# Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks

**Quan Zhang**
McCombs School of Business
The University of Texas at Austin
Austin, TX 78712
quan.zhang@mccombs.utexas.edu

**Mingyuan Zhou**
McCombs School of Business
The University of Texas at Austin
Austin, TX 78712
mingyuan.zhou@mccombs.utexas.edu

## Abstract

We propose Lomax delegate racing (LDR) to explicitly model the mechanism of survival under competing risks and to interpret how the covariates accelerate or decelerate the time to event. LDR explains non-monotonic covariate effects by racing a potentially infinite number of sub-risks, and consequently relaxes the ubiquitous proportional-hazards assumption which may be too restrictive. Moreover, LDR is naturally able to model not only censoring, but also missing event times or event types. For inference, we develop a Gibbs sampler under data augmentation for moderately sized data, along with a stochastic gradient descent maximum a posteriori inference algorithm for big data applications. Illustrative experiments are provided on both synthetic and real datasets, and comparison with various benchmark algorithms for survival analysis with competing risks demonstrates distinguished performance of LDR.

## 1 Introduction

In survival analysis, one can often use nonparametric approaches to flexibly estimate the survival function from lifetime data, such as the Kaplan–Meier estimator [1], or to estimate the intensity of a point process for event arrivals, such as the isotonic Hawkes process [2] and neural Hawkes process [3] that can be applied to the analysis of recurring events. When exploring the relationship between the covariates and time to events, existing survival analysis methods often parameterize the hazard function with a weighted linear combination of covariates. One of the most popular ones is the Cox proportional hazards model [4], which is semi-parametric in that it assumes a non-parametric baseline hazard rate to capture the time effect. These methods are often applied to population-level studies that try to unveil the relationship between the risk factors and hazard function, such as to what degree a unit increase in a covariate is multiplicative to the hazard rate. However, the interpretability is often obtained by sacrificing model flexibility, because the proportional-hazards assumption is violated when the covariate effects are non-monotonic. For example, both very high and very low ambient temperature were related to high mortality rates in Valencia, Spain, 1991-1993 [5], and a significantly increased mortality rate is associated with both underweight and obesity [6].

To accommodate nonlinear covariate effects such as non-monotonicity, existing (semi-)parametric models often expand the design matrix with transformed data, like the basis functions of smoothing splines [7, 8] and other transformations guided by subjective knowledge. Instead of using hand-designed data transformations, there are several recent studies in machine learning that model complex covariate dependence with flexible functions, such as deep exponential families [9], neural networks [10–12] and Gaussian processes [13]. With enhanced flexibilities, these recent approaches are often good at assessing individual risks, such as predicting a patient's hazard function or survival time. However, except for the Gaussian process whose results are not too difficult to interpret for

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

low-dimensional covariates, they often have difficulty in explaining how the survival is impacted by which covariates, limiting their use in survive analysis where interpretability plays a critical role. Some approaches discretize the real-valued survival time and model the surviving on discrete time points or intervals [14–17]. They transform the time-to-event modeling problem into regression, classification, or ranking ones, at the expense of losing continuity information implied by the survival time and potentially having inconsistent categories between training and testing.

In survival analysis, it is very common to have competing risks, in which scenario the occurrence of an event under a risk precludes events under any other risks. For example, if the event of interest is death, then all possible causes of death are competing risks to each other, since a subject that died of one cause would never die of any other cause. Apart from modeling the time to event, in the presence of competing risks, it is also important to model the event type, or under which risk the event is likely to occur first. Though one can censor subjects with an occurrence of the event under a competing risk other than the risk of special interest, so that every survival model that can handle censoring is able to model competing risks, it is problematic to violate the principle of non-informative censoring [18,19]. The analysis of competing risks should be carefully designed and people often model two types of hazard functions, cause specific [20,21] and subdistribution [20–22] hazard functions. The former applies to studying etiology of diseases, while the latter is favorable when developing prediction models and risk-censoring systems [19].

In the analysis of competing risks, there is also a trade-off between interpretability and flexibility. The aforementioned cause specific and subdistribution hazard functions use a Cox model with competing risk [19,23] and a Fine-Gray subdistribution model [22], respectively, which are both proportional hazard models. Both models are semi-parametric, and assume that the hazard rate is proportional to the exponential of the inner product of the covariate and regression coefficient vectors, along with a nonparametric baseline hazard function. However, the existence of non-monotonic covariate effects can easily challenge and break the proportional-hazards assumption inherited from their corresponding single-risk model. This barrier has been surmounted by nonparametric approaches, such as random survival forests [24], Gaussian processes with a single layer [25] or two [26], and classification-based neural networks that discretize the survival time [27]. These models are designed for competing risks, using the covariates as input and the survival times (or their monotonic transformation) or probabilities as output. Though having good model fit, the non-parametric approaches are specifically used for studies at an individual level, such as predicting the survival time, but not able to tell how the covariates affect the survival or cumulative incidence functions [22,28]. Moreover, it might be questionable for Alaa and van der Schaar [26] to assume a normal distribution on survival times which are positive almost surely and asymmetric in general.

To this end, we construct Lomax delegate racing (LDR) survival model, a gamma process based nonparametric Bayesian hierarchical model for survival analysis with competing risks. The LDR survival model utilizes the race of exponential random variables to model both the time to event and event type and subtype, and uses the summation of a potentially countably infinite number of covariate-dependent gamma random variables as the exponential distribution rate parameters. It is amenable to not only censoring data, but also missing event types or event times. Code for reproducible research is available at https://github.com/zhangquan-ut/Lomax-delegate-racing-for-survival-analysis-with-competing-risks.

## 2 Exponential racing and survival analysis

Let $t \sim \text{Exp}(\lambda)$ represent an exponential distribution, with probability density function (PDF) $f(t \,|\, \lambda) = \lambda e^{-\lambda t},\ \ t \in \mathbb{R}_+$, where $\mathbb{R}_+$ represents the nonnegative side of the real line, and $\lambda > 0$ is the rate parameter such that $\mathbb{E}[t] = \lambda^{-1}$ and $\text{Var}[t] = \lambda^{-2}$. Shown below is a well-known property that characterizes a race among independent exponential random variables [29,30].

**Property 1** (Exponential racing ). *If $t_j \sim \text{Exp}(\lambda_j)$, where $j = 1, \dots, J$, are independent to each other, then $t = \min\{t_1, \dots, t_J\}$ and the argument of the minimum $y = \text{argmin}_{j \in \{1,\dots,J\}} t_j$ are independent, satisfying*

$$t \sim \text{ Exp}\left(\sum_{j=1}^{J} \lambda_j\right),\ y \sim \text{Categorical}\left(\lambda_1 \Big/ \sum_{j=1}^{J} \lambda_j, \cdots, \lambda_J \Big/ \sum_{j=1}^{J} \lambda_j\right). \tag{1}$$

Suppose there is a race among teams $j = 1, \cdots, J$, whose completion times $t_j$ follow $\text{Exp}(\lambda_j)$, with the winner being the team with the minimum completion time. Property 1 shows the winner's completion time $t$ still follows an exponential distribution and is independent of which team wins the race. In the context of survival analysis, if we consider a competing risk as a team and the latent survival time under this risk as the completion time of the team, then $t$ will be the observed time to event (or failure time) and $y$ the event type (or cause of failure). Exponential racing not only describes a natural mechanism of competing risks, but also provides an attractive modeling framework amenable to Bayesian inference, as conditioning on $\lambda_j$'s, the joint distribution of the event type $y$ and time to event $t$ becomes fully factorized as

$$P(y, t \,|\, \{\lambda_j\}_{1,J}) = \lambda_y e^{-t \sum_{j=1}^{J} \lambda_j}. \tag{2}$$

In survival analysis, it is rarely the case that both $y$ and $t$ are observed for all observations, and one often needs to deal with missing data and right or left censoring. We write $t \sim \text{Exp}_\Psi(\lambda)$ as a truncated exponential random variable defined by PDF $f_\Psi(t \,|\, \lambda) = \lambda e^{-\lambda t} / \int_\Psi \lambda e^{-\lambda u} du$, where $t \in \Psi$ and $\Psi$ is an open interval on $\mathbb{R}_+$ representing censoring. Concretely, $\Psi$ can be $(T_{r.c.}, \infty)$, indicating right censoring with censoring time $T_{r.c.}$, can be $(0, T_{l.c.})$, indicating left censoring with censoring time $T_{l.c.}$, or can be a more general case $(T_1, T_2)$, $T_2 > T_1$.

If we do not observe $y$ or $t$, or there exists censoring, we have the following two scenarios, for both of which it is necessary to introduce appropriate auxiliary variables to achieve fully factorized likelihoods: 1) If we only observe $y$ (or $t$), then we can draw $t$ (or $y$) shown in (1) as an auxiliary variable, leading to the fully factorized likelihood as in (2); 2) If we do not observe $t$ but know $t \in \Psi$ with $P(t \in \Psi \,|\, \{\lambda_j\}_{1,J}) = \int_\Psi (\sum_j \lambda_j) e^{-\sum_j \lambda_j u} du$, then we draw $t \sim \text{Exp}_\Psi(\sum_j \lambda_j)$, resulting in the likelihood

$$P\left(t, t \in \Psi \,|\, \textstyle\sum_j \lambda_j\right) = f_\Psi\left(t \,|\, \textstyle\sum_j \lambda_j\right) P\left(t \in \Psi \,|\, \textstyle\sum_j \lambda_j\right) = \left(\textstyle\sum_j \lambda_j\right) e^{-t \sum_j \lambda_j}. \tag{3}$$

Together with $y$, which can be drawn by (1) if it is missing, the likelihood $P(y, t, t \in \Psi \,|\, \{\lambda_j\}_{1,J})$ becomes the same as in (2). The procedure of sampling $t$ and/or $y$, generating fully factorized likelihoods under different censoring conditions, plays a crucial role as a data augmentation scheme that will be used for Bayesian inference of the proposed LDR survival model.

In survival analysis with competing risks, one is often interested in modeling the dependence of the event type $y$ and failure time $t$ on covariates $\boldsymbol{x} = (1, x_1, \ldots, x_V)'$. Under the exponential racing framework, one may simply let $\lambda_j = e^{\boldsymbol{x}'\boldsymbol{\beta}_j}$, where $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{jV})'$ is the regression coefficient vector for the $j$th competing risk or event type. However, the hazard rate for the $j$th competing risk, expressed as $\lambda_j = e^{\boldsymbol{x}'\boldsymbol{\beta}_j}$, is restricted to be log-linear in the covariates $\boldsymbol{x}$. This clear restriction motivates us to generalize exponential racing to Lomax racing, which can have a time-varying hazard rate for each competing risk, and further to Lomax delegate racing, which can use the convolution of a potentially countably infinite number of covariate-dependent gamma distributions to model each $\lambda_j$.

## 3 Lomax and Lomax delegate racings

In this section, we generalize exponential racing to Lomax racing, which relates survival analysis with competing risks to a race of conditionally independent Lomax distributed random variables. We further generalize Lomax racing to Lomax delegate racing, which races the winners of conditionally independent Lomax racings. Below we first briefly review Lomax distribution.

