples, including a sample size fraction of 0.632, are drawn without replacement. The prediction error for each bootstrap sample is calculated using the observations left out of the sample. The final prediction error is obtained by a weighted average between the apparent prediction error (calculated on training data)

Table 3. *Performance of different penalties in high dimensions for model selection*

			Logit					cll				
	cens	cor	sn	sp	% all	Mean est	Median num	sn	sp	% all	Mean est	Median num
L	20%	iid	0.996	0,908	0,926	0,238	107	0.999	0.902	0.974	0.118	77
Boosting			0,989	0,961	0,822	0,181	57	0.999	0.953	0.974	0.112	52
RL			0.934	0.985	0.400	0.32	28	0.964	0.985	0.520	0.440	29
AL			0.997	0.888	0.934	0.393	128	0.999	0.900	0.952	0.240	126
L		cor	0.996	0.925	0.931	0.304	90	0.999	0.914	0.994	0.404	102
Boosting			0.994	0.971	0.887	0.233	47	1.000	0.962	1.000	0.339	58
RL			0.942	0.997	0.370	0.428	20	0.979	0.999	0.670	0.593	20
AL			0.998	0.917	0.967	0.461	101	0.999	0.931	0.995	0.599	86
L		Binder	1.000	0.915	0.994	0.289	101	1.000	0.897	1.000	0.415	120
Boosting			0.998	0.971	0.959	0.278	47	1.000	0.960	1.000	0.472	63
RL			0.990	0.980	0.820	0.377	39	1.000	0.984	1.000	0.580	36
AL			0.999	0.893	0.983	0.398	124	1.000	0.903	1.000	0.561	113
L	50%	iid	0.979	0,915	0,693	0,215	102	0.989	0.913	0.786	0.073	50
Boosting			0,973	0,957	0,621	0,176	61	0.988	0.952	0.788	0.077	44
RL			0.895	0.974	0.270	0.274	31	0.914	0.971	0.310	0.343	34
AL			0.982	0.893	0.715	0.370	125	0.984	0.906	0.736	0.189	112
L		cor	0.991	0.931	0.838	0.290	85	0.999	0.919	0.974	0.382	98
Boosting			0.994	0.968	0.810	0.235	50	0.999	0.962	0.992	0.328	58
RL			0.908	0.998	0.170	0.406	19	0.958	0.998	0.430	0.558	20
AL			0.996	0.923	0.917	0.452	95	0.999	0.934	0.988	0.577	83
L		Binder	0.999	0.915	0.985	0.281	100	1.000	0.900	1.000	0.403	116
Boosting			0.997	0.969	0.942	0.281	49	1.000	0.955	1.000	0.467	64
RL			0.984	0.978	0.720	0.373	38	0.999	0.983	0.980	0.565	36
AL			0.998	0.893	0.969	0.393	122	1.000	0.901	1.000	0.551	117

Application of the `glmnet` function to fit LASSO, adaptive LASSO, and relaxed LASSO penalty with binomial error function and logit link. The function `coxboost` was used to apply the boosted Fine & Gray's model. Three scenarios are simulated: (1) one thousand independent covariates (i.i.d.); (2) one thousand correlated covariates (cor), and (3) one thousand covariates with block correlations mimicking the correlation found in high-throughput studies of Binder. Only 16 covariates have a true impact on the CIF of interest. The average percentage of selection of covariates with true non-0 coefficient (sn) and the average percentage of non-selection of covariates with true 0 coefficient (sp) is reported. Moreover, the percentage of selected models including all of the 16 covariates with true non-0 coefficients is reported (% all), together with the mean value of the parameters selected with impact on the CIF (mean est) and the median number of selected covariates (median num). Data were simulated from the proportional odds model with 20% and 50% censoring.

and the bootstrap cross-validated prediction error estimate. As expected, the boosted Fine & Gray's model and relaxed LASSO presented the larger training error with respect to adaptive LASSO. The situation is reversed with the bootstrap cross-validation procedure. The final estimated prediction error curves are reported in Figure 2. According to this internal validation procedure, adaptive LASSO has an advantage in terms of prediction error with respect to boosting and relaxed LASSO, which show a very similar performance. This is probably connected to the high probability of inclusion of all relevant covariates, which appears as a characteristic of Adaptive LASSO from the simulations.

