Journal of Machine Learning Research 24 (2023) 1-43 Submitted 7/22; Revised 2/23; Published 8/23

Weibull Racing Survival Analysis with Competing Events, Left Truncation, and Time-Varying Covariates

Quan Zhang QUAN.ZHANG@BROAD.MSU.EDU
Department of Accounting and Information Systems
Michigan State University
East Lansing, MI 48824, U.S.A.

Yanxun Xu YANXUN.XU@JHU.EDU
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218, U.S.A.

Mei-Cheng Wang MCWANG@JHU.EDU
Department of Biostatistics
Johns Hopkins University
Baltimore, MD 21205, U.S.A.

Mingyuan Zhou MINGYUAN.ZHOU@MCCOMBS.UTEXAS.EDU
Department of Information, Risk, and Operations Management
Department of Statistics and Data Sciences
The University of Texas at Austin
Austin, TX 78712, U.S.A.

Editor: Sanmi Koyejo

Abstract

We propose Bayesian nonparametric Weibull delegate racing (WDR) to fill the gap in interpretable nonlinear survival analysis with competing events, left truncation, and time-varying covariates. We set a two-phase race among a potentially infinite number of sub-events to model nonlinear covariate effects, which does not rely on transformations or complex functions of the covariates. Using gamma processes, the nonlinear capacity of WDR is parsimonious and data-adaptive. In prediction accuracy, WDR dominates cause-specific Cox and Fine-Gray models and is comparable to random survival forests in the presence of time-invariant covariates. More importantly, WDR can cope with different types of censoring, missing outcomes, left truncation, and time-varying covariates, on which other nonlinear models, such as the random survival forests, Gaussian processes, and deep learning approaches, are largely silent. We develop an efficient MCMC algorithm based on Gibbs sampling. We analyze biomedical data, interpret disease progression affected by covariates, and show the potential of WDR in discovering and diagnosing new diseases.

Keywords: Bayesian nonparametrics, censoring and missing outcomes, interpretable nonlinearity, MCMC

1. Introduction

In survival analysis, it is common to consider competing events (also known as competing risks) that are mutually exclusive. In other words, the occurrence of one event precludes

the occurrence of another. For example, when studying the time to death of a patient, all possible causes of death are competing events since a patient who died of one cause would never die of others. In the presence of competing events, people model both the event time and the event type or which one of the competing events occurs first. One may argue that every survival model can handle competing events if each competing event is analyzed separately and meanwhile, subjects having suffered from other events are treated as right-censored[1]. However, this approach can be problematic because it violates the assumption of independent or non-informative censoring (Kalbfleisch and Prentice, 2011), which means the censoring time is stochastically independent of the length of survival time, and thus often leads to biased estimation of the cumulative incidence function (Austin et al., 2016).

Left truncation and time-varying covariates often exist in biomedical studies and should be accommodated in survival analysis. Left truncation occurs when subjects who have already experienced a failure event or who have passed some milestone are not eligible to be recruited. Restricting the analysis to the recruited subjects without accounting for left truncation results in an immortal time bias (Lévesque et al., 2010; McGough et al., 2021) because these subjects cannot have the failure event prior to entering the study. For example, in Alzheimer's disease studies, the failure time can be the age at the onset of symptoms of mild cognitive impairment (MCI), and only subjects who are disease-free at baseline are recruited. In this setting, a patient having MCI at the time of recruitment is excluded from the study, leading to an underestimated risk at her age if the analysis does not account for left truncation. Time-varying covariates occur when a covariate value changes over time during the follow-up period. In Alzheimer's disease studies, biomarkers such as cerebrospinal fluid variables and cognitive measures can be collected longitudinally and are informative of the disease progression. In this paper, we consider covariate measurements at discrete points in time, which is often the case in practice.

