A review on statistical and machine learning competing risks methods

Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

A review on statistical and machine learning competing risks methods

Karla Monterrubio-Gómez[*1], **Nathan Constantine-Cooke** [1,2], and **Catalina A. Vallejos**[**1,3]

[1] MRC Human Genetics Unit, University of Edinburgh, Edinburgh, United Kingdom

[2] Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

[3] The Alan Turing Institute, London, United Kingdom

When modelling competing risks survival data, several techniques have been proposed in both the statistical and machine learning literature. State-of-the-art methods have extended classical approaches with more flexible assumptions that can improve predictive performance, allow high dimensional data and missing values, among others. Despite this, modern approaches have not been widely employed in applied settings. This article aims to aid the uptake of such methods by providing a condensed compendium of competing risks survival methods with a unified notation and interpretation across approaches. We highlight available software and, when possible, demonstrate their usage via reproducible R vignettes. Moreover, we discuss two major concerns that can affect benchmark studies in this context: the choice of performance metrics and reproducibility.

Key words: Competing risks; Survival analysis; Time-to-event data; Risk prediction.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX

1 Introduction

Survival analysis comprises a collection of methods to model the time until an event of interest occurs. Usually, the goal is to estimate the risk of observing the event by a given time or to quantify the relationship between event risk and known covariates. Survival methods are widely used in several fields; including medicine, social sciences, engineering and economics. Survival methods have been reviewed by Cox and Oakes (1984), Carpenter (1997), Klein and Moeschberger (2006) and, more recently, Wang et al. (2019).

A typical element of survival data is *censoring*, where event times are unknown. This can occur for several reasons, e.g. lost of follow-up. Survival methods such as the popular Cox proportional hazards (CPH) model (Cox, 1972) often assume independent censoring: those who were censored at a specific time are representative of all those who remained at risk.

In some cases, a subject can experience more than one type of mutually exclusive events — typically referred to as *competing risks* (CR). For instance, a patient can die from different causes (e.g. cancer or non-cancer death). If the main focus is a specific event type, others could be recorded as censored observations. However, the independent censoring assumption does not hold in this setting: if one event occurs, the others are no longer possible. This can lead to biased estimates in standard models (Austin et al., 2016).

The development of CR survival models is an active area of research (e.g. Ng and McLachlan, 2003; Ishwaran et al., 2014; Lee et al., 2018; Nemchenko et al., 2018; Dauda et al., 2019; Sparapani et al., 2020), but state-of-the-art approaches have not been widely adopted in applied settings. This may be because papers are not aimed for practitioners, or due to lack of clear benchmarks that highlight the strengths and drawbacks of each method. The lack of (open-source) software can also prevent wide adoption. As a result,

*Corresponding author: e-mail: karla.monterrubio-gomez@ed.ac.uk

**Corresponding author: e-mail: catalina.vallejos@ed.ac.uk

real-world CR applications have primarily made use of long-established methods (e.g. Fine and Gray, 1999), leaving the application of more modern methodologies often limited only to academic exercises.

The purpose of this review is to summarise the current landscape of CR approaches, including methods developed by two overlapping but still distinctive communities; namely, statistics and machine learning. We aim to unify the notation and interpretation across methods, facilitating their comparison. To aid the uptake of state-of-the-art tools, we highlight available software and, when possible, demonstrate their use via reproducible `R` vignettes (see `www.github.com/VallejosGroup/CompRisksVignettes`). We also discuss common issues encountered when evaluating new CR methods; such as reproducibility and the choice of performance metrics. We highlight that this review focuses solely in traditional CR methods for the purpose of risk prediction or to quantify the association between covariates and event risk. Other important topics such as joint modelling of CR time-to-event and longitudinal data (e.g. Williamson et al., 2008; Andrinopoulou et al., 2014; Hickey et al., 2018), and causal inference in the presence of CR (e.g. Rudolph et al., 2020; Syriopoulou et al., 2022) are not covered here.

2 Background

Consider a continuous random variable, $T \geq 0$, defined as the time until which an event of interest occurs. Let $f(t)$ be the probability density function for T. Often, survival models are specified via the survival function $S(t) = \Pr(T > t) = \int_t^\infty f(t)\, dt$ or the hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}, \tag{1}$$

i.e. the instantaneous rate, given that no event has occurred by time t. A variety of parametric and non-parametric methods exist when a single event type can occur. In the presence of multiple event types (e.g. cancer/non-cancer death), a composite event (e.g. all cause mortality) can be defined at the cost of reduced data granularity. Instead, CR survival models can explicitly capture different event types. Here, we focus on mutually exclusive events: any event prevents the others. If one event prevents others but not vice-versa (e.g. myocardial infarction and death), If one event prevents others but not vice-versa (e.g. myocardial infarction and death), semi-CR (Fine et al., 2001; Peng and Fine, 2007; Hsieh et al., 2008) or illness-death models (Andersen et al., 2002; Meira-Machado and Sestelo, 2019; Xu et al., 2010) may be used. More generally, multi-state approaches (Hougaard, 1999) may be required when multiple events (as defined by the transition between different states, e.g. healthy to ill to relapse to death) may be observed.

2.1 Competing risks survival models

Assume K event types and let $Z \in \{1, \ldots, K\}$ be a random variable representing the observed type of event (as a convention, $Z = 0$ is also typically used to denote censoring). Different frameworks have been used to define CR survival models. First, using the cause-specific (CS) hazard function, which quantifies the instantaneous rate for the k-th event type for subjects that have not experienced *any* event:

$$h_k^{\text{CS}}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t, Z = k \mid T > t)}{\Delta t}. \tag{2}$$

The overall hazard in (1) is the sum across all CS hazards, i.e. $h(t) = \sum_{k=1}^{K} h_k^{\text{CS}}(t)$.

Alternatively, CR survival models can also be defined via the cumulative incidence function (CIF):

$$\text{CIF}_k(t) = \Pr(T \leq t, Z = k), \tag{3}$$

i.e. the probability of observing the k-th event type before time t (and prior to other events) or the sub-distribution hazard function (often referred to as the Fine-Gray hazard, Gray, 1988)

$$h_k^{\text{FG}}(t) = -\frac{d \log(1 - \text{CIF}_k(t))}{dt} = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t, Z = k \mid T > t \cup (T < t \cap Z \neq k))}{\Delta t},$$

(4)

quantifying the instant rate of the k-th event for subjects that have not had *that* event by time t, but including those who experienced a competing event. Note that, the key difference between (2) and (4) is the risk set used to define the probability. In addition, the one-to-one correspondence between $h_k^{\text{FG}}(t)$ and $\text{CIF}_k(\cdot)$ (see equation (4)) is not valid for the k-th CS hazard in (2). Indeed, $\text{CIF}_k(\cdot)$ depends on the CS hazard functions for *all* event types. This relationship is given by:

$$\text{CIF}_k(t) = \int_0^t h_k^{\text{CS}}(s) S(s)\, ds, \quad \text{where} \quad S(t) = \exp\left[-\sum_{k=1}^{K}\left(\int_0^t h_k^{\text{CS}}(s)\, ds\right)\right] \tag{5}$$

Finally, latent failure times CR models assume $T = \min\{T_1, \ldots, T_K\}$ and $Z = \text{argmin}_k\{T_k\}$, where T_k is an event-specific time which is unobserved, unless $Z = k$. Such models are typically defined through the joint survival function $S_{T_1,\ldots,T_K}(t_1, \ldots, t_K) = \Pr(T_1 > t_1, \ldots T_K > t_K)$. However, the marginal distributions of the latent times T_k are non-identifiable, unless non testable assumptions (e.g. independence between T_k's (Cox, 1962; Tsiatis, 1975) or that dependency arises through a known copula (Zheng and Klein, 1995)) are made. This non-identifiability and the usage of non-testable assumptions can make this class of models difficult to interpret (see e.g. Andersen and Keiding, 2012).

2.2 Regression models for CR survival data

Often the aim is to quantify how a set of covariates (features) affects CR outcomes, or to use such covariates in order to predict the risk associated to different event types. Assume we have observations $\{(T_i, Z_i), i = 1, \ldots, n\}$, where T_i and Z_i represent the event time and event type for the i-th subject. Let $\mathbf{x}_i \in \mathbb{R}^p$ be a p-dimensional vector of features for subject i. As in Cox (1972), a regression model can be defined via the CS hazard functions in (2) as

$$h_k^{\text{CS}}(t_i \mid \mathbf{x}_i) = h_{k0}^{\text{CS}}(t_i) \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}_k\right), \quad k = 1, \ldots K, \tag{6}$$

where $h_{k0}^{\text{CS}}(\cdot)$ is a CS baseline hazard and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots \beta_{kp})^\top$ a vector of covariate effects such that $\exp(\beta_{kj})$ is the relative change in the CS hazard linked to a unit change in the j-th covariate. Inference can be done by fitting K separate CPH models where competing events are treated as censored observations. As in Cox (1972), an estimate of $h_{k0}^{\text{CS}}(\cdot)$ is not required to infer $\boldsymbol{\beta}_k$ (which can be derived from the partial likelihood; **Appendix A**). However, $h_{k0}^{\text{CS}}(\cdot)$ is required to perform prediction. For this purpose, the estimator in Breslow (1972) or a parametric model (e.g. Weibull) can be used.

Fine and Gray (1999) developed an alternative approach based on the sub-distribution hazard function in (4). Analogous to (6), this is defined by

$$h_k^{\text{FG}}(t_i \mid \mathbf{x}_i) = h_{k0}^{\text{FG}}(t_i) \exp\left(\mathbf{x}_i^\top \boldsymbol{\gamma}_k\right), \quad k = 1, \ldots K, \tag{7}$$

where $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kp})^\top$ is a vector of covariate effects estimated using the inverse probability weighting (Robins and Rotnitzky, 1992). The sign of γ_{kj} indicates whether an increase in the j-th covariate is associated with an increase/decrease in the incidence of the event, but γ_{kj} does not measure effect sizes on the probability of the occurrence of the event. Note that $\boldsymbol{\gamma}_k$ is not in the same scale as $\boldsymbol{\beta}_k$ (which quantifies how differences in covariate values translate to differences in the CS hazard functions); thus, one should be cautious when comparing their values (Austin and Fine, 2017). Moreover, the proportional hazards assumption cannot simultaneously hold in (6) and (7) (Grambauer et al., 2010).

Due to (4), (7) is often referred as a CIF regression model and can be re-written as:

$$\log\left[-\log\{1 - \text{CIF}_k(t_i \mid \mathbf{x}_i)\}\right] = \log\left[-\log\{1 - \text{CIF}_{k0}(t_i)\}\right] + \mathbf{x}_i^\top \boldsymbol{\gamma}_k, \quad k = 1, \ldots, K, \tag{8}$$

where $\text{CIF}_{k0}(\cdot)$ is the baseline CIF for the k-th event (all covariate values equal to zero). This can be interpreted as a Generalized Linear Model (GLM) with a complementary log-log link function. The CIF regression approach is better suited than (6) when developing risk prediction models (Austin et al., 2016). However, one limitation is that, for certain covariate and time specifications, the sum of the K estimated CIFs may exceed 1 (Austin et al., 2021).

Regression models based on latent failure times also exist. For example, under independence, an *accelerated failure time* (AFT) model (Kalbfleisch and Prentice, 2011) can be used for each latent time. Let $\log(T_{ik})$ be the k-th latent time for subject i. The AFT model can be defined as

$$\log(T_{ik}) = \mathbf{x}_i^\top \boldsymbol{\nu}_k + \varepsilon_{ik}, \tag{9}$$

where $\boldsymbol{\nu}_k = (\nu_{k1}, \dots \nu_{kp})^\top$ is a vector of regression parameters and ε_{ik} are independent and identically distributed errors. Depending on the error distribution, several parametric models can be obtained (e.g. Weibull or log-Normal). In addition, in the case of dependent latent failure times, Heckman and Honoré (1989) provide identifiability conditions for both PH and AFT models.

3 Recent advances on competing risks survival models

Whilst the approaches described in Section 2.2 have been successfully used in a wide range of applications, they do not always provide the flexibility required for specific use cases. More recent, CR methods have introduced flexibility in terms of non-linear covariate effects, time varying covariates, variable selection, missing data, and scalability, among others. Here, we summarise such approaches. Previous reviews in this area (e.g. Zhang et al., 2008; Haller et al., 2013) have primarily focused on the statistics literature. Instead, we provide a more comprehensive survey which covers recent contributions by the machine learning community. As the boundary between these disciplines is diffuse (Bzdok et al., 2018), we do not explicitly distinguish them. Instead, methods are grouped based on the specifications discussed in Section 2.2.

Table 1 summarises the methods included in this review where we highlight differences in terms of:

1. **Type**: is the model defined by a finite (small) number of parameters (parametric)? Or is it semi-parametric (e.g. non-parametric baseline hazard and parametric covariate effects)? Or non-parametric?
2. **Proportional Hazards (PH)**: does the method assume a PH specification in terms of either the CS hazard function (6) or the sub-distribution hazard function (7)?
3. **High dimensional**: can the method be applied to datasets with high-dimensional covariates ($n < p$), performing either feature selection (or regularization) or dimensionality reduction?
4. **Missing data**: does the method support missing covariate values and/or missing outcomes (e.g. unknown event type)? If so, under what assumptions?

The use of a (semi-)parametric model does not guarantee a straightforward interpretation for the regression coefficients; that depends on the model specification. If the model is defined via the CS hazard, regression coefficients may be interpreted in a CS hazard scale, but their interpretation in a CIF scale is generally complex due to the relationship in (5). The opposite is true for CIF-based models. More generally, unless a PH specification is assumed, the actual values for the regression coefficients are often difficult to interpret in a meaningful way (e.g. only their sign may be interpreted as increases/decreases in risk).