Let $\lambda \sim \text{Gamma}(r, 1/b)$ represent a gamma distribution with $\mathbb{E}[\lambda] = r/b$ and $\text{Var}[\lambda] = r/b^2$. Mixing the rate parameter of an exponential distribution with $\lambda \sim \text{Gamma}(r, 1/b)$ leads to a Lomax distribution [31] $t \sim \text{Lomax}(r, b)$, with shape $r > 0$, scale $b > 0$, and PDF

$$f(t \,|\, r, b) = \textstyle\int_0^\infty \text{Exp}(t; \lambda) \text{Gamma}(\lambda; r, 1/b) d\lambda = r b^r (t + b)^{-(r+1)}, \;\; t \in \mathbb{R}_+.$$

When $r > 1$, we have $\mathbb{E}[t] = b/(r-1)$, and when $r > 2$, we have $\text{Var}[t] = b^2 r / [(r-1)^2 (r-2)]$. The Lomax distribution is a heavy-tailed distribution. Its hazard rate and survival function can be expressed as $h(t) = r/(t+b)$ and $S(t) = (t + b^{-1})^{-r}$, respectively.

### 3.1 Covariate-dependent Lomax racing

We generalize covariate-dependent exponential racing by letting

$$t_j \sim \text{Exp}(\lambda_j), \; \lambda_j \sim \text{Gamma}(r, e^{\boldsymbol{x}'\boldsymbol{\beta}_j}).$$

Marginalizing out $\lambda_j$ leads to $t_j \sim \text{Lomax}(r, e^{-\boldsymbol{x}'\boldsymbol{\beta}_j})$. Lomax distribution was initially introduced to study business failures [31] and has since then been widely used to model the time to event in survival analysis [32–35]. Previous research on this distribution [36–38], however, has been mainly focused on point estimation of parameters, without modeling covariate dependence and performing Bayesian inference. We define Lomax racing as follows.

**Definition 1.** *Lomax racing models the time to event $t$ and event type $y$ given covariates $\boldsymbol{x}$ as*

$$t = t_y,\ y = \text{argmin}_{j\in\{1,\dots,J\}}\, t_j,\ t_j \sim \text{Lomax}(r, e^{-\boldsymbol{x}'\boldsymbol{\beta}_j}). \tag{4}$$

To explain the notation, suppose a patient has both diabetes ($j = 1$) and cancer ($j = 2$), then $t_1$ will be the patient's latent survival time under diabetes and $t_2$ under cancer. The patient's observed survival time is $\min(t_1, t_2)$. Note Lomax racing can also be considered as an exponential racing model with multiplicative random effects, since $t_j$ in (4) can also be generated as

$$t_j \sim \text{Exp}(\epsilon_j e^{\boldsymbol{x}'\boldsymbol{\beta}_j}),\ \epsilon_j \sim \text{Gamma}(r, 1).$$

There are two clear benefits of Lomax racing over exponential racing. The first benefit is that given $\boldsymbol{x}$ and $\boldsymbol{\beta}_j$, the hazard rate for the $j$th competing risk, expressed as $r/(t_j + e^{-\boldsymbol{x}'\boldsymbol{\beta}_j})$, is no longer a constant as $e^{\boldsymbol{x}'\boldsymbol{\beta}_j}$. The second benefit is that closed-form Gibbs sampling update equations can be derived, as will be described in detail in Section 4 and the Appendix.

For competing risk $j$, we can also express $t_j \sim \text{Exp}(\epsilon_j e^{\boldsymbol{x}'\boldsymbol{\beta}_j}),\ \epsilon_j \sim \text{Gamma}(r, 1)$ as

$$\ln(t_j) = -\boldsymbol{x}'\boldsymbol{\beta}_j + \varepsilon_j,\ \varepsilon_j = \ln(\varepsilon_{j1}/\varepsilon_{j2}),\ \varepsilon_{j1} \sim \text{Exp}(1),\ \varepsilon_{j2} \sim \text{Gamma}(r, 1).$$

Thus Lomax racing regression uses an accelerated failure time model [18] for each of its competing risks. More specifically, with $S_0(t_j) = (t_j + 1)^{-r}$ and $h_0(t_j) = \frac{r}{t_j+1}$, we have

$$S_j(t_j) = (e^{\boldsymbol{x}'\boldsymbol{\beta}_j} t_j + 1)^{-r} = S_0(e^{\boldsymbol{x}'\boldsymbol{\beta}_j} t_j),\ \ h_j(t_j) = r(t_j + e^{-\boldsymbol{x}'\boldsymbol{\beta}_j})^{-1} = e^{\boldsymbol{x}'\boldsymbol{\beta}_j} h_0(e^{\boldsymbol{x}'\boldsymbol{\beta}_j} t_j), \tag{5}$$

and hence $e^{-\boldsymbol{x}'\boldsymbol{\beta}_j}$ can be considered as the accelerating factor for competing risk $j$. Considering all $J$ risks, we can express survival function $S(t)$ and hazard function $h(t)$ as

$$S(t) = \prod_{j=1}^{J} S_j(t) = \prod_{j=1}^{J} (e^{\boldsymbol{x}'\boldsymbol{\beta}_j} t + 1)^{-r} = \prod_{j=1}^{J} S_0(e^{\boldsymbol{x}'\boldsymbol{\beta}_j} t), \quad h(t) = \frac{-dS(t)/dt}{S(t)} = \sum_{j=1}^{J} \frac{r}{t + e^{-\boldsymbol{x}'\boldsymbol{\beta}_j}}. \tag{6}$$

The nosology of competing risks is often subjected to human knowledge, diagnostic techniques, and patient population. Diseases with the same phenotype, categorized into one competing risk, might have distinct etiology and different impacts on survival, and thus require different therapies. For example, for a patient with both diabetes and cancer, it can be unknown whether the patient has Type 1 or Type 2 diabetes, where Type 1 is ascribed to insufficient production of insulin from pancreas whereas Type 2 arises from the cells' failure in responding properly to insulin [39]. In this regard, it is often necessary for a model to identify sub-risks within a pre-specified competing risk, which may not only improve the fit of survival time, but also help diagnose new disease subtypes. We develop Lomax delegate racing, assuming that a risk consists of several sub-risks, under each of which the latent failure time is accelerated by the exponential of a weighted linear combination of covariates.

### 3.2 Lomax delegate racing

Based on the idea of Lomax racing that an individual's observed failure time is the minimum of latent failure times under competing risks, we further propose *Lomax delegate racing* (LDR), assuming a latent failure time under a competing risk is the minimum of the failure times under a number of sub-risks appertaining to this competing risk. In particular, let us first denote $G_j \sim \Gamma\text{P}(G_{0j}, 1/c_{0j})$ as a gamma process defined on the product space $\mathbb{R}^+ \times \Omega$, where $\mathbb{R}^+ = \{x : x > 0\}$, $G_{0j}$ is a finite and continuous base measure over a complete separable metric space $\Omega$, and $1/c_{0j}$ is a positive scale parameter, such that $G_j(A) \sim \text{Gamma}(G_{0j}(A), 1/c_{0j})$ for each Borel set $A \subset \Omega$. A draw from the gamma process consists of countably infinite non-negatively weighted atoms, expressed as $G_j = \sum_{k=1}^{\infty} r_{jk}\delta_{\boldsymbol{\beta}_{jk}}$. Now we formally define LDR survival model as follows.

**Definition 2** (Lomax delegate racing)**.** *Given a random draw of a gamma process $G_j \sim \Gamma\text{P}(G_{0j}, 1/c_{0j})$, expressed as $G_j = \sum_{k=1}^{\infty} r_{jk}\delta_{\boldsymbol{\beta}_{jk}}$, for each $j \in \{1, \dots, J\}$, Lomax delegate racing models the time to event $t$ and event type $y$ given covariates $\boldsymbol{x}$ as*

$$t = t_y,\ y = \underset{j\in\{1,\dots,J\}}{\text{argmin}}\ t_j,\ t_j = t_{j\kappa_j},\ \kappa_j = \underset{k\in\{1,\dots,\infty\}}{\text{argmin}}\ t_{jk},\ t_{jk} \sim \text{Lomax}(r_{jk}, e^{-\boldsymbol{x}'\boldsymbol{\beta}_{jk}}). \tag{7}$$

In contrast to specifying a fixed number of competing risks $J$, the gamma process not only admits a race among a potentially infinite number of sub-risks, but also parsimoniously shrinks toward zero the weights of negligible sub-risks [40, 41], so that the non-monotonic covariate effects on the failure time under a competing risk can be interpreted as the *minimum*, which is a nonlinear operation, of failure times under sub-risks whose accelerating factor is log-linear in $\boldsymbol{x}$. As shown in the following Corollary, LDR can also be considered as a generalization of exponential racing, where the exponential rate parameter of each competing risk $j$ is a weighted summation of a countably infinite number of gamma random variables with covariate-dependent weights.

**Corollary 1.** *Lomax delegate racing survival model can also be expressed as*

$$t = t_y,\ y = \operatorname{argmin}_{j\in\{1,\dots,J\}} t_j,\ t_j \sim \text{Exp}\left(\sum\nolimits_{k=1}^{\infty} e^{\boldsymbol{x}'\boldsymbol{\beta}_{jk}}\tilde{\lambda}_{jk}\right),\ \tilde{\lambda}_{jk} \sim \text{Gamma}(r_{jk}, 1). \tag{8}$$

We provide in the Appendix the marginal distribution of $t$ in LDR for situations where predicting the failure time is of interest. The survival and hazard functions of LDR, which generalize those of Lomax racing in (6), can be expressed as

$$S(t) = \prod_{j=1}^{J}\prod_{k=1}^{\infty} P(T_{jk} > t_j) = \prod_{j=1}^{J}\prod_{k=1}^{\infty}(e^{\boldsymbol{x}'\boldsymbol{\beta}_{jk}}t_j + 1)^{-r_{jk}}, \quad h(t) = \sum_{j=1}^{J}\sum_{k=1}^{\infty}\frac{r_{jk}}{t_j + e^{-\boldsymbol{x}'\boldsymbol{\beta}_{jk}}}. \tag{9}$$

LDR survival model can be considered as a two-phase racing, where in the first phase, for each of the $J$ pre-specified competing risk there is a race among countably infinite sub-risks, and in the second phase, $J$ risk-specific failure times race with each other to eventually determine both the observed failure time $t$ and event type $y$. Moreover, Corollary 1, representing LDR as a single-phase exponential racing, more explicitly explains non-monotonic covariate effects on $t_j$ by writing the exponential rate parameter of $t_j$ as the aggregation of $\{e^{\boldsymbol{x}'\boldsymbol{\beta}_{jk}}\}_{k=1}^{\infty}$ weighted by gamma random variables with the shape parameters as the atom weights of the gamma process $G_j$.