When studying competing events in biomedical research, interpreting model inference is often of interest. Two of the most popular models are the cause-specific Cox model (Prentice et al., 1978; Lunn and McNeil, 1995; Putter et al., 2007) and the Fine-Gray proportional subdistribution hazards model (Fine and Gray, 1999), which are both semi-parametric. The former models the cause-specific hazard function and is often applied to studying the etiology of diseases, while the latter models the subdistribution function and is favorable when developing prediction models and risk-censoring systems (Austin et al., 2016). Both models assume proportional hazards and use a linear function of covariates to interpret how much a unit increase in a covariate is multiplicative to the hazard functions. We refer the readers to Austin et al. (2020) for a detailed review on interpreting the Fine-Gray model in various applications.

However, semi-parametric and parametric models for competing events often achieve interpretability by sacrificing flexibility. Specifically, the Cox and Fine-Gray models and other semi-parametric approaches depend on the linear function of covariates in partial likelihoods, and thus the model fit can be undermined in the presence of non-monotonic covariate effects. Data transformation and stratification can alleviate such a problem but often require expert opinions and/or an excessive number of parameters. One can also replace the linear function of covariates with complex functions, such as splines (Danieli and

1. Right-censored data, or right censoring, means that we do not observe the exact event time of a subject but know it is larger than some value, and this value is defined as the right censoring time.

Abrahamowicz, 2019) and neural networks (Kvamme et al., 2019), but this practice leads to difficulty in model interpretation and may overburden practitioners who try to explain a covariate effect. Other (semi-)parametric methods for competing event analysis have been categorized by Haller et al. (2013) into mixture models (Ng and McLachlan, 2003; Lau et al., 2011), vertical modeling (Nicolaie et al., 2010, 2019), and the pseudo-observation approach (Klein and Andersen, 2005; Rahman et al., 2021); they are also incapable of nonlinear modeling unless making data transformations, and/or they cannot account for left truncation or time-varying covariates.

Nonparametric or deep learning approaches can enhance model flexibility, but their interpretation is not straightforward, and how to deal with left truncation or time-varying covariates by these approaches has not been sufficiently investigated. For example, the random survival forests for competing events (Ishwaran et al., 2014) has simplified the modeling of nonlinear covariate effects, and Bayesian models have been developed, such as Gaussian processes (Barrett and Coolen, 2013; Alaa and Schaar, 2017) and Lomax racing (Zhang and Zhou, 2018). However, no extensions of these models have been made to accommodate left truncation or time-varying covariates. Shi et al. (2021) propose a dependent Dirichlet process mixture model for competing events in the presence of a binary time-varying covariate for treatment switching, but it is limited to the scenario where only one binary time-varying covariate exists. In addition to replacing the linear function of covariates in a (semi-)parametric model by neural networks (Kvamme et al., 2019; Nagpal et al., 2021; Rahman et al., 2021), another stream of deep learning methods is to discretize the time and transform competing-event survival analysis to a classification problem (Lee et al., 2018, 2019; Tjandra et al., 2021). These classification-based methods cannot easily handle left truncation and are sensitive to time discretization. Specifically, a fine-grained discretization causes imbalanced labels, while a coarse-grained one leads to inaccurate survival estimation. Moreover, a potential overfitting on smaller data and a general lack of interpretation in neural networks can concern practitioners.

To address the aforementioned challenges, we construct Weibull delegate racing (WDR), a gamma process-based nonparametric Bayesian hierarchical model, for survival analysis with competing events. It achieves both interpretability and flexibility without transformations or complex functions of covariates. WDR assumes non-informative censoring and uses the race among Weibull random variables to jointly model event times, event types, and potential subtypes. It enables data-adaptive nonlinear modeling and has the interpretation as a race among latent sub-events. We propose an efficient MCMC algorithm based on Gibbs sampling and slice sampling for posterior inference. The MCMC can handle different types of censoring and impute missing event types. To the best of our knowledge, WDR is the first approach for interpretable nonlinear modeling of competing events in the presence of left truncation and time-varying covariates. Moreover, it delivers outstanding performance in prediction accuracy.