4 Approaches based on a cause-specific hazard specification

4.1 Penalised regression

As mentioned in Section 2.2, (6) can be estimated using available software for CPH models. If $p < n$, but large with respect to n, this could lead to overfitting. Moreover, this is not possible in high-dimensional

settings ($p > n$). Penalisation can alleviate these problems by shrinking regression coefficients towards zero. This is achieved by when maximising the penalised partial likelihood, i.e. :

$$\hat{\boldsymbol{\beta}}_k = \underset{\boldsymbol{\beta}_k}{\operatorname{argmax}}\, \mathcal{L}^{\mathrm{CS}}(\boldsymbol{\beta}_k) - \lambda \pi_\theta(\boldsymbol{\beta}_k), \tag{10}$$

where $\mathcal{L}^{\mathrm{CS}}(\boldsymbol{\beta}_k)$ is the partial likelihood (**Appendix A**), $\pi_\theta(\cdot)$ is a penalty function (which may depend on a parameter θ) and λ is a weight that controls the shrinkage strength, leading to models with varying levels of sparsity (the penalty can comprise two of more terms and corresponding weights). Typically, cross-validation is used to choose optimal weights (Tibshrani, 1997). Often, the value of λ that minimises a pre-specified loss function is selected (typically referred to as $\lambda_{\min}$). However, in some cases, several values of λ may result in similar cross-validated performance. Hence, choosing $\lambda_{\min}$ may not be appropriate. An alternative, more parsimonious, choice can be selected by using the so-called $\lambda_{1\mathrm{se}}$ (Hastie et al., 2009). The latter represents the highest value of λ (i.e. the strongest regularisation) such that the chosen performance metric is within 1 standard deviation of the one associated to $\lambda_{\min}$. Several types of penalty have been proposed, e.g. *lasso* (Tibshrani, 1997), *adaptive lasso* (Zhang and Lu, 2007), *elastic net* (Engler and Li, 2009) and *scad* (Fan and Li, 2002). Each of these leads to different properties of the algorithm (e.g. some introduce sparsity, setting some regression coefficients to be exactly equal to zero). Related Bayesian methods have also been proposed (e.g. Ibrahim et al., 1999).

4.2 Boosting

Boosting (Breiman, 1998; Friedman et al., 2000) methods aim to convert a *weak learner* (typically a simple algorithm) into a *strong learner* (a composite algorithm that combines several weak learners) through iterative optimisation. These ideas have been adapted to derive survival methods suitable to high-dimensional settings. For the CPH model, boosting can iteratively infer regression coefficients whilst introducing sparsity; this can be also applied to CR data under the model in (6). Cox model-based boosting (Ridgeway, 1999) and Cox likelihood-based boosting (Binder and Schumacher, 2008) follow a similar procedure. At each iteration $b = 1, \ldots, B$, all possible *marginal* regression coefficient updates (i.e. one at the time) are explored. An optimality criteria is then used to select the j^*-th covariate and to update the corresponding coefficient as $\beta_{kj^*}^{(b)} = \beta_{kj^*}^{(b-1)} + a_{j^*}^{(b)}$, where $\beta_{kj^*}^{(b-1)}$ is the previous value. However, the methods differ on the criteria that is used to select j^* and on how $a_{j^*}^{(b)}$ is calculated: for Ridgeway (1999) these are based on the gradient of the partial log-likelihood, instead Binder and Schumacher (2008) uses the L_2-norm penalised partial log-likelihood (see De Bin, 2016, for more details). In both cases, if $p < n$, the final estimate converges to the Cox (1972) estimator as $B \to \infty$ (De Bin, 2016). The optimal number of boosting iterations B can be tuned e.g. using cross-validation (Verweij and Van Houwelingen, 1993). Similar to the previous section, B may be chosen to minimise a pre-specified metric (more parsimonious choices may also be selected using a similar rationale to $\lambda_{1\mathrm{se}}$ in the previous section).

The methods above introduce sparsity by permitting both mandatory and optional covariates. The first approach incorporates mandatory features through an offset term (Boulesteix and Hothorn, 2010), where the regression coefficients for the mandatory covariates are not updated during the boosting procedure. In contrast, the method by Binder and Schumacher (2008) permits updating the regression coefficients of both, mandatory and optional covariates, but the optional features may be excluded through penalisation. De Bin (2016) pointed out how these different strategies can be reformulated in an equivalent manner.

4.3 Lunn-McNeil

Instead of modelling each event type separately, Lunn and McNeil (1995) proposed a joint model. This can allow a more parsimonious model specification by (e.g. sharing model parameters across different event types) which may, in turn, lead to lower sample size requirements. Inference can be performed using standard survival analysis software after converting the data into an augmented layout which is constructed

as follows. For individuals that experienced one of the K event types, the data is duplicated $K-1$ times, setting the event as censored for the duplicates. For censored observations, K rows are added with the event marked as censored (see example in **Appendix B**). Note that the augmented data layout also allows the use of the penalised regression and boosting algorithms described in the previous sections.

Two frameworks for inference are introduced. First, a *stratified* approach, which is equivalent to separate CPH models for each event type and where the k-th hazard is given by

$$h_k^{\text{LM1}}(t_i \mid \mathbf{x}_i) = h_{k0}^{\text{LM1}}(t_i) \exp\left(\sum_{k=1}^{K} \delta_{ik} \mathbf{x}_i^\top \boldsymbol{\beta}_k\right). \tag{11}$$

In (11), $\delta_{ik} = \mathbb{1}\{Z_i = k\}$ are event type indicators, $h_{k0}^{\text{LM1}}(\cdot)$ is a CS baseline hazard and $\boldsymbol{\beta}_k$ denotes a cause-specific$p$-dimensional vector of regression coefficients. Unlike cases in which separate models are fit for each event type, this approach permits the use ofsimpler models, e.g. where covariate effects are shared across different event types ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$).

The second, *unstratified*, framework is defined as

$$h_k^{\text{LM2}}(t_i \mid \mathbf{x}_i) = h_0^{\text{LM2}}(t_i) \exp\left(\mathbf{x}_i^\top \boldsymbol{\theta}' + \sum_{k=2}^{K} \alpha_k \delta_{ik} + \sum_{k=2}^{K} \delta_{ik} \mathbf{x}_i^\top \boldsymbol{\beta}_k'\right), \tag{12}$$

where the baseline hazard $h_0^{\text{LM2}}(\cdot)$ and the p-dimensional vector of coefficients $\boldsymbol{\theta}'$ relate to the first event type ($k = 1$), which is used as a reference. In (12), α_k and β_k' $(k = 2, \ldots, K)$ capture event-specificdeviations (baseline hazards and covariate effects) with respect to the reference event.Inference on those parameters can therefore be used to compare the behaviour of different event types. However, one disadvantage of this approach is that it assumes the shape of the baseline hazard to be the same across different event types (except for a proportionality constant). This is not generally appropriate for all applications.

5 Approaches based on the CIF

5.1 Penalised regression

Kuk and Varadhan (2013) developed a stepwise approach (forwards and backwards) to perform variable selection under the model in (7). Alternatively, similar to the methods described in Section 4.1, penalised regression approaches that adapt (7) to high-dimensional scenarios ($p > n$) have also been proposed. In particular, Fu et al. (2017) introduced a general penalised regression framework using a coordinate descent algorithm that permits individual and grouped variable selection. The authors implemented four types of penalties: *lasso* (Tibshrani, 1997), *adaptive lasso* (Zhang and Lu, 2007), *scad* (Fan and Li, 2002) and *mcp* (Zhang, 2010), as well as their grouped variations. A related method by Ha et al. (2014) was developed under a *frailty* model specification with shared or correlated random effects. More recently, Sun and Wang (2023) introduced Random Approximate Elastic Net (RAEN) based on an split-and-merge strategy. First, variables are split into several sets of correlated variables ($p < n$). Within each set, an elastic net penalised version of (7) is used to pre-select variables based on boostrap samples. Finally, pre-selected variables are merged into a single group prior to a final selection step (also using elastic net and boostrap).

5.2 Boosting

As in Section 4.2, Binder et al. (2009) developed a boosting approach to iteratively estimate the regression coefficients in (7), whilst supporting high-dimensional covariate settings. More concretely, they proposed a *sub-distribution hazard boosting* approach. To enable feature selection, covariates are divided into a set of mandatory ($\mathcal{I}^{\text{mand}}$) and a set of optional ($\mathcal{I}^{\text{opt}}$) features. At each boosting iteration, $b = 1, \ldots B$, regression coefficients for mandatory features $\gamma_{kl} (\forall l \in \mathcal{I}^{\text{mand}})$ are estimated jointly by maximising the

partial likelihood. Then, for optional covariates, only one regression parameter is updated. Tha latter is selected based on penalised partial log-likelihood estimates for all possible models:

$$h_k^{\text{FG}}(t_i \mid \mathbf{x}_i) = h_{k0}^{\text{FG}}(t_i) \exp\left(\zeta_{ki}^{(b-1)} + x_{ij}\eta_{kj}^{(b)}\right), \quad \zeta_{ki}^{(b-1)} = \mathbf{x}_i^\top \boldsymbol{\gamma}_k^{(b-1)}, \quad j \in \mathcal{I}^{\text{opt}}, \tag{13}$$

where $\zeta_{ki}^{(b-1)}$ is treated as an offset. Regression coefficients are then updated as $\gamma_{kj}^{(b)} = \gamma_{kj}^{(b-1)} + \eta_{kj}^{(b)}$ for the selected covariate and $\gamma_{kj}^{(b)} = \gamma_{kj}^{(b-1)}$ otherwise.

5.3 Pseudo-values

Following Fine (2001) and Andersen et al. (2003), Klein and Andersen (2005) propose a method based on the jackknife (leave-one-out) CIF estimator and GLMs. A regression model is directly specified in terms of the CIF, using an arbitrary link function. Given a pre-specified time point grid $\tau_1, \ldots, \tau_M$ (the authors recommend five to ten equally spaced points for this purpose), Andersen et al. (2003) define *pseudo-values* for the CIF of the i-th individual at time point τ_m for the k-th event type. These are given by

$$\theta_{imk} = n\text{CIF}_k(\tau_m) - (n-1)\text{CIF}_k^{-i}(\tau_m), \tag{14}$$

where $\text{CIF}_k(\tau_m)$ is the Aalen-Johansen (Aalen and Johansen, 1978) CIF estimator evaluated using all the data and $\text{CIF}_k^{-i}(\tau_m)$ is the corresponding estimate after removing the i-th observation. Then, based on these pseudo-values, a GLM is used to estimate covariate effects on the CIF:

$$g(\theta_{imk}) = \alpha_{mk} + \mathbf{x}_i^\top \boldsymbol{\gamma}_k, \tag{15}$$

where α_{mk} and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kp})^\top$ are regression coefficients estimated via generalised estimating equations (Liang and Zeger, 1986) and $g(\cdot)$ is a link function. If $g(\cdot)$ is a complementary log-log link, then (8) is recovered (i.e. a PH specification is assumed for the sub-distribution hazard) and $\boldsymbol{\gamma}_k$ can be interpreted in the same way as for the Fine and Gray (1999) method. However, this is not the case for more general link functions which induce a different (parametric) relationship between covariates and the sub-distribution hazard. In such cases, as $g(\cdot)$ is monotonic, the sign of the regression coefficients is associated with increases/decreases in the CIF, but the actual values are harder to interpret.

Finally, (15) can be extended to include time-varying covariate effects can be added as:

$$g(\theta_{imk}) = \alpha_{mk} + \mathbf{v}_i^\top \boldsymbol{\eta}_k(t_i) + \mathbf{u}_i^\top \boldsymbol{\gamma}_k, \tag{16}$$

where observed covariates, $\mathbf{x}_i$, are split into those with time varying effects ($\mathbf{v}_i$) and those with constant effects ($\mathbf{u}_i$), whose corresponding regression coefficients are $\boldsymbol{\eta}_k(t_i)$ and $\boldsymbol{\gamma}_k$, respectively.

5.4 Direct binomial

Similar to Klein and Andersen (2005), Scheike et al. (2008) propose a semi-parametric strategy that does not rely on a PH assumption. Their approach extends (8) to a more general class that enables both, time-varying and constant covariate effects. This includes a goodness-of-fit test to check if time-varying effects are required. The regression model is defined as

$$\text{CIF}_k(t_i \mid \mathbf{x}_i) = g^{-1}(\boldsymbol{\eta}_k(t_i), \boldsymbol{\gamma}_k, \mathbf{x}_i), \tag{17}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\eta}_k(t_i)$ are time varying parameters, and $\boldsymbol{\gamma}_k$ captures constant covariate effects. Both, $\boldsymbol{\gamma}_k$ and $\boldsymbol{\eta}_k(t_i)$ are estimated through score equations. More precisely, Scheike et al. (2008) studied an additive and multiplicative specification, defined respectively as

$$g\{\text{CIF}_k(t_i \mid \mathbf{x}_i)\} = \mathbf{v}_i{}^\top \boldsymbol{\eta}_k(t_i) + f(\boldsymbol{\gamma}_k, \mathbf{u}_i, t_i), \qquad \text{and} \tag{18}$$

$$g\{\text{CIF}_k(t_i \mid \mathbf{x}_i)\} = \left[\mathbf{v}_i{}^\top \boldsymbol{\eta}_k(t_i)\right] f(\boldsymbol{\gamma}_k, \mathbf{u}_i, t_i), \tag{19}$$

where $f(\cdot)$ is a known function and $\mathbf{x}_i$ is split as in (16). As in Section 5.3, the interpretation of $\boldsymbol{\gamma}_k$ is not straightforward for arbitrary choices of $g(\cdot)$ and $f(\cdot)$. However, the sign of $\boldsymbol{\gamma}_k$ can be interpreted in terms of whether changes in covariate values are linked to increases or decreases in the CIF.

