Received: 29 November 2016 | Revised: 18 March 2017 | Accepted: 8 May 2017

DOI: 10.1002/bimj.201600242

RESEARCH PAPER

Biometrical Journal

Prediction accuracy and variable selection for penalized cause-specific hazards models

Maral Saadati[1] | **Jan Beyersmann**[2] | **Annette Kopp-Schneider**[1] | **Axel Benner**[1]

[1]Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Institute of Statistics, University of Ulm, Ulm, Germany

Correspondence
Maral Saadati, Division of Biostatistics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
Email: m.saadati@dkfz.de

We consider modeling competing risks data in high dimensions using a penalized cause-specific hazards (CSHs) approach. CSHs have conceptual advantages that are useful for analyzing molecular data. First, working on hazards level can further understanding of the underlying biological mechanisms that drive transition hazards. Second, CSH models can be used to extend the multistate framework for high-dimensional data. The CSH approach is implemented by fitting separate proportional hazards models for each event type (iCS). In the high-dimensional setting, this might seem too complex and possibly prone to overfitting. Therefore, we consider an extension, namely "linking" the separate models by choosing penalty tuning parameters that in combination yield best prediction of the incidence of the event of interest (penCR). We investigate whether this extension is useful with respect to prediction accuracy and variable selection. The two approaches are compared to the subdistribution hazards (SDH) model, which is an established method that naturally achieves "linking" by working on incidence level, but loses interpretability of the covariate effects. Our simulation studies indicate that in many aspects, iCS is competitive to penCR and the SDH approach. There are some instances that speak in favor of linking the CSH models, for example, in the presence of opposing effects on the CSHs. We conclude that penalized CSH models are a viable solution for competing risks models in high dimensions. Linking the CSHs can be useful in some particular cases; however, simple models using separately penalized CSH are often justified.

KEYWORDS
competing risks, high-dimensional data, penalization, prediction

1 | INTRODUCTION

In clinical studies it is often necessary to analyze time-to-event data, where patients may encounter different event types. Consider, for example, patients diagnosed with bladder cancer as studied by Dyrskjøt et al. (2007). In the course of time, the disease may progress for some patients, while other patients may die without prior progression. These events can be seen as different causes of failure. Dyrskjøt et al. (2007) use the composite clinical endpoint "progression-free survival", that is, time from diagnosis to progression or death. However, such analyses suffer from a number of drawbacks. As pointed out by Mell and Jeong (2010), effect estimates from composite endpoint analyses represent a mixture of the effects associated with the causes of failure, their interpretation can be complex and unintuitive. This issue becomes even more problematic if the different endpoints are not of comparable importance. Furthermore, the use of a composite endpoint may result in a loss of power to detect effects, particularly if they differ across component endpoints or if they are in opposite directions.

Since combined endpoints do not distinguish different event types, covariate effects cannot be related to a specific event of interest. This is in stark contrast to the fact that variables are measured in high resolution (e.g., microarray features), but the endpoint is left rather unspecified. Competing risks models, in particular cause-specific hazards (CSHs) regression, allow us to disentangle the effects of covariables on different event types. Thus, it is becoming more and more popular to analyze such data using competing risks methods, which are easily implemented in standard software (Putter, Fiocco, & Geskus, 2007). Yet a lot of care is needed in correctly interpreting the results due to the inherent complexity of competing risks (Beyersmann, Dettenkofer, Bertz & Schumacher, 2007).

When considering a single time-to-event endpoint, there is as a clear correspondence between the survival probability and cumulative hazard function, which gives rise to hazard-based regression models in survival analysis. In a competing risks setting however, there is no longer a 1-to-1 relationship between the (CSH) hazard and event probability. The effect of a covariable on the CSH does not immediately translate to an effect on the cumulative incidence. Users might struggle to decide which approach to use, namely regression of hazard functions (via CSHs models) or regression of the cumulative incidence (using, e.g., a subdistribution hazard [SDH] model); each comes with its own merits and pitfalls.

We argue that CSHs are "the quantities of choice" for competing risks models for several reasons. First and foremost, CSHs represent "the basic estimable quantities in the competing risks framework" (Prentice et al., 1978). They allow to analyze the effects of covariates on hazards level and thereby model biological aspects of disease progression. This is particularly interesting in the presence of molecular information, where researchers aim at identifying and understanding which microarray features contribute to which events. Furthermore, we demonstrate that CSH models can also be used for prediction of incidence probabilities. Finally, CSH can be extended to multistate models, which allow us to distinguish events in even greater detail.

A central aspect of this work is to study competing risks models with high-dimensional covariates, where the sample size n is far smaller than the number of variables p. In particular, we examine the effects of penalization in the context of hazard-based competing risks models. Using CSH models as a starting point, we investigate the performance of lasso-penalized CSH models fit separately for each event based on a maximal partial likelihood criterion. Driven by the question of whether we can improve model performance by linking the CSH models for different causes, we consider an extension, namely by maximizing with respect to best prediction for the event of interest. This suggestion is compared to the established approach to achieve synthesis (or linking) of the CSHs, namely the SDH model.

The current manuscript is organized as follows. After a brief introduction to competing risks models in the classical, low-dimensional setting, we discuss in Section 2 established methods for the analysis of high-dimensional covariates and introduce penalized CSH models coupled with respect to prediction error. Sections 3 and 4 cover simulations with different proportionality assumptions. We close with an application example on bladder cancer (Section 5) and a brief discussion in Section 6.

1.1 | Cause-specific hazards regression models

A competing risks process can be viewed as a multistate process with one transient state (initial state) and K absorbing states, which represent all possible outcomes (cf. Andersen and Keiding, 2002, for a detailed introduction to multistate models). All patients start in the initial state 0 and may experience one of K different failures types as time progresses. Let T denote the failure time, at which a patient leaves the initial state, i.e. $T = \inf\{t > 0 | Z_t \neq 0)\}$, and $(Z_t)_{t\geq 0}$ the competing risks process with $Z_t \in \{0, 1, \dots, K\}$. The CSH for cause k is given by

$$\alpha_k(t) = \lim_{\Delta t \searrow 0} \frac{P(T \in [t, t + \Delta t), Z_T = k | T \geq t)}{\Delta t}.$$

The CSHs represent transition intensities from state 0 to states $k = 1, \dots, K$ and thereby completely determine the stochastic behavior of the competing risks process $(Z_t)_{t\geq 0}$.

Furthermore, CSH regression models enable the use of familiar approaches, such as proportional hazards (PHs) regression. In particular, the proportional CSH model for cause $k = 1, \dots, K$ is given by

$$\alpha_k(t; X) = \alpha_{0k}(t) \exp(X\boldsymbol{\beta}_k), \tag{1}$$

where $\alpha_{0k}(t)$ is the baseline hazard for cause k at time t, $X \in \mathbb{R}^{n\times p}$ is the design matrix, and $\boldsymbol{\beta}_k \in \mathbb{R}^p$ the vector of cause-specific regression coefficients. The beauty of CSH is that it is possible to fit a model for each cause separately, where each cause has its

own baseline hazard and covariate effect estimate. In particular, the partial likelihood can be written in terms of a stratified Cox model as a product of partial likelihoods for each cause:

$$L(\beta_1, \ldots, \beta_K) = \prod_{k=1}^{K} L_k(\beta_k). \tag{2}$$

Consequently, the partial likelihood is maximized when all factors are maximized. In practice, this means that we can fit separate/"independent" models for each cause using standard methods. The reason is that censoring by an observed competing risk is independent in the sense that it preserves the correct form of the intensity of the cause-specific counting process; see Andersen, Borgan, Gill, and Keiding (1993, Chapter 3), for a formal proof and Allignol, Beyersmann, and Schmoor (2016), for a more heuristic explanation.

Yet, while CSH models allow to conveniently fit hazards regression models, one needs to analyze all CSHs to understand the impact of different effects on the absolute risks. To be more precise, the cumulative incidence function (CIF) describing the absolute risk of an individual with observed covariate matrix X to experience event k before time t given by

$$F_k(t|X) = p(T \leq t, Z_T = k|X) = \int_0^t \exp\left(-\int_0^u \sum_{j=1}^{K} \alpha_j(v, X)dv\right)\alpha_k(u, X)du. \tag{3}$$

From the above formula, we can see that the CIF for event k not only depends on the CSH of event k, but also on the CSH of all other events. Therefore, it is not sufficient to only analyze the CSH of the event of interest as it does not fully express the effects on the expected proportion of individuals to experience event k before time t.

1.2 | Subdistribution hazards models

In an attempt to reestablish the 1-to-1 correspondence between hazard and incidence Fine and Gray (1999) used the notion of the SDH. The SDH $h_k(t)$ for the event of interest (say event k) is designed as the hazard associated with the cumulative incidence, that is,

$$F_k(t) = p(T \leq t, Z_T = k) = 1 - \exp\left(-\int_0^t h_k(u)du\right).$$

Therefore, Fine and Gray define the SDH $h_k(t) = \frac{dF_k(t)/dt}{1-F_k(t)}$, which mimics the well-known hazard definition in survival analysis and thereby allows to fit regression models assuming proportional SDH:

$$h_k(t; X) = h_{0k}(t)exp(X\boldsymbol{\gamma}), \tag{4}$$

where $h_{0k}(t)$ denotes the baseline SDH for event k at time t and $\boldsymbol{\gamma}$ the vector of regression coefficients describing the influence of corresponding covariates on the cumulative incidence. Note that the baseline hazards $\alpha_{0k}(t)$ and $h_{0k}(t)$ are in general not the same and also $\boldsymbol{\gamma} \neq \boldsymbol{\beta}_k$ in general. An advantage of the SDH framework is that only *one* model needs to be fitted if interest focuses on F_k, as opposed to the CSH approach, where K models are employed. The CIFs of the competing events are not modeled in the SDH approach, whereas they are naturally obtained as a byproduct in the CSH framework. An issue of the Fine and Gray model is that one cannot interpret the estimated effect of a covariate on the hazard level, but only on the cumulative incidence. Andersen and Keiding (2012) discuss interpretability in this context. Fine and Gray (1999) demonstrated that γ can be consistently estimated by restructuring the event times and status. In the presence of censoring, it is necessary to retrieve the censoring distribution for patients with competing events. One approach is to use the censoring distribution of the observed data and to weight the patients who experienced the competing events, using inverse probability of censoring weighting.