The paper proceeds as follows. In Section 2, we propose Weibull racing and Weibull delegate racing. Section 3 shows the Bayesian inference and how WDR deals with time-varying covariates, censoring, and missing event times or types, inducing the connection between WDR and discrete choice models for classification. In Section 4, we use synthetic data to showcase WDR's parsimonious nonlinear modeling capacity and outstanding performance. In Section 5, we analyze real data of lymphoma and Alzheimer's disease to understand how

hazards of competing events are influenced by covariates and show the potential of WDR in discovering and diagnosing new diseases. Section 6 concludes the paper. We defer to the Appendices the proofs, some technical details of WDR including the inference algorithms, some experiment settings, and supplementary results.

2. Weibull Racing and Weibull Delegate Racing

We first define the left-truncated Weibull distribution and introduce the Weibull racing property that can be directly used for competing event analysis with left truncation. Then we propose the Weibull racing survival model assuming monotonically accelerating or decelerating covariate effects on event times. Finally, we extend it to the Weibull delegate racing model that allows monotonic or non-monotonic covariate effects.

2.1 Left-Truncated Weibull Distribution and Weibull Racing

Let $T \sim \text{Weibull}(a, \lambda)$ denote a Weibull random variable T with the density function equal to $f(t \,|\, a, \lambda) = a\lambda t^{a-1} \exp(-\lambda t^a),\ t \in \mathbb{R}_+$, and the survival function (that is $\Pr(T > t)$, the probability that $T > t$) $S(t \,|\, a, \lambda) = \exp(-\lambda t^a),\ t \in \mathbb{R}_+$ with $\mathbb{R}_+$ representing the nonnegative side of the real line. $a > 0$ is the Weibull shape parameter, and $\lambda > 0$ such that the expectation $\mathbb{E}(T) = \lambda^{-1/a}\Gamma(1 + 1/a)$. We introduce the left-truncated Weibull distribution in Definition 1 and show its survival and hazard[2] functions in Corollary 1.

Definition 1 *A left-truncated Weibull distribution, Weibull$_\tau(a, \lambda)$ with $a > 0$ and $\lambda > 0$ and the left truncation threshold $\tau \geq 0$, is defined by the density function $f_\tau(t \,|\, a, \lambda) = f(t \,|\, a, \lambda)/S(\tau \,|\, a, \lambda) = a\lambda t^{a-1} \exp(-\lambda(t^a - \tau^a))$ if $t \geq \tau$ and otherwise 0, where $f(\cdot \,|\, a, \lambda)$ and $S(\cdot \,|\, a, \lambda)$ are the density and survival functions of Weibull(a, λ).*

Corollary 1 *A left-truncated Weibull distribution, Weibull$_\tau(a, \lambda)$, has the survival function $S_\tau(t \,|\, a, \lambda) = \exp(-\lambda(t^a - \tau^a))$ and the hazard function $h_\tau(t \,|\, a, \lambda) = a\lambda t^{a-1}$ for $t \geq \tau$. If $t < \tau$, $S_\tau(t \,|\, a, \lambda) = 1$ and $h_\tau(t \,|\, a, \lambda) = 0$.*

Assuming that a subject's event time follows an untruncated Weibull(a, λ) distribution, $S_\tau(t \,|\, a, \lambda)$ is essentially the conditional probability of surviving over t given that the subject has already survived τ. Consequently, $f_\tau(t \,|\, a, \lambda)$ (or $S_\tau(t \,|\, a, \lambda)$) is the likelihood of the subject with a left truncation time τ and an event (or right censoring) time t. Weibull$_\tau(a, \lambda)$ is reduced to the untruncated Weibull(a, λ) if $\tau = 0$. Property 1 characterizes a race among independent left-truncated Weibull random variables.

Property 1 *If $t_j \sim \text{Weibull}_\tau(a, \lambda_j)$, $j = 1, \ldots, J$, are independent to each other, the minimum $t = \min\{t_1, \ldots, t_J\}$ and the argument of the minimum $y = \text{argmin}_{j \in \{1,\ldots,J\}}\{t_j\}$ are independent and satisfy*

$$t \sim \ \text{Weibull}_\tau(a, \textstyle\sum_{j=1}^J \lambda_j),\ y \sim \text{Categorical}(\lambda_1/\textstyle\sum_{j=1}^J \lambda_j, \cdots, \lambda_J/\textstyle\sum_{j=1}^J \lambda_j). \tag{1}$$