More recently, Ambrogi and Scheike (2016) extended this approach to enable its use in high-dimensional settings. Similar to the models in Section 5.1, they proposed the used of a penalised regression framework.

5.5 Parametric constrained CIF

A related approach was proposed by Shi et al. (2013), which extended (8) as

$$g_k\{\text{CIF}_k(t_i \mid \mathbf{x}_i)\} = g_k\{\text{CIF}_{k0}(t_i)\} + \mathbf{x}_i^\top \boldsymbol{\gamma}_k, \quad k = 1, 2;$$

where the link functions $g_k\{\cdot\}$ are the generalised odds rate model by Jeong and Fine (2006):

$$g_k(u) = \log\left[\{(1-u)^{-\alpha_k} - 1\}/\alpha_k\right], \quad \text{with} \quad 0 < \alpha_k < \infty. \tag{20}$$

To ensure that $\text{CIF}_1(t \mid \mathbf{x}_i) + \text{CIF}_2(t \mid \mathbf{x}_i) = 1$ as $t \to \infty$, Shi et al. (2013) treat both events differently. For the primary event ($k = 1$), $\text{CIF}_1(t)$ is set using a modified three-parameter logistic function (Cheng, 2009) for the baseline hazard:

$$\text{CIF}_{10}(t_i) = \frac{p_1[\exp\{b_1(t_i - c_1) - \exp(-b_1 c_1)\}]}{1 + \exp\{b_1(t_i - c_1)\}}, \tag{21}$$

where p_k is the log-term probability of the k-th event ($\text{CIF}_k(t) \to p_k$ as $t \to \infty$), $b_k > 0$ dictates how fast $\text{CIF}_k(t)$ approaches p_k, and $c_k \in \mathbb{R}$. Instead, for the competing event ($k = 2$), they do not specify direct covariate effects and the CIF is given by:

$$\text{CIF}_2(t_i \mid \mathbf{x}_i) = \frac{p_2(\mathbf{x}_i)[\exp\{b_2(t_i - c_2) - \exp(-b_2 c_2)\}]}{1 + \exp\{b_2(t_i - c_2)\}}, \tag{22}$$

with $p_2(\mathbf{x}_i) = (1 - p_1)^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_1)}$ and, where b_2 and c_2 as in (21). Inference is performed via maximum likelihood, and can be extended to allow for right, interval and left censoring. In this context, $\boldsymbol{\gamma}_k$ can be interpreted in a similar way as discussed in Sections 5.3 and 5.4.

5.6 Dependent Dirichlet processes (DDP)

For $K = 2$, Shi et al. (2021) introduced a Bayesian non-parametric approach based on infinite mixtures of Weibull distributions (Kottas, 2006). For each mixture component, $\text{CIF}_1(t \mid \mathbf{x}_i)$ is defined as in (8), i.e. it assumes a PH specification for the sub-distribution hazard. In turn, the baseline CIF is parametrized in terms of a *normalised* baseline CIF $D_{01}(t)$, such that $\text{CIF}_{10}(t) = c \times D_{01}(t)$ and $c = \lim_{t\to\infty} \text{CIF}_{10}(t)$. The CIF for the second event type follows the specification by Fan (2008), which ensures that $\text{CIF}_1(t \mid \mathbf{x}_i) + \text{CIF}_2(t \mid \mathbf{x}_i) = 1$ as $t \to \infty$. This leads to

$$\text{CIF}_2(t_i \mid \mathbf{x}_i) = (1-c)^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_1)}\left[1 - \{1 - D_{02}(t_i)\}^{\exp(\mathbf{x}_i^\top \boldsymbol{\gamma}_2)}\right], \; D_{02}(t) = \text{CIF}_{20}(t)/(1-c). \tag{23}$$

Finally, assuming that $D_{01}(t)$ and $D_{02}(t)$ correspond to Weilbull distributions, the DDP model defines the i-th subject likelihood contribution as a Dirichlet Process mixture model (Escobar and West, 1995) which permits clustering of observations. As in Gelman et al. (2008), a weekly informative Cauchy prior is assigned to regression coefficients assuming that covariates are standardized to have mean equal to zero (with 0.5 standard deviation for continuous covariates; binary variables are coded such that there is a difference of 1 the levels). The DDP approach scales linearly with the sample size and with the number of features. Moreover, it permits inference with interval censored data and time-dependent covariates.

Note that, although a PH assumption is specified for the sub-distribution hazard of the first event type within each mixture component, the latter does not hold for the overall mixture. As such DDP may be applied to datasets for which the PH assumption is not appropriate.

5.7 Survival Multitask Boosting (SMTBoost)

SMTBoost (Bellot and van der Schaar, 2018a) is a non-parametric method that combines boosting (Breiman, 1998; Friedman et al., 2000) and multi-task learning (Caruana, 1993) to jointly estimate the CIF associated to all event types, assuming they share a common structure. The aim is to minimize the difference between the observed and predicted survival status via the following loss function:

$$L = \frac{1}{K}\sum_{k=1}^{K} L_k, \quad \text{with} \quad L_k = \tfrac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\tfrac{1}{\tau}\int_0^{\tau}(\mathbb{1}\{T_i \leq t, Z_i = k\} - \text{CIF}_k(t \mid \mathbf{x}_i))^2 dt\right] \quad (24)$$

SMTBoost uses binary partitioned trees as weak learners to recursively split individuals into homogeneous groups (nodes) with similar time-to-event outcomes. Following Ishwaran et al. (2014), the splitting rule is based on the Gray's log-rank test (Gray, 1988): this compares CIF_k between nodes and defines a composite rule via a weighted sum of the test's statistic across all event types. At each terminal node m, the Aalen-Johansen estimator (Aalen and Johansen, 1978) for the k-th CIF, $\text{CIF}^{\text{AJ}}_{k,m}(t_i)$, is calculated. Let $\mathcal{C}_m$ be the index set of observations in node m, the CIF is then computed as

$$\text{CIF}_k(t_i \mid \mathbf{x}_i) = \sum_m \mathbb{1}\{i \in \mathcal{C}_m\}\text{CIF}^{\text{AJ}}_{k,m}(t_i). \quad (25)$$

The boosting procedure is used to iteratively construct the trees based on weighted versions of the training data, were a higher weight is given to samples with higher prediction error in the previous iteration. Final predictions are calculated as a weighted average across all trees.

To quantify the influence of each feature, SMTBoost uses a variable importance measure (computed per event type). The authors demonstrated the performance of SMTBoost for event types with low incidence, in datasets with a large number of observations as well as cases in which not all covariates were informative.

5.8 Derivative-based neural network modelling (DeSurv)

Danks and Yau (2022) proposed a flexible, non-parametric approach that can be seen as a continuous time version of the work by Lee et al. (2018) (DeepHit) or as a CR generalisation of DeepSurv (Katzman et al., 2018). Similar to the mixture models in Section 7.1, DeSurv factorises the CIF as

$$\begin{aligned}\text{CIF}_k(t_i \mid \mathbf{x}_i) &= \Pr(T \leq t_i \mid Z_i = k, \mathbf{x}_i)\Pr(Z_i = k \mid \mathbf{x}_i) \\ &\equiv \tilde{F}_k(t_i \mid \mathbf{x}_i)\pi_{ik}(\mathbf{x}_i), \quad \text{with} \quad \sum_{k=1}^{K}\pi_{ik}(\mathbf{x}_i) = 1. \end{aligned} \quad (26)$$

In turn, $\tilde{F}_k(t_i \mid \mathbf{x}_i)$ (a proper cumulative density function) is defined as $\tilde{F}_k(t_i \mid \mathbf{x}_i) = \tanh(u_k(t_i \mid \mathbf{x}_i))$, where $u_k(t_i \mid \mathbf{x}_i)$ is a strictly monotonic function whose derivative is parametrised as a neural network with positive output range (the authors note that any strictly monotonic cumulative density function can be use in place of $\tanh$). $\pi_{ik}(\mathbf{x}_i)$ is also modelled as a neural network, using a softmax activation function to satisfy the sum constrain in (26). In this setting, training is performed using the log-likelihood as a loss function and the Adam optimisation algorithm (Kingma and Ba, 2015).

6 Approaches based on a latent survival times specification

6.1 Deep Multi-task Gaussian Processes (DMGPs)

Alaa and van der Schaar (2017) proposed a Bayesian non-parametric method using deep Gaussian Processes (GPs) (Damianou and Lawrence, 2013), which provide a flexible approach to capture complex relationships between covariates and outputs. DMGPs build upon a hierarchical construction which resembles a two-layers neural network, but within a fully probabilistic model. Given known covariate values,

DMGPs can be used to infer a posterior distribution for the survival times and to estimate an individual-specific CIF. Let $\mathbf{T}_i = (T_{i1}, \ldots, T_{iK})^\top$ be a vector of latent survival times for subject i. DMGPs assume $\mathbf{T}_i = g(\mathbf{x}_i) + \epsilon_i$, where $g(\cdot)$ is a multi-output *random* function and ϵ_i is an error term. In turn, $g(\cdot)$ is defined via a two-layer hierarchical model which enables non-Gaussian outputs:

$$\mathbf{T}_i|\boldsymbol{\zeta}_i \sim \mathrm{N}\left(f_T(\boldsymbol{\zeta}_i), \omega_T^2\mathbf{I}_K\right), \tag{27}$$

$$\boldsymbol{\zeta}_i \sim \mathrm{N}\left(f_\zeta(\mathbf{x}_i), \omega_\zeta^2\mathbf{I}_q\right). \tag{28}$$

where $\mathbf{I}_q$ is a q-dimensional identity matrix and $\boldsymbol{\zeta}_i$ a q-dimensional latent variable ($q = 3$ was used in the original publication). Moreover, $f_T(\cdot) : \mathbb{R}^q \to \mathbb{R}^K$ and $f_\zeta(\cdot) : \mathbb{R}^p \to \mathbb{R}^q$ are independent zero centred vector-valued GPs whose covariance functions (kernels) $C_{\theta_\zeta}(\cdot,\cdot)$ and $C_{\theta_T}(\cdot,\cdot)$ depend on parameters θ_ζ and θ_T, respectively. These are defined using the intrinsic coregionalization model, which has been previously used in the context of multi-task learning (Álvarez et al., 2012, Section 4.2). These control the smoothness of $f_T(\cdot)$ and $f_\zeta(\cdot)$. Inference is obtained via a variational framework (Blei et al., 2017), combined with the inducing points approach of Titsias (2009) to derive a tractable algorithm. The implementation allows the presence of right censored observations. Note that, conditional on $f_T(\boldsymbol{\zeta}_i)$, (27) assumes independence among the latent survival times. This cannot be verified.

6.2 Deep Survival machines (DSM)

Introduced by Nagpal et al. (2021), DSM combines a parametric model for the survival times with a deep learning framework. The approach is initially introduced for a single event type and then extended to a CR setting. For each event type, the latent survival times are modelled as a finite mixture of L distributions, whose parameters (and mixture weights) are linked to covariates via a neural network:

$$T_{ik}|Z_{ik} = l \sim \mathrm{P}(\tilde{\lambda}_{kl} + h(\Phi_\theta(\mathbf{x}_i)^T\boldsymbol{\lambda}_k), \tilde{\alpha}_{kl} + h(\Phi_\theta(\mathbf{x}_i)^T\boldsymbol{\alpha}_k)), \tag{29}$$

$$Z_{ik} \sim \mathrm{Discrete}(\mathrm{softmax}(\Phi_\theta(\mathbf{x}_i)^T\boldsymbol{\omega}_k)), \tag{30}$$

where $\mathrm{P}(\cdot,\cdot)$ denotes either a Weibull or log-normal distribution; $h(\cdot)$ is an activation function; $\boldsymbol{\lambda}_k, \boldsymbol{\alpha}_k$ and $\boldsymbol{\omega}_k$ are cause-specific vectors of parameters, and $\tilde{\lambda}_{kl}, \tilde{\alpha}_{kl}$ act as component/cause-specific intercepts. $\Phi_\theta(\cdot)$ is modelled as a multilayer perceptron (Hastie et al., 2009) which creates a non-linear map between input covariates and a low-dimensional space. The latter is shared across all event types.

All model parameters are jointly learned during training,optimising a loss function which down-weights censored observations to reduce potential biases towards long-tails in the survival distribution. In their experiments, the authors use cross-validation to inform hyperparameter choices (e.g. L).