1.3 | Direct modeling

Another approach for competing risks regression is to directly model the CIF while adjusting for covariates. This can be done, for example, as suggested by Klein and Andersen (2005), by generating pseudo-values for each subject over a grid of prespecified time points. Using a link function, one can then model the conditional CIF of the event of interest and estimate regression parameters from a psuedo-score equation. Another direct modeling approach was proposed by Scheike, Zhang, and Gerds (2008),

called direct binomial regression, where the CIF is modeled via a link function and a weighted response based on inverse probability of censoring.

While direct modeling has many merits and is very general by allowing for different link functions, we wish to stay within the hazard-based framework of Cox-like models. Therefore, pseudo-value regression and direct binomial modeling are not considered in this paper.

1.4 | Prediction accuracy and variable selection for competing risks

There are often two main aspects to consider in the context of model building. One is prediction performance and the other is variable selection/interpretation. We now clarify these objectives in the competing risks scenario as they will serve as evaluation criteria in our simulation studies. Schoop, Beyersmann, Schumacher, and Binder (2011) defined prediction accuracy (a.k.a. Brier score) for the event of interest k as

$$PE_k(s) = E\left(I(T \leq s, Z_T = k) - \pi_k(s|X)\right)^2, \tag{5}$$

where $\pi_k(s|X)$ denotes the predicted CIF of type k. Loosely speaking, one could say that prediction accuracy is incidence-based. This leads us to a question we will investigate in a simulation study: how well do CSH models perform with respect to predicting the cumulative incidence of an event of interest?

Variable selection, on the other hand, is expressed in terms of the true positive rate (TPR) and the false discovery rate (FDR) with respect to the event of interest (say event 1). In particular, the TPR is given by the number of correctly identified relevant variables (a.k.a. true positives) divided by the total number of relevant variables. The FDR is defined as the number of unrelated variables chosen (a.k.a. false positives) divided by the total number of variables chosen. We define "relevant variables" as variables that contribute to the CIF of event 1, that is, variables associated with *any* of the K CSHs. This definition of true positives can be considered rather lenient, because it is more in the spirit of SDH (count a variable as a true positive if it has a parameter estimate $\neq 0$ for *any* of the CSHs). However, we emphasize that the CSH framework offers more! With CSHs, we can distinguish the effects of variables on each hazard separately, thereby disentangling the effects on the CIF into effects on the hazards level. Therefore, one could also argue for a second, stricter definition of true positives, namely as variables associated with the CSH of the event of interest rather than its CIF.

We clarify this difference in an example. Suppose variable X_1 has parameters $\beta_{11} \neq 0 \neq \beta_{21}$. In the lenient version, X_1 is evaluated as a true positive if $\hat{\beta}_{11} \neq 0$ or $\neq \hat{\beta}_{21}$ in the CSH model. However, in the strict version, X_1 is evaluated as a true positive if and only if $\hat{\beta}_{11} \neq 0$, irrespective of $\hat{\beta}_{21}$. For SDH methods there exists only one version: X_1 is a true positive if $\hat{\gamma}_1 \neq 0$.

Even though it can be argued that the strict definition is more accurate for evaluating the variable selection performance for CSH models, we will use the first, more lenient definition in our simulation experiments to ensure comparability of results between CSH and SDH approaches.

2 | METHODS FOR HIGH-DIMENSIONAL COMPETING RISKS MODELS

2.1 | Cause-specific hazards based approaches

To the best of our knowledge, there has been little research on penalized CSH models. As mentioned in the introduction, CSHs describe the driving forces behind observed transitions and offer interpretable regression coefficients, discussed by Andersen and Keiding (2012). In this way, CSH methods provide a powerful tool to further our understanding of biological mechanisms that govern each of the transition hazards. This is particularly desirable when analyzing molecular data, where we hope to understand, for example, the role of "driving" genes.

CSH models can be fitted separately for each cause, as discussed in Section 1.1. In this work, we explore two versions for fitting penalized models: the first is penalization with the aim of maximizing the penalized partial likelihood, the second is maximizing the prediction accuracy in an attempt to link the separate models.

Consider the likelihood for the CSH model given in Section 1.1. It is easily possible to add a penalty term to the log-partial likelihood $l(\boldsymbol{\beta})$, thereby maximizing the penalized log-partial likelihood: $\max_\beta l(\boldsymbol{\beta}) - Pen(\boldsymbol{\beta})$, where $Pen(\boldsymbol{\beta})$ denotes a penalty function on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$. We choose to implement a lasso penalty (cf. Tibshirani, 1996), thereby maximizing $l(\boldsymbol{\beta}) - \lambda \parallel \boldsymbol{\beta} \parallel_1$. This choice is mainly due to the lasso's simplicity and variable selection ability, as well as its common usage in biomedical models. This is not to say that the lasso is always an appropriate form of penalization, but we use it an exemplary penalization scheme. Naturally, the penalty type should be chosen corresponding to the assumptions on the data structure (sparsity,

correlation, etc.) as well as the research goal (interpretation of selected variables, best prediction, etc.). The lasso assumes sparsity of the underlying data structure and has some shortcomings in handling highly correlated variables. Furthermore, it does not yield asymptotically unbiased estimates. Other penalties (e.g., elastic net and smoothly clipped absolute deviation (SCAD)) have been proposed to address these and other issues. However, these issues are present in all contexts of high-dimensional data and not unique to competing risks settings. There has been much work on comparing different penalty types for different purposes (Benner, Zucknick, Hielscher, Ittrich, & Mansmann, 2010), but this topic is out of the scope of our current work, where we wish to focus on the issues arising from multiple time-to-event outcomes.

Consider the lasso tuning parameter λ of the objective function $\max_\beta l(\boldsymbol{\beta}) - \lambda \parallel \boldsymbol{\beta} \parallel_1$. As noted in connection with equation (2), it is possible to maximize the log-partial likelihood l_k for each cause separately. However, using a global tuning parameter λ on all components of L_k may not be optimal. An idea would therefore be to allow a different tuning parameter λ_k for each cause k when fitting the separate CSH models as explained in the following.

2.1.1 | Independent CSH models with penalty (iCS)

We allow for different tuning parameters λ_k in each penalized CSH model. In particular, we maximize the penalized log-partial likelihood for each cause k by

$$\max_{\beta_k} \quad \left\{ l_k(\boldsymbol{\beta}_k) - \lambda_k \parallel \boldsymbol{\beta}_k \parallel_1 \right\}. \tag{6}$$

Equation (6) results in estimates $\hat{\boldsymbol{\beta}}_k^{pen}$, which are "shrunk" compared to the estimates $\hat{\boldsymbol{\beta}}_k$ from (1). Indeed the lasso penalty can estimate some components of $\boldsymbol{\beta}_k$ to be exactly 0, thereby achieving variable selection. Note that in this approach, the tuning parameters λ_k are obtained via cross-validation with respect to minimal deviance (i.e., the penalized log-partial likelihood) as defined by Simon, Friedman, Hastie, and Tibshirani (2011). We will refer to this model as "iCS" in our simulations.

As can be seen in equation (5), the prediction accuracy of competing risks models is defined based on the CIF. We discussed earlier that it is necessary to analyze the CSH models for all K causes in order to express the effects on the cumulative incidence. This observation gives rise to the question of whether it would be useful to "couple" or "link" the models to achieve better prediction performance, especially in high-dimensional settings. Therefore, as an extension of the independently penalized CSH models, we consider "linking" CSH models by choosing the optimal tuning parameters λ_k with respect to best prediction of the cumulative incidence.

2.1.2 | Linked CSH models with penalty (penCR)

Suppose we have $K = 2$ event types, then we can fit two "independently" penalized CSH models as described above. These two models each maximize the penalized log-partial likelihood for their cause k using the tuning parameters λ_k from cross-validation on that cause. However, in many applications the competing endpoints are not of equal importance in the sense that there is a prespecified event of interest (e.g., event 1). Fitting two completely separate penalized models based on maximizing the penalized partial likelihood might not necessarily have the best accuracy for predicting the CIF of event 1.

Therefore, we attempt to link the two independently penalized CSH models by choosing the optimal tuning parameters λ_1 and λ_2 with respect to minimal prediction error of the event of interest at a fixed time s. The suggested algorithm works as follows:

1. Set up a grid of lambda values that ranges from the smallest (full model) to the largest (empty model): $\lambda_{kr}, r = 1, \ldots, R$.
2. Partition the data into, for example, 10 folds. For each fold
   - **(a)** use 9 of 10 folds to fit cause-specific penalized regression model from (6) for the cause of interest (e.g. event 1) for all $r = 1, \ldots, R$;
   - **(b)** predict for each patient in the 10th fold the probability of event 1.
3. Calculate the prediction error $PE_1(s)$, in other words, the Brier score for event 1 at time point s (e.g., see Fig. 1).
4. Select the pair $(\lambda_{1r_1^*}, \lambda_{1r_2^*})$ with the smallest average prediction error and fit the final CSH model using the optimal tuning parameters $(\lambda_{1r_1^*}, \lambda_{1r_2^*})$.