Intuitively, suppose there is a race among teams $j = 1, \cdots, J$, each of whose completion time t_j follows Weibull$_\tau(a, \lambda_j)$, and the winner is the team that has the minimum completion

2. The hazard function is defined as the density function over the survival function.

time. Property 1 implies that the winner's completion time t still follows a left-truncated Weibull distribution and is independent of which team wins. In the context of survival analysis with left truncation, a team j is a competing event with the requirement that the latent survival time t_j exceeds τ, t is the observed time to event (or the observed failure time), and y is the event type (or the cause of failure). We refer to Property 1 as Weibull racing. It not only describes a natural mechanism of surviving under competing events, but also provides an attractive modeling framework amenable to Bayesian inference. Conditioning on a and λ_j's, the joint likelihood of the event type y and the event time t is

$$p(y, t \,|\, a, \{\lambda_j\}_j) = a\lambda_y t^{a-1} \exp(-(t^a - \tau^a) \textstyle\sum_{j=1}^J \lambda_j), \tag{2}$$

which is fully factorized and thus facilitates Bayesian inference by MCMC.

2.2 Linear Weibull Racing Survival Model

We model linear covariate effects on event times in the Weibull racing framework by introducing a gamma-mixed Weibull distribution. Let $\lambda \sim \text{Gamma}(r, 1/b)$ denote a gamma random variable with $\mathbb{E}(\lambda) = r/b$ and $\text{var}(\lambda) = r/b^2$. If $t \sim \text{Weibull}_\tau(a, \lambda)$ and $\lambda \sim \text{Gamma}(r, 1/b)$, we have the marginal density of t given a, r, b, and τ as

$$f_\tau(t \,|\, a, r, b) = \textstyle\int_0^\infty \text{Weibull}_\tau(t; a, \lambda)\text{Gamma}(\lambda; r, 1/b)d\lambda = arb^r t^{a-1}/(b + t^a - \tau^a)^{r+1}.$$

If $a \leq 1$, this gamma-mixed Weibull distribution has a decreasing density $f_\tau(t \,|\, a, r, b)$. If $a > 1$, the shape of $f_\tau(t \,|\, a, r, b)$ depends on the values of a, r, b, and τ. Specifically, the gamma-mixed Weibull distribution has the mode at $[(a-1)(b-\tau^a)/(1+ar)]^{1/a}$ if this value is greater than τ; otherwise $f_\tau(t \,|\, a, r, b)$ is monotonically decreasing. Leveraging Property 1 and the gamma-mixed Weibull distribution, we define the Weibull racing survival model as follows.

Definition 2 *Suppose subject i survives competing events $\{j \,|\, j = 1, \ldots, J\}$ with latent event times $\{t_{ij}\}_j$. Given its left truncation time τ_i and time-invariant covariates $\boldsymbol{x}_i$, under the Weibull racing survival model, the subject's observed event time t_i and event type y_i are generated by*

$$t_i = t_{iy_i},\ y_i = \text{argmin}_{j \in \{1,\ldots,J\}}\, t_{ij},\ t_{ij} \sim \text{Weibull}_{\tau_i}(a, \lambda_{ij}),\ \lambda_{ij} \sim \text{Gamma}(r_j, \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j)).$$

We explain the notations using an Alzheimer's disease example where there are two competing events, the onset of mild cognitive impairment (MCI) ($j = 1$) and death ($j = 2$). Given subject i entering the study at age τ_i that is her left truncation time, we assume her latent age at onset of MCI is t_{i1} and latent age at death is t_{i2}, where t_{ij}, $j = 1, 2$, depends on the linear function $\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j$ of the time-invariant covariates $\boldsymbol{x}_i$ and the parameters a and r_j. If the onset of MCI happens before death ($t_{i1} < t_{i2}$), we observe the onset of MCI ($y_i = 1$) at time $t_i = t_{i1}$. Otherwise, we observe death ($y_i = 2$) at time $t_i = t_{i2}$. Note that t_{ij}'s, $j = 1, \ldots, J$, are independent only if $\{\lambda_{ij}\}_j$ are given. In fact, they are marginally dependent as we only know $\boldsymbol{x}_i$ and infer $\{r_j, \boldsymbol{\beta}_j\}_j$.