6.3 Bayesian Lomax delegate racing (LDR)

LDR (Zhang and Zhou, 2018) can be seen as a generalisation of *exponential racing* in which latent times are assumed to be independent and exponentially distributed, leading to $T = \min\{T_1, \ldots, T_K\}$ also being exponentially distributed. LDR extends exponential racing in two ways. First, a Lomax distribution (Lomax, 1954) is assigned to the latent survival times. This has heavier tails and can be interpreted as a scale mixture of exponential distributions. Subsequently, Zhang and Zhou (2018) assumed that each latent time is determined by a potentially infinite number of sub-risks (e.g. different etiologies of a disease), leading to a generalisation of a Gamma process (Wolpert and Ickstadt, 1998). To facilitate implementation, Zhang and Zhou (2018) truncated the Gamma process to L sub-risks ($L = 10$ was used in their experiments). The L sub-risks play a similar role to CRs, but nested within each event type. Let $T_{ik} = \min\{T_{ik1}, \ldots, T_{ikL}\}$ and $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ip})^\top$, the LDR model is based on the following hierarchical formulation:

$$\begin{aligned} T_{ikl}|\lambda_{ikl} &\sim \mathrm{Exp}(\lambda_{ikl}\exp(\mathbf{x}_i'^\top\boldsymbol{\nu}_{kl})) \\ \lambda_{ikl}|\alpha_{kl}, \boldsymbol{\nu}_{kl} &\sim \mathrm{Gamma}\left(\alpha_{kl}, 1\right), \end{aligned} \tag{31}$$

where $\boldsymbol{\nu}_{kl} = (\nu_{kl0}, \ldots, \nu_{klp})^\top$ are regression coefficients specific to the l-th sub-risk within the k-th event type. Prior distributions for α_{kl} and $\boldsymbol{\nu}_{kl}$ are discussed in Zhang and Zhou (2018, Appendix B). After marginalisation of λ_{ikl}, inference uses a Gibbs sampler (Geman and Geman, 1984) for moderate n, and maximum a posteriori through stochastic gradient descent (Kiefer et al., 1952) for larger datasets.

LDR assumes independence across cause-specific latent survival times, which cannot be verified. In the presence of censored observations or missing event types, LDR uses a data augmentation strategy impute these values whilst performing inference. The latter assumes a missing-at-random mechanism.

7 Other approaches for continuous time CR data

In this section we introduce methods that do not fall in any of the previous categories or that can accommodate more than one specification. For instance, Ishwaran et al. (2014) enable, both, a CS hazard and a CIF formulation for covariate effects.

7.1 Mixture models

These models decompose the joint distribution of the event time and event type into marginal probabilities $\pi_{ik}(\mathbf{x}_i) = \Pr(Z_i = k \mid \mathbf{x}_i)$ of each event type (mixing proportions) and the conditional survival distribution $S_k(t_i \mid \mathbf{x}_i) = \Pr(T > t_i \mid Z_i = k, \mathbf{x}_i)$. This assumes that each individual will experience an specific event type, which is chosen randomly (Larson and Dinse, 1985). Mixture models typically need large sample sizes and long follow ups to avoid identifiability issues (Haller et al., 2013). In addition, these methods have been pointed out as difficult to interpret due to the number of parameters (Haller et al., 2013), and its reliance on conditioning on the future when decomposing the joint distribution (Andersen and Keiding, 2012). In general, the CR model is set as a K-component mixture

$$S(t_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) S_k(t_i \mid \mathbf{x}_i), \qquad \text{with} \qquad \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) = 1. \tag{32}$$

In this context, several model specifications have been proposed. In particular, Larson and Dinse (1985) assumed the number of events across types to be multinomial with probabilities:

$$\pi_{ik}(\mathbf{x}_i) = \frac{\exp\left(a_k + \mathbf{x}_i^\top \mathbf{b}_k\right)}{1 + \sum_{l=1}^{K-1} \exp\left(a_l + \mathbf{x}_i^\top \mathbf{b}_l\right)}, \qquad k = 1, \ldots, K, \tag{33}$$

where a_k and $\mathbf{b}_k$ are regression coefficients. In addition, they assume

$$S_k(t_i \mid \mathbf{x}_i) = \exp\left\{ - \int_0^{t_i} h_{k0}(u) \exp(\mathbf{x}_i^\top \boldsymbol{\theta}_k)\, du \right\}, \tag{34}$$

where the baseline hazards, $h_{k0}(t_i)$, are piecewise constant functions within L disjoint intervals. This formulation implicitly introduces covariate effects via the CIF via $\text{CIF}_k(t_i \mid \mathbf{x}_i) = \pi_{ik}(\mathbf{x}_i) \{1 - S_k(t_i \mid \mathbf{x}_i)\}$.

Maximum likelihood estimates can be obtained using expectation-maximisation (EM, Dempster et al., 1977). One challenge is to select L: a large L may lead to an overparametrised model; a small L may cause poor fitting (Kuk, 1992). To overcome this, Kuk (1992) propose a semi-parametric model with arbitrary baseline hazards and suggests to infer a_k, $\mathbf{b}_k$ and $\boldsymbol{\theta}_k$ using a Monte Carlo approximation of the marginal likelihood. These estimates are subsequently used within EM to infer $h_{k0}(t_i)$. Similarly, Ng and McLachlan (2003) propose a semi-parametric approach that uses expectation-conditional maximisation (Meng and Rubin, 1993). In this case, multiple initialisations may be required to ensure convergence of the algorithm. Furthermore, Chang et al. (2007) propose a different algorithm for maximum likelihood estimation along with asymptotic properties of the estimators.

7.2 Tree-based mixture models

Using the same decomposition as in Section 7.1, Bellot and van der Schaar (2018b) introduced a Bayesian semi-parametric tree-based mixture model. It uses generalised gamma distributions (Cox et al., 2007) to model the conditional survival distributions in (32), such that $S_k(t_i \mid \mathbf{x}_i) := \text{GGamma}(t_i \mid \theta_{ik}, \sigma_i, \lambda_i)$, where θ_{ik} is scale parameter and σ_i, λ_i control the shape of the distribution (all strictly positive). Note that the generalised gamma distribution contains the Weibull ($\lambda_i = 1$), Gamma ($\sigma_i = \lambda_i$) and log-normal ($\lambda_i = 1$) distributions as specific cases. Let $\boldsymbol{\theta}_i := (\theta_{i1}, \dots \theta_{iK})^\top$ and $\boldsymbol{\pi}_i := (\pi_{i1}, \dots \pi_{iK})^\top$ be subject-specific parameter vectors. Dependency on covariate values $\mathbf{x}_i$ is introduced as follows:

$$\begin{aligned} \boldsymbol{\theta}_i \mid \mathbf{x}_i &= g_\theta(\mathbf{x}_i) + \epsilon_{\beta i}, \quad \epsilon_{\theta i} \sim \text{N}(0, \omega_\theta^2) \\ \boldsymbol{\pi}_i \mid \mathbf{x}_i &= l(g_\pi(\mathbf{x}_i) + \epsilon_{\pi i}), \quad \epsilon_{\pi i}, \sim \text{N}(0, \omega_\pi^2), \end{aligned} \tag{35}$$

where $g_\theta(\cdot)$ and $g_\pi(\cdot)$ are $\mathbb{R}^p \to \mathbb{R}^K$ functions defined by Multivariate Random Forests (Segal and Xiao, 2011), $l(x_i) = x_i / \sum_i x_i$, and $(\omega_\theta^2, \omega_\pi^2)$ are fixed hyperparameters. Shape parameters (σ_i, λ_i) are constrained to be shared by subjects that experience the same event type, capturing cause specific risk profiles. Gamma and Gaussian priors are then assigned to the associated cause-specific parameters. Inference is performed via an adaptive Metropolis-within-Gibbs scheme (Hasting, 1970; Roberts and Rosenthal, 2009). Moreover, a permutation approach (Ishwaran, 2007) is used to obtain a measure of variable importance for the absolute risk of observing a given event type and for the cause specific conditional survival distribution.

7.3 Vertical modelling

Nicolaie et al. (2010) propose to decompose the joint distribution of the event time and type to first estimate the overall probability of event occurrence and then the probability of a specific event type given that the event occurred at a given time. This decomposition is unlike the one used by the models in Section 7.1, which are formulated in the opposite manner. The vertical modelling approach requires to fit two models, one for the overall hazard function and one for the relative CS hazard defined as $r_k(t_i) = h_k(t_i)/h(t_i)$.

To estimate the overall hazard function $h(t)$, all event types are considered as events, regardless of their cause. It can be estimated using a PH approach or with a Nelson-Aalen estimator for a single categorical covariate. Instead, the CS relative hazard can be fitted via a multinomial logistic regression, with spline basis functions to smooth the function over time. Nicolaie et al. (2010) discussed two specifications for $r_k(t_i \mid \mathbf{x}_i)$, one that incorporates interaction effects and one with an additive structure which is given by

$$r_k(t_i \mid \mathbf{x}_i) = \frac{\exp\left(\mathbf{b}(t_i)^\top \boldsymbol{\eta}_k + \mathbf{x}_i^\top \boldsymbol{\theta}_k\right)}{\sum_{l=1}^K \exp(\mathbf{b}(t_i)^\top \boldsymbol{\eta}_l + \mathbf{x}_i^\top \boldsymbol{\theta}_l)}, \tag{36}$$

where $\boldsymbol{\theta}_k$ denotes a p-dimensional vector of covariate effects, $\mathbf{b}(t_i)$ introduces smooth dependency on t via q spline basis functions and $\boldsymbol{\eta}_k$ represents the regression coefficients associated to them. For identifiability, all entries of $\boldsymbol{\theta}_1$ and $\boldsymbol{\eta}_1$ are set equal to one. The model in (36) can be extended to allow for interactions between covariates and splines. While interpretability of the regression coefficients in the relative CS hazard can be challenging, a graphical representation of the estimated relative hazards over time can provide relevant insights. For instance, one can infer the contribution of the different event types to the overall rate of failure along time. Note that this approach is implicitly modelling covariate effects via a CS hazard, $h_k(t)$; however, the method does not follow a PH assumption.

Nicolaie et al. (2015) have extended this approach to deal with missing event types under a missing-at-random mechanism. In such case, all observations are used when inferring the overall hazard function $h(t)$ but observations with missing event types are ignored when estimating $h_k(t)$.

7.4 Random survival forests (RSF)

Ishwaran et al. (2014) introduced a non-parametric approach using an ensemble of random forests (Breiman, 2001). RSF can handle right censored observations, high-dimensional and large data problems, several

competing events, and permits non-linear/interaction covariate effects. RSF uses a pre-specified number of boostrap samples (B) to grow B trees using a random selection of covariates when deciding how to split each node. The tree is iteratively grown until a stopping rule is satisfied (e.g. the terminal nodes have at least n_0 individuals). To divide each node, the j-th covariate is selected and subjects are divided into daughter nodes based on its value ($x_j \leq c$ or $x_j > c$). Different *splitting rules* can be used to select j and c. Two event-specific splitting rules were proposed by Ishwaran et al. (2014): one based on the log-rank test (Mantel et al., 1966) and one related to a modification of the Gray's test (Gray, 1988). The former aims to maximise the difference of CS hazard rates between daughter nodes and its therefore better suited to identify covariates that influence the CS hazard. Instead, the second rule aims to maximise CIF differences for the selected event. In addition, a combined splitting rule was proposed for cases where the objective is to select covariates that affect any cause or when the goal is to predict the CIF of all causes. The splitting rules are detailed in Ishwaran et al. (2014, Section 3.3).

For each individual i, RSF computes a tree-specific estimate for the k-th CIF using the Aalen-Johansen estimator. This is denoted as $\text{CIF}_k^{(b)\text{AJ}}(t \mid \mathbf{x}_i)$ and the calculation is based on the terminal node to which the individual was assigned based on the value of their covariates. Alternatively, RSF also report a measure of the expected amount of time (e.g. life years) lost due to the k-th cause before time τ:

$$M_k^{(b)}(\tau \mid \mathbf{x}_i) = \int_0^\tau \text{CIF}_k^{(b)\text{AJ}}(t_i \mid \mathbf{x}_i)\, dt, \tag{37}$$

where τ is such that the probability of being uncensored is bounded away from zero. Final estimates for each individual are calculated as an average across all trees.

RSF can perform variable selection based a measure of variable importance (VIMP) and a minimal depth metric. For each variable, VIMP quantifies the change in predictive accuracy after adding random noise to the variable. VIMP is calculated in an event-specific or non-event-specific manner using a random node assignment strategy (Ishwaran et al., 2008). Instead, minimal depth (Ishwaran et al., 2010) is non-event-specific and measures the depth of the first node in which a variable was selected by the splitting rule. As illustrated by Ishwaran et al. (2014), both metrics can be combined to select a final set of variables.

In the presence of missing values (covariates or outcomes), the adaptive tree imputation algorithm (Ishwaran et al., 2008) can be applied. The latter, iteratively splits the nodes whilst imputing missing values by randomly drawing from non-missing observations within their node. Tang and Ishwaran (2017) shown that the performance of this approach (and related ones) depends on a variety of factors, including the missingness mechanism (e.g. missing-at-random) and the correlation amongst covariates.

8 Competing risks survival models for discrete time-to-event data

Note that Janitza and Tutz (2015) proposed another approach based on random forests using a discrete scale for the survival times.

So far, the models included in this review focus on continuous survival times. However, time-to-event outcomes are often recorded in a discrete scale (e.g. weeks, months). Recently, Schmid and Berger (2021) provided an overview for approaches developed in this context. The predominant method is a CR extension for the *proportional odds* model (Cox, 1972). This introduces covariate effects through a discrete-time version of the cause-specific hazard function:

$$h_k^{\text{D}}(t) = \frac{\Pr(T = t, Z = k)}{\Pr(T \geq t)}. \tag{38}$$

The CR proportional odds model (Tutz, 1995) is then defined as:

$$\log\left(\frac{h_k^{\text{D}}(t_i \mid \mathbf{x}_i)}{h_0^{\text{D}}(t_i \mid \mathbf{x}_i)}\right) = \lambda_{kt_i} + \mathbf{x}_i^\top \boldsymbol{\Omega}_k, \qquad k = 1, \ldots, K, \tag{39}$$

where $h_0^{\mathrm{D}}(t_i \mid \mathbf{x}_i) = 1 - \sum_{k=1}^{K} h_k^{\mathrm{D}}(t_i \mid \mathbf{x}_i)$, λ_{kt_i} are baseline log-odds (k-th event versus no event) and $\boldsymbol{\Omega}_k = (\omega_{k1}, \ldots, \omega_{kp})^\top$ regression coefficients. This model can be estimated in most statistical software as a multinomial logistic regression. For this purpose, the data is transformed into a person-period format (Scott and Kennedy, 2005), using binary indicators $Y_{itk} = \mathbb{1}\{T_i = t, Z_i = k\}$ to capture whether an event of type k is observed at time t for subject i. In this context, the k-th CIF can be then estimated as

$$\mathrm{CIF}_k(t_i \mid \mathbf{x}_i) = \sum_{t=0}^{t_i} q(t, k \mid \mathbf{x}_i), \tag{40}$$

where $q(t, k \mid \mathbf{x}_i) = \Pr(T_i = t, Z_i = k \mid \mathbf{x}_i)$.