We will refer to this model as "penCR." Note that penCR is a far more computationally expensive algorithm than iCS: over an $R \times R$ grid of λ values, iCS fits $2R$ lasso models, whereas penCR fits R^2 models. Figure 1 shows an example of how the Brier score may look for different values of λ_1, λ_2 over a 20-by-20 grid. We observe a region in the upper left corner, where prediction accuracy seems highest for that particular constellation of tuning parameters λ_1, λ_2. In this example, penCR would

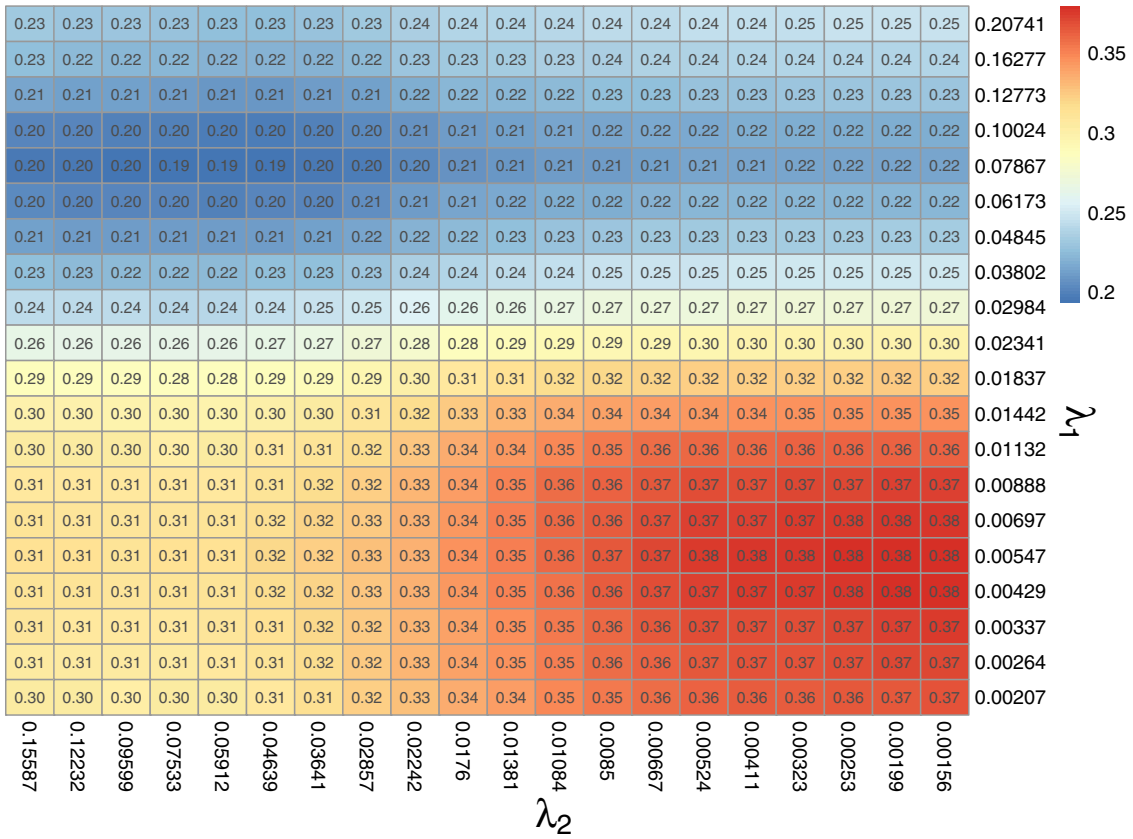

FIGURE 1 Demonstrating the basic idea of the penCR approach: Over a grid of tuning parameters for each CSH model, calculate the prediction error and choose the pair of tuning parameters with best prediction performance. Depicted is an example of the estimated Brier score for event 1 in a simulated competing risks data set with n = 300, p = 506 of which 6 variables are associated with the cause-specific hazards.

choose $\lambda_{1r_1^*} \approx 0.079$ and $\lambda_{1r_2^*} \approx 0.059$ because they together yield the best prediction performance (iCS would choose $\lambda_1 *\approx 0.1$ and $\lambda_2 *\approx 0.096$ in this example). Furthermore, prediction errors are very high in the lower right corner. This is due to the fact that for small values of λ_1, λ_2 less penalization is applied, which leads to larger models that overfit the data.

In addition, one needs to consider the choice of the time point s, with respect to which optimal tuning parameters are searched. It can also be argued that s is itself a tuning parameter for this algorithm and needs to be optimized as well. In this case, some researchers might be tempted to try different values for s and report the best results. This is, of course, not good practice and the choice of s needs to be included in the validation process. However, we advocate to choose s as a clinically/biologically relevant time point (e.g., relapse-free survival within three years from remission). In the absence of such a time point related to the clinical context, it is also possible to use the integrated Brier score rather than choosing an arbitrary time point s. In our simulation experiments, we observed very similar conclusions for penCR using the integrated Brier score (results not shown).

2.2 | Subdistribution hazards based approaches

Several researchers have suggested the use of penalized SDH-based methods to handle competing risks problems in high dimensions. The advancement of the SDH framework is mainly due to the fact that it requires only fitting one model, namely for the event of interest. In a sense, SDH methods achieve "linking" automatically by working directly with the CIF. Therefore, it is natural to compare the performance of iCS and penCR to a lasso-penalized SDH model. However, first, we briefly touch upon currently available approaches in the SDH literature.

Binder, Allignol, Schumacher, and Beyersmann (2009) were the first to suggest SDH models for high-dimensional data using likelihood-based boosting with focus on prediction performance. They establish that boosting is a viable approach to fit proportional SDH models for high-dimensional data while allowing for mandatory covariates (e.g., clinically established predictors). Likelihood-based boosting for competing risks data has been implemented in the `CoxBoost` package by Binder (2013). It has been shown that component-wise updates of the parameter vector (component-wise boosting) result in models, which are very similar to lasso-penalized fits (Bühlmann & Hothorn, 2007; Efron et al., 2004).

In their recent work, Fu, Parikh, and Zhou (2016) examine penalized SDH models with different penalizations such as convex penalties (e.g., lasso, adaptive lasso) as well as nonconvex penalties (e.g., SCAD, MCP). They establish asymptotic properties of the proposed methods including oracle properties (cf. Fan & Li, 2001). These suggestions have been implemented by the same authors in the `crrp` package. Unfortunately, the current version 1.0 is designed for low-dimensional settings and does not work in our simulation scenarios.

In an effort to use a penalized SDH approach comparable to the proposed penalized CSH models (Section 2.1), we employ SDH models with a lasso penalty in our simulations. As mentioned in Section 1.2, one can use standard software to obtain consistent estimates of $\boldsymbol{\gamma}$ by restructuring the event times and status. This has been implemented by Therneau (2015) in the survival package for R. In our simulations we do not have any censoring, therefore it is particularly simple to restructure the data and obtain risk sets as required by Fine and Gray without the need for weights. We can then apply a lasso penalty to the Cox-like SDH model for the restructured data and fit a penalized Fine and Gray model in this manner using standard software, for example, `glmnet` package. In the application example, where censoring is indeed present, we use an imputation approach as suggested by Ruan and Gray (2008). Censoring times are imputed for individuals who experienced the competing event, using the `kmi` package by Allignol and Beyersmann (2010).

In all simulations, we use boosting and lasso penalization on SDH. The results of the two methods are extremely similar in all aspects. Therefore, we do not show results for boosting in our simulations.

3 | SIMULATIONS - PART 1: ASSUMING PROPORTIONAL CSH

Most of the work presented in this paper is based on a simulation study, therefore, the data-generating process determines to great degree the obtained results. One issue when simulating competing risks data is that, in general, we cannot fulfill the PHs assumption for CSH and SDH simultaneously as shown by Beyersmann and Schumacher (2007). This is due to the relationship between CSH and SDH:

$$a_1(t) = \left(1 + \frac{F_2(t)}{P(T > t)}\right) h_1(t). \tag{7}$$

The fact that the SDH can be formulated as a time-dependent proportion of the CSH means that if PH holds for one, it cannot hold for the other, in general. In our simulation design, we need to specify the data-generating process and decide whether the proportionality assumption holds for CSH or SDH, thereby giving a disadvantage to the misspecified model. Grambauer, Schumacher, and Beyersmann (2010) showed that in the case of proportional CSHs, one can still gain useful information from the SDH model in terms of a time-averaged effect on the CIF even though the SDH model is misspecified.

In the first part of the simulation study, we choose to formulate the models using (proportional) CSH. Only in this way, is it possible to formulate and distinguish effect types of different signs and magnitudes on the different transition hazards (see simulation settings for details). In the second part, we generate data using proportional SDH to examine the effects of misspecification. Finally, the third part of the simulation study concerns the scenario, where proportionality holds for the SDH and CSH of cause 1 simultaneously, but is violated for CSH of cause 2.

3.1 | Data-generating process, parameter settings, and evaluation criteria

Competing risks data can be simulated using the approach by Beyersmann, Allignol, and Schumacher (2011). The idea is to generate failure times T from the all-cause hazard $\alpha_{\cdot}(t) = \alpha_1(t) + \alpha_2(t)$ and to run a binomial experiment to decide with probability $\alpha_k(T)/\alpha_{\cdot}(T)$ on event type $k = 1, 2$. We generate competing risks data assuming proportional CSH for $n = 300$ patients with $p \approx 500$ or 2000 variables for $K = 2$ event types. The two events have equal baseline transition hazards $\alpha_{01} = \alpha_{02} = 0.05$, resulting in equal event distributions for both types. This simplification, which might not be reasonable in real-life applications, helps us to get a clearer picture of the strengths and weaknesses of the proposed methods by avoiding issues that arise from unbalanced event frequencies. See Supporting Information for detailed simulation results using unequal baseline hazards.

The cause-specific covariate effects $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix} \underbrace{*, \ldots, *}_{\text{related}}, & \overbrace{0, \ldots, 0}^{500 \text{ or } 2000} \\ \underbrace{*, \ldots, *}_{\text{related}}, & \underbrace{0, \ldots, 0}_{\text{unrelated}} \end{pmatrix}$ are different for each setting. For simplicity, the data contain no censoring and no correlation structure among covariates $X \sim N(0, 1) \in R^{n \times p}$. Results are averaged over 200 simulation runs. In our simulations we use a grid (of manageable size) for penCR, namely 30×30, as the range of possible values for (λ_1, λ_2) and $s = 36$ as time for optimal prediction accuracy. It is generally recommended to choose s as a biologically relevant time point for the clinical data at hand. However, in our simulated data sets, we do not have a biological framework to guide us.