Given $\boldsymbol{x}_i$, τ_i, a, r_j, and $\boldsymbol{\beta}_j$ for $j = 1, \ldots, J$, the survival function $S(t)$ and hazard function $h(t)$ at t, $t > \tau_i$, for the observed event time t_i in the Weibull racing survival model

are

$$S(t) = \Pr(t_i > t) = \prod_{j=1}^{J} [\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j)(t^a - \tau_i^a) + 1]^{-r_j},\ h(t) = \sum_{j=1}^{J} \frac{a r_j t^{a-1}}{t^a - \tau_i^a + \exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j)}.$$

We can rewrite the latent event time of competing event j as $t_{ij} \sim \text{Weibull}_{\tau_i}(a, \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j)\tilde{\lambda}_j)$ with $\tilde{\lambda}_j \sim \text{Gamma}(r_j, 1)$. Particularly, when $\tau_i = 0$, $\log t_{ij} = -\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j/a + \log \tilde{t}_j$ with $\tilde{t}_j \sim \text{Weibull}(a, \tilde{\lambda}_j)$. From this perspective, time to each competing event j in Weibull racing is characterized by an accelerated failure time model (Kalbfleisch and Prentice, 2011) in that the covariates $\boldsymbol{x}_i$ accelerate or decelerate the baseline event time $\tilde{t}_j$ by $|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j/a|$. Furthermore, given covariates $\boldsymbol{x}_i$ and τ_i, we can write the survival and hazard functions of t_{ij} for event j as

$$S_j(t) = \Pr(t_{ij} > t) = [\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j)(t^a - \tau_i^a) + 1]^{-r_j},\ h_j(t) = \frac{a r_j t^{a-1}}{t^a - \tau_i^a + \exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j)}. \tag{3}$$

Both S_j and h_j are monotonic in each coordinate of $\boldsymbol{x}_i$.

In the Weibull racing survival model, the latent time to a competing event is linearly accelerated by covariates. However, a predefined competing event may be further categorized into subtypes, and each subtype has a different dependence on $\boldsymbol{x}_i$. Finding the subtypes is important but can be difficult. In medical research where competing diseases are of interest, the nosology of a disease is often subject to human knowledge, diagnostic techniques, and patient population. Multiple diseases of the same phenotype may have been recognized as one disease (competing event), and their distinct etiology and different impacts on patients' survival can be identified only if the difficulties are overcome. For instance, diabetes is categorized into type I and type II; diffuse large B-cell lymphoma has been known to have three subtypes or arguably more, and each subtype is attributed to a different genotype (Rosenwald et al., 2002). In this regard, the Weibull racing survival model is restrictive in that it requires the competing events to be so well defined that their latent times linearly depend on covariates. To circumvent a fine-grained specification of competing events, which is not always feasible, we further develop Weibull delegate racing, assuming that a competing event consists of a potentially infinite number of sub-events, to each of which the latent time is linearly accelerated by the covariates. Decomposing events into sub-events not only improves the model fit but also helps to explore the underlying mechanisms of disease progression.

2.3 Weibull Delegate Racing

Based on the idea of Weibull racing that a subject's observed event time is the minimum of latent times to the predefined competing events, we propose Weibull delegate racing (WDR) survival analysis in the nonparametric Bayesian framework, assuming that the time to a competing event is the minimum of latent times to a number of sub-events appertaining to this competing event. In particular, we denote a gamma process defined on the product space $\mathbb{R}_+ \times \Omega$ by $G_j \sim \Gamma\text{P}(G_{0j}, 1/c_{0j})$, where G_{0j} is a finite and continuous base measure over a complete separable metric space Ω, 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 a countably infinite number of non-negatively weighted atoms, expressed as $G_j = \sum_{k=1}^{\infty} r_{jk}\delta_{\boldsymbol{\beta}_{jk}}$. We define the WDR model as follows.