The person-period representation of discrete time CR datasets has enabled several extensions for the model in (39) based on statistical and machine learning approaches developed for binary or multinomial outcomes (see Schmid and Berger, 2021, for an overview). For example, to perform feature selection, a penalised multinomial logistic regression (e.g. as implemented in glmnet) could be employed. Other approaches specifically developed for discrete time CR datainclude SSPN (Nemchenko et al., 2018) and DeepHit (Lee et al., 2018), both using neural networks. Another recent approach, by Sparapani et al. (2020), is based on Bayesian additive regression trees (BART, Hill et al., 2020). BART permits non-linear/interaction effects, non-proportional hazards, missing data and uses a sparse prior for high-dimensional covariate spaces. For completeness, as the method by Sparapani et al. (2020) is implemented within the popular BART R package (Sparapani et al., 2021), we decided to include it in this review.

Sparapani et al. assume that the binary indicators Y_{itk} follow a multinomial distribution with event probabilities $\pi_{itk} = \Pr(T_i = t, Z_i = k \mid T_i \geq t, \mathbf{x}_i)$, which can be seen as a discrete hazard (if the survival times are not discrete, a discretised scale is adopted with each observed/censored time treated as a distinct time-point). As multinomial implementations of BART are not widely available, the authors propose two formulations using BART probit models, focusing on $K = 2$. In the fist formulation, one model is used for the time until *any* event occurs and a second model for the conditional probability of the event being of type $k = 1$ given that an event occurred. In contrast, the second formulation employs one model for the conditional probability of experiencing event type $k = 1$ at time t given that the subject is still at risk. A second model is then used for the conditional probability of a type $k = 2$ event at time t given that the subject is still at risk and that it has not experience a type $k = 1$ event. Prior distributions for the required parameters in the models are discussed in detail in Sparapani et al. (2020, Section 2).

In the presence of missing covariate values, the existing BART implementation (Sparapani et al., 2021) enables the use of record-level hot-decking imputation (De Waal et al., 2011). The latter imputes missing values by randomly sampling from non-missing values, regardless of their event time or type. Such approach may not be appropriate if the number of missing values is high.

9 Software and reproducibility

Provision of open-source and well documented software is critical to ensure wide adoption of new statistical or machine learning methods. Towards this goal, Sonabend et al. (2021) developed the `mlr3proba` `R` library, providing a common interface for several survival models, including some of the CR approaches here presented (removed from CRAN on May 2022, but actively maintained and available in GitHub). Another software resource was implemented by Mahani and Sharabiani (2019), supporting Bayesian and non-Bayesian inference for cause-specific hazard models.

Here, we summarise available software for the methods described in the previous Sections. While some implementations are available as `R` or `Python` packages, other methods are only accessible through *ad hoc* source code in public repositories or, in the worse case scenario, there is no code available for the method's implementation. Table 2 summarises this. Note that some methods (e.g. Lunn and McNeil, 1995) can be applied using standard survival analysis software (e.g. the `survival` R package Therneau, 2023), without the need for bespoke implementations. To facilitate adoption, for the methods which have

Table 1 Summary of the available methods for survival regression with CR.

Model	Type	Proportional hazards (PH)	High dimensions (p)	Missing data
Approaches based on a cause-specific hazard specification				
Cox proportional CS hazard	Semi-parametric[1]	✓	✗	✗
Lunn-McNeil	Semi-parametric	✓	✗	✗
Penalised Cox PH	Semi-parametric	✓	✓	✗
Cox model-based boosting	Semi-parametric	✓	✓	✗
Cox likelihood-based boosting	Semi-parametric	✓	✓	✗
Approaches based on the CIF				
Fine-Gray	Semi-parametric	✓	✗	✗
Penalised proportional sub-distribution hazard	Semi-parametric	✓	✓	✗
Sub-distribution hazard boosting	Semi-parametric	✓	✓	✗
Pseudo-values	Semi-parametric	✗[2]	✗	✗
Direct binomial	Semi-parametric	✗[2]	✓[9]	✗
Parametric constrained CIF	Parametric	✗[2]	✗	✗
Dependent DP	Non-parametric	✗	✗	✗
SMTBoost	Non-parametric	✗	✓[3]	✗
DeSurv	Non-parametric	✗	✗	✗
Approaches based on a latent survival times specification				
Deep multi-task GPs	Non-parametric	✗	✗	✗
DSM	Non-parametric[7]	✗	✓	✗
Bayesian LDR	Non-parametric	✗	✗	✓[4]
Others approaches for continuous time-to-event data				
Mixture models	Several	✗	✗	✗
Tree-based Bayesian mixture model	Semi-parametric	✗	✓[3]	✗
Vertical modelling	Semi-parametric	✗	✗	✓[5]
RSF	Non-parametric	✗	✓	✓[6]
CR survival models for discrete time-to-event data				
BART	Non-parametric	✗	✓	✓[8]

[1] This and related methods can be parametric if e.g. a Weibull model is used for the baseline hazard.

[2] Depends on the choice of link function (PH holds for complementary log-log link).

[3] Reports a variable importance measure. Original publication only considered $n > p$ cases.

[4] A data augmentation scheme (e.g. within a Gibbs sampler) is used in the presence of censoring (unknown event time) or missing event types (missing-at-random). Missing covariate values are not permitted.

[5] Nicolaie et al. (2010) does not permit missing data. The extension by Nicolaie et al. (2015) allows missing event types (missing-at-random). Missing covariate values are not permitted.

[6] Adaptive tree imputation (Ishwaran et al., 2008) can be used to impute missing covariates or outcomes. See Tang and Ishwaran (2017) for an evaluation in different settings (e.g. missing-not-at-random).

[7] Parametric survival model but a neural network learns a lower-dimensional covariate representation.

[8] Missing covariate values are imputed using record-level hot-decking imputation (De Waal et al., 2011). Only recommended when the number of missing values is small.

[9] Not in the original model, but supported when using the extension by Ambrogi and Scheike (2016).

available R libraries, we provide vignettes to illustrate their usage using publicly available data (Pintilie, 2006). Vignettes are available at `www.github.com/VallejosGroup/CompRisksVignettes`.

Even when there are software packages accompanied with documentation and when analysis code is publicly available, reproducibility of an existing analysis is not guaranteed e.g. due to differences in the computational environment (Beaulieu-Jones and Greene, 2017). Moreover, static vignettes or code included as part of a paper are not always updated as the associated software changes. This may introduce challenges when applying or benchmarking new methods. To ensure reproduciblity of the vignettes provided here, we also prepared a Docker image (Boettiger, 2015) with all software requirements. The latter is available at: `https://github.com/VallejosGroup/CompRisksVignettes/pkgs/container/comprisksvignettes`.

10 Practical considerations

When deciding on a suitable CR method to employ, the user needs to carefully consider several aspects. First, from a practical point of view, it is likely that only those approaches with available and well documented software can be efficiently adopted by practitioners (see software availability in Table 2). The scalability of such software is also critical, as some (e.g. those that use Markov Chain Monte Carlo algorithms such as Roberts and Rosenthal, 2009, to perform inference) may only be suitable for moderate size data sets. Second, as there is no single approach that works well for all applications, the specific research question at hand plays a key role in the choice. For instance, Austin et al. (2016) highlight that methods based on a CIF formulation are more appropriate for the development of risk prediction models; whereas, CS formulations are better suited to resolve etiological questions. Moreover, in some cases, applying both types of methods can provide useful insights of the covariate effects on, both, the incidence and the rate of occurrence of the event — despite model misspecification (Grambauer et al., 2010; Latouche et al., 2013).

The specific characteristics of the available data can help also to guide the user to make a suitable choice. The summary presented in Table 1 can help to inform this choice. For instance, in the presence of high-dimensional covariate spaces some methods perform feature selection or dimensionality reduction; while others require the user to provide a reduced low-dimensional set of covariates that should be pre-selected. For example, Austin et al. (2017) reported that the number of (primary) events can substantially affect estimation performance for the Fine and Gray (1999) model, with data requirements varying depending on the type of covariates (e.g. 10 events per covariate may be enough for continuous covariates; but 40-50 events may be required in more general cases). More flexible methods likely have higher data requirements but, in the absence of systematic benchmarks, the guidelines above may provide a rough reference.

If the goal is to understand how different covariates affect an event's risk, methods that infer parametric (generally linear in terms of a log-hazard function) covariate effects may be preferable, despite their reduced flexibility. Alternatively, post-hoc variable importance metrics such as Shapley values (Lundberg and Lee, 2017) may be used for more complex approaches (e.g. those based on neural networks). When selecting a model, one should also consider the bias-variance trade off between parametric, semi-parametric and non-parametric approaches (see e.g. Wey et al., 2015). Finally, in the context of risk prediction, ensemble approaches which combine the predictions generated by different models (e.g. van der Laan et al. (2007)) can be used to bypass the need to select a single method and improve predictive performance.

11 Application to cancer data

Here, we use a publicly available dataset to demonstrate two common tasks performed in the context of CR analyses: (i) inference on hazard ratios (cause-specific or sub-distribution hazards) and (ii) risk prediction. We focus solely on those methods that have well documented and up-to-date `R` software (Table 2). All analysis code is available at `www.github.com/VallejosGroup/CompRisksVignettes`.

Table 2 Software available for survival regression with CR.

Model	CRAN	mlr3proba	Github (username/repository)
Approaches based on a cause-specific hazard specification			
Cox proportional CS hazard	`riskRegression`, `survival`, `rms`	✓	
Lunn-McNeil	Same as above	✓	
Penalised Cox PH	`glmnet`	✓	
Cox model-based boosting	`mboost`	✓	
Cox likelihood-based boosting	`CoxBoost`[1]	✓	binderh/CoxBoost
Approaches based on the CIF			
Fine-Gray	`riskRegression`, `cmprsk`, `crrstep`[2]	✓	
Penalised proportional sub-distribution hazard	`RAEN`[3]	✗	
Sub-distribution hazard boosting	`CoxBoost`[1]	✓	binderh/CoxBoost
Pseudo-values	`pseudo`+`GEEPACK`[4]	✗	
Direct binomial	`timereg`	✗	
Parametric constrained CIF[5]	✗	✗	✗
Dependent DP	`DPWeibull`[6]	✗	✗
SMTBoost	✗	✗	alexisbellot/SurvBoost[7]
DeSurv	✗	✗	djdanks/DeSurv (`Python`)
Approaches based on a latent survival times specification			
Deep multi-task GPs	✗	✗	
DSM	✗	✗	autonlab/auton-survival (`Python`)
Bayesian LDR	✗	✗	zhangquan-ut/ Lomax-delegate-racing-for-survival-analysis-with-competing-risks
Others approaches for continuous time-to-event data			
Mixture models[8]	`NPMLEcmprsk`[9]	✗	
Tree-based Bayesian mixture model	✗	✗	alexisbellot/HBM[7]
Vertical modelling[10]	`splines`+`survival`	✗	
RSF	`randomForestSRC`	✓	
CR survival models for discrete time-to-event data			
BART	`BART`	✗	

[1] Removed from CRAN on Nov 11, 2020 (`https://CRAN.R-project.org/package=CoxBoost`)
[2] To perform forwards/backwards stepwise variable selection.
[3] Removed from CRAN on Jan 25, 2023 (`https://CRAN.R-project.org/package=RAEN`
[4] $K = 2$ only.
[5] Example code as supplementary material in Shi et al. (2013).
[6] Removed from CRAN on April 26, 2022 (`https://CRAN.R-project.org/package=DPWeibull`)
[7] Bespoke R functions only.
[8] See example `R` code in Haller et al. (2013). For Ng and McLachlan (2003), Fortran code is available upon request.
[9] This library implements the approach by Chang et al. (2007).
[10] Example code as supplementary material in Haller et al. (2013)

The data corresponds to the Hodgkin's disease (HD) study described in Pintilie (2006), and which is also available in the `randomForestSRC` package. Hodgkin's disease is a type of cancer in the lymphatic system that is often diagnosed in early adulthood. When diagnosed early, the disease is categorised either in stage I or stage II depending on how much the disease has spread. In both cases, the treatment includes radiation and/or chemotherapy. Relapse is considered as the primary event of interest. Naturally, some subjects will die before relapse, and therefore death from any cause constitutes a competing event for relapse. Table 3 summarises the available data. The latter was randomly split into training and testing sets (80% / 20% split), using stratified sampling to ensure a similar number of events is present in both sets.

First, we considered the (semi-)parametric approaches listed in Sections 4 and 5 and compared parameter estimates obtained for the training set (Figure 1). Generally, all approaches agreed in terms of the sign of the effects: e.g. radiation therapy increases the hazard of relapse (both in the CS-hazard and sub-distribution hazard scale). However, different methods disagreed in terms of the magnitude of effect sizes. In particular, for approaches based on a CS specification, lasso penalised regression and model-based boosting approaches led to lower estimates (in absolute value) than the Cox model in (6). This can be expected due to the shrinkage/sparsity that is introduced by such methods. Cox-lasso with λ set as λ_{1se} led to the most parsimonious model with several effect sizes set to be exactly equal to zero.