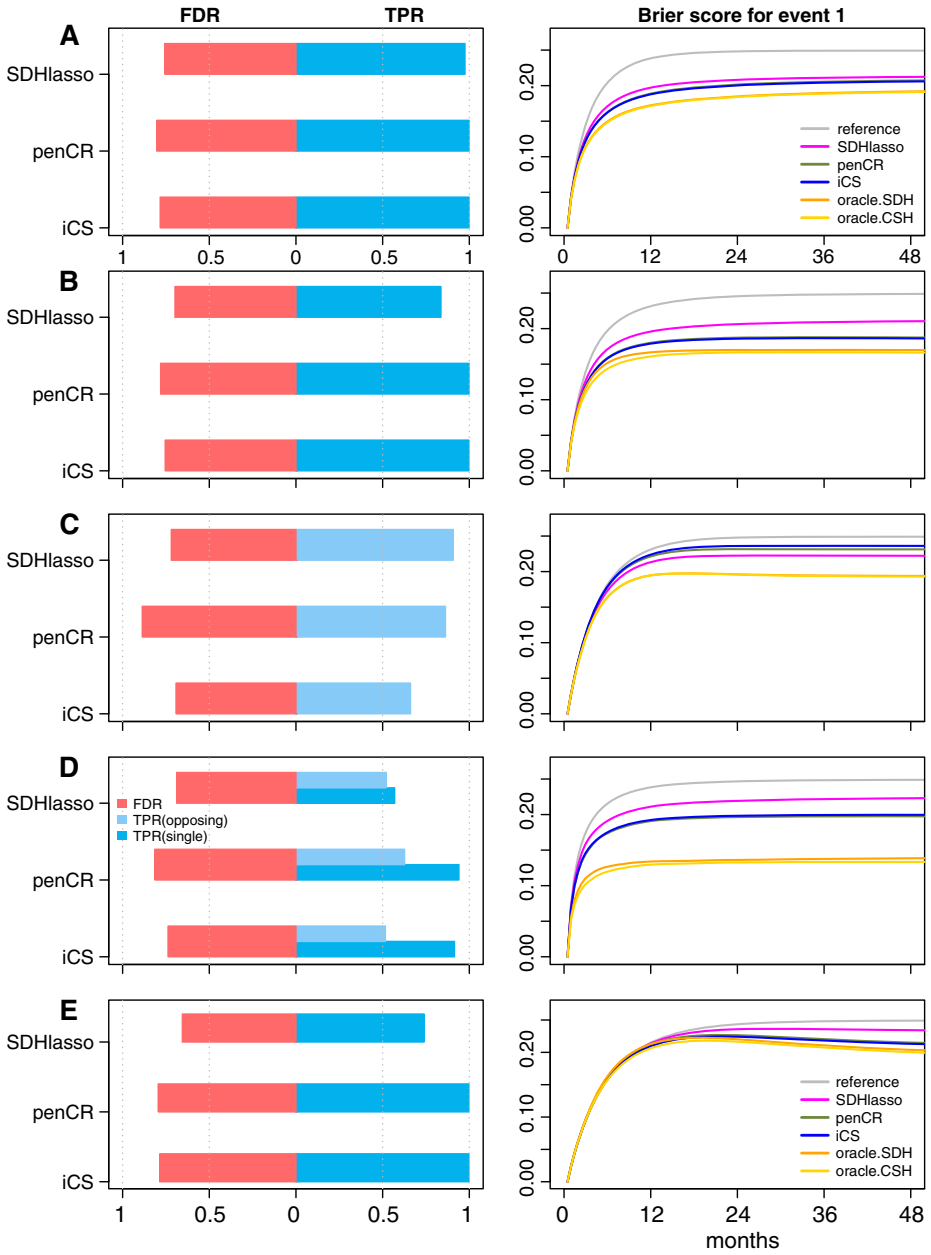

FIGURE 2 Variable selection results (left panel) and prediction accuracy (right panel) assuming proportional CSH for different scenarios: (A) single effects on endpoint of interest, (B) single effects on both endpoints, (C) opposing effects and (D) mixture of effects, and (E) no effects on endpoint of interest

The proposed methods are evaluated with respect to variable selection and prediction accuracy for the event of interest (e.g., event 1) as defined in Section 1.4. Prediction accuracy (a.k.a. prediction of CIF of event 1) is expressed in terms of the observed Brier score on a simulated test data set (again $n = 300$) averaged over 200 simulation runs. In all simulations we can use the "reference" model (see, e.g., right panel of Fig. 2) as an upper bound and the oracle models as a lower bound for prediction accuracy. The "reference" model denotes the nonparametric Aalen–Johansen estimate for the cumulative incidence without any covariates. If a model has prediction error higher than the reference, it overfits the data.

There are two so-called oracle models, one based on CSH and another on SDH. The oracle models represent a lower bound for prediction error in the sense that they use the correct set of relevant variables. However, it is important to note that even though both oracle models use the true set of relevant variables, only one of them correctly models the underlying data-generating process and the other is misspecified. In Section 3, data are generated according to proportional CSH, therefore, the Fine and Gray model is misspecified with a regression parameter that now represents a time-averaged effect (see Beyersmann & Schumacher,

2007). In Section 4, it is the other way around: data are generated according to proportional SDH, thereby violating proportionality of CSHs. Recall that Graf, Schmoor, Sauerbrei, and Schumacher (1999) showed how the Brier score can be decomposed into 'imprecision' (i.e. bias in prediction of the true event probability) and 'inseparability' (variability of event status). Thus, by comparing the oracle models, we can quantify the effect of misspecification in different situations: the imprecision of the two approaches differ due to bias introduced by misspecification, but the inseparability term is equal for both.

Note that in our simulation studies, we fit both oracle models on a large data set ($n = 3000$). Recall from Grambauer et al. (2010) that the estimated coefficients from the Fine and Gray model are asymptotically consistent for the so-called least false parameter (LFP), which is "least false by giving the best approximation within the misspecified model class." Thus, fitting the SDH oracle models on a larger data set is done as a means to ensure near convergence to the LFP.

In order to simulate appropriate situations, we can use the CSH machinery. As mentioned earlier, CSHs describe the driving forces behind observed transitions. This is of particular interest in the analysis of molecular data, where we wish to understand the biological mechanisms that drive the observed transition hazards (e.g., regulation of gene expression or epigenetic alterations). These concepts can be easily studied in CSH framework, where one can observe different types of effect on the CSH for each variable. In the following simulation, we discuss several possible effect types and examine the weaknesses and strengths of CSH and SDH methods in these situations.

Assuming proportional CSH, we consider the following four scenarios:

(a) *Single effects on endpoint of interest.*
In the first scenario, we generate data where six variables are associated with the CSHs of event 1 and no variables associated with event 2, that is, $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\overbrace{0.5,\dots,0.5,}^{6} & \overbrace{0,\dots,0}^{500}\\ 0,\dots,0, & 0,\dots,0\end{pmatrix}$. This situation might arise in a study of cancer patients in remission who may experience relapse or die in remission. Here, we expect some biomarkers to be associated with relapse, but there might be no biomarkers related to death.

(b) *Single effects on both endpoints.*
The second simulation setting is similar to scenario (a) in the sense that there are again six variables associated with event 1. However, we now also have another six variables associated with event 2, that is, $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\overbrace{0.5,\dots,0.5,}^{6} & \overbrace{0,\dots,0,}^{6} & \overbrace{0,\dots,0}^{500}\\ 0,\dots,0, & 0.5,\dots,0.5, & 0,\dots,0\end{pmatrix}$. In this scenario, different sets of relevant variables influence the two hazards separately. This could be the case in the presence of distinct biological pathways related to each endpoint.

(c) *Opposing effects.*
In the third scenario we consider six variables that have moderate, but opposing effects on the two event types, that is, $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\overbrace{+0.25,-0.25,\dots,-0.25,}^{6} & \overbrace{0,\dots,0}^{500}\\ -0.25,+0.25,\dots,+0.25, & 0,\dots,0\end{pmatrix}$. This situation might occur, for example, when considering a cancer therapy with severe side effects and toxicity. Treating patients with such a therapy might decrease their risk of relapse, but may also be associated with higher non-relapse mortality. Note that the chosen effect sizes of ± 0.25 are only of moderate size on each transition, but the absolute difference between transitions is 0.5, similar to setting (a) and (b).

(d) *Mixture of effects.*
In the fourth scenario, we consider a simulation setting where different effect types are present, namely opposing as well as single effects: $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\overbrace{+0.25,-0.25,\dots,-0.25,}^{6} & \overbrace{0.5,\dots,0.5,}^{12} & \overbrace{0,\dots,0,}^{6} & \overbrace{0,\dots,0}^{2000}\\ \underbrace{-0.25,+0.25,\dots,+0.25,}_{\text{opposing}} & \underbrace{0,\dots,0,}_{\text{single variables}} & 0.5,\dots,0.5, & 0,\dots,0\end{pmatrix}$. This scenario aims to mimic the mix of effects present in a molecular data set, while assuming the underlying biological model to be sparse. Here we choose $p = 2000$, which is larger than in the previous settings in order to test the proposed methods for slightly more realistic dimensions. Also, there is more signal in the simulated data in this case (i.e., larger number of relevant variables). Furthermore, we choose more variables to affect the CSH of the event of interest than that of the competing event. This seems plausible in the context of molecular information when the competing event is death, which is rarely associated with many biomarkers.

(e) *No effects on endpoint of interest.*
Finally, we consider a scenario in which there are actually no variables affecting the CSH of the event of interest, but six covariates that influence the CSH of the competing event: $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}\overbrace{0,\dots,0,}^{6} & \overbrace{0,\dots,0}^{500}\\ 0.5,\dots,0.5, & 0,\dots,0\end{pmatrix}$. This situation is similar to scenario (a) when the event of interest is death rather than relapse.

3.2 | Simulation results for part 1

Results from simulations assuming proportional CSH are depicted in Figure 2. The left panel shows the results in terms of variable selection for the proposed methods, the right panel depicts the average prediction error over 200 simulation runs.

(a) In the first scenario (single effects on endpoint of interest) all penalized methods perform similarly well in terms of variable selection, we observe TPRs close to 100% and FDRs of approximately 75%. Regarding prediction accuracy, we can see that the misspecified SDH oracle model predicts as well as the CSH oracle model, which represents the best possible model for the simulated data. This is indication that in this scenario the violation of proportionality in the SDH models is negligible (see Grambauer et al., 2010, and Appendix B for further discussion). The prediction error curves for iCS and penCR are very close to each other and somewhat lower than that of the penalized SDH model.