Definition 3 *Suppose subject i survives competing events $\{j \,|\, j = 1, \ldots, J\}$ with latent event times $\{t_{ij}\}_j$. Given its left truncation time τ_i and time-invariant covariates $\boldsymbol{x}_i$ and a random draw of a gamma process $G_j \sim \Gamma\mathrm{P}(G_{0j}, 1/c_{0j})$, expressed as $G_j = \sum_{k=1}^{\infty} r_{jk}\delta_{\boldsymbol{\beta}_{jk}}$ for $j = 1, \ldots, J$, Weibull delegate racing models subject i's observed event time t_i and event type y_i as*

$$t_i = t_{iy_i},\ y_i = \mathrm{argmin}_{j\in\{1,\ldots,J\}}\, t_{ij},\ t_{ij} = t_{ij\kappa_{ij}},\ \kappa_{ij} = \mathrm{argmin}_{k\in\{1,\ldots,\infty\}}\, t_{ijk},$$
$$t_{ijk} \sim \mathrm{Weibull}_{\tau_i}(a, \lambda_{ijk}),\ \lambda_{ijk} \sim \mathrm{Gamma}(r_{jk}, \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})).$$

We assume in WDR an infinite number of latent sub-events under a prespecified competing event j, and each sub-event k has a latent event time t_{ijk} for subject i. WDR can be considered as a two-phase race among latent sub-events for each subject. In the first phase, within each competing event j there is a race among its countable sub-events $\{k \,|\, k = 1, \ldots, \infty\}$, and the winner, namely sub-event κ_{ij} whose time $t_{ij\kappa_{ij}} = \min_k t_{ijk}$, will represent event j in the second phase of the race by letting t_{ij} equal to $t_{ij\kappa_{ij}}$. In the second phase, J competing events associated with the event times $\{t_{ij}\}_j$ compete with each other, and eventually, the winner's event time and type are observed as $t_i = \min_j t_{ij}$ and $y_i = \mathrm{argmin}_j\, t_{ij}$. Although WDR assumes a potentially infinite number of sub-events within each competing event j, the total weights of these sub-events $G_j = \sum_{k=1}^{\infty} r_{jk}\delta_{\boldsymbol{\beta}_{jk}}$ is finite by the gamma process. Consequently, the weights of negligible sub-events are parsimoniously and data-adaptively shrunk towards zero (Zhou et al., 2016). Therefore, the nonlinearity of WDR is fulfilled by the racing of a finite number of significant sub-events, and each sub-event time is linearly accelerated by the covariates $\boldsymbol{x}_i$.

Intuitively, the nonlinear modeling capacity of WDR is fulfilled by taking the minimum among J minima, which is a two-step nonlinear operation. In mathematics, the event time t_{ij} and its survival function S_j and hazard function h_j for competing event j are no longer monotonic in $\boldsymbol{x}_i$ as shown by Corollary 2.

Corollary 2 *Weibull delegate racing is equivalent to*

$$t_i = t_{y_i},\ y_i = \mathrm{argmin}_j\, t_{ij},\ t_{ij} \sim \mathrm{Weibull}_{\tau_i}(a, \textstyle\sum_{k=1}^{\infty} \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})\tilde{\lambda}_{jk}),\ \tilde{\lambda}_{jk} \sim \mathrm{Gamma}(r_{jk}, 1).$$

The survival function and the hazard function of t_{ij} for event j are

$$S_j(t) = \Pr(t_{ij} > t) = \textstyle\prod_{k=1}^{\infty}[\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})(t^a - \tau_i^a) + 1]^{-r_{jk}},\ h_j(t) = \sum_{k=1}^{\infty} \frac{a r_{jk} t^{a-1}}{t^a - \tau^a + \exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})}.$$