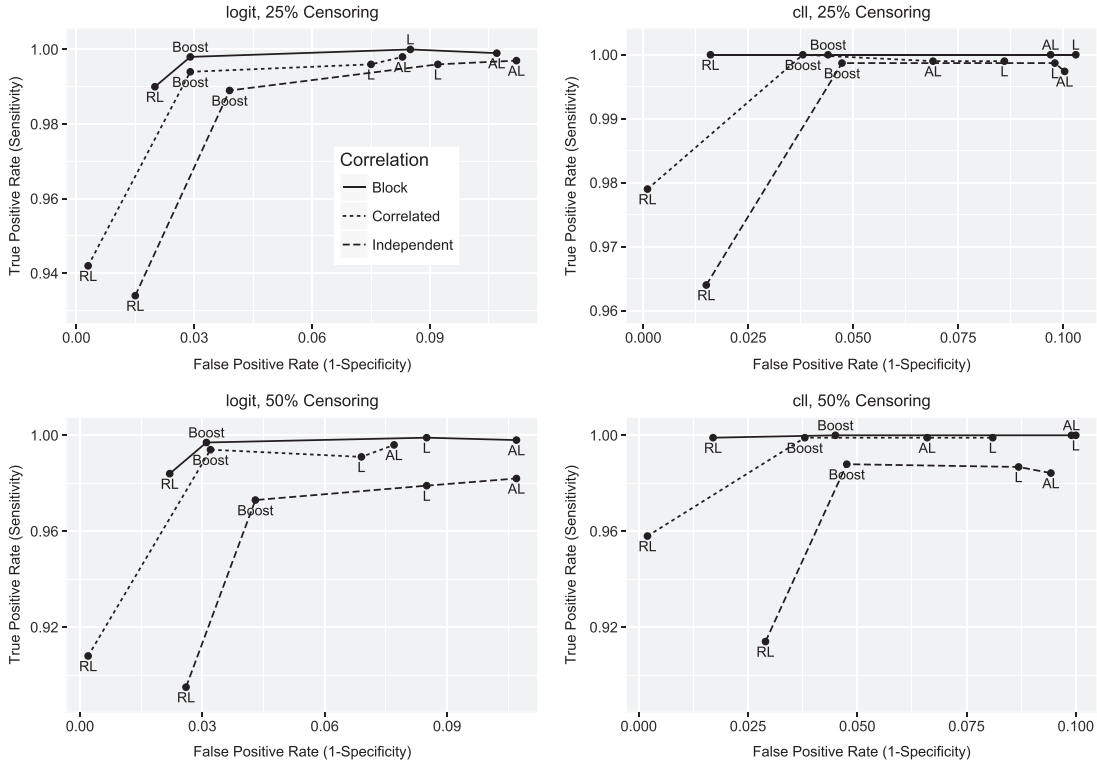

Fig. 1. Sensitivity (mean percentage of selection for variables with impact on the CIF) and 1-specificity (mean percentage of selection for unimportant covariates) for the different methods under comparison: L is binomial regression with LASSO penalty; AL is binomial regression with adaptive LASSO penalty; RL is binomial regression with relaxed LASSO penalty; Boost is boosting of Fine & Gray's model. Three simulation scenarios are used with 1000 covariates ($N(0, 1)$) and 16 covariates with true impact (coefficient 0.5): independent, correlated, and block correlated.

4. DISCUSSION

The topic of identification of significant transcripts with a survival outcome was investigated using different approaches (Sinnott and Cai, 2013). In presence of competing risks relevant analyses could consider cause-specific hazards or the cumulative incidence. For the cause-specific hazard analysis, which is sensible to investigate disease dynamics, the tools to be used borrow from standard survival analysis. It is to be remarked that all the regression models involved in the different cause-specific hazards considered must be evaluated in order to have a complete picture of the covariate impact.

For cumulative incidence, which is more sensible for clinical decision-making, specific methods should be developed. Present proposals deal with the possible application with high-dimensional covariates of the subdistribution approach, which is widely used in clinical applications and is now available in many statistical softwares.

In particular, Binder *and others* (2009) proposed a shrinkage approach based on boosting applied to the Fine & Gray's model.