(b) In the second scenario (single effects on both endpoints), iCS and penCR have very similar FDRs around 75% and are able to identify basically all variables, which affect the CSH of event 1. SDH lasso with an FDR of 70% identifies approximately 84% of true positives that influence the CIF of event 1. The prediction error curves in the right panel indicate a much larger Brier score for the penalized SDH model. Furthermore, in this scenario there is a considerable difference between the oracle models, indicating rather severe consequences of the violation of the proportional hazards assumption of the SDH models (see Appendix B for details). As in scenario (a), penCR shows no advantage in prediction accuracy and variable selection compared to iCS.

(c) In the third scenario (opposing effects), we can see that the iCS model struggles with correctly identifying this type of variables. iCS has a TPR of 66%, whereas penCR performs better with 86%. The SDH lasso method does remarkably well in identifying such variables to have an effect on the cumulative incidence level as shown by the TPR of 91%. In this situation where the effect on each CSH is moderate, but their difference is large our results speak in favor of "linking" the two CSHs, which is naturally the case for the SDH approach and as we have attempted with the penCR suggestion. However, the price paid by penCR is a considerably high FDR of approximately 88%.
Contrary to scenarios (a) and (b), SDH lasso now has a lower prediction error. This gain in prediction accuracy of SDH lasso is even more remarkable considering that it is misspecified! The violation of proportionality for SDH shows no tangible consequences: the prediction error curves of the two oracle models are very close to each other (cf. Appendix B). In some simulation runs penCR showed a slight advantage to iCS, due to the fact that it recognized more relevant variables. However, this advantage does not prevail on average over 200 runs. We conclude that also in this setting penCR and iCS perform similarly in predicting incidence probabilities for event 1. This is likely due to the higher false positive rate of penCR and the small magnitude of effects in this scenario.

(d) In the fourth scenario (mixture of effects) we observe, as expected, a mixture of what we gathered previously. Regarding variable selection: the SDH lasso performs weakest for single effects with a TPR around 57%. Opposing effects are recognized equally poorly by iCS and SDH approaches, TPR around 52%, whereas penCR does slightly better with 62%, but this comes at the price of a higher FDR. Regarding prediction accuracy we observe a picture very similar to scenario (b). It seems that the effects of the single variables outweigh those of the (somewhat weaker) opposing variables such that SDH lasso has considerably poorer prediction performance than the CSH models.

(e) The fifth scenario is one where we would expect SDH lasso to struggle because the CIF of the event of interest is only affected by variables that influence the hazard of the competing event. It is indeed the case that SDH lasso has lower TPR and performs only slightly better than the reference model with respect to prediction accuracy. The CSH approaches, on the other hand, choose larger models (higher TPR and higher FDR), but perform reasonably well (similar to oracle models) in predicting event 1, which is really only governed by the occurrence of the competing event.

For the case of unequal baseline hazards, we observe roughly the same tendencies as in the balanced case. When $\alpha_{02} >> \alpha_{01}$, that is, the competing event is far more prevalent than the event of interest, the SDH approach performs slightly worse than in the balanced case. In particular, it no longer shows an advantage in scenario (c) regarding opposing effects. See Supporting Information figures for detailed results of unbalanced scenarios.

3.3 | "Common" effects

The observant reader may have noticed that we did not consider any effects of the type $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, for example, $\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0.5, \ldots, 0.5, \ 0, \ldots, 0 \\ 0.5, \ldots, 0.5, \ 0, \ldots, 0 \end{pmatrix}$, which we will refer to as "common" effects. There are two reasons, why this simulation setting

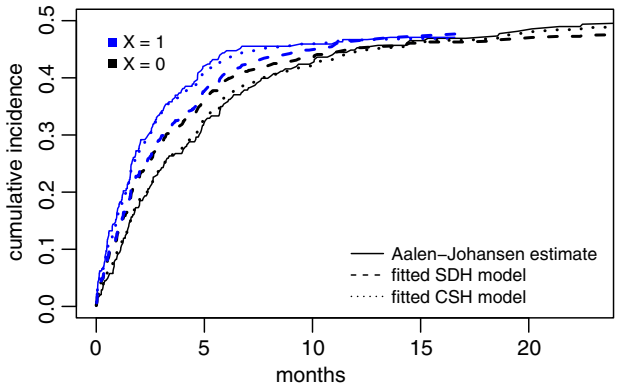

FIGURE 3 Cumulative incidence of event 1 for the two groups of the binary variable X with common effects on both CSHs as fitted non-parametrically using Aalen–Johansen estimator and semiparametrically based on SDH and CSH—Data were generated assuming proportional CSH, leading to a severe violation of proportionality of the SDH model

was excluded. The first has to do with the data-generating process for proportional CSH data, which severely disadvantages the SDH models in the case of "common" effects. Recall, the data are generated by first drawing failure times T from the all-cause hazard $\alpha_{\cdot}(t) = \alpha_1(t) + \alpha_2(t)$, then a binomial experiment decides whether event 1 occurred with probability

$$\frac{\alpha_1(T)}{\alpha_1(T) + \alpha_2(T)} = \frac{\alpha_{01}(T)\exp(X\boldsymbol{\beta}_1)}{\alpha_{01}(T)\exp(X\boldsymbol{\beta}_1) + \alpha_{02}(T)\exp(X\boldsymbol{\beta}_2)} \overset{\beta_1=\beta_2}{=} \frac{\alpha_{01}(T)}{\alpha_{01}(T) + \alpha_{02}(T)}.$$

In other words, common effects (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$) have no effect on *event type* in the binomial step of the data generation, but only on *time of event occurrence*. While this is easily handled by CSH approaches as they work on hazard scales, SDH models have a severe disadvantage by working on incidence probability scales. This is illustrated in Figure 3 , where a single binomial variable X is generated assuming proportional CSH with $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = 0.65$ and constant baseline hazards $\alpha_{01}(t) = \alpha_{02}(t) = 0.1$. We can observe that the CIFs of event 1 for the two groups ($X = 0$ and $X = 1$) start off close to each other and separate rapidly with time. However, they come together again after about five months, indicating that for large t the two groups are no longer distinguishable. The CSH model seems to fit the data quite well as its predicted cumulative incidences are very close to the observed data expressed by the Aalen–Johansen estimate. The SDH model on the other hand has considerable issues in predicting the CIFs.

Naturally, models assuming proportional SDH are at a disadvantage when the assumption of proportionality is not met. But the setting with data generated assuming proportional CSHs and common effects seems to reinforce this issue even more. Therefore this setting was not included in our simulations.

The second reason for not considering common effects further is that they seem rather unusual in clinical settings in the context of molecular data. We were hard pressed to find a real-life example for such an effect in discussions with clinical partners. Naturally, the interpretation and effect types in a competing risks data set are very context-dependent. Such "common" effects could occur when two competing risks endpoints are similar or closely related (e.g., cancer patients in remission may experience local relapse of the tumor or distant metastasis.) However, one possible approach would be to combine two competing endpoints that are very similar.

4 | SIMULATIONS—PART 2: ASSUMING PROPORTIONAL SDH

In order to get a more complete picture of the effects of misspecification and nonproportionality, we also simulate data that are proportional with respect to SDH. Recall that the SDH model considers only the effects on the CIF of the event of interest (here event 1). In this situation, we no longer express the notion of "opposing effects" or "single effects" on the CSHs, that is, the driving forces of the competing risks process, because these do not exist in a SDH-context, but only in a CSH framework. All that is modeled is the effect on the CIF of event 1. This is done as proposed by Fine and Gray (1999), where we assume the CIF for event 1 to follow.

$$P(T_i \leq t, Z_{T_i} = 1 | X_i = x) = 1 - \left(1 - q * (1 - e^{-t})\right)^{\exp(X\gamma_1)}. \quad (8)$$

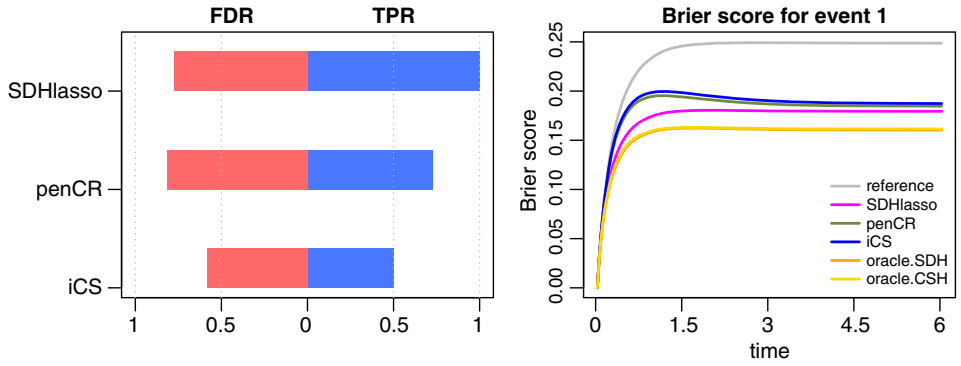

FIGURE 4 Variable selection results (left panel) and prediction accuracy (right panel) for simulations assuming proportional SDH

Let $q \in (0, \min(1, \frac{1}{1-e^{-t}}))$ and let $1-(1-q)^{\exp(X\gamma_1)} = P(Z_{T_i} = 1|X_i = x)$ denote the probability for an individual with covariate x to experience event 1. Furthermore, we assume the following structure for CIF of event 2:

$$P(T_i \leq t, Z_{T_i} = 2|X_i = x) = (1-q)^{exp(X\gamma_1)} \left(1 - e^{-t\exp(X\gamma_2)}\right). \tag{9}$$

Assuming proportional SDH, we generate competing risks data for $n = 300$ patients with $p = 506$ variables for $K = 2$ event types.

The covariate effects vector as used in formulas (8) and (9) is set to $\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = \begin{pmatrix}\overbrace{0.5,\ldots,0.5}^{6}, & \overbrace{0,\ldots,0}^{500}\\ -0.5,\ldots,-0.5, & 0,\ldots,0\end{pmatrix}$.