# 4 Bayesian inference

LDR utilizes a gamma process [42] to support countably infinite regression-coefficient vectors for each pre-specified risk. The gamma process $G_j \sim \Gamma\text{P}(G_{0j}, 1/c_{0j})$ has an inherent shrinkage mechanism in that the number of atoms whose weights are larger than a positive constant $\epsilon$ is finite almost surely and follows a Poisson distribution with mean $\int_{\epsilon}^{\infty} r^{-1}e^{-c_{0j}r}dr$. For the convenience of implementation, as in Zhou et al. [43], we truncate the total number of atoms of a gamma process to be $K$ by choosing a finite and discrete base measure as $G_{0j} = \sum_{k=1}^{K}\frac{\gamma_{0j}}{K}\delta_{\boldsymbol{\beta}_{jk}}$. Let us denote $\boldsymbol{x}_i$ and $y_i$ as the covariates and the event type, respectively, for individual $i \in \{1, \dots, n\}$. We express the full hierarchical model of the (truncated) gamma process LDR survival model as

$$y_i = \underset{j\in\{1,\dots,J\}}{\operatorname{argmin}}\ t_{ij},\ \ t_i = \min_j t_{ij},\ \ t_{ij} = t_{ij\kappa_{ij}},\ \ \kappa_{ij} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmin}}\ t_{ijk},\ \ t_{ijk} \sim \text{Exp}(\lambda_{ijk}),$$
$$\lambda_{ijk} \sim \text{Gamma}(r_{jk}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}}),\ \ r_{jk} \sim \text{Gamma}(\gamma_{0j}/K, 1/c_{0j}),\ \ \boldsymbol{\beta}_{jk} \sim \prod\nolimits_{v=0}^{V}\mathcal{N}(0, \alpha_{vjk}^{-1}), \tag{10}$$

where we further let $\alpha_{vjk} \sim \text{Gamma}(a_0, 1/b_0)$. The joint probability given $\{\lambda_{ijk}\}_{jk}$ is

$$P(t_i, \kappa_{iy_i}, y_i \,|\, \{\lambda_{ijk}\}_{jk}) = P(t_i \,|\, \{\lambda_{ijk}\}_{jk})P(\kappa_{iy_i}, y_i \,|\, \{\lambda_{ijk}\}_{jk}) = \lambda_{iy_i\kappa_{iy_i}}e^{-t_i\sum_{j=1}^{S}\sum_{k=1}^{K}\lambda_{ijk}},$$

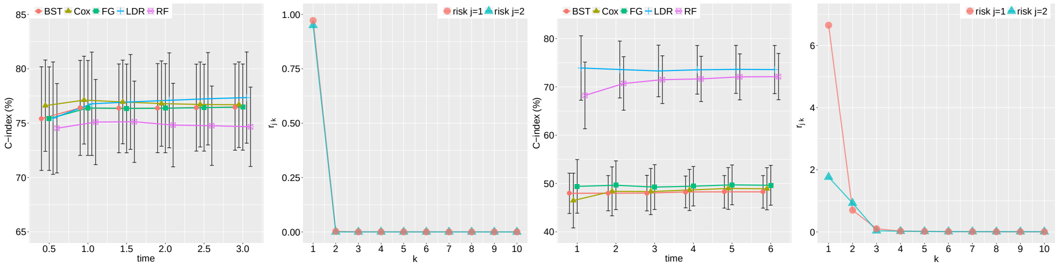
which is amenable to posterior simulation for $\lambda_{ijk}$. Let us denote $\text{NB}(x; r, p) = \frac{\Gamma(x+r)}{x!\Gamma(r)}p^x(1-p)^r$ as the likelihood for negative binomial distribution and $\sigma(x) = 1/(1 + e^{-x})$ as the sigmoid function. Further marginalize out $\lambda_{ijk} \sim \text{Gamma}(r_{jk}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}})$ leads to a fully factorized joint likelihood as

$$P(t_i, \kappa_{iy_i}, y_i \,|\, \boldsymbol{x}_i, \{\boldsymbol{\beta}_{jk}\}_{jk}) = t_i^{-1}\prod\nolimits_j\prod\nolimits_k \text{NB}\left(\mathbf{1}(\kappa_{iy_i} = k, y_i = j); r_{jk}, \sigma(\boldsymbol{x}_i'\boldsymbol{\beta}_{jk} + \ln t_i)\right), \tag{11}$$

which is amenable to posterior simulation using the data augmentation based inference technique for negative binomial regression [44, 45]. The augmentation schemes of $t_i$ and/or $y_i$ discussed in Section 2 are used to achieve (11) in the presence of censoring or as a remedy for missing data. We describe in detail both Gibbs sampling and maximum a posteriori (MAP) inference in the Appendix.

Table 1: Synthetic data generating process.

| Synthetic data 1 | Synthetic data 2 |
|---|---|
| $t_i = \min(t_{i1}, t_{i2}, 3.5)$,<br>$t_{i1} \sim \text{Exp}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_1})$, $t_{i2} \sim \text{Exp}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_2})$ | $t_i = \min(t_{i1}, t_{i2}, 6.5)$,<br>$t_{i1} \sim \text{Exp}(1/\cosh(\boldsymbol{x}_i'\boldsymbol{\beta}_1))$, $t_{i2} \sim \text{Exp}(1/\lvert\sinh(\boldsymbol{x}_i'\boldsymbol{\beta}_2)\rvert)$ |



(a) C-index of risk 1 for synthetic data 1.

(b) $r_{jk}$ by descending order for synthetic data 1.

(c) C-index of risk 1 for synthetic data 2.

(d) $r_{jk}$ by descending order for synthetic data 2.

Figure 1: Cause-specific C-indices and shrinkage of $r_{jk}$ by LDR for synthetic data 1 and 2.

## 5 Experimental results

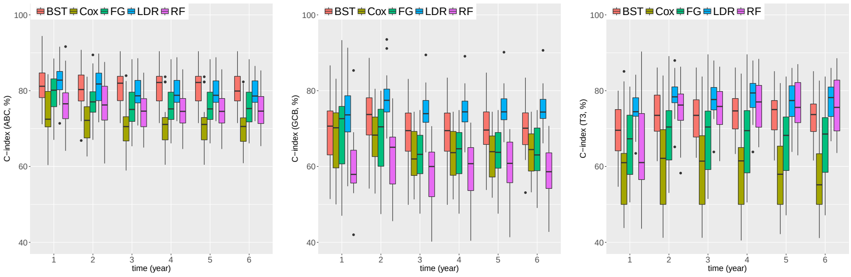
In this section, we validate the proposed LDR model by a variety of experiments using both synthetic and real data. Some data description, implementation of benchmark approaches, and experiment settings are deferred to the Appendix for brevity. In all experiments we exclude from the testing data the observations that have unknown failure times or event types. We compare the proposed LDR survival model, cause-specific Cox proportional hazards model (Cox) [19,23], Fine-Gray proportional subdistribution hazards model (FG) [22] and its boosting algorithm (BST) which is more stable for high-dimensional covariates [46], and random survival forests (RF) [24], which are all designed for survival analysis with competing risks. We show that LDR performs uniformly well regardless of whether the covariate effects are monotonic or not. Moreover, LDR is able to infer the missing cause of death and/or survival time of an observation, both of which in general cannot be handled by these benchmark methods. The model fits of LDR by Bayesian inference via Gibbs sampling and MAP inference via stochastic gradient descent (SGD) are comparable. We will report the results of Gibbs sampling, as it provides an explicit criterion to prune unneeded model capacity (Steps 1 and 8 of Appendix B), avoiding the need of model selection and parameter tuning. For large scale data, performing MAP inference via SGD is recommended if Gibbs sampling takes too long to run a sufficiently large number of iterations. We quantify model performance by cause-specific concordance index (C-index) [23], where the C-index of risk $j$ at time $\tau$ in this paper is computed as

$$\mathcal{C}_j(\tau) = P\left(Score_j(\boldsymbol{x}_i, \tau) > Score_j(\boldsymbol{x}_{i'}, \tau) \,|\, y_i = j \text{ and } [t_i < t_{i'} \text{ or } y_{i'} \neq j]\right),$$

where $i \neq i'$ and $Score_j(\boldsymbol{x}_i, \tau)$ is a prognostic score at time $\tau$ depending on $\boldsymbol{x}_i$ such that its higher value reflects a higher risk of cause $j$. Intuitively, for cause $j$, if patient $i$ died of this cause ($i.e.$, $y_i = j$), and patient $i'$ either died of another cause ($i.e.$, $y_{i'} \neq j$) or died of this cause but lived longer than patient $i$ ($i.e.$, $t_i < t_{i'}$), then it is likely that $Score_j(\boldsymbol{x}_i, \tau)$ for patient $i$ is higher than $Score_j(\boldsymbol{x}_{i'}, \tau)$ for patient $i'$, and the ranking of risks for this pair of patients is concordant. C-index measures such concordance, and a higher value indicates better model performance. Wolbers et al. [23] write C-index as a weighted average of time-dependent AUC that is related to sensitivity, specificity, and ROC curves for competing risks [47]. So a C-index around $0.5$ implies a model failure. A good choice of the prognostic score is the cumulative incidence function, $i.e$, $Score_j(\boldsymbol{x}_i, \tau) = \text{CIF}_j(i, \tau) = P(t_i \leq \tau, y_i = j)$ [18, 22, 28]. Distinct from a survival function that measures the probability of surviving beyond some time, CIF estimates the probability that an event occurs by a specific time in the presence of competing risks. For LDR given $\{r_{jk}\}$ and $\{\boldsymbol{\beta}_{jk}\}$,

$$\text{CIF}_j(i, \tau) = P(t_i \leq \tau, y_i = j) = \mathbb{E}\left[\frac{\sum_k \lambda_{ijk}}{\sum_{j',k} \lambda_{ij'k}}\left(1 - e^{-\tau \sum_{j',k} \lambda_{ij'k}}\right)\right],$$

where the expectation is taken over $\{\lambda_{jk}\}_{j,k}$, where $\lambda_{ijk} \sim \text{Gamma}(r_{jk}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}})$. The expectation can be evaluated by Monte-Carlo estimation if we have a point estimate or a collection of post-burn-in MCMC samples of $r_{jk}$ and $\boldsymbol{\beta}_{jk}$.



(a) C-index of ABC. (b) C-index of GCB. (c) C-index of T3.

Figure 2: Cause-specific C-indices for DLBCL data.

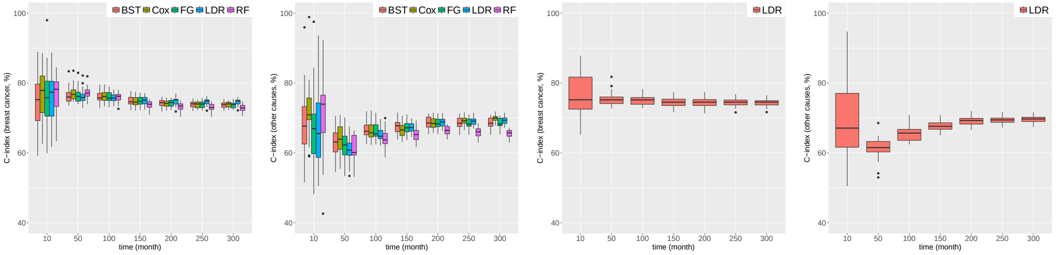
### 5.1 Synthetic data analysis

We first simulate two datasets following Table 1, where $\boldsymbol{x}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_3)$, to illustrate the unique nonlinear modeling capability of LDR. In Table 1 $t_{ij}$ denotes the latent survival time under risk $j$, $j = 1, 2$ and $t_i$ is the observed time to event. The event type $y_i = \arg\min_j t_{ij}$ if $t_i < T_{r.c.}$, with $y_i = 0$ indicating right censoring if $t_i = T_{r.c.}$, where the censoring time $T_{r.c.} = 3.5$ for data 1 and 6.5 for data 2. We simulate 1,000 random observations, and use 800 for training and the remaining 200 for testing. We randomly take 20 training/testing partitions, on each of which we evaluate the testing cause-specific C-index at time $0.5, 1, 1.5, \cdots, 3$ for data 1 and at time $1, 2, \cdots, 6$ for data 2. The sample mean ± standard deviation of the estimated cause-specific C-indices of risks 1, and the estimated $r_{jk}$'s of both risks by LDR (from one random training/testing partition but without loss of generality) for data 1 are displayed in panels (a) and (b) of Figure 1, respectively. Analogous plots for data 2 are shown in panels (c) and (d). The testing C-indices of risk 2 are analogous to those of risk 1 for both datasets, thus shown in Figure 5 in the Appendix for brevity.

For data 1 where the survival times under both risks depend on the covariates monotonically, LDR has comparable performance with Cox, FG, and BST, and all these four models slightly outperform RF in terms of the mean values of C-indices. The underperformance of RF in the case of monotonic covariate effects has also been observed in its original paper [24]. For data 2 where the survival time and covariates are not monotonically related, LDR and RF at any time evaluated significantly outperform the other three approaches, all of which fail on this dataset as their C-indices are around $0.5$ for both risks. Panels (b) and (d) of Figure 1 show $r_{jk}$ inferred on data 1 and 2, respectively. More specifically, both risks consist of only one sub-risk for data 1. By contrast, two sub-risks of the two respective risks can approximate the complex data generating process of data 2.