Finally, for those approaches that enable risk prediction, we estimated the risk of observing a relapse before $t = 5$ years for all individuals in the test set (Figure 2). Overall, the predictions obtained by the different approaches were highly correlated. As it may be expected, predictions were nearly identical when comparing the more traditional approaches (CS-Cox, Fine and Gray and direct binomial). With respect to those methods, the strongest discrepancies were observed for RSF and dependent DP. To further explore these discrepancies, we considered the behaviour of the predicted probabilities as a function of age, sex and treatment (Figure 3). Generally, we observed that the predicted risk was higher for older individuals. However, whilst the Fine and Gray approach led to a clear separation of risk scores according to sex and treatment, a more complex pattern was observed for RSF.

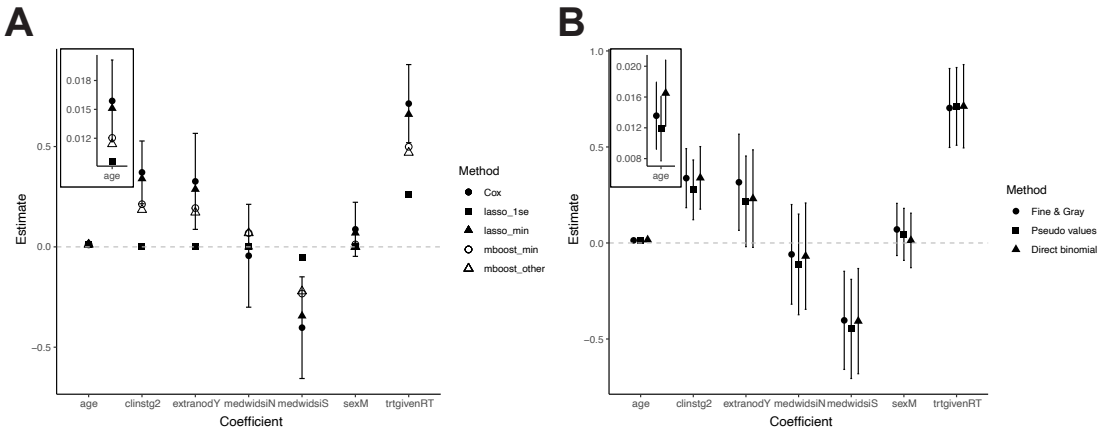

Figure 1 Parameter estimation. Parameter estimates for covariate-specific effects for the primary event type (relapse) are compared across methods. **A** shows estimates obtained for models defined under a CS-hazards specification: cause-specific Cox (Cox), Cox-lasso using λ_{1se} (lasso 1se; $\lambda = 0.034$), Cox-lasso using λ_{min} (lasso min; $\lambda = 0.002$), model-based Cox boosting where B is chosen to minimise the negative cross-validated likelihood (mboost min, $B = 292$) and model-based Cox boosting with $B = 250$ (mboost other). **B** shows estimates obtained for models defined under a CIF specification. In both cases, due to differences in scale, an inset shows parameter estimates associated to age.

Table 3 Descriptive statistics for the HD dataset. Age is given in years. For sex, F=female and M=Male. For treatment given (trtgiven), RT=Radiation, CMT=Chemotherapy and radiation. Fos size of mediastinum involvement (medwidsi), N=No, S=Small, L=Large. For extranodal disease (extranod), Y=Extranodal disease, N=Nodal disease. For clinical stage (clinstg), 1=Stage I, 2=Stage II. Continuous variables are described by using mean ± SD, and categorical variables are described as frequencies and percentages.

Characteristic		Overall $n = 865$	Relapse $n = 291$	Dead $n = 135$	Censored $n = 439$
Age		35.3 (15.5)	37.6 (17.2)	47.5 (17.3)	30.0 (10.4)
Sex					
	F	402 (46.5%)	132 (45.4%)	47 (34.8%)	223 (50.8%)
	M	463 (53.5%)	159 (54.6%)	88 (65.2%)	216 (49.2%)
Treatment (trtgiven)					
	CMT	249 (28.8%)	61 (21.0%)	42 (31.1%)	146 (33.3%)
	RT	616 (71.2%)	230 (79.0%)	93 (68.9%)	293 (66.7%)
Mediastinum involvement (medwids)					
	L	113 (13.1%)	36 (12.4%)	10 (7.4%)	67 (15.3%)
	N	464 (53.6%)	171 (58.8%)	92 (68.1%)	201 (45.8%)
	S	288 (33.3%)	84 (28.9%)	33 (24.4%)	171 (39.0%)
Extranodal disease (extranod)					
	N	786 (90.9%)	263 (90.4%)	125 (92.6%)	398 (90.7%)
	Y	79 (9.1%)	28 (9.6%)	10 (7.4%)	41 (9.3 %)
Clinical stage (clinstg)					
	1	296 (34.2%)	97 (33.3%)	58 (43.0%)	141 (32.1%)
	2	469 (65.8%)	184 (66.7%)	77 (57.0%)	298 (67.9%)

12 Evaluating performance

When proposing a new method, researchers are often interested in evaluating and comparing its performance. For example, for (semi-)parametric models, one may use synthetic data to assess whether parameter estimates are unbiased. For approaches that include variable selection, one may evaluate their ability to identify a correct set of input variables. When the goal is to perform risk prediction, the emphasis is on evaluating how well a method is able to predict *whether* and/or *when* specific event types will occur. To evaluate predictive performance, an external (or test) dataset that was not used to fit the model could be used. However, internal validation (e.g. via boostrapping or cross-validation) is also important, particularly for small datasets or when the number of observed events is small (Steyerberg and Harrell, 2016).

Recently, Van Geloven et al. (2022) discussed how to evaluate predictive performance in competing risks settings, providing examples in `R` (see `https://github.com/survival-lumc/ValidationCompRisks`). They focused on cases in which the goal is to predict *whether* the event of interest will occur within a given time-frame (e.g. 5-year survival). Van Geloven et al. (2022) emphasised the need to evaluate different aspects of predictive performance including *calibration*, something that is often overlooked when developing risk prediction models (Van Calster et al., 2019). A well calibrated model will assign the correct event probability at all levels of predicted risk. In practice, calibration is often assessed graphically, but numerical summaries are also available (see e.g. Van Calster et al., 2019; Huang et al., 2020). Another important aspects are *discrimination*, i.e. whether the model assigns a higher risk to individuals who experience the event earlier. Measures for discrimination include the concordance index (Wolbers et al., 2014; Ahuja and der Schaar, 2019), and time-dependant receiver operating characteristic (ROC) curves (Blanche et al., 2019; Saha and Heagerty, 2010). Here, we briefly describe some of the metrics that can be used to evaluate these aspects.

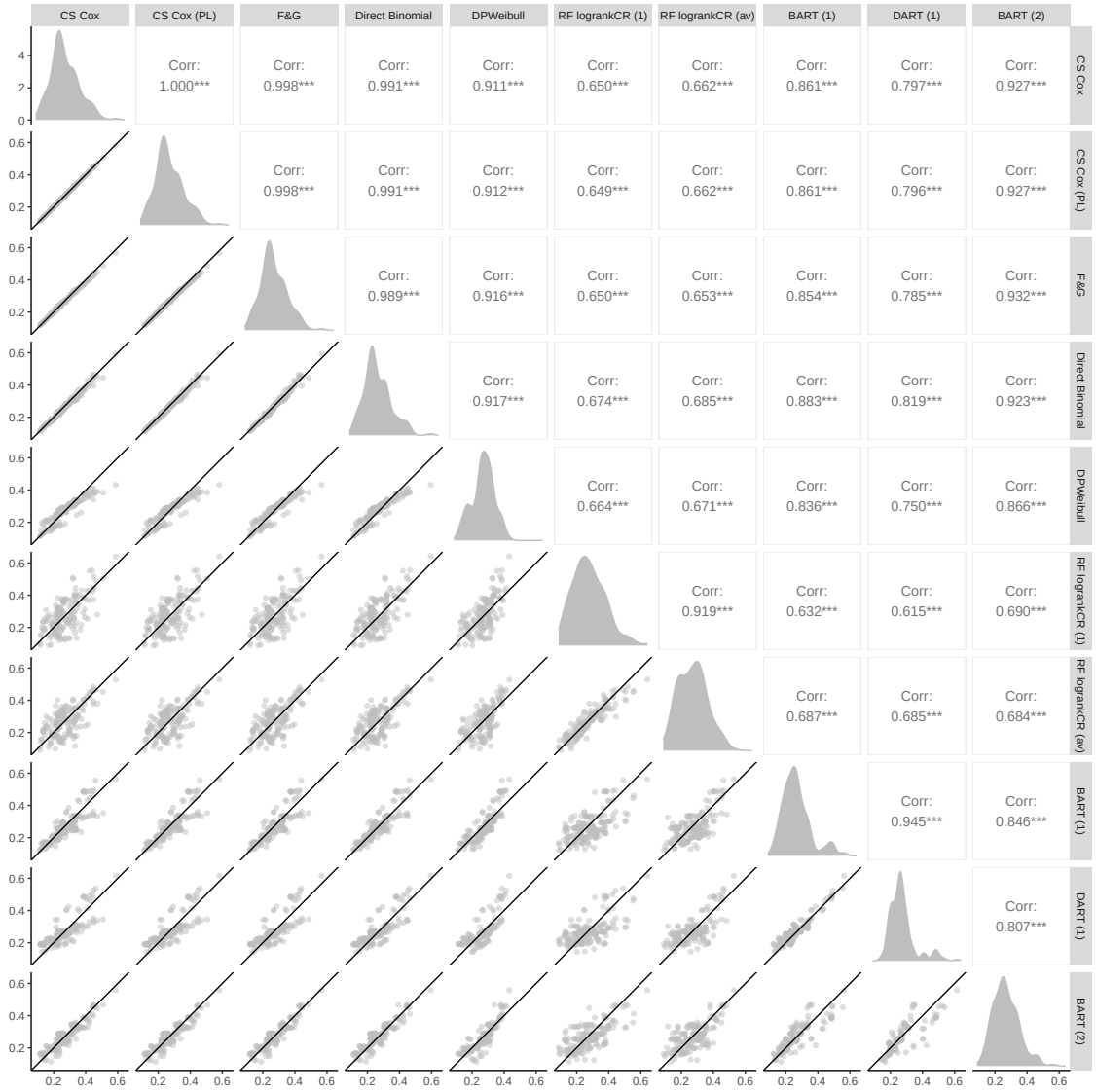

Figure 2 Out-of-sample predictions. Comparison of individual-level predictions (test set) for the risk of observing a relapse (primary event of interest) before $t = 5$ years. For the CS-Cox model, predictions were calculated using the `riskRegression` R package which takes into account competing events when calculating predictions. The first column uses an exponential approximation, such that $S(t) = exp(-H(t)_1 - H(t)_2)$, where $H(t)_j$ denotes the cumulative hazard for cause j at time t. The second column uses a product limit estimator that ensures $\text{CIF}_1(t|X) + \text{CIF}_2(t|X) = 1$, as $t \to \infty$. For random forests (RF), two splitting rules were considered: a modified Gray's criteria focusing on the first event type (relapse), and considering the average between the first and second (death) event types. DPWeibull denotes dependent DP. BART (1) and BART (2) denote the first and second formulation of the BART model, respectively. DART (1) is a variation of the first BART formulation which enables feature selection.

Concordance. A popular metric to assess discrimination in the context of survival models is via a *concordance index* (also referred to as C-index, Harrell et al., 1982). Generally, higher C-index indicates better

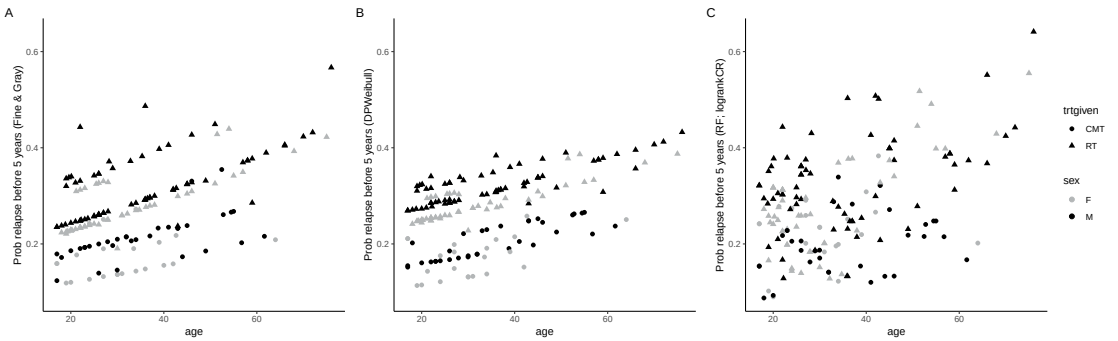

Figure 3 Out-of-sample predictions with respect to age, sex and treatment. Individual-level predictions (test set) for the risk of observing a relapse (primary event of interest) before $t = 5$ years as a function of age, sex and treatment. **A**: Fine and Gray, **B**: dependent DP, **C**: RSF with a splitting rule based on a modified Gray's criteria focusing on the first event type (relapse).

discrimination (and a value equal to 0.5 indicates no discrimination ability). Several definitions are available, including some that have been adapted to CR settings. For example, if the aim is to predict whether kth event type is observed prior to a pre-specified time τ, Wolbers et al. (2014) proposed the following cause-specific time-dependent C-index:

$$C_k(\tau) = \Pr\left(\text{CIF}_k(\tau \mid \mathbf{x}_i) > \text{CIF}_k(\tau \mid \mathbf{x}_j) \mid \{Z_i = k\} \wedge \{T_i \leq \tau\} \wedge \{T_i \leq T_j \vee Z_j \neq k\}\right), \quad (41)$$

for a random pair of individuals (i and j). This metric quantifies if model is able to correctly rank the risk of observing the . More recently, Ahuja and der Schaar (2019) proposed a joint concordance index to evaluate the model's ability to correctly predict both the event type and time. Their approach may be of interest in cases where more than one event type is of interest. If the interest is to assess discrimination across the whole follow-up period rather than at a specific time-point τ, a weighted average of $C_k(\tau)$ could be used (see e.g. the approach proposed by Antolini et al. (2005) for a single event type).