As before, the data contain no censoring and no correlation structure among covariates $X \sim N(0,1) \in \mathbb{R}^{n\times p}$. Results are averaged over 200 simulation runs.

Once again we compare the proposed models with respect to prediction accuracy and variable selection. Note that in this situation, we have generated effects on the CIFs and not on the CSHs.

Variable selection results are shown in the left panel of Figure 4. We can see that the SDH lasso model has highest TPR, followed by penCR. iCS only identifies 50% of the six relevant variables and has the lowest FDR by choosing smaller models. penCR's gain in TPR is countered by its higher FDR.

In the right panel of Figure 4, we can observe the prediction accuracy in the setting with proportional SDH. As expected, the SDH lasso model has best prediction performance. The CSH-based models are now at a disadvantage due to the violation of proportionality. We observe a slight advantage of penCR in prediction accuracy compared to iCS.

5 | SIMULATIONS—PART 3: ASSUMING PROPORTIONAL SDH AND PROPORTIONAL CSH OF EVENT 1

In previous simulations, the proportionality assumptions were discussed as a crucial factor because misspecification naturally leads to lower performance. In an effort to make "fair" comparisons, we now generate data that satisfy proportionality for SDH as well as for the CSH model for event 1. This can be achieved simultaneously by choosing as particular form for the CSH for event 2, which is not proportional. Following Beyersmann et al. (2011), we generate failure times T from the all-cause hazard $\alpha_{\cdot}(t) = \alpha_1(t) + \alpha_2(t)$ and run a binomial experiment to decide with probability $\alpha_k(T)/\alpha_{\cdot}(T)$ on event type $k = 1, 2$. Using the following CSH models ensures the desired proportionality for SDH as well as for CSH of event 1:

$$\alpha_1(t; X) = \alpha_{01}(t)\exp(X\boldsymbol{\beta})$$

$$\alpha_2(t; X) = \alpha_{02}(t) + \{1 - \exp(X\boldsymbol{\beta})\}(\alpha_{01}(t) - h_{01}(t)).$$

Note that $\boldsymbol{\beta}$ denotes the covariate effects vector for CSH of event 1. In this particular situation, $\boldsymbol{\beta}$ is also the effect on the SDH (see Appendix A for derivation and details). Yet the CSH for event 2 does not satisfy the proportionality assumption, meaning that the CSH approach is still at a slight disadvantage for prediction of the CIF of event 1.

As before, we generate competing risks data for $n = 300$ patients with $p = 506$ variables for $K = 2$ event types. The baseline hazards were chosen to be time-constant: $\alpha_{02}(t) = 0.5, \alpha_{01}(t) = 0.1, h_0(t) = 0.05$ and covariate effects $\boldsymbol{\beta} =$

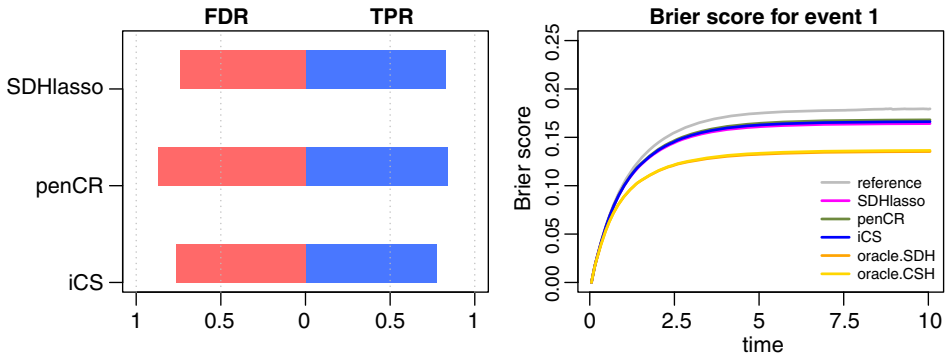

FIGURE 5 Variable selection results (left panel) and prediction accuracy (right panel) for simulations assuming proportional SDH and proportional CSH for event 1

$(.5, .5, .5, -.5, -.5, -.5, \overbrace{0, \ldots, 0}^{500})$. Again the design matrix contains no censoring and no correlation structure $X \sim N(0, 1) \in \mathbb{R}^{n \times p}$. Results are averaged over 200 simulation runs.

Figure 5 shows variable selection results and prediction performance. All methods show comparable TPR and Brier scores. The penCR method has higher FDR as observed in previous simulation experiments. All in all, we can conclude that when we establish "as much proportionality as possible" regarding the event of interest, the iCS approach gives very similar results to the penalized SDH method and performs better than penCR. This indicates that in this case it is not necessary to link the separate models using SDH or penCR.

6 | APPLICATION TO BLADDER CARCINOMA DATA

We apply the proposed methods to a real-life data set of bladder carcinoma patients investigated by Dyrskjøt et al. (2007). This data set has already been used as an example of competing risks data with high-dimensional covariates by Binder, Allignol, Schumacher, and Beyersmann (2009) and Ambrogi and Scheike (2016). Following the data cleaning proposed by Binder et al. (2009), we finally have clinical information for $n = 301$ patients diagnosed with nonmuscle-invasive bladder cancer (and genomic information available for $n = 300$). Of the 300 patients considered in this study, 83 experience progression or death from bladder cancer (event of interest), 33 patients experience death from other causes (competing event), and 184 patients remain event-free. Established clinical risk factors are age, gender, grading (PUNLMP/low vs. high), and stage (pT_a vs. pT_1). These variables are included in all models as mandatory covariates (i.e., in an unpenalized fashion), as is treatment. Preprocessed genomic information is publicly available at GEO (accession no. GSE5479) in the form of 1381 custom platform microarray features, which are included as penalized covariates in the models.

Using this data set, we fit the proposed iCS and penCR models as well as the SDH lasso and the coxboost model as suggested by Binder et al. (2009). In our implementation of lasso-penalized SDH models, we use multiple imputation to account for censoring following Ruan and Gray (2008). Prediction performance is averaged over all imputation runs.

Coxboost is run here with maximal number of boosting steps = 200 and SDH lasso with 50 imputations. The penCR method uses $s = 60$ months as the time with respect to which prediction error is minimized. This choice is based on guidelines of the European Association of Urology that considers, among others, the probability of recurrence at five years. As suggested by Gerds and Schumacher (2007), we calculate the prediction error using the 0.632+ estimate on $B = 500$ bootstrap runs as is depicted in Figure 6. We are able to reproduce the results observed by Binder et al. (2009) and can compare them to results from penalized CSH models. With respect to prediction accuracy, we observe a slight improvement by the regularized models compared to the model including only the clinical variables age, gender, grading, stage, and treatment. The regularized models are very close in terms of prediction error, no method stands out in particular. Also in terms of variable selection, all methods obtained very similar results. Based on the idea of "inclusion frequencies," we ran all analyses 200 times with different seeds and compared the set of variables that were selected at least 100 times. Seven genes were consistently found by all methods. As expected, penCR chose the largest model (33 genes) and contains all but three variables selected by all other methods together. The CSH approaches share six variables that were not selected by the SDH methods.

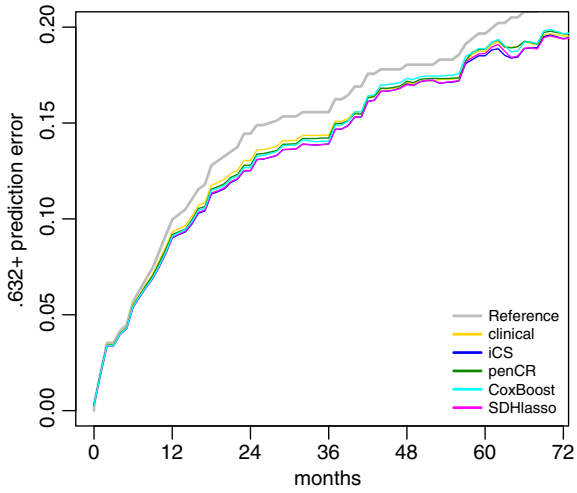

FIGURE 6 0.632+ prediction error curves for bladder carcinoma data—All regularized methods include clinical variables as mandatory covariates and genomic information in a penalized fashion

7 | DISCUSSION

We set out to study penalized competing risks models based on CSHs. CSH methods have been understudied as most of the current literature in this field focuses on SDH. We think this a relevant question, because CSH have some conceptual advantages. As pointed out by Prentice et al. (1978), CSHs are the natural building blocks for competing risks modeling. CSHs can be regarded as the quantities of choice for analyzing microarray data with the goal of understanding underlying biological mechanisms. This is because CSHs allow us to differentiate between effects on different event types. Finally, penalized CSHs can offer extensions for multistate modeling in high dimensions.

The downside of the CSH approach is that it requires fitting several models, namely one for each of the K causes. This is probably why SDH approaches are popular, particularly in high dimensions. They only require fitting one model and in a sense naturally link the CSHs by working on incidence level. Furthermore, SDH approaches have proven useful for prediction purposes, but they suffer from interpretational challenges (cf. Andersen and Keiding, 2012).

When working in high dimensions one could use penalization techniques, such as lasso or boosting. In the context of penalized CSH models, one would then fit K separate penalized CSH models (referred to as iCS). One of our main concerns was that fitting these separate models in high dimensions might be prone to overfitting or possibly be inefficient with respect to prediction. To address these concerns, we also investigated an extension of independently penalized CSH models in an effort to "link" the models. We chose to do this with respect to prediction accuracy, because prediction error is defined based on the cumulative incidence of the event of interest. In our extension, named penCR, tuning parameters were chosen as the pair that minimizes Brier score at time s. We compared the performance of the independent and the linked CSH model to the natural competitor, namely penalized SDH.

Our choice to focus on prediction accuracy for the event of interest at a clinically relevant time s is motivated by research questions relating to cancer (e.g., three-year progression-free survival). However, one could certainly envision other versions. If there is no clinically important time, one could consider using the integrated Brier score. This version gave very similar findings in our simulation studies (results not shown) and can be used as an alternative. In the absence of an event of interest, one could also consider optimizing the prediction error with respect to both/all CIFs. The proposed version of penCR is therefore only one of many ways to "link" the independent CSH models.