### 5.2 Real data analysis

We analyze a microarray gene-expression profile [48] to assess our model performance on real data. The dataset contains a total of 240 patients with diffuse large B-cell lymphoma (DLBCL). Multiple unsuccessful treatments to increase the survival rate suggest that there exist several subtypes of DLBCL that differ in responsiveness to chemotherapy. In the DLBCL dataset, Rosenwald et al. [48] identify three gene-expression subgroups, including activated B-cell-like (ABC), germinal-center B-cell-like (GCB), and type 3 (T3) DLBCL, which may be related to three different diseases as a result of distinct mechanisms of malignant transformation. They also suspect that T3 may be associated with more than one such mechanism. In our analysis, we treat the three subgroups and their potential malignant transformation mechanisms as competing risks from which the patients suffer. As the total number of patients is small which is often the case in survival data, we consider 434 genes that have no missing values across all the patients. Seven of the 434 genes have been reported to be related to clinical phenotypes and four of the seven to have non-monotonic effects on the risk of death [7]. Since some gene expressions may be highly correlated, we follow the same selection procedure of Li and Luan [7] to include as covariates the seven genes, together with another 33 genes having the highest Cox partial score statistic, so that both Cox proportional model and FG subdistribution model for competing risks do not collapse for computational singularity. We use 200 observations for training and the remaining 40 for testing. We take 20 random training/testing partitions and report in Figure 2 boxplots of the testing C-indices evaluated at year $1, 2, \cdots, 6$, by the same five approaches used in the analysis of synthetic datasets.



(a) BC. (b) OC. (c) BC, with unknown event types. (d) OC, with unknown event types.

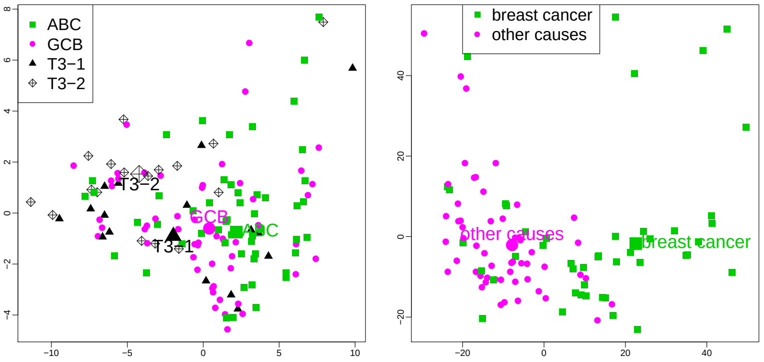
Figure 3: C-indices for SEER breast cancer data.

The boxplots of BST and LDR are roughly comparable for ABC, but the median of LDR is slightly higher than those of BST until year 2, and hereafter slightly lower. For GCB and T3, LDR results in higher median C-indices than all the other benchmarks do at any time evaluated, indicating LDR provides a big improvement in predicting lymphoma CIFs. Interestingly noted is that RF has low performance in both ABC and GCB, but outperforms Cox, FG, and BST and is comparable to LDR in T3. This implies that the gene expressions may have monotonic effects on survival under ABC or GCB, but it is not the case for T3, which can be validated by the fact that LDR learns one sub-risk for ABC and GCB, respectively, and two sub-risks for T3. To better show the improvements of LDR over existing approaches, we calculate the difference of C-indices between LDR and each of the other four benchmarks within each training-testing partition, and report the sample mean and standard deviation across partitions in Table 11 in the Appendix. On average, the improvements of LDR over Cox, FG, and BST are bigger for T3 than those for ABC or GCB, whereas LDR outperforms RF by a larger margin for ABC and GCB than for T3. This shows another advantage of LDR that it fits consistently well regardless of whether the covariate effects are monotonic or not.

We further analyze a publicly accessible dataset from the *Surveillance, Epidemiology, and End Results* (SEER) Program of National Cancer Institute [49]. The SEER dataset we use contains two risks: one is breast cancer and the other is “other causes,” which we denote as BC and OC, respectively. It also contains some incomplete observations, each of which with an unknown cause of death but observed uncensored time to death, that can be handled by LDR. The individual covariates include the patients’ personal information, such as age, gender, race, and diagnostic and therapy information. More details are deferred to the Appendix.

We first eliminate all observations with unknown causes of death, so we can make comparison between LDR, Cox, FG, BST, and RF. We take 20 random training/testing partitions of the dataset, in each of which $80\%$ of observations are used as training and the remaining $20\%$ as testing. In Figure 3, panels (a) and (b) show the boxplots of C-indices for BC and OC, respectively, obtained from the 20 testing sets by the five models at months $10, 50, 100, \cdots, 300$. For BC the C-indices by LDR are comparable to those by the other four approaches until month 150 and slightly higher afterwards. For the OC the C-indices by LDR are slightly lower than those by Cox, FG, and BST, but become similar after month 100. Also note that RF underperforms the other four approaches since month 100 for BC and month 50 for OC. Comparable C-indices from LDR, Cox, FG, and BST imply monotonic impacts of covariates on survival times under both risks. In fact, for either risk we learn a sub-risk which dominates the others in terms of weights. Furthermore, we analyze the SEER dataset by LDR using the same training/testing partitions, but additionally including the observations having missing causes of death into the 20 training sets, and show the testing C-indices in panels (c) and (d) of Figure 3. We see the testing C-indices are very similar to those in (a) and (b). More importantly, LDR provides a probabilistic inference on missing time to event or missing causes during the model training procedure.

In Appendix E we further provide the Brier scores [50, 51] of each risk in all data sets over time. Brier score quantifies the deviation of predicted CIF’s from the actual outcomes and a smaller value implies a better model performance [52]. Tables 2-10 in Appendix E show Brier scores by the models compared on the four data sets, indicating the model out-of-sample prediction performance is basically consistent with those quantified by C-indices. Specifically, for the cases of synthetic data 1, SEER, and both ABC and GCB of DLBCL, where C-indices imply linear covariate effects, the Brier scores are comparable for Cox, FG, BST, and LDR, and slightly smaller than those of RF. For synthetic data 2 and T3 of DLBCL where C-indices imply nonlinear covariate effects, the Brier



(a) DLBCL.

(b) SEER.

Figure 4: Isomap visualization of the observations and inferred sub-risk representatives.

scores by LDR and RF are smaller than those by Cox, FG, and BST. Moreover, the Brier scores by LDR are slightly larger than those of RF for synthetic data 2 but smaller for T3 of DLBCL.

To show the interpretability of LDR, we visualize representative individuals, each of which suffers from an inferred sub-risk. Specifically, for each inferred sub-risk $k$ under risk $j$, we find the representative by evaluating a weighted average of all uncensored observations as $\sum_i w_{ijk}\boldsymbol{x}_i / \sum_i w_{ijk}$, where $w_{ijk} = \mathbb{E}\left(\frac{\lambda_{ijk}}{\sum_{j'}\sum_{k'}\lambda_{ij'k'}}\right)$, $\lambda_{ijk} \sim \text{Gamma}(\hat{r}_{jk}, e^{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{jk}})$, and $\hat{r}_{jk}$ and $\hat{\boldsymbol{\beta}}_{jk}$ are the estimated values of $r_{jk}$ and $\boldsymbol{\beta}_{jk}$, respectively. The weight $w_{ijk}$ extracts the component of $\boldsymbol{x}_i$ that is likely to make the event of sub-risk $k$ under risk $j$ first occur. Then we implement an Isomap algorithm [53] and visualize in Figure 4 the representatives along with uncensored observations in both DLBCL and SEER. Details are provided in the Appendix.

In Figure 4, small symbols denote uncensored observations and large ones the representatives. Panels (a) and (b) show the representatives suffering from sub-risks in the DLBCL and SEER dataset, respectively. In panel (a), we use green for ABC, pink for GCB, and black for T3. The only representative suffering from ABC (GCB) is surrounded by small green (pink) symbols, indicating they signify a typical gene expression profile that may result in the corresponding malignant transformation. There are two representatives suffering from the two sub-risks of T3, denoted by a large triangle and a large diamond, respectively. They approximately lie in the center of the respective cluster of small triangles and diamonds, which denote patients suffering from the corresponding sub-risks of T3 with an estimated probability greater than $0.5$. The two sub-risks of T3 and the representatives verify the heterogeneity of gene expressions under this risk, and strengthen the belief that T3 consists of more than one type of DLBCL [48]. For the SEER data, we randomly select 100 of the 2088 uncensored observations with known event types for visualization. In panel (b), we use green for BC and pink for OC. LDR learns only one sub-risk for each of these two risks, and place for each risk a representative approximately at the center of the cluster of patients who died of that risk.

## 6 Conclusion

We propose Lomax delegate racing (LDR) for survival analysis with competing risks. LDR models the survival times under risks as a two-phase race of sub-risks, which not only intuitively explains the mechanism of surviving under competing risks, but also helps model non-monotonic covariate effects. We use the gamma process to support a potentially countably infinite number of sub-risks for each risk, and rely on its inherent shrinkage mechanism to remove unneeded model capacity, making LDR be capable of detecting unknown event subtypes without pre-specifying their numbers. LDR admits a hierarchical representation that facilities the derivation of Gibbs sampling under data augmentation, which can be adapted for various practical situations such as missing event times or types. A more scalable (stochastic) gradient descent based maximum a posteriori inference algorithm is also developed for big data applications. Experimental results show that with strong interpretability and outstanding performance, the proposed LDR survival model is an attractive alternative to existing ones for various tasks in survival analysis with competing risks.

### Acknowledgments

The authors acknowledge the support of Award IIS-1812699 from the U.S. National Science Foundation, and the computational support of Texas Advanced Computing Center.

## References

[1] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[2] Y. Wang, B. Xie, N. Du, and L. Song, "Isotonic Hawkes processes," in *International conference on machine learning*, pp. 2226–2234, 2016.

[3] H. Mei and J. M. Eisner, "The neural Hawkes process: A neurally self-modulating multivariate point process," in *NIPS*, pp. 6754–6764, 2017.

[4] D. R. Cox, "Regression models and life-tables," in *Breakthroughs in statistics*, pp. 527–541, Springer, 1992.

[5] F. Ballester, D. Corella, S. Pérez-Hoyos, M. Sáez, and A. Hervás, "Mortality as a function of temperature. A study in Valencia, Spain, 1991-1993.," *International journal of epidemiology*, vol. 26, no. 3, pp. 551–561, 1997.

[6] K. M. Flegal, B. I. Graubard, D. F. Williamson, and M. H. Gail, "Cause-specific excess deaths associated with underweight, overweight, and obesity," *Jama*, vol. 298, no. 17, pp. 2028–2037, 2007.

[7] H. Li and Y. Luan, "Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2403–2409, 2005.

[8] W. Lu and L. Li, "Boosting method for nonlinear transformation models with censored survival data," *Biostatistics*, vol. 9, no. 4, pp. 658–667, 2008.

[9] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei, "Deep survival analysis," in *Machine Learning for Healthcare Conference*, pp. 101–114, 2016.

[10] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, p. 24, 2018.

[11] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pp. 544–547, IEEE, 2016.

[12] P. Chapfuwa, C. Tao, C. Li, C. Page, B. Goldstein, L. Carin, and R. Henao, "Adversarial time-to-event modeling," in *ICML*, 2018.