Although some works thus exist, it is, however important to have different methods at hand to compare the results. Here we point out that most competing risks methods are IPCW and as such often directly can utilize the existing software for penalized estimation.

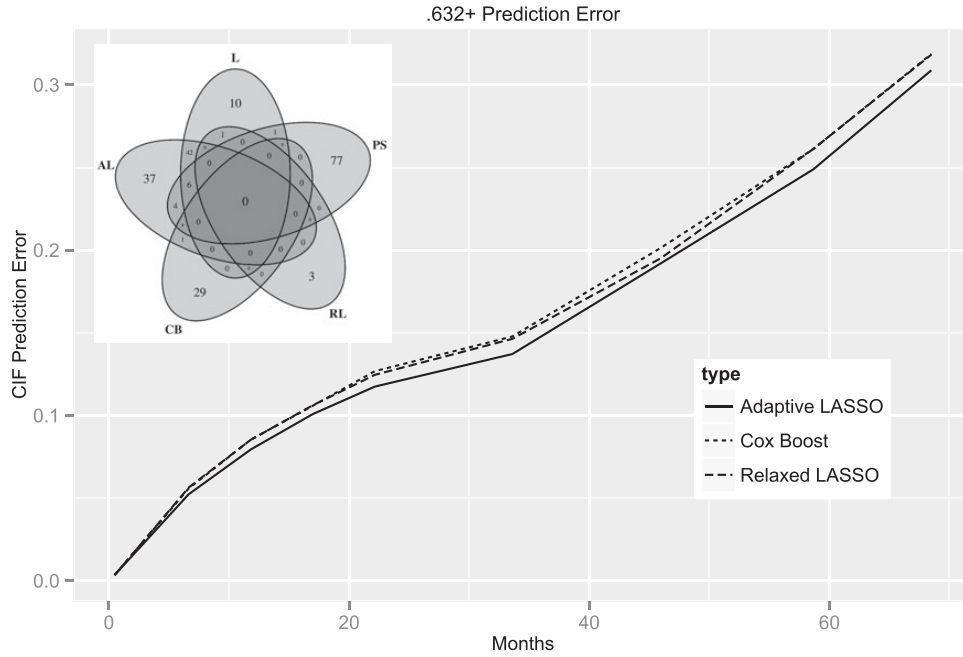

Fig. 2. Prediction error curves are calculated for 10 selected times at quantiles of the unique failure time distribution (considering the cause of interest). Two hundred bootstrap samples were drawn without replacement and of size equal to the fraction 0.632 of the sample size (about 190 patients). For each bootstrap sample the entire procedure of model building was repeated, i.e. the selection of the penalty parameter using cross-validation or the optimal number of boosting steps. Predictions were collected for each bootstrap samples using data out of the bootstrap sample itself. The final prediction error is a weighted mean of the apparent prediction error (that is calculated on training data) and of the prediction error obtained using bootstrap out-of-sample predictions. The weights are time-dependent and the calculation of prediction error uses IPCW weights. On the top left corner, Venn Diagram showing the overlap between the probes selected by the different methods. L, LASSO; AL, adaptive LASSO, RL, relaxed LASSO; CB, CoxBoost; PS, progression signature from Dyrskjøt *and others* (2007).

In this work a modification of the estimating equations of Scheike *and others* (2008) is proposed with the use of binomial error and logit link function. The possibility to extend the procedure in a high-dimensional setting is then immediate using standard software for regularized regression.

Specifically, software for penalized generalized linear models can be used, provided there is an option for `weights`. The free statistical software `R`, R Core Team (2015), has many possibilities. We used here the `glmnet` function (Friedman *and others*, 2010), which is efficient and has options to exclude some parameters from the penalty and an option for `weights`. Unfortunately, it is not possible at present to change the link function. In particular, it would be interesting to use a *cloglog* link to compare the results with the available approaches using the subdistribution approach.