Our simulation studies indicate that in many aspects the iCS approach is competitive to penCR and lasso-penalized SDH, in particular when baseline hazards of both event types are of similar magnitude or the competing event is more prevalent. Some particular instances speak in favor of linking the CSH models via penCR or using SDH. For example, if there is particular interest in detecting variables with effects on both endpoints (such as opposing effects). Such variables might have interesting biological meaning and can be better detected by linking the CSHs. While penCR has the advantage of interpretability in terms of transition hazards, it often shows a higher FDR. The SDH approach, on the other hand, frequently choses smaller models, but lacks interpretability on hazard level. We also discussed misspecification, violation of proportionality assumptions and the consequences in different scenarios. In the interesting setting where proportionality is fulfilled for SDH and CSH of event 1

simultaneously, we observe very similar performance of the different methods. This was also the case in our application example for the analysis of gene expression levels of bladder carcinoma patients.

In summary, our results lead us to believe that penalized CSH models are a viable solution for competing risks models in high dimensions because they allow for interpretation of the transition hazards and have prediction accuracy similar to the SDH method. Linking the CSHs might be useful in some particular cases, but simple models using separately penalized CSH are often justified. These finding can be relevant in future work when high-dimensional competing risks models are extended to multistate models.

REFERENCES

Allignol, A., & Beyersmann, J. (2010). Software for fitting nonstandard proportional subdistribution hazards models. *Biostatistics*, *11*.

Allignol, A., Beyersmann, J., & Schmoor, C. (2016). Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Statistics*, *15*, 297–305.

Ambrogi, F., & Scheike, T. H. (2016). Penalized estimation for competing risks regression with applications to high-dimensional covariates. *Biostatistics*, *17*, 708–721.

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

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*, 91–115.

Andersen, P. K., & Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, *31*, 1074–1088.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., & Mansmann, U. (2010). High-dimensional Cox models: The choice of penalty as part of the model building process. *Biometrical Journal*, *52*, 50–69.

Beyersmann, J., Allignol, A., & Schumacher, M. (2011). *Competing risks and multistate models with R*. Springer. Springer Verlag New York.

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

Beyersmann, J., & Schumacher, M. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine*, *26*, 1649–1652.

Binder, H. (2013). CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks. R package Version 1.4. https://CRAN.R-project.org/package=CoxBoost.

Binder, H., Allignol, A., Schumacher, M., & Beyersmann, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, *25*, 890–896.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *4*, 477–505.

Dyrskjøt, L., Zieger, K., Real, F. X., Malats, N., Carrato, A., (…) Orntoft TF. (2007). Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: A multicenter validation study. *Clinical Cancer Research*, *13*, 3545–3551.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

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

Fu, Z., Parikh, C. R., & Zhou, B. (2016). Penalized variable selection in competing risks regression. *Lifetime Data Analysis*, *23*: 353–376.

Gerds, T. A., & Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, *63*, 1283–1287.

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

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

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

Mell, L. K., & Jeong, J.-H. (2010). Pitfalls of using composite primary end points in the presence of competing risks. *Journal of Clinical Oncology*, *28*, 4297–4299.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr., A. V., Flournoy, N., Farewell, V., & Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, *34*, 541–554.

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

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

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

Schoop, R., Beyersmann, J., Schumacher, M., & Binder, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, *53*, 88–112.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*, 1–13.

Therneau, T. M. (2015). A package for survival analysis in S. R package Version 2.38. https://CRAN.R-project.org/package=survival.

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

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Saadati M, Beyersmann J, Kopp-Schneider A, Benner A. Prediction accuracy and variable selection for penalized cause-specific hazards models. *Biometrical Journal*. 2018;60:288–306. https://doi.org/10.1002/bimj.201600242

APPENDIX A

We investigate in which situation proportionality can be fulfilled for SDH and the CSH of event 1 simultaneously.

Assume two competing events, where event 1 is of primary interest. The time-dependent relationship between SDH and CSH for event 1, which can be written as

$$h_1(t) = \underbrace{\frac{S(t)}{1 - F_1(t)}}_{=:r(t)} \alpha_1(t),$$

where $S(t) = P(T > t) = 1 - F_1(t) - F_2(t)$ denotes the survival function. Assuming proportional SDH and proportional CSH for event 1, we can fit the following models:

$$h_1(t; X) = h_{01}(t) \exp(X\gamma) \quad \text{and} \quad \alpha_1(t; X) = \alpha_{01}(t) \exp(X\beta_1).$$

In this case,

$$r(t; X) = \frac{\exp(-A(t; X))}{\exp(-H(t; X))},$$

where $A(t; X) = A_1(t; X) + A_2(t; X)$ denotes the all-cause hazard and $H(t; X)$ the cumulative SDH.

The log-hazard ratio for CSH of event 1 is thus

$$\exp(X\boldsymbol{\beta}_1) = \frac{\alpha_1(t; X)}{\alpha_{01}(t)} = \frac{h_1(t; X)/r(t; X)}{h_1(t; 0)/r(t; 0)} = \frac{r(t; 0)}{r(t; X)} \exp(X\boldsymbol{\gamma})$$

for any t and any X.

Consider $t = 0$ then $r(t = 0; X) = 0$ for any X and therefore $\boldsymbol{\beta}_1 = \boldsymbol{\gamma}$, meaning the covariate effect on the CSH of event 1 is equal the effect on the SDH! This is an interesting observation, which leads us to consider how $\alpha_2(t)$ would look if we set $\frac{r(t;0)}{r(t;X)} = 1$ in general while assuming proportional SDH and proportional CSH for event 1.

Let $\frac{r(t;0)}{r(t;X)} = 1$ for all any t and any X, then $\boldsymbol{\beta}_1 = \boldsymbol{\gamma} =: \boldsymbol{\beta}$ and

$$\begin{aligned} 0 &= -\log\left(\frac{r(t;X)}{r(t;0)}\right) = \\ &= A(t;X) - H(t;X) - A(t;0) + H(t;0) = \\ &= A_2(t;X) + A_{01}(t;X) - A_2(t;0) - A_{01}(t;0) - \{H(t;X) - H(t;0)\} = \\ &= A_2(t;X) - A_2(t;0) + A_{01}(t;0)\{\exp(X\boldsymbol{\beta}) - 1\} - H(t;0)\{\exp(X\boldsymbol{\beta}) - 1\} \end{aligned}$$

Note that the above equation is 0 for all t, therefore it is constant in t, and we conclude that the derivative with respect to t equals 0. Thus

$$0 = \alpha_2(t;X) - \alpha_2(t;0) + \alpha_{01}(t)\{\exp(X\boldsymbol{\beta}) - 1\} - h_0\{\exp(X\boldsymbol{\beta}) - 1\}.$$

We conclude that when

$$\alpha_2(t;X) = \alpha_{02}(t) + \{1 - \exp(X\boldsymbol{\beta})\}(\alpha_{01}(t) - h_{01}(t))$$

both proportional SDH and proportional CSH for event 1 are fulfilled. This observation gives us the tool required to generate competing risks data that fulfill proportionality for the Fine and Gray model and the CSH for event 1 simultaneously. This comes at the price of "sacrificing" the proportionality of CSH for the competing event. We also note that to ensure positive values for $\alpha_2(t;X)$ a complex interplay of baseline hazards and covariate effects arises. For example, one could use a rather large $\alpha_2(t;0)$ (baseline hazard for event 2) or small effect sizes. In our simulations in Section 5, we used baseline hazards $\alpha_2(t;0) = .5, \alpha_1(t;0) = .1, \gamma(t;0) = .05$ and covariate effects $\boldsymbol{\beta} = (.5, .5, .5, -.5, -.5, -.5, \overbrace{0, \ldots, 0}^{500})$ Another interesting observation concerns the hazard ratio for event 2:

$$\frac{\alpha_2(t;X)}{\alpha_{02}(t)} = 1 + \frac{\{1 - \exp(X\boldsymbol{\beta})\}(\alpha_{01}(t) - h_{01}(t))}{\alpha_{02}(t)}.$$

Thus $\frac{\alpha_2(t;X)}{\alpha_{02}(t)} \approx 1$ when the second summand is small. In our simulations from Section 3 in setting (a), we modeled single effects on the endpoint of interest. In essence, we set $\frac{\alpha_2(t;X)}{\alpha_{02}(t)} = 1$ and observed very similar performance of the penalized Fine and Gray and the proportional CSH model. This can be explained by the fact that the proportionality was nearly fulfilled for SDH and CSH of event 1 simultaneously with similar effect size as we now understand.

APPENDIX B

In this section, we explore the estimate obtained by the Fine and Gray model in the presence of misspecification. As was done in the first part of our simulations, we again assume the true CSHs to be proportional. We know that the corresponding SDH are then time-dependent. As pointed out by Grambauer et al. (2010), the Fine and Gray model fitted to this type of data estimates a time-averaged effect of the SDH, which converges in probability to the least false parameter. We will examine this issue in a situation that was not studied by Grambauer et al., namely in the presence of "single effects on both endpoints," see next.