Brier score. Schoop et al. (2011) adapted the proper scoring score introduced by Graf et al. (1999) to competing risks settings. For a given prediction time τ, the Brier score for cause k is defined as the weighted average of the squared differences between the cause-specific event indicators and the predicted cause-specific survival probabilities:

$$\text{BS}_k(\tau) = \tfrac{1}{n} \sum_{i=1}^{n} w_i \left[\mathbb{1}\{T_i \leq \tau, Z_i = k\} - \Pr(T_i \leq \tau, Z_i = k \mid \mathbf{x}_i)\right]^2, \quad (42)$$

where the weights w_i capture right censoring (Schoop et al., 2011, , Theorem 4.1). This can be interpreted as a metric of overall performance, as it encompasses both calibration and discrimination. To summarise performance across a range of time-points, an integrated Brier Score can be used (Graf et al., 1999).

The lower the value of (42), the better. However, the absolute value of (42) is difficult to interpret as its scale depends on the number of observed events. As an alternative, an scaled version of (42) can be used. The scaled Brier score can be computed as follows (Van Geloven et al., 2022):

$$\text{BS}_k(\tau)^{\text{scaled}} = 1 - \frac{\text{BS}_k(\tau)}{\text{BS}_k(\tau)^{\text{null}}}, \quad (43)$$

where $\text{BS}_k(\tau)^{\text{null}}$ denotes the Brier score under the null model (no covariates) and which can be computed using the Aalen-Johansen estimator (Aalen and Johansen, 1978). The later lies between 0 and 1, where 1 indicates perfect predictions.

13 Discussion

We summarised a broad range of competing risks modelling techniques, covering both the statistical and machine learning literature. Our objective is to provide a synthesised catalogue, with unified notation and interpretation. We also briefly review the metrics that are most commonly used to evaluate and compare predictive performance. Furthermore, we discuss some practical considerations that may help practitioners to decide which method is more appropriate to address their scientific question using their available data.

In order to promote the usage of state-of-the-art approaches, we point out to available software and, demonstrate its practical implementation through reproducible `R` vignettes. Emphasis on reproducibility is critical when developing and evaluating new methods. While making the implementation of the method publicly available using version control hosting tools; such as GitHub or BitBucket, helps towards this goal; this is not enough. The code must be well documented and, when possible, accompanied with the raw data (synthetic or real) that was used to assess performance. It is also important provide details on how such data was generated or processed, as well as a clear description of any *ad hoc* choices made (e.g. inclusion/exclusion criteria). For instance, the Surveillance, Epidemiology, and End Results (SEER) Program[1] datasets have been employed to showcase several CR methods (e.g. Zhang and Zhou, 2018; Alaa and van der Schaar, 2017; Bellot and van der Schaar, 2018b; Nemchenko et al., 2018; Bellot and van der Schaar, 2018a). However, detailed information on how the dataset used was preprocessed is usually not provided (in some cases, authors do not even provide the full list of covariates used in the analysis). Similar issues have been reported when using the MIMIC database (Johnson et al., 2017). This highlights an urgent need for more systematic and reproducible benchmark pipelines (Mangul et al., 2019) for competing risks methods which will help to reduce the gap between developers and users. Such benchmark would ideally help users to better understand the advantages and limitations of each method, including sample size requirements (both in terms of total sample size and the number of events) and scalability, among others. For this purpose, the inclusion of a wide range of data sets (with varying characteristics) is important.

Inevitably, as new methods are developed, this review will be outdated. To address this, we aim to have our GitHub repository (`www.github.com/VallejosGroup/CompRisksVignettes`) as a living resource in which others can contribute additional vignettes via pull requests. We hope this will help to improve accessibility to novel competing risks approaches as they are developed.

Acknowledgements KMG was supported by an MRC University Unit grant to the MRC Human Genetics Unit. NC-C was supported by the Medical Research Council and University of Edinburgh via a Precision Medicine PhD studentship (MR/N013166/1). CAV was supported by a Chancellor's Fellowship provided by The University of Edinburgh. CAV was also supported by a British Heart Foundation-Turing Cardiovascular Data Science Award (BCDSA/100003). For the purpose of open access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The authors would like to acknowledge the support of Rodney Sparapani, Shu-Kay Angus Ng and Geoffrey McLachlan. They kindly provided insight about their methods, and shared and/or pointed out to code for their implementation.

Conflict of Interest
The authors have declared no conflict of interest.

Appendix

A. Parameter estimation for the Cox proportional CS hazard model

Separate CPH models are used for each event type, treating competing events as censored observations. Regression coefficients in (6) can be estimated without the need to infer the corresponding baseline hazards. For the k-th event type, the corresponding regression coefficients, $\boldsymbol{\beta}_k$, in (6) can be estimated by

[1] https://seer.cancer.gov

maximising the partial likelihood:

$$\mathcal{L}^{\text{CS}}(\boldsymbol{\beta}_k) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}{\sum_{l \in R_i} \exp(\mathbf{x}_l^\top \boldsymbol{\beta}_k)} \right)^{\mathbb{1}\{Z_i = k\}}, \tag{44}$$

where R_i denotes the set of observations at risk at time t_i, i.e. subjects that are not censored or that have not experienced a competing event by time t_i.

B. Lunn-McNeil augmented layout

Assume we have two event types, $K = 2$. The following table shows the observed data for 3 subjects. The first, experienced event type 2 at time 10, the second is assumed to be censored by time 70, and the third experienced event type 1 at time 14.

Table 4 Original layout

Individual	Event time (T)	Event type (Z)	Covariates
1	10	2	$\mathbf{x}_1^\top$
2	70	0	$\mathbf{x}_2^\top$
3	14	1	$\mathbf{x}_3^\top$

The augmented layout required for LM approach necessitates to have 2 rows per subject, one for each competing event. In addition, we add event type indicators δ_{ik}.

Table 5 Augmented layout for LM

Individual	Event time (T)	Event type (Z)	Event type indicator (δ)		Covariates
			$Z = 1$	$Z = 2$	
1	10	2	0	1	$\mathbf{x}_1^\top$
1	10	0	1	0	$\mathbf{x}_1^\top$
2	70	0	1	0	$\mathbf{x}_2^\top$
2	70	0	0	1	$\mathbf{x}_2^\top$
3	14	1	1	0	$\mathbf{x}_3^\top$
3	14	0	0	1	$\mathbf{x}_3^\top$

A.3. R vignettes

Vignettes showcasing the usage of some methods are available online at: `https://github.com/VallejosGroup/CompRisksVignettes`.

References

Odd O. Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978. ISSN 03036898, 14679469. URL `http://www.jstor.org/stable/4615704`.

Kartik Ahuja and Mihaela van der Schaar. Joint concordance index. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 2206–2213, 2019. doi: 10.1109/IEEECONF44664.2019.9048941.

Ahmed M Alaa and Mihaela van der Schaar. Deep multi-task Gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334, 2017.

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266, 2012. ISSN 1935-8237. doi: 10.1561/2200000036. URL https://doi.org/10.1561/2200000036.

Federico Ambrogi and Thomas H Scheike. Penalized estimation for competing risks regression with applications to high-dimensional covariates. *Biostatistics*, 17(4):708–721, 2016.

Per Kragh Andersen and Niels Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12):1074–1088, 2012. doi: https://doi.org/10.1002/sim.4385. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4385.

Per Kragh Andersen, Steen Z Abildstrom, and Susanne Rosthøj. Competing risks as a multi-state model. *Statistical Methods in Medical Research*, 11(2):203–215, 2002. doi: 10.1191/0962280202sm281ra. URL https://doi.org/10.1191/0962280202sm281ra. PMID: 12040697.

Per Kragh Andersen, John P. Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003. ISSN 00063444. URL http://www.jstor.org/stable/30042016.

Eleni-Rosalina Andrinopoulou, Dimitris Rizopoulos, Johanna J. M. Takkenberg, and Emmanuel Lesaffre. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine*, 33(18):3167–3178, 2014. doi: https://doi.org/10.1002/sim.6158. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6158.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. doi: 10.1002/sim.2427. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2427.

Peter C. Austin and Jason P. Fine. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Statistics in Medicine*, 36(27):4391–4400, 2017. doi: 10.1002/sim.7501. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7501.

Peter C. Austin, Douglas S. Lee, and Jason P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016. doi: 10.1161/CIRCULATIONAHA.115.017719. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.115.017719.

Peter C Austin, Arthur Allignol, and Jason P Fine. The number of primary events per variable affects estimation of the subdistribution hazard competing risks model. *Journal of clinical epidemiology*, 83:75–84, 2017.

Peter C. Austin, Ewout W. Steyerberg, and Hein Putter. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Statistics in Medicine*, 40(19):4200–4212, 2021. doi: https://doi.org/10.1002/sim.9023. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9023.

Brett K Beaulieu-Jones and Casey S Greene. Reproducibility of computational workflows is automated using continuous analysis. *Nature biotechnology*, 35(4):342–346, 2017.

Alexis Bellot and Mihaela van der Schaar. Multitask boosting for survival analysis with competing risks. In *Advances in Neural Information Processing Systems 31*, pages 1390–1399. 2018a. URL http://papers.nips.cc/paper/7413-multitask-boosting-for-survival-analysis-with-competing-risks.pdf.

Alexis Bellot and Mihaela van der Schaar. Tree-based Bayesian mixture model for competing risks. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 910–918, 2018b. URL http://proceedings.mlr.press/v84/bellot18a.html.

Harald Binder and Martin Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics*, 9(1):14, 2008.

Harald Binder, Arthur Allignol, Martin Schumacher, and Jan Beyersmann. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25(7):890–896, 02 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp088. URL https://doi.org/10.1093/bioinformatics/btp088.

Paul Blanche, Michael W Kattan, and Thomas A Gerds. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Carl Boettiger. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review*, 49(1): 71–79, 2015.

Anne-Laure Boulesteix and Torsten Hothorn. Testing the additional predictive value of high-dimensional molecular data. *BMC bioinformatics*, 11(1):1–11, 2010.

Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998. ISSN 00905364. URL http://www.jstor.org/stable/120055.

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

Norman E Breslow. Discussion of professor Cox's paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34:216–217, 1972.

Danilo Bzdok, Naomi Altman, and Martin Krzywinski. Statistics versus machine learning. *Nature Methods*, 15(5): 233–234, 2018.

Mark Carpenter. Survival analysis: A self-learning text. *Technometrics*, 39(2):228–229, 1997. doi: 10.1080/00401706.1997.10485091.

Richard Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann, 1993.

I-Shou Chang, Chao A. Hsiung, Chi-Chung Wen, Yuh-Jenn Wu, and Che-Chi Yang. Non-parametric maximum-likelihood estimation in a semiparametric mixture model for competing-risks data. *Scandinavian Journal of Statistics*, 34(4):870–895, 2007. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/41548585.

Yu Cheng. Modeling cumulative incidences of dementia and dementia-free death using a novel three-parameter logistic function. *The International Journal of Biostatistics*, 5(1), 2009. doi: doi:10.2202/1557-4679.1183. URL https://doi.org/10.2202/1557-4679.1183.

Christopher Cox, Haitao Chu, Michael F. Schneider, and Alvaro Muñoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26(23):4352–4374, 2007. doi: 10.1002/sim.2836. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2836.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34 (2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x.

David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

D.R. Cox. *Renewal theory*. Methuen, 1962.

Andreas Damianou and Neil Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, 2013. URL http://proceedings.mlr.press/v31/damianou13a.html.

Dominic Danks and Christopher Yau. Derivative-based neural modelling of cumulative distribution functions for survival analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 7240–7256. PMLR, 2022.

Kazeem Adesina Dauda, Biswabrata Pradhan, B. Uma Shankar, and Sushmita Mitra. Decision tree for modeling survival data with competing risks. *Biocybernetics and Biomedical Engineering*, 39(3):697 – 708, 2019. ISSN 0208-5216. doi: https://doi.org/10.1016/j.bbe.2019.05.001. URL http://www.sciencedirect.com/science/article/pii/S0208521619300245.

Riccardo De Bin. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computational Statistics*, 31(2):513–531, 2016. ISSN 1613-9658. doi: 10.1007/s00180-015-0642-2. URL https://doi.org/10.1007/s00180-015-0642-2.

Ton De Waal, Jeroen Pannekoek, and Sander Scholtus. *Handbook of statistical data editing and imputation*, volume 563. John Wiley & Sons, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

David Engler and Yi Li. Survival analysis with high-dimensional covariates: An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009. doi: doi:10.2202/1544-6115.1423. URL https://doi.org/10.2202/1544-6115.1423.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291069.

Jianqing Fan and Runze Li. Variable Selection for Cox's proportional Hazards Model and Frailty Model. *The Annals of Statistics*, 30(1):74 – 99, 2002. doi: 10.1214/aos/1015362185. URL https://doi.org/10.1214/aos/1015362185.

X Fan. *Bayesian nonparametric inference for competing risks data*. PhD thesis, Medical College of Wiscosin, 2008.