In stark contrast to (3) of the linear Weibull racing survival model, S_j and h_j of WDR for event j are non-monotonic in $\boldsymbol{x}_i$. The countable gamma-mixed Weibull survival time t_{ij} has enhanced flexibility, relaxing the parametric restrictions of conventional Weibull survival models (Commenges et al., 1998; Sparling et al., 2006). One can verify that if $a \leq 1$, h_j is decreasing in t, and otherwise, h_j can be increasing (for a long enough time) then decreasing or arbitrarily non-monotonic in t. More importantly, the flexibility of WDR is data-adaptively parsimonious. By Corollary 2, the truncated Weibull distribution of t_{ij} is parameterized by a weighted sum of an infinite number of covariate-dependent functions $\{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})\}_k$ with gamma-distributed weights $\{\tilde{\lambda}_{jk}\}_k$. With the gamma process regularizing $\{r_{jk}\}_k$, the weight $\tilde{\lambda}_{jk}$ of a negligible sub-event k approaches to zero. Eventually, a relatively small

number of sub-events have considerable weights, and their covariate dependence $\{\boldsymbol{\beta}_{jk}\}_k$ may suggest different mechanisms of disease progression of event j. The marginal density of t_i given a, $\{r_{jk}, \boldsymbol{\beta}_{jk}\}_{j,k}$, and $t_i > \tau_i$ is shown by Theorem 4 in Appendix A.

3. Time-Varying Covariates and Bayesian Inference

We formulate survival analysis with time-varying covariates in Section 3.1 by assuming piecewise Weibull hazards and transforming the problem into dealing with left truncation and right censoring. Section 3.2 provides data augmentation schemes for Weibull racing in the presence of censoring and missing outcome. The schemes also apply to WDR. We interpret WDR as a discrete choice model for classification if all the event times are missing. Section 3.3 shows the hierarchical model of WDR that facilitates an efficient MCMC algorithm with the weights of unnecessary sub-events shrunk towards zero. The inference accommodates various types of censoring and missing outcome imputation by sampling the event times or types as auxiliary variables. In addition, we propose for big data analysis a maximum a posteriori estimation that admits optimization by stochastic gradient descent; the details are provided in Appendix B.

3.1 Weibull Delegate Racing with Time-Varying Covariates

We consider time-varying covariates in WDR. Suppose subject i has V-dimensional, time-varying covariates $\boldsymbol{X}_i(t) = (X_{i1}(t), X_{i2}(t), \ldots, X_{iV}(t))^{\mathrm{T}}$. Without loss of generality, some covariates $X_{iv}(t)$ can be time-invariant such that $X_{iv}(t) = x_v$ for all t. We consider, as is often the case, intermittent covariate measurements. Concretely, subject i enters the study at time $\tau_i^{(0)}$ with covariates $\boldsymbol{x}_i^{(0)}$ and then is observed at times $\tau_i^{(1)}, \ldots, \tau_i^{(L_i)}$ with covariates $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(L_i)}$, respectively, for followup visits before a failure event or censoring. A positive $\tau_i^{(0)}$ represents a left truncation, and $\tau_i^{(0)}, \tau_i^{(1)}, \ldots, \tau_i^{(L_i)}$ and the number of updates L_i are subject-specific. We assume constant covariates between followup visits, that is, $\boldsymbol{X}_i(t) = \boldsymbol{x}_i^{(l)}$ for $t \in [\tau_i^{(l)}, \tau_i^{(l+1)})$, $l = 0, 1, \ldots, L_i - 1$, and $\boldsymbol{X}_i(t) = \boldsymbol{x}_i^{(L_i)}$ for $t \geq \tau_i^{(L_i)}$. We use WDR to predict survival probabilities by the (time-varying) covariates, but not vice versa; how the time-varying covariates are influenced by the survival status or disease progression is out of our scope.

We model subjects surviving competing events $j = 1, \ldots, J$ with time-varying covariates in the framework of WDR by assuming piecewise parametric hazard functions (Sparling et al., 2006). Specifically, for subject i, we assume a Weibull$(a, \sum_{j,k} \lambda_{ijk}^{(l)})$ hazard in the lth interval $[\tau_i^{(l)}, \tau_i^{(l+1)})$ for $l = 0, 1, \ldots, L_i - 1$ and $[\tau_i^{(l)}, \infty)$ for $l = L_i$, namely $h^{(l)}(t) = a \sum_{j,k} \lambda_{ijk}^{(l)} t^{a-1}$ where $\lambda_{ijk}^{(l)} \sim \mathrm{Gamma}(r_{jk}, \exp(\boldsymbol{x}_i^{(l)\mathrm{T}} \boldsymbol{\beta}_{jk}))$. Consequently, the survival function of subject i at time t, $t > \tau_i^{(L_i)}$, after the last covariate measurement is