Recall from formula (7) the time-dependent relationship between SDH and CSH, which can be equivalently written as

$$h_1(t) = \frac{S(t)}{1 - F_1(t)}\alpha_1(t),$$

where $S(t) = P(T > t) = 1 - F_1(t) - F_2(t)$ denotes the survival function. Both hazards notions are modeled in a Cox-like fashion, namely

$$h_1(t) = h_{01}(t)\exp(X\boldsymbol{\gamma}) \quad \text{and} \qquad \alpha_1(t) = \alpha_{01}(t)\exp(X\boldsymbol{\beta}_1).$$

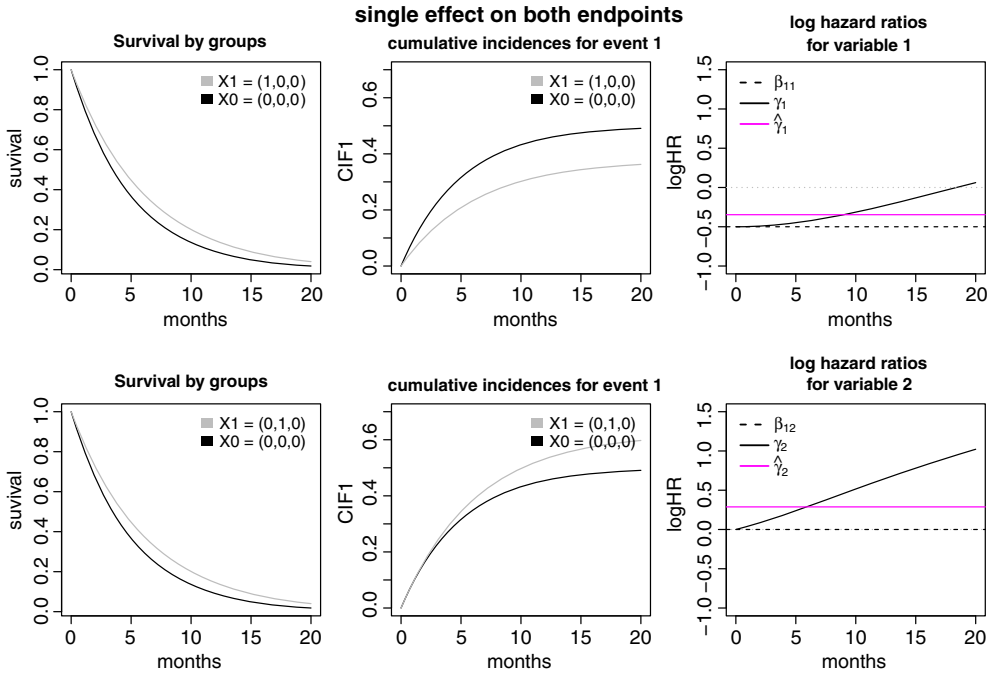

FIGURE 7 Situation depicting results for single effects on both endpoints, that is, $\beta = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \end{pmatrix}$—top row: survival curves, CIF for event 1 by groups $\boldsymbol{x}_1 = (1,0,0)$ versus $\boldsymbol{x}_0 = (0,0,0)$ and log-hazard ratios for variable 1; second row: now considering groups $\boldsymbol{x}_1 = (0,1,0)$ versus $\boldsymbol{x}_0 = (0,0,0)$ and log-hazard ratios for variable 2

Therefore, the subdistribution hazard ratio can be expressed in terms of the CSH ratio:

$$\frac{h_1(t, X = \boldsymbol{x}_1)}{h_1(t, X = \boldsymbol{x}_0)} = \frac{S(t, X = \boldsymbol{x}_1)}{S(t, X = \boldsymbol{x}_0)} \frac{1 - F_1(t, X = \boldsymbol{x}_0)}{1 - F_1(t, X = \boldsymbol{x}_1)} \frac{\exp(\boldsymbol{x}_1 \boldsymbol{\beta}_1)}{\exp(\boldsymbol{x}_0 \boldsymbol{\beta}_1)}.$$

Suppose we have $p = 3$ variables with effect sizes $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix} \in \mathbb{R}^{2 \times p}$ and the corresponding effects on the SDH expressed in terms of $\boldsymbol{\gamma} \in \mathbb{R}^p$. The following examples use binary covariates $X \in \{0,1\}^p$ because they allow for simple interpretation of observed hazard ratios. Furthermore, we consider complete data (i.e., no censoring), as a simplification of the work done by Grambauer.

- We can express the SDH ratio for the first component of $\boldsymbol{\gamma}$ in terms of β_{11} by letting $\boldsymbol{x}_1 = (1,0,0)$ and $\boldsymbol{x}_0 = (0,0,0)$:

$$\exp(\gamma_1) = \frac{h_1(t, X = \boldsymbol{x}_1)}{h_1(t, X = \boldsymbol{x}_0)} = \frac{S(t, X = \boldsymbol{x}_1)}{S(t, X = \boldsymbol{x}_0)} \frac{F_1(t, X = \boldsymbol{x}_0)}{F_1(t, X = \boldsymbol{x}_1)} \exp(\beta_{11}).$$

- On the other hand, choosing $\boldsymbol{x}_1 = (0,1,0)$ and $\boldsymbol{x}_0 = (0,0,0)$ allows us to express the SDH ratio for the second component of $\boldsymbol{\gamma}$ in terms of β_{12}:

$$\exp(\gamma_2) = \frac{h_1(t, X = \boldsymbol{x}_1)}{h_1(t, X = \boldsymbol{x}_0)} = \frac{S(t, X = \boldsymbol{x}_1)}{S(t, X = \boldsymbol{x}_0)} \frac{F_1(t, X = \boldsymbol{x}_0)}{F_1(t, X = \boldsymbol{x}_1)} \exp(\beta_{12}).$$

Our goal is to examine situations in which the time-averaged SDH ratio gives an adequate summary of the true SDH ratio. To obtain $\hat{\boldsymbol{\gamma}}$, we generate a simulated data set ($n = 10{,}000$) according to the desired CSHs and fit the Fine and Gray model. On this large data set, we expect $\hat{\boldsymbol{\gamma}}$ to be very close to the least false parameter, which is "the best approximation within the misspecified model class (here: proportional SDH model)" (Grambauer et al., 2010).

Consider the following three situations for different effect types on the CSH:

(1) *single effects on both endpoints:* $\beta = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \end{pmatrix}$;

(2) *single effect on only the endpoint of interest:* $\beta = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$;

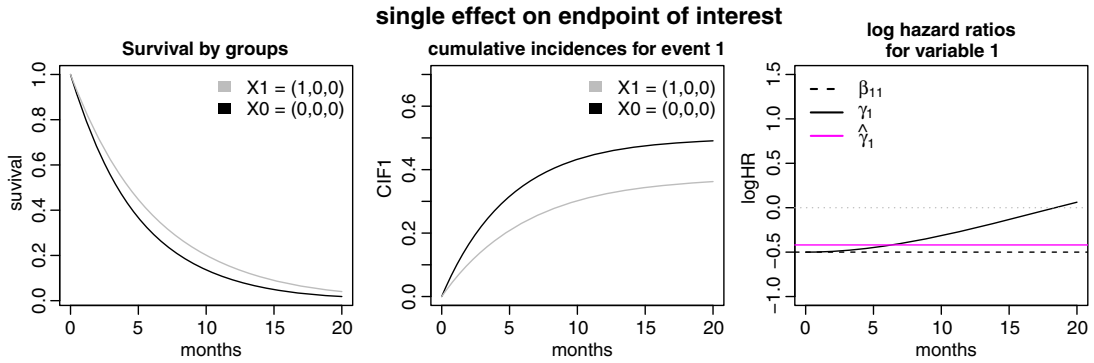

FIGURE 8 Reproduction of results from Grambauer et al. for situation with a single effect only on the endpoint of interest: $\beta = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

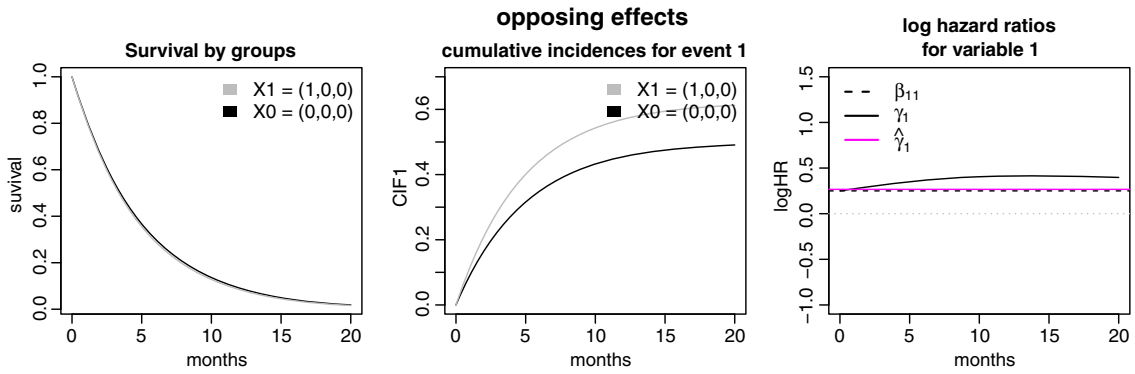

FIGURE 9 Reproduction of results from Grambauer et al. for situation with opposing effects: $\beta = \begin{pmatrix} 0.25 & 0 & 0 \\ -0.25 & 0 & 0 \end{pmatrix}$

(3) *opposing effects:* $\beta = \begin{pmatrix} 0.25 & 0 & 0 \\ -0.25 & 0 & 0 \end{pmatrix}$.

Grambauer et al. studied, among other scenarios, situations (2) and (3). We reproduce their results in Figures 8 and 9 for illustrative comprehensiveness, however, we are mainly interested in setting (1).

Figure 7 depicts the results for setting (1), where there are single effects on both endpoints. Observe that there is a clear increase in the log-SDH ratio for larger times t in both variables. The estimated log-SDH ratio for variable 1, denoted as $\hat{\gamma}_1$, is very close to the log-CSH ratio β_{11} in the early time interval $t \in [0, 7.5]$. This time span is where the a large proportion of observed events occur (overall survival rates ≈30%).

By contrast, the estimated log-SDH ratio for variable 2, denoted by $\hat{\gamma}_2$, does not remain close to the log-CSH ratio β_{12} during this time interval, where the bulk of the time-to-event information is located. We observe that the time-averaged effect $\hat{\gamma}_2$ estimated in the Fine and Gray model is not a good approximation of γ_2. It is therefore understandable that the Fine and Gray model has worse prediction accuracy than the CSH model in this situation.

Figures 8 and 9 reproduce part of the results published by Grambauer et al., where there is only one variable associated with the CSHs. We observe that the time-averaged estimate of $\boldsymbol{\gamma}$ produced by the Fine and Gray model is a reasonable approximation to the true log-SDH ratio γ_1, particularly in the early time interval where most events take place. In these situations, we expect similar prediction accuracies for the CSH model and the (misspecified) SDH model.