[13] T. Fernández, N. Rivera, and Y. W. Teh, "Gaussian processes for survival analysis," in *NIPS*, pp. 5021–5029, 2016.

[14] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, and Y. Bengio, "Deep learning for patient-specific kidney graft survival analysis," *arXiv preprint arXiv:1705.10245*, 2017.

[15] Y. Li, J. Wang, J. Ye, and C. K. Reddy, "A multi-task learning formulation for survival analysis," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1715–1724, ACM, 2016.

[16] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," in *NIPS*, pp. 1845–1853, 2011.

[17] X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath, "Deep survival analysis: Nonparametrics and missingness," in *Machine Learning for Healthcare Conference*, 2018.

[18] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, vol. 360. John Wiley & Sons, 2011.

[19] P. C. Austin, D. S. Lee, and J. P. Fine, "Introduction to the analysis of survival data in the presence of competing risks," *Circulation*, vol. 133, no. 6, pp. 601–609, 2016.

[20] H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in biostatistics: Competing risks and multi-state models," *Statistics in medicine*, vol. 26, no. 11, pp. 2389–2430, 2007.

[21] B. Lau, S. R. Cole, and S. J. Gange, "Competing risk regression models for epidemiologic data," *American journal of epidemiology*, vol. 170, no. 2, pp. 244–256, 2009.

[22] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American statistical association*, vol. 94, no. 446, pp. 496–509, 1999.

[23] M. Wolbers, P. Blanche, M. T. Koller, J. C. Witteman, and T. A. Gerds, "Concordance for prognostic models with competing risks," *Biostatistics*, vol. 15, no. 3, pp. 526–539, 2014.

[24] H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau, "Random survival forests for competing risks," *Biostatistics*, vol. 15, no. 4, pp. 757–773, 2014.

[25] J. E. Barrett and A. C. Coolen, "Gaussian process regression for survival data with competing risks," *arXiv preprint arXiv:1312.1591*, 2013.

[26] A. M. Alaa and M. v. d. Schaar, "Deep multi-task gaussian processes for survival analysis with competing risks," in *NIPS*, 2017.

[27] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks," AAAI, 2018.

[28] M. J. Crowder, *Classical competing risks*. CRC Press, 2001.

[29] S. M. Ross, *Introduction to Probability Models*. Academic Press, 10th ed., 2006.

[30] F. Caron and Y. W. Teh, "Bayesian nonparametric models for ranked data," in *NIPS*, pp. 1520–1528, 2012.

[31] K. Lomax, "Business failures: Another example of the analysis of failure data," *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 847–852, 1954.

[32] J. Myhre and S. Saunders, "Screen testing and conditional probability of survival," *Lecture Notes-Monograph Series*, pp. 166–178, 1982.

[33] H. A. Howlader and A. M. Hossain, "Bayesian survival estimation of Pareto distribution of the second kind based on failure-censored data," *Computational statistics & data analysis*, vol. 38, no. 3, pp. 301–314, 2002.

[34] E. Cramer and A. B. Schmiedt, "Progressively Type-II censored competing risks data from Lomax distributions," *Computational Statistics & Data Analysis*, vol. 55, no. 3, pp. 1285–1303, 2011.

[35] F. Hemmati and E. Khorram, "On adaptive progressively Type-II censored competing risks data," *Communications in Statistics-Simulation and Computation*, pp. 1–23, 2017.

[36] S. Al-Awadhi and M. Ghitany, "Statistical properties of Poisson-Lomax distribution and its application to repeated accidents data," *Journal of Applied Statistical Science*, vol. 10, no. 4, pp. 365–372, 2001.

[37] A. Childs, N. Balakrishnan, and M. Moshref, "Order statistics from non-identical right-truncated Lomax random variables with applications," *Statistical Papers*, vol. 42, no. 2, pp. 187–206, 2001.

[38] D. E. Giles, H. Feng, and R. T. Godwin, "On the bias of the maximum likelihood estimator for the two-parameter Lomax distribution," *Communications in Statistics-Theory and Methods*, vol. 42, no. 11, pp. 1934–1950, 2013.

[39] R. Varma, N. M. Bressler, Q. V. Doan, M. Gleeson, M. Danese, J. K. Bower, E. Selvin, C. Dolan, J. Fine, S. Colman, *et al.*, "Prevalence of and risk factors for diabetic macular edema in the United States," *JAMA ophthalmology*, vol. 132, no. 11, pp. 1334–1340, 2014.

[40] M. Zhou, Y. Cong, and B. Chen, "Augmentable gamma belief networks," *Journal of Machine Learning Research*, vol. 17, no. 163, pp. 1–44, 2016.

[41] M. Zhou, "Softplus regressions and convex polytopes," *arXiv:1608.06383*, 2016.

[42] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, 1973.

[43] M. Zhou and L. Carin, "Negative binomial process count and mixture modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 307–320, 2015.

[44] M. Zhou, L. Li, D. Dunson, and L. Carin, "Lognormal and gamma mixed negative binomial regression," in *ICML*, pp. 1343–1350, 2012.

[45] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya–Gamma latent variables," *J. Amer. Statist. Assoc.*, vol. 108, no. 504, pp. 1339–1349, 2013.

[46] H. Binder, A. Allignol, M. Schumacher, and J. Beyersmann, "Boosting for high-dimensional time-to-event data with competing risks," *Bioinformatics*, vol. 25, no. 7, pp. 890–896, 2009.

[47] P. Saha and P. Heagerty, "Time-dependent predictive accuracy in the presence of competing risks," *Biometrics*, vol. 66, no. 4, pp. 999–1011, 2010.

[48] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *New England Journal of Medicine*, vol. 346, no. 25, pp. 1937–1947, 2002.

[49] S. R. P. S. S. B. National Cancer Institute, DCCPS, *Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1973-2014)*, Released April 2017, based on the November 2016 submission. Released April 2017, based on the November 2016 submission.

[50] T. A. Gerds, T. Cai, and M. Schumacher, "The performance of risk prediction models," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 50, no. 4, pp. 457–479, 2008.

[51] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: A framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, p. 128, 2010.

[52] H. Van Houwelingen and H. Putter, *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

[53] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[54] P. G. Moschopoulos, "The distribution of the sum of independent gamma random variables," *Annals of the Institute of Statistical Mathematics*, vol. 37, no. 1, pp. 541–544, 1985.

[55] T. A. Gerds, *pec: Prediction Error Curves for Risk Prediction Models in Survival Analysis*, 2017. R package version 2.5.4.

[56] T. A. Gerds and T. H. Scheike, *riskRegression: Risk Regression Models for Survival Analysis with Competing Risks*, 2015. R package version 1.1.7.

[57] B. Gray, *cmprsk: Subdistribution Analysis of Competing Risks*, 2014. R package version 2.2-7.

[58] H. Binder, *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*, 2013. R package version 1.4.

[59] H. Ishwaran and U. Kogalur, *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2018. R package version 2.6.0.

[60] T. H. Cormen, *Introduction to algorithms*. MIT press, 2009.

[61] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

# Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks: Appendix

Quan Zhang and Mingyuan Zhou

## A Marginal distribution of failure time in LDR

**Theorem 1.** *If $t_i \sim$ Gamma$(1, 1/\lambda_{i\bullet\bullet})$ with $\lambda_{i\bullet\bullet} = \sum_{j,k} \lambda_{ijk}$ and $\lambda_{ijk} \sim$ Gamma$(r_{jk}, 1/b_{ijk})$, the PDF of $t_i$ given $\{r_{jk}\}$ and $\{b_{ijk}\}$ is*

$$f(t_i \,|\, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) = c_i \sum_{m=0}^{\infty} \frac{(\rho_i + m)\delta_{im} b_{i(1)}^{\rho_i+m}}{(t_i + b_{i(1)})^{1+\rho_i+m}},$$

*and the cumulative density function (CDF) is*

$$P(t_i < q \,|\, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) = 1 - c_i \sum_{m=0}^{\infty} \frac{\delta_{im} b_{i(1)}^{\rho_i+m}}{(q + b_{i(1)})^{\rho_i+m}}, \tag{12}$$

*where $c_i = \prod_{j,k} \left(\frac{b_{ijk}}{b_{i(1)}}\right)^{r_{jk}}$, $b_{i(1)} = \max_{j,k} b_{ijk}$, $\rho_i = \sum_{j,k} r_{jk}$, $\delta_{i0} = 1$, $\delta_{im+1} = \frac{1}{m+1} \sum_{h=1}^{m+1} h\gamma_{ih}\delta_{im+1-h}$ for $m \geq 1$, and $\gamma_{ih} = \sum_{j,k} \frac{r_{jk}}{h} \left(1 - \frac{b_{ijk}}{b_{i(1)}}\right)^h$.*

It is difficult to utilize the PDF or CDF of $t_i$ in the form of series, but we can use a finite truncation to approximate (12). Concretely, as $P(t_i < \infty \,|\, n_i = 1, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) = c_i \sum_{m=0}^{\infty} \delta_{im} = 1$, we find an $M$ so large that $c_i \sum_{m=0}^{M} \delta_{im}$ close to 1 (say no less than 0.9999), and use $1 - c_i \sum_{m=0}^{M} \frac{\delta_{im} b_{i(1)}^{\rho_i+m}}{(q+b_{i(1)})^{\rho_i+m}}$ as an approximation. Consequently, sampling $t_i$ is feasible by inverting the approximated CDF for general cases. We have tried prediction by finite truncation on some synthetic data and found $M$ is mostly between 10 and 30 which is computationally acceptable.

*Proof.* We first study the distribution of gamma convolution. Specifically, if $\lambda_t \overset{ind}{\sim}$ Gamma$(r_t, 1/b_t)$ with $r_t, b_t \in \mathbb{R}_+$, then the PDF of $\lambda = \sum_{t=1}^{T}$ can be written in a form of series [54] as

$$f(\lambda \,|\, r_1, b_1, \cdots, r_T, b_T) = \begin{cases} c \sum_{m=0}^{\infty} \frac{\delta_m \lambda^{\rho+m-1} e^{-\lambda b_{(1)}}}{\Gamma(\rho+m)/b_{(1)}^{\rho+m}} & \text{if } \lambda > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $c = \prod_{t=1}^{T} \left(\frac{b_t}{b_{(1)}}\right)^{r_t}$, $b_{(1)} = \max_t b_t$, $\rho = \sum_{t=1}^{T} r_t$, $\delta_0 = 1$, $\delta_{m+1} = \frac{1}{m+1} \sum_{h=1}^{m+1} h\gamma_h \delta_{m+1-h}$ and $\gamma_h = \sum_{t=1}^{T} r_t \left(1 - \frac{b_t}{b_{(1)}}\right)^h /h$. [54] proved that $0 < \gamma_{ih} \leq \rho_i b_{i0}^h/h$ and $0 < \delta_{im} \leq \frac{\Gamma(\rho_i+m) b_{i0}^m}{\Gamma(\rho_i)m!}$ where $b_{i0} = max_{j,k}(1 - \frac{b_{ijk}}{b_{i(1)}})$. With $n_i \equiv 1$, we want to show the PDF of $t_i$,