$$S(t \,|\, \{\tau_i^{(l)}, \lambda_{ijk}^{(l)}\}_{l=0}^{L_i}) = S_{\tau_i^{(L_i)}}(t \,|\, a, \textstyle\sum_{j,k} \lambda_{ijk}^{(L_i)}) \prod_{l=0}^{L_i-1} S_{\tau_i^{(l)}}(\tau_i^{(l+1)} \,|\, a, \textstyle\sum_{j,k} \lambda_{ijk}^{(l)}) \tag{4}$$

and the density function is

$$f(t \,|\, \{\tau_i^{(l)}, \lambda_{ijk}^{(l)}\}_{l=0}^{L_i}) = f_{\tau_i^{(L_i)}}(t \,|\, a, \textstyle\sum_{j,k} \lambda_{ijk}^{(L_i)}) \prod_{l=0}^{L_i-1} S_{\tau_i^{(l)}}(\tau_i^{(l+1)} \,|\, a, \textstyle\sum_{j,k} \lambda_{ijk}^{(l)}) \tag{5}$$

where $S_{\tau_i^{(l)}}$ and $f_{\tau_i^{(l)}}$ are the survival and density functions, respectively, of a left-truncated Weibull distribution, as shown in Definition 1 and Corollary 1.

Equations (4) and (5) solve the problem of time-varying covariates by dealing with left truncation and right censoring. By Equation (5), the likelihood of subject i having an event at time t with covariate values $\boldsymbol{x}_i^{(0)}, \ldots, \boldsymbol{x}_i^{(L_i)}$ measured at $\tau_i^{(0)}, \ldots, \tau_i^{(L_i)}$ is equivalent to the joint likelihood of $L_i + 1$ pseudo subjects: One pseudo subject has the left truncation time $\tau_i^{(L_i)}$ and the event time t with fixed covariates $\boldsymbol{x}_i^{(L_i)}$ and the other L_i have the left truncation time $\tau_i^{(l)}$ and the right censoring time $\tau_i^{(l+1)}$ with fixed covariates $\boldsymbol{x}_i^{(l)}$ for $l = 0, \ldots, L_i - 1$. Analogously, Equation (4) implies that a right censoring of the subject with time-varying covariates is equivalent to L_i+1 pseudo subjects with a left truncation and a right censoring. The derivation of (4) and (5) is deferred to Appendix A. Hereby, WDR handles time-varying covariates by equivalently modeling left truncation and right censoring of pseudo subjects with time-invariant covariates. Since left truncation is intrinsically accommodated by WDR, we focus on censoring and Bayesian inference in the presence of time-invariant covariates in the remainder of this section.

3.2 Censoring, Missing Outcomes, and WDR Classification

We denote by Ψ a subject-specific censoring condition on a left-truncated event time $t_i \sim \text{Weibull}_{\tau_i}(a, \lambda)$. Specifically, Ψ can be $(T_{r.c.}, \infty)$ indicating a right censoring at $T_{r.c.} > \tau_i$, $(\tau_i, T_{l.c.})$ a left censoring at $T_{l.c.} > \tau_i$, or (T_1, T_2) an interval censoring with $T_2 > T_1 > \tau_i$. The density of t_i with the constraint $t_i \in \Psi$ is $f_{\tau_i,\Psi}(t \,|\, a, \lambda) = f_{\tau_i}(t \,|\, a, \lambda) / \int_\Psi f_{\tau_i}(s \,|\, a, \lambda) ds$.

If y_i or t_i is missing or there exists censoring, we have two situations where auxiliary variables can be introduced to achieve a factorized likelihood of the Weibull racing model: (a) If we only observe y_i (or t_i) without censoring, then we can draw t_i (or y_i) by (1) as an auxiliary variable, leading to the factorized likelihood