The method presented requires the correct specification of the censoring model in order to estimate IPCW weights. In the simulations we used the Kaplan–Meier estimator for this purpose and this is consistent if the censoring is independent also from the covariates. It is possible to use Cox regression or Aalen additive regression in more complex applications (Scheike *and others*, 2008). It is noteworthy that the approach can readily be used in more complex frameworks such as the direct modeling of covariate effects in transition probabilities in multistage models. The direct modeling instead of the modeling of

all transition intensities involved in state transitions is particularly simple to interpret and easily implementable (Scheike and Zhang, 2007).

5. SOFTWARE

The R function `prep.glm.comprisk` is available in package `timereg` (Scheike and Martinussen, 2006; Scheike and Zhang, 2011) to prepare stacked data for performing binomial regression.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

REFERENCES

AMBROGI, F., BIGANZOLI, E. AND BORACCHI, P. (2008). Estimates of clinically useful measures in competing risks survival analysis. *Statistics in Medicine* **27**(30), 6407–6425.

BINDER, H. (2013). *CoxBoost: Cox Models by Likelihood Based Boosting for a Single Survival Endpoint or Competing Risks*. R package version 1.4.

BINDER, H., ALLIGNOL, A., SCHUMACHER, M. AND BEYERSMANN, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* **25**(7), 890–896.

BINDER, H. AND SCHUMACHER, M. (2008). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology* **7**(1), Article 12.

BÜHLMANN, P. AND VAN DE GEER, S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer.

DYRSKJØT, L., ZIEGER, K., REAL, F. X., MALATS, N., CARRATO, A., HURST, C., KOTWAL, S., KNOWLES, M. AND MALMSTROM, P. U. *and others* (2007). Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. *Clinical Cancer Research* **13**(12), 3545–3551.

ERIKSSON, F., LI, J., SCHEIKE, T. AND ZHANG, M. J. (2015). The proportional odds cumulative incidence model for competing risks. *Biometrics* **71**(3), 687–695.

FAN, J., FENG, Y., SAMWORTH, R. AND WU, Y. (2010). *SIS: Sure Independence Screening*. R package version 0.6.

FINE, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics* **2**(1), 85–97.

FINE, J. P. AND GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**(446), 496–509.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1–22.

GERDS, T. A., SCHEIKE, T. H. AND ANDERSEN, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine* **31**(29), 3921–3930.

GERDS, T. A. AND SCHUMACHER, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63**(4), 1283–1287.

Geskus, R. B. (2011). Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics* **67**(1), 39–49.

Johnson, B. A., Lin, D. Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**(482), 672–680.

Klein, J. P. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61**(1), 223–229.

Latouche, A., Allignol, A., Beyersmann, J., Labopin, M. and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology* **66**(6), 648–653.

Meijer, R. J. and Goeman, J. J. (2013). Model selection for high-dimensional models. In: Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (editors), *Handbook of Survival Analysis*. London: Chapman & Hall/CRC.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52**(1), 374–393.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**(3), 1436–1462.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

Scheike, T. H. and Martinussen, T. (2006) *Dynamic Regression Models for Survival Data*. New York: Springer.

Scheike, T. H. and Zhang, M. J. (2007). Direct modelling of regression effects for transition probabilities in multistate models. *Scandinavian Journal of Statistics* **34**(1), 17–32.

Scheike, T. H. and Zhang, M. J. (2008). Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Analysis* **14**(4 SPEC. ISS.), 464–483.

Scheike, T. H. and Zhang, M. J. (2011). Analyzing competing risk data using the R timereg package. *Journal of Statistical Software* **38**(2), 1–15.

Scheike, T. H., Zhang, M. J. and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**(1), 205–220.

Simon, R., Lam, A., Li, M., Ngan, M., Menenzes, S. and Zhao, Y. (2007). Analysis of gene expression data using BRB-array tools. *Cancer Informatics* **3**, 11–17.

Sinnott, J. A. and Cai, T. (2013). High-dimensional regression models. In: Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (editors), *Handbook of Survival Analysis*. London: Chapman & Hall/CRC.

Sun, L. and Zhang, Z. (2009). A class of transformed mean residual life models with censored survival data. *Journal of the American Statistical Association* **104**(486), 803–815.

Zhang, M. J. and Fine, J. (2008). Summarizing differences in cumulative incidence functions. *Statistics in Medicine* **27**(24), 4939–4949.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

[*Received June 11, 2015; revised March 11, 2016; accepted for publication March 12, 2016*]