$$\begin{aligned} & f(t_i \,|\, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) \\ =& \int_0^{\infty} f(t_i \,|\, \lambda_{i\bullet\bullet}) f(\lambda_{i\bullet\bullet} \,|\, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) d\lambda_{i\bullet\bullet} \\ =& \int_0^{\infty} \sum_{m=0}^{\infty} \frac{c_i \delta_{im} t_i^{n_i-1} \lambda_{i\bullet\bullet}^{n_i+\rho_i+m-1} \exp(-t_i\lambda_{i\bullet\bullet} - b_{i(1)}\lambda_{i\bullet\bullet})}{\Gamma(n_i)\Gamma(\rho_i+m)} d\lambda_{i\bullet\bullet} \\ =& \sum_{m=0}^{\infty} \int_0^{\infty} \frac{c_i \delta_{im} t_i^{n_i-1} \lambda_{i\bullet\bullet}^{n_i+\rho_i+m-1} \exp(-t_i\lambda_{i\bullet\bullet} - b_{i(1)}\lambda_{i\bullet\bullet})}{\Gamma(n_i)\Gamma(\rho_i+m)} d\lambda_{i\bullet\bullet} \qquad (13) \\ =& \frac{c_i t_i^{n_i-1}}{\Gamma(n_i)} \sum_{m=0}^{\infty} \frac{\Gamma(n_i+\rho_i+m)\delta_{im} b_{i(1)}^{\rho_i+m}}{\Gamma(\rho_i+m)(t_i+b_{i(1)})^{n_i+\rho_i+m}}, \end{aligned}$$

which suffices to prove the equality in (13). Note that

$$\begin{aligned}
&f(t_i\,|\,n_i,\lambda_{i\bullet\bullet})f(\lambda_{i\bullet\bullet}\,|\,\{r_{jk}\}_{j,k},\{b_{ijk}\}_{j,k})\\
=&\frac{c_i}{\Gamma(n_i)}t_i^{n_i-1}\lambda_{i\bullet\bullet}^{n_i+\rho_i-1}b_{i(1)}^{\rho_i}\exp(-t_i\lambda_{i\bullet\bullet}-b_{i(1)}\lambda_{i\bullet\bullet})\sum_{m=0}^{\infty}\frac{\Gamma(\rho_i+m)}{\delta_{im}b_{i(1)}^{m}\lambda_{i\bullet\bullet}^{m}}\\
\leq&\frac{c_i}{\Gamma(n_i)}t_i^{n_i-1}\lambda_{i\bullet\bullet}^{n_i+\rho_i-1}b_{i(1)}^{\rho_i}\exp(-t_i\lambda_{i\bullet\bullet}-b_{i(1)}\lambda_{i\bullet\bullet})\sum_{m=0}^{\infty}\frac{(b_{i0}b_{i(1)}\lambda_{i\bullet\bullet})^m}{\Gamma(\rho_i)m!}\\
=&\frac{c_i}{\Gamma(n_i)}t_i^{n_i-1}\lambda_{i\bullet\bullet}^{n_i+\rho_i-1}b_{i(1)}^{\rho_i}\exp(-t_i\lambda_{i\bullet\bullet}-b_{i(1)}\lambda_{i\bullet\bullet}+b_{i0}b_{i(1)}\lambda_{i\bullet\bullet}),
\end{aligned}$$

which shows the uniform convergence of $f(t_i\,|\,n_i,\lambda_{i\bullet\bullet})f(\lambda_{i\bullet\bullet}\,|\,\{r_{jk}\}_{j,k},\{b_{ijk}\}_{j,k})$. So the integration and countable summation are interchangeable, and consequently, (13) holds. Next we want to show the CDF of $t_i$,

$$\begin{aligned}
P(t_i<q\,|\,n_i,\{r_{jk}\}_{j,k},\{b_{ijk}\}_{j,k})&=\int_0^q\frac{c_it_i^{n_i-1}}{\Gamma(n_i)}\sum_{m=0}^{\infty}\frac{\Gamma(n_i+\rho_i+m)\delta_{im}b_{i(1)}^{\rho_i+m}}{\Gamma(\rho_i+m)(t_i+b_{i(1)})^{n_i+\rho_i+m}}dt_i\\
&=\sum_{m=0}^{\infty}\int_0^q\frac{c_it_i^{n_i-1}}{\Gamma(n_i)}\frac{\Gamma(n_i+\rho_i+m)\delta_{im}b_{i(1)}^{\rho_i+m}}{\Gamma(\rho_i+m)(t_i+b_{i(1)})^{n_i+\rho_i+m}}dt_i. \qquad (14)
\end{aligned}$$

It suffices to show (14). Note that

$$\begin{aligned}
&\frac{c_it_i^{n_i-1}}{\Gamma(n_i)}\sum_{m=0}^{\infty}\frac{\Gamma(n_i+\rho_i+m)\delta_{im}b_{i(1)}^{\rho_i+m}}{\Gamma(\rho_i+m)(t_i+b_{i(1)})^{n_i+\rho_i+m}}\\
\leq&\frac{c_it_i^{n_i-1}}{\Gamma(n_i)}\sum_{m=0}^{\infty}\frac{\Gamma(n_i+\rho_i+m)b_{i(1)}^{\rho_i+m}\Gamma(n_i+\rho_i+m)}{\Gamma(\rho_i+m)(t_i+b_{i(1)})^{n_i+\rho_i+m}\Gamma(\rho_i)m!}\\
=&\frac{c_it_i^{n_i-1}}{\Gamma(n_i)}\frac{\Gamma(\rho_i+n_i)b_{i(1)}^{\rho_i}}{\Gamma(\rho_i)(t_i+b_{i(1)})^{n_i+\rho_i}}\sum_{m=0}^{\infty}\left[\frac{\Gamma(n_i+\rho_i+m)}{\Gamma_{n_i+\rho_i}m!}\left(\frac{b_{i(1)}}{t_i+b_{i(1)}}\right)^m\right]\\
=&\frac{c_it_i^{n_i-1}\Gamma(\rho_i+n_i)b_{i(1)}^{\rho_i}t_i^{n_i+\rho_i}}{\Gamma(n_i)\Gamma(\rho_i)(t_i+b_{i(1)})^{2(n_i+\rho_i)}}.
\end{aligned}$$

The last equation holds because the summation of a negative binomial probability mass function is 1. So $f(t_i\,|\,n_i,\{r_{jk}\}_{j,k},\{b_{ijk}\}_{j,k})$ is uniformly convergent and (14) holds. Plugging in $n_i=1$ and calculating the integration, we obtain the CDF of $t_i$. □

## B Bayesian inference of LDR

With $\boldsymbol{x}_i$ denoting the covariates, $y_i$ event type, and $t_i$ the time to event of observation $i$, we express the full hierarchical form of LDR defined in (7), as

$$\begin{aligned}
&t_i=t_{iy_i},\ y_i=\underset{j\in\{1,\ldots,J\}}{\operatorname{argmin}}\ t_{ij},\ t_{ij}=t_{ij\kappa_{ij}},\ \kappa_{ij}=\underset{k\in\{0,\ldots,K\}}{\operatorname{argmin}}\ t_{ijk},\\
&t_{ijk}\sim\text{Exp}(\lambda_{ijk}),\ \lambda_{ijk}\sim\text{Gamma}(r_{jk},e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}}),\ k=1,\cdots,K,\\
&\boldsymbol{\beta}_{jk}\sim\prod_{g=1}^{P}\mathcal{N}(0,\alpha_{gjk}^{-1}),\ \alpha_{gjk}\sim\text{Gamma}(a_0,1/b_0),\ r_{jk}\sim\text{Gamma}(\gamma_{0j}/K,1/c_{0j}),
\end{aligned}$$

where $k=1,\cdots,K$, $i=1,\cdots,n$, and $j=1,\cdots,J$. We further let $\gamma_{0j}\sim\text{Gamma}(e_0,1/f_0)$, $c_{0j}\sim\text{Gamma}(e_1,1/f_1)$, $r_0\sim\text{Gamma}(e_0,1/f_0)$, and set $e_0=f_0=e_1=f_1=0.01$. Let us denote $T_i$ and $T_{ic}$ as the observed failure time and right censoring time, respectively, for observation $i$. Since left censoring is uncommon and not shown in the real datasets analyzed, we only consider right censoring in our inference and leave to readers other types of censoring which can be analogously done. A Gibbs sampler accommodating missing event times or missing event types proceeds by iterating the following steps.

1. If $y_i$ is observed, we first sample $\kappa_{iy_i}$ by
$$P(\kappa_{iy_i} = k \,|\, y_i, \cdots) = \frac{\lambda_{iy_ik}}{\sum_{k'=1}^{K} \lambda_{iy_ik'}}.$$
If $y_i$ is unobserved which means a missing event type, we sample $(y_i, \kappa_{iy_i})$ by
$$P(y_i = j, \kappa_{iy_i} = k \,|\, \cdots) = \frac{\lambda_{ijk}}{\sum_{j'=1}^{S} \sum_{k'=1}^{K} \lambda_{ij'k'}}.$$
We then denote $m_{jk} = \sum_{i:y_i=j} \mathbf{1}(\kappa_{iy_i} = k)$. Define $n_{ijk} = 1$ if $y_i = j$ and $\kappa_{iy_i} = k$, and otherwise $n_{ijk} = 0$. The above sampling procedure means that given the event type $y_i$, we sample the index of the sub-risk that has the minimum survival time.
2. Update $t_i$ for $i = 1, \cdots, n$, $j = 1, \cdots, J$ and $k = 1, \cdots, K$.
   (a) If the failure time $T_i$ is observed, we set $t_i = T_i$.
   (b) Otherwise, we let $t_i = T_{ic} + \tilde{t}_i$, where $(\tilde{t}_i \,|\, -) \sim \text{Exp}(\sum_{j=1}^{S} \sum_{k=1}^{K} \lambda_{ijk})$ and $T_{ic}$ is the right censoring. Note $T_{ic} = 0$ if both event time and censoring time are missing for observation $i$.
3. Sample $(\lambda_{ijk} \,|\, -) \sim \text{Gamma}\left(r_{jk} + n_{ijk}, \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}}}{1+t_i e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}}}\right)$, for $i = 1, \cdots, n$, $j = 1, \cdots, J$ and $k = 1, \cdots, K$.
4. Sample $\boldsymbol{\beta}_{jk}$, for $j = 1, \cdots, J$ and $k = 1, \cdots, K$, by Pólya Gamma (PG) data augmentation. First Sample $(\omega_{ijk} \,|\, -) \sim \text{PG}(r_{jk} + n_{ijk}, \boldsymbol{x}_i'\boldsymbol{\beta}_{jk} + \log t_i)$. Then sample $(\boldsymbol{\beta}_{jk} \,|\, -) \sim \text{MVN}(\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ where $\boldsymbol{\Sigma}_{jk} = (V_{jk} + \boldsymbol{X}'\Omega_{jk}\boldsymbol{X})^{-1}$, $\boldsymbol{X} = [\boldsymbol{x}_1', \cdots, \boldsymbol{x}_N']'$, $\Omega_{jk} = \text{diag}(\omega_{1jk}, \cdots, \omega_{njk})$ and $\boldsymbol{\mu}_{jk} = \boldsymbol{\Sigma}_{jk}\left[-\sum_{i=1}^{N}\left(\omega_{ijk}\log t_i + \frac{r_{jk}-n_{ijk}}{2}\right)\boldsymbol{x}_i\right]$. Note to sample from the Pólya-Gamma distribution, we use a fast and accurate approximate sampler of Zhou [41] that matches the first two moments of the original distribution; we set the truncation level of that sampler as five.
5. Sample $(\alpha_{vjk} \,|\, -) \sim \text{Gamma}\left(a_0 + 0.5, 1/(b_0 + 0.5\beta_{vjk}^2)\right)$ for $v = 0, \cdots, V$, $j = 1, \cdots, J$ and $k = 1, \cdots, K$.
6. Sample $r_{jk}$ and $\gamma_{0j}$, for $j = 1, \cdots, J$ and $k = 1, \cdots, K$, by Chinese restaurant table (CRT) data augmentation [43].
   First sample $(n_{ijk}^{(2)} \,|\, -) \sim \text{CRT}(n_{ijk}, r_{jk})$, and $(l_{jk} \,|\, -) \sim \text{CRT}(\sum_{i=1}^{N} n_{ijk}^{(2)}, \gamma_{0j}/K)$. Then sample $(r_{jk} \,|\, -) \sim \text{Gamma}\left(\sum_{i=1}^{N} n_{ijk}^{(2)} + \gamma_{0j}/K, \frac{1}{c_{0j}+\sum_{i=1}^{N}\log(1+t_i e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}})}\right)$, and $(\gamma_{0j} \,|\, -) \sim \text{Gamma}\left(e_0 + \sum_{k=1}^{K} l_{jk}, \frac{1}{f_0 - \frac{1}{K}\sum_{k=1}^{K}\log(1-p_{jk})}\right)$, where $p_{jk} = \frac{\sum_{i=1}^{N}\log(1+t_i e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}})}{c_{0j}+\sum_{i=1}^{N}\log(1+t_i e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}})}$.
7. Sample $(c_{0j} \,|\, -) \sim \text{Gamma}\left(e_1 + \gamma_{0j}, \frac{1}{f_1+\sum_{k=1}^{K} r_{jk}}\right)$ for $j = 1, \cdots, J$.
8. For $j = 1, \cdots, J$ and $k = 1, \cdots, K$, prune sub-risk $k$ of risk $j$ for all observations if $m_{jk} = 0$, by setting $\lambda_{ijk} \equiv 0$ and $t_{ijk} \equiv \infty$ for $\forall i$.