Jason P. Fine. Regression modeling of competing crude failure probabilities. *Biostatistics*, 2(1):85–97, 03 2001. ISSN 1465-4644. doi: 10.1093/biostatistics/2.1.85. URL https://doi.org/10.1093/biostatistics/2.1.85.

Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. doi: 10.1080/01621459.1999.10474144. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144.

Jason P Fine, Hongyu Jiang, and Rick Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, 28(2):337–407, 04 2000. doi: 10.1214/aos/1016218223. URL https://doi.org/10.1214/aos/1016218223.

Zhixuan Fu, Chirag R Parikh, and Bingqing Zhou. Penalized variable selection in competing risks regression. *Lifetime data analysis*, 23(3):353–376, 2017.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999. doi: https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2529::AID-SIM274⟩3.0.CO;2-5. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819990915/30%2918%3A17/18%3C2529%3A%3AAID-SIM274%3E3.0.CO%3B2-5.

Nadine Grambauer, Martin Schumacher, and Jan Beyersmann. Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in medicine*, 29(7-8):875–884, 2010.

Robert J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3):1141–1154, 1988. ISSN 00905364. URL http://www.jstor.org/stable/2241622.

Il Do Ha, Minjung Lee, Seungyoung Oh, Jong-Hyeon Jeong, Richard Sylvester, and Youngjo Lee. Variable selection in subdistribution hazard frailty models with competing risks data. *Statistics in Medicine*, 33(26):4590–4604, 2014. doi: https://doi.org/10.1002/sim.6257. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6257.

Bernhard Haller, Georg Schmidt, and Kurt Ulm. Applying competing risks regression models: an overview. *Lifetime data analysis*, 19(1):33–58, 2013.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

W. K. Hasting. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

James J. Heckman and Bo E. Honoré. The identifiability of the competing risks model. *Biometrika*, 76(2):325–330, 1989. ISSN 00063444. URL http://www.jstor.org/stable/2336666.

Graeme L. Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. A comparison of joint models for longitudinal and competing risks data, with application to an epilepsy drug randomized controlled trial. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 181(4):pp. 1105–1123, 2018. ISSN 09641998, 1467985X. URL https://www.jstor.org/stable/48547194.

Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, 2020. doi: 10.1146/annurev-statistics-031219-041110. URL https://doi.org/10.1146/annurev-statistics-031219-041110.

Philip Hougaard. Multi-state models: a review. *Lifetime data analysis*, 5:239–264, 1999.

Jin-Jian Hsieh, Weijing Wang, and A. Adam Ding. Regression analysis based on semicompeting risks data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):3–20, 2008. doi: 10.1111/j.1467-9868.2007.00621.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00621.x.

Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

Joseph G Ibrahim, Ming-Hui Chen, and Steven N MacEachern. Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics*, 27(4):701–717, 1999.

Hemant Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1: 519–537, 2007.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.

Hemant Ishwaran, Udaya B. Kogalur, Eiran Z. Gorodeski, Andy J. Minn, and Michael S. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010. doi: 10.1198/jasa.2009.tm08622. URL https://doi.org/10.1198/jasa.2009.tm08622.

Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 04 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxu010. URL https://doi.org/10.1093/biostatistics/kxu010.

Silke Janitza and Gerhard Tutz. Prediction models for time discrete competing risks. Technical report, Department of Statistics, Ludwig-Maximilians-Universität München, 2015.

Jong-Hyeon Jeong and Jason Fine. Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):187–200, 2006. doi: https://doi.org/10.1111/j.1467-9876.2006.00532.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2006.00532.x.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. Reproducibility in critical care: a mortality prediction case study. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 361–376. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/johnson17a.html.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *In Proc. 3rd International Conference on Learning Representations (ICLR)*, 2015.

John P. Klein and Per Kragh Andersen. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1):223–229, 2005. doi: 10.1111/j.0006-341X.2005.031209.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2005.031209.x.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

Athanasios Kottas. Nonparametric bayesian survival analysis using mixtures of weibull distributions. *Journal of Statistical Planning and Inference*, 136(3):578–596, 2006. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2004.08.009. URL https://www.sciencedirect.com/science/article/pii/S0378375804003465.

Anthony Y.C. Kuk. A semiparametric mixture model for the analysis of competing risks data. *Australian Journal of Statistics*, 34(2):169–180, 1992. doi: 10.1111/j.1467-842X.1992.tb01351.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1992.tb01351.x.

Deborah Kuk and Ravi Varadhan. Model selection in competing risks regression. *Statistics in medicine*, 32(18): 3077–3088, 2013.

Martin G. Larson and Gregg E. Dinse. A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3):201–211, 1985. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2347464.

Aurelien Latouche, Arthur Allignol, Jan Beyersmann, Myriam Labopin, and Jason P. Fine. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, 66(6):648–653, 2013. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2012.09.017. URL https://www.sciencedirect.com/science/article/pii/S0895435612003484.

Changhee Lee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1): 13–22, 04 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.1.13. URL https://doi.org/10.1093/biomet/73.1.13.

Kenneth S Lomax. Business failures: Another example of the analysis of failure data. *Journal of the American statistical association*, 49(268):847–852, 1954.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Mary Lunn and Don McNeil. Applying Cox regression to competing risks. *Biometrics*, 51(2):524–532, 1995. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2532940.

Alireza S. Mahani and Mansour T. A. Sharabiani. Bayesian, and non-bayesian, cause-specific competing-risk analysis for parametric and nonparametric survival functions: The r package cfc. *Journal of Statistical Software, Articles*, 89(9):1–29, 2019. ISSN 1548-7660. doi: 10.18637/jss.v089.i09. URL https://www.jstatsoft.org/v089/i09.

Serghei Mangul, Lana S Martin, Brian L Hill, Angela Ka-Mei Lam, Margaret G Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. Systematic benchmarking of omics computational tools. *Nature communications*, 10 (1):1–11, 2019.

Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966.

Luís Meira-Machado and Marta Sestelo. Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal*, 61(2):245–263, 2019. doi: https://doi.org/10.1002/bimj.201700200. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700200.

Xiao-Li Meng and Donald B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2):267–278, 1993. ISSN 00063444. URL http://www.jstor.org/stable/2337198.

C. Nagpal, X. R. Li, and A. Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2021. doi: 10.1109/JBHI.2021.3052441.

Anton Nemchenko, Trent Kyono, and Mihaela Van Der Schaar. Siamese survival analysis with competing risks. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 260–269, 2018. ISBN 978-3-030-01424-7.

SK Ng and GJ McLachlan. An em-based semi-parametric mixture model approach to the regression analysis of competing-risks data. *Statistics in Medicine*, 22(7):1097–1111, 2003.

M. A. Nicolaie, Hans C. van Houwelingen, and H. Putter. Vertical modeling: a pattern mixture approach for competing risks modeling. *Statistics in Medicine*, 29(11):1190–1205, 2010. doi: 10.1002/sim.3844. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3844.

MA Nicolaie, HC van Houwelingen, and H Putter. Vertical modelling: Analysis of competing risks data with missing causes of failure. *Statistical Methods in Medical Research*, 24(6):891–908, 2015. doi: 10.1177/0962280211432067. URL https://doi.org/10.1177/0962280211432067. PMID: 22179822.

Limin Peng and Jason P. Fine. Regression modeling of semicompeting risks data. *Biometrics*, 63(1):96–108, 2007. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/4541304.

Melania Pintilie. *Competing risks: a practical perspective*. John Wiley & Sons, 2006.

Greg Ridgeway. The state of boosting. *Computing science and Statistics*, pages 172–181, 1999.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive mcmc. *Journal of computational and graphical statistics*, 18(2):349–367, 2009.

James M. Robins and Andrea Rotnitzky. *Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers*, pages 297–331. Birkhäuser Boston, 1992. ISBN 978-1-4757-1229-2. doi: 10.1007/978-1-4757-1229-2_14. URL https://doi.org/10.1007/978-1-4757-1229-2_14.

Jacqueline E Rudolph, Catherine R Lesko, and Ashley I Naimi. Causal inference in the face of competing events. *Current epidemiology reports*, 7:125–131, 2020.

P Saha and PJ Heagerty. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66(4): 999–1011, 2010.

Thomas H. Scheike, Mei-Jie Zhang, and Thomas A. Gerds. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220, 02 2008. ISSN 0006-3444. doi: 10.1093/biomet/asm096. URL https://doi.org/10.1093/biomet/asm096.

Matthias Schmid and Moritz Berger. Competing risks analysis for discrete time-to-event data. *WIREs Computational Statistics*, 13(5):e1529, 2021. doi: https://doi.org/10.1002/wics.1529. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1529.

Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, 2011. doi: https://doi.org/10.1002/bimj.201000073. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201000073.

Marc A. Scott and Benjamin B. Kennedy. Pitfalls in pathways: some perspectives on competing risks event history analysis in education research. *Journal of Educational and Behavioral Statistics*, 30(4):413–442, 2005. ISSN 10769986, 19351054. URL http://www.jstor.org/stable/3701297.

Mark Segal and Yuanyuan Xiao. Multivariate random forests. *WIREs Data Mining and Knowledge Discovery*, 1(1): 80–87, 2011. doi: 10.1002/widm.12. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.12.

Haiwen Shi, Yu Cheng, and Jong-Hyeon Jeong. Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal*, 55(1):82–96, 2013. doi: https://doi.org/10.1002/bimj.201200011. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201200011.

Yushu Shi, Purushottam Laud, and Joan Neuner. A dependent Dirichlet process model for survival data with competing risks. *Lifetime Data Analysis*, 27(1):156–176, 2021.

Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: An R Package for machine learning in survival analysis. *Bioinformatics*, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab039. URL https://doi.org/10.1093/bioinformatics/btab039.

Rodney Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric competing risks analysis using Bayesian additive regression trees. *Statistical Methods in Medical Research*, 29(1):57–77, 2020. doi: 10.1177/0962280218822140. URL https://doi.org/10.1177/0962280218822140. PMID: 30612519.

Rodney Sparapani, Charles Spanbauer, and Robert McCulloch. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1): 1–66, 2021. doi: 10.18637/jss.v097.i01.

Ewout W Steyerberg and Frank E Harrell. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology*, 69:245–247, 2016.

Han Sun and Xiaofeng Wang. High-dimensional feature selection in competing risks modeling: A stable approach using a split-and-merge ensemble algorithm. *Biometrical Journal*, 65(2):2100164, 2023. doi: https://doi.org/10.1002/bimj.202100164. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202100164.

Elisavet Syriopoulou, Sarwar I Mozumder, Mark J Rutherford, and Paul C Lambert. Estimating causal effects in the presence of competing events using regression standardisation with the stata command standsurv. *BMC Medical Research Methodology*, 22(1):1–16, 2022.

Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

Terry M Therneau. *A Package for Survival Analysis in R*, 2023. URL https://CRAN.R-project.org/package=survival. R package version 3.5-5.

Robert Tibshrani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997. doi: https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4⟨385::AID-SIM380⟩3.0.CO;2-3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%281997 0228%2916%3A4%3C385%3A%3AAID-SIM380%3E3.0.CO%3B2-3.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, 16–18 Apr 2009. URL http://proceedings.mlr.press/v5/titsias09a.html.

A Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20—22, 1975. ISSN 0027-8424. doi: 10.1073/pnas.72.1.20. URL https://europepmc.org/articles/PMC432231.

Gerhard Tutz. Competing risks models in discrete time with nominal or ordinal categories of response. *Quality and Quantity*, 29(4):405–420, 1995.

Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7, 2019.

Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. doi: doi:10.2202/1544-6115.1309. URL https://doi.org/10.2202/1544-6115.1309.

Nan Van Geloven, Daniele Giardiello, Edouard F Bonneville, Lucy Teece, Chava L Ramspek, Maarten van Smeden, Kym I E Snell, Ben van Calster, Maja Pohar-Perme, Richard D Riley, Hein Putter, and Ewout Steyerberg. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ*, 377, 2022. doi: 10.1136/bmj-2021-069249. URL https://www.bmj.com/content/377/bmj-2021-069249.

Pierre JM Verweij and Hans C Van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24): 2305–2314, 1993.

Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6):1–36, February 2019. ISSN 0360-0300.

Andrew Wey, John Connett, and Kyle Rudser. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3):537–549, 2015.

P. R. Williamson, R. Kolamunnage-Dona, P. Philipson, and A. G. Marson. Joint modelling of longitudinal and competing risks data. *Statistics in Medicine*, 27(30):6426–6438, 2008. doi: https://doi.org/10.1002/sim.3451. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3451.

Marcel Wolbers, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, 02 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxt059. URL https://doi.org/10.1093/biostatistics/kxt059.

Robert L. Wolpert and Katja Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2): 251–267, 06 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.2.251. URL https://doi.org/10.1093/biomet/85.2.251.

Jinfeng Xu, John D. Kalbfleisch, and Beechoo Tai. Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725, 2010. doi: https://doi.org/10.1111/j.1541-0420.2009.01340.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2009.01340.x.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894 – 942, 2010. doi: 10.1214/09-AOS729. URL https://doi.org/10.1214/09-AOS729.

Hao Helen Zhang and Wenbin Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 05 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm037. URL https://doi.org/10.1093/biomet/asm037.

Mei-Jie Zhang, XU Zhang, and Thomas H Scheike. Modeling cumulative incidence function for competing risks data. *Expert review of clinical pharmacology*, 1(3):391–400, 2008.

Quan Zhang and Mingyuan Zhou. Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks. In *Advances in Neural Information Processing Systems*, pages 5002–5013, 2018.

Ming Zheng and John P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995. ISSN 00063444. URL http://www.jstor.org/stable/2337633.