## C Maximum a posteriori estimation

With the reparameterization that $\lambda_{ijk} = \tilde{\lambda}_{ijk} e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{jk}}$ where $\tilde{\lambda}_{ijk} \overset{iid}{\sim} \text{Gamma}(r_{jk}, 1)$ we first find $p_i$, the likelihood of observation $i$ having event type $y_i$ at event time $t_i$.

$$p_i = \mathbb{E}\left(P(t_i, y_i \,|\, \boldsymbol{\lambda}_i)\right) \equiv \int (p_{t_i} \times p_{y_i})\, p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i$$

where $\tilde{\boldsymbol{\lambda}}_i = \{\tilde{\lambda}_{ijk}\}_{j,k}$, $p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) = \prod_{j,k} \text{Gamma}(r_{jk}, 1)$, $\boldsymbol{r} = \{r_{jk}\}_{j,k}$, Gamma$(r_{jk}, 1)$ is the pdf of a gamma distribution with shape $r_{jk}$ and scale 1, and

$$p_{t_i} = \begin{cases} (\sum_{j,k} \tilde{\lambda}_{ijk} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{jk}}) \exp\left\{-t_i \sum_{jk} \tilde{\lambda}_{ijk} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{jk}}\right\} & \text{if } t_i \text{ is uncensored and observed,} \\ \exp\left\{-T_{ic} \sum_{jk} \tilde{\lambda}_{ijk} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{jk}}\right\} & \text{if } t_i \text{ is right censored at } T_{ic}, \text{ i.e., } t_i > T_{ic}, \\ 1 & \text{if } t_i \text{ is missing, but } y_i \text{ is not,} \end{cases}$$

$$p_{y_i} = \begin{cases} \frac{\sum_k \tilde{\lambda}_{i y_i k} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{y_i k}}}{\sum_{j,k} \tilde{\lambda}_{ijk} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_{jk}}} & \text{if } y_i \text{ is not missing,} \\ 1 & \text{if } y_i \text{ is missing, but } t_i \text{ is not.} \end{cases}$$

Note that we do not define $P(t_i, y_i \,|\, \boldsymbol{\lambda}_i)$ if both $t_i$ and $y_i$ are missing and remove such observations from data. We write $p_{t_i} \equiv p_t(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r})$ and $p_{y_i} \equiv p_y(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r})$.

Imposing a prior $p(\boldsymbol{\beta}_{jk})$ on $\boldsymbol{\beta}_{jk}$ and $p(r_{jk})$ on $r_{jk}$, the log posterior is

$$\log P = \sum_i \log p_i + \sum_{j,k} \log p(\boldsymbol{\beta}_{jk}) + \sum_{j,k} \log p(r_{jk}) + C \tag{15}$$

where $C$ is a constant function of $\{\boldsymbol{\beta}_{jk}\}$ and $\{r_{jk}\}$. In practice we assume a Student's $t$ distribution with degrees of freedom $a$ on each element of $\boldsymbol{\beta}_{jk}$ and a Gamma$(0.01/K, 1/0.01)$ prior on $r_{jk}$. We also found a Gamma$(1/K, 1)$ prior on $r_{jk}$ or an $l^2$-regularizer, $0.001||\boldsymbol{r}||_2$, is more numerically stable. Then we have

$$\log P = \sum_i \log p_i + \sum_{v,j,k} -\frac{a+1}{2} \log\left(1 + \beta_{vjk}^2/a\right) + \sum_{j,k} \left[(0.01/K - 1)\log r_{jk} - 0.01 r_{jk}\right] + c$$

where $c$ is also a constant function of $\{\boldsymbol{\beta}_{jk}\}$ and $\{r_{jk}\}$. For simplicity, we define $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{jk}\}_{j,k}$. We want to maximize $\log P$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{r}$. The difficulty lies in $p_i$ being the expectation of $p_{t_i} \times p_{y_i}$ over $\tilde{\boldsymbol{\lambda}}_i$ which is a random variable parameterized by $\boldsymbol{r}$. Now we show how to approximate the derivatives of $\log p_i$ by Monte-Carlo simulation and score function gradients. Specifically,

$$\nabla_{\boldsymbol{\beta}} \log p_i = \frac{\int \left[\nabla_{\boldsymbol{\beta}} \left(p_{t_i} \times p_{y_i}\right)\right] p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i}{\int \left(p_{t_i} \times p_{y_i}\right) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i} \approx \frac{\frac{1}{M} \sum_{m=1}^M \nabla_{\boldsymbol{\beta}} \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})\right]}{\frac{1}{M} \sum_{m=1}^M \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})\right]} \tag{16}$$

where $M$ is a reasonably large number, say 10, $\tilde{\boldsymbol{\lambda}}_i^{(m)} = \{\tilde{\lambda}_{ijk}^{(m)}\}_{jk}$ and $\tilde{\lambda}_{ijk}^{(m)} \overset{iid}{\sim}$ Gamma$(r_{jk}, 1)$, $\forall i = 1, \cdots, n$ and $m = 1, \cdots, M$. With the fact that $\nabla_{\boldsymbol{r}} p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) = p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r})$,

$$\begin{aligned} \nabla_{\boldsymbol{r}} \log p_i &= \frac{\int \nabla_{\boldsymbol{r}} \left[\left(p_{t_i} \times p_{y_i}\right) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r})\right] d\tilde{\boldsymbol{\lambda}}_i}{\int \left(p_{t_i} \times p_{y_i}\right) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i} \\ &= \frac{\int \left(p_{t_i} \times p_{y_i}\right) \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i}{\int \left(p_{t_i} \times p_{y_i}\right) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i} \\ &\approx \frac{\frac{1}{M} \sum_{m=1}^M p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})}{\frac{1}{M} \sum_{m=1}^M \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})\right]} \\ &= \sum_{m=1}^M \frac{p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})}{\sum_{m'=1}^M \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m')} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m')} \,|\, \boldsymbol{r})\right]} \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}). \end{aligned} \tag{17}$$

Therefore, we can approximate the derivatives of $\log P$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{r}$ by plugging in (16) and (17), respectively, and maximize $-\log P$ by (stochastic) gradient descent.

## D Description of SEER data and experiment settings

### D.1 SEER data for survival analysis

We use breast cancer data from Surveillance, Epidemiology, and End Results Program (SEER) of National Cancer Institute between 1973 and 2003. There are two causes of death; the first is breast cancer and the second is *other causes* treated as a whole. Explanatory variables include age of diagnosis, gender, race, marital status, historic stage, behavior type, tumor size, tumor extension, number of malignant tumors, number of regional nodes containing tumor, number of regional nodes that are examined or removed, confirmation type and surgery type. We use dummies for all categorical variables and select a subset of patient collected from the hospital *C503* so that we do not have to consider site effects. We exclude observations with any missing values in explanatory variables. Finally, there are 2647 and 4166 observations in our data if we exclude and include observations with a missing cause of death, respectively.

### D.2 Experiment settings

We run $10,000$ interations of Gibbs sampler for LDR with the gamma process truncated at $K = 10$ for all experiments, take the first $8,000$ as burn-in, and estimate CIF by averaging its estimators from the last $2,000$ iterations. For random survial forests, we set the number of trees equal to 100 and the number of splits equal to 2 as suggested by Ishwaran et al. [24]. We use `R` for all the analysis: C-indices are estimated by package `pec` [55], the Cox model by `riskRegression` [56], FG by `cmprsk` [57], BST by `CoxBoost` [58], and RF by `randomForestSRC` [59].

Isomap algorithm is often used for nonlinear dimensionality reduction. We first find five nearest neighbors of each observation, and then construct a neighborhood graph where an observation is connected to another with the edge length equal to the Euclidean distance if it is a 5-nearest neighbor. We calculate the shortest path between two nodes of the graph by Floyd–Warshall algorithm [60] and obtain a geodesic distance matrix with which we compute two-dimensional embeddings by classical multidimensional scaling [61].

## E Additional experimental results

We first show in Table 2 through Table 8 the Brier score at the evaluation time for each risk of the synthetic data sets, SEER and DLBCL data, respectively. Brier score (BS) for risk $j$ at time $\tau$ can be estimated by $\mathrm{BS}_j(\tau) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}(t_i \leq \tau, y_i = j) - P(t_i \leq \tau, y_i = j)\right]^2$, with a smaller value indicating a better model fit. Note that the model performance quantified by Brier score is basically consistent with quantified by C-indices. For the cases like synthetic data 1, SEER and ABC and GCB of DLBCL, where covariates are believed to be linearly influential by C-indices, the Brier scores are comparable for Cox, FG, BST and LDR, and slightly smaller than those of RF. For synthetic data 2 and T3 of DLBCL where C-indices imply nonlinear covariate effects, the Brier scores of LDR and RF are smaller than those of Cox, FG and BST. Moreover, the Brier score of LDR is slightly larger than those of RF for synthetic data 2 but smaller for T3 of DLBCL.

Table 2: Brier score for risk 1 of synthetic data 1.

| | $\tau = 0.5$ | $\tau = 1$ | $\tau = 1.5$ | $\tau = 2$ | $\tau = 2.5$ | $\tau = 3$ |
|---|---|---|---|---|---|---|
| Cox | .165±.012 | **.166**±.010 | .165±.010 | .166±.012 | **.164**±.012 | **.162**±.012 |
| FG | .168±.010 | .167±.010 | .166±.009 | .166±.012 | **.164**±.013 | **.162**±.012 |
| BST | .167±.010 | **.166**±.010 | .166±.010 | .166±.010 | .166±.011 | .165±.010 |
| RF | .173±.013 | .175±.012 | .171±.012 | .172±.014 | .172±.014 | .170±.014 |
| LDR | **.164**±.014 | **.166**±.011 | **.164**±.010 | **.165**±.012 | **.164**±.013 | **.162**±.013 |

Table 3: Brier score for risk 2 of synthetic data 1.

| | $\tau = 0.5$ | $\tau = 1$ | $\tau = 1.5$ | $\tau = 2$ | $\tau = 2.5$ | $\tau = 3$ |
|---|---|---|---|---|---|---|
| Cox | **.152**±.011 | **.158**±.014 | **.158**±.015 | .157±.015 | **.157**±.014 | .159±.014 |
| FG | .157±.012 | .159±.014 | .159±.015 | .158±.015 | .158±.014 | .159±.014 |
| BST | .158±.013 | **.158**±.013 | **.158**±.013 | .158±.013 | .158±.013 | .158±.013 |
| RF | .164±.012 | .166±.015 | .166±.016 | .164±.015 | .165±.014 | .165±.014 |
| LDR | **.152**±.012 | **.158**±.014 | **.158**±.016 | **.156**±.015 | **.157**±.014 | **.158**±.014 |

We show in Figure 5 the C-indices of risk 2 for synthetic data 1 and 2 used in Section 5.1. The C-indices of risk 2 for data 1 are very similar to those of risk 1 as in panel (a) of Figure 1; LDR, Cox,

Table 4: Brier score for risk 1 of synthetic data 2.

| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ | $\tau = 4$ | $\tau = 5$ | $\tau = 6$ |
|---|---|---|---|---|---|---|
| Cox | .206±.008 | .235±.006 | .241±.005 | .242±.005 | .243±.005 | .243±.005 |
| FG | .206±.008 | .235±.006 | .241±.006 | .242±.005 | .243±.005 | .243±.005 |
| BST | .234±.005 | .234±.005 | .234±.005 | .234±.005 | .234±.005 | .234±.005 |
| RF | **.186**±.010 | **.193**±.011 | **.188**±.011 | **.186**±.010 | **.184**±.010 | **.183**±.010 |
| LDR | .193±.007 | .194±.007 | .191±.006 | .191±.006 | .191±.006 | .191±.006 |

Table 5: Brier score for risk 2 of synthetic data 2.

| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ | $\tau = 4$ | $\tau = 5$ | $\tau = 6$ |
|---|---|---|---|---|---|---|
| Cox | .251±.002 | .247±.003 | .245±.004 | .244±.004 | .244±.004 | .244±.004 |
| FG | .251±.002 | .247±.003 | .245±.004 | .244±.004 | .244±.004 | .244±.005 |
| BST | .245±.003 | .245±.003 | .245±.003 | .245±.003 | .245±.003 | .245±.003 |
| RF | **.178**±.011 | **.182**±.010 | **.181**±.010 | **.182**±.010 | **.182**±.010 | **.183**±.010 |
| LDR | .204±.006 | .199±.005 | .197±.005 | .198±.005 | .197±.005 | .199±.005 |

Table 6: Brier score for ABC of DLBCL.

| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ | $\tau = 4$ | $\tau = 5$ | $\tau = 6$ |
|---|---|---|---|---|---|---|
| Cox | .162±.056 | .190±.055 | .196±.058 | .202±.054 | .196±.053 | .202±.054 |
| FG | .159±.057 | .185±.058 | .198±.058 | .196±.057 | .196±.056 | .199±.055 |
| BST | .136±.045 | .146±.045 | **.163**±.044 | **.154**±.044 | **.150**±.045 | **.152±.044** |
| RF | .156±.052 | .173±.055 | .196±.051 | .198±.051 | .198±.051 | .200±.051 |
| LDR | **.131**±.050 | **.143**±.050 | **.163**±.047 | .158±.045 | .155±.043 | .156±.041 |

Table 7: Brier score for GCB of DLBCL.

| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ | $\tau = 4$ | $\tau = 5$ | $\tau = 6$ |
|---|---|---|---|---|---|---|
| Cox | .138±.048 | .212±.051 | .266±.061 | 268±.062 | .265±.062 | .277±.063 |
| FG | .137±.046 | .206±.064 | .268±.059 | .265±.062 | .267±.063 | .273±.064 |
| BST | .133±.046 | .204±.056 | .262±.042 | .252±.036 | .253±.048 | .257±.041 |
| RF | .137±.038 | .197±.054 | .248±.050 | .247±.046 | .253±.050 | .262±.053 |
| LDR | **.129**±.035 | **.179**±.052 | **.242**±.053 | **.236**±.050 | **.237**±.052 | **.244**±.052 |

Table 8: Brier score for T3 of DLBCL.

| | $\tau = 1$ | $\tau = 2$ | $\tau = 3$ | $\tau = 4$ | $\tau = 5$ | $\tau = 6$ |
|---|---|---|---|---|---|---|
| Cox | .193±.053 | .190±.061 | .206±.069 | .220±.071 | .233±.068 | .245±.072 |
| FG | .183±.051 | .186±.062 | .195±.067 | .212±.069 | .230±.070 | .234±.069 |
| BST | .169±.046 | .172±.044 | .177±.049 | .185±.046 | .185±.047 | .193±.048 |
| RF | .117±.045 | .151±.046 | .157±.043 | .169±.049 | .180±.051 | .185±.052 |
| LDR | **.111**±.035 | **.137**±.038 | **.142**±.036 | **.151**±.041 | **.165**±.044 | **.171**±.046 |

Table 9: Brier score for breast cancer of SEER.

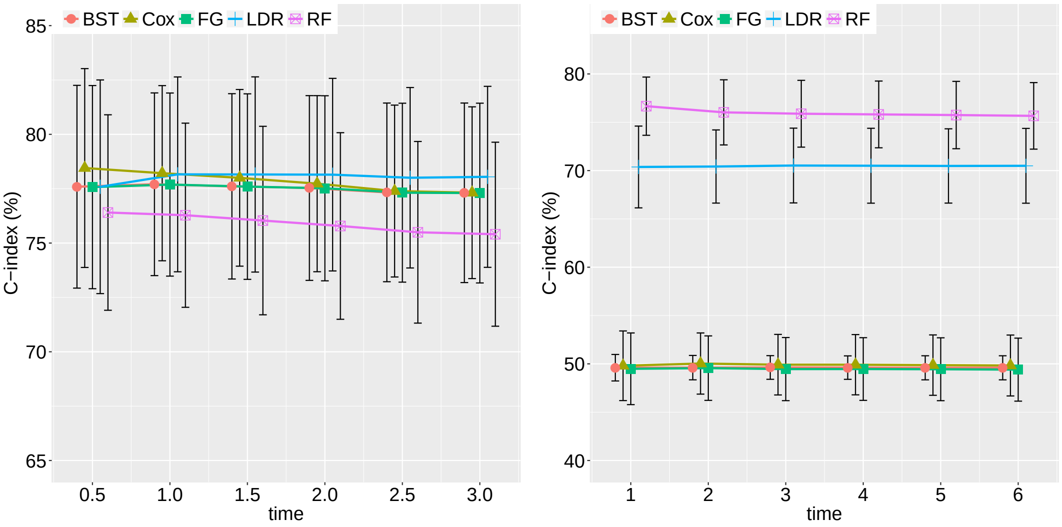
| | $\tau = 10$ | $\tau = 50$ | $\tau = 100$ | $\tau = 150$ | $\tau = 200$ | $\tau = 250$ | $\tau = 300$ |
|---|---|---|---|---|---|---|---|
| Cox | **.014**±.003 | **.106**±.006 | **.150**±.006 | .169±.006 | .177±.007 | **.180**±.006 | **.179**±.005 |
| FG | .016±.003 | .112±.011 | .156±.009 | .170±.006 | .177±.011 | .186±.013 | .189±.010 |
| BST | **.014**±.004 | .114±.008 | .154±.007 | **.168**±.004 | **.174**±.009 | .184±.009 | .184±.008 |
| RF | .015±.003 | **.106**±.007 | .151±.007 | .174±.008 | .182±.008 | .185±.008 | .187±.007 |
| LDR | .018±.003 | .107±.006 | .153±.006 | .173±.006 | .182±.007 | .186±.006 | .185±.006 |

FG and BST are comparable and all slightly outperform RF in terms of mean values. The C-indices of risk 2 for data 2 are also analogous to those of risk 1 as in panel (c) of Figure 1 except that LDR

Table 10: Brier score for other causes of SEER.

| | $\tau = 10$ | $\tau = 50$ | $\tau = 100$ | $\tau = 150$ | $\tau = 200$ | $\tau = 250$ | $\tau = 300$ |
|---|---|---|---|---|---|---|---|
| Cox | **.008**±.003 | **.073**±.011 | **.141**±.010 | .195±.010 | **.204**±.010 | **.193**±.009 | **.178**±.007 |
| FG | **.008**±.003 | .076±.010 | .161±.013 | .241±.018 | .290±.029 | .302±.035 | .301±.040 |
| BST | **.008**±.003 | .074±.009 | .142±.011 | .201±.010 | .213±.016 | .203±.006 | .228±.018 |
| RF | **.008**±.003 | **.073**±.010 | .145±.011 | .200±.010 | .213±.009 | .207±.009 | .199±.008 |
| LDR | .009±.003 | .083±.008 | .148±.008 | **.193**±.008 | .205±.009 | .199±.008 | .194±.008 |

slightly underperforms RF in terms of mean values. But they both significally outperforms the other three approaches which completely fail.



(a) C-index of risk 2 for synthetic data 1.

(b) C-index of risk 2 for synthetic data 2.

Figure 5: Cause-specific C-indices of risk 2 for synthetic data 1 and 2.

Since we have random partitions in the analysis of DLBCL dataset, improvements of LDR can be underrated for the overlaps of boxplots across the five approaches in Figure 2. Therefore, we calculate the difference of C-indices between LDR and each of the other four benchmarks within each random partition, and report the mean and standard deviation in Table 11 where $\Delta_{\mathrm{X}}$, $\mathrm{X} \in \{\mathrm{Cox},\ \mathrm{FG},\ \mathrm{BST},\ \mathrm{RF}\}$, denotes the C-index of LDR minus that of approach X. In terms of mean difference, LDR outperforms all the other benchmarks for all the three risks at any time evaluated except for BST under risk ABC.

Table 11: Difference of C-indices between LDR and other benchmarks.

| | ABC | | | | GCB | | | | T3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| year | $\Delta_{\mathrm{COX}}$ | $\Delta_{\mathrm{FG}}$ | $\Delta_{\mathrm{BST}}$ | $\Delta_{\mathrm{RF}}$ | $\Delta_{\mathrm{COX}}$ | $\Delta_{\mathrm{FG}}$ | $\Delta_{\mathrm{BST}}$ | $\Delta_{\mathrm{RF}}$ | $\Delta_{\mathrm{COX}}$ | $\Delta_{\mathrm{FG}}$ | $\Delta_{\mathrm{BST}}$ | $\Delta_{\mathrm{RF}}$ |
| 1 | .09±.08 | .03±.05 | .01±.06 | .06±.08 | .07±.09 | .06±.09 | .07±.06 | .16±.12 | .16±.15 | .11±.12 | .06±.05 | .10±.12 |
| 2 | .09±.06 | .03±.04 | .00±.07 | .04±.08 | .11±.08 | .10±.08 | .05±.06 | .17±.13 | .20±.17 | .10±.08 | .05±.05 | .03±.08 |
| 3 | .09±.05 | .04±.05 | -.01±.06 | .05±.06 | .12±.07 | .12±.06 | .05±.06 | .16±.09 | .20±.17 | .10±.09 | .05±.05 | .03±.08 |
| 4 | .09±.05 | .04±.05 | -.01±.06 | .05±.06 | .11±.07 | .12±.06 | .05±.06 | .15±.10 | .21±.15 | .11±.09 | .04±.05 | .02±.08 |
| 5 | .09±.05 | .04±.05 | -.01±.06 | .05±.06 | .12±.07 | .12±.06 | .05±.06 | .15±.09 | .21±.16 | .11±.08 | .04±.05 | .01±.08 |
| 6 | .09±.05 | .03±.05 | -.01±.06 | .04±.06 | .11±.07 | .12±.06 | .05±.06 | .16±.09 | .23±.14 | .11±.08 | .04±.05 | .02±.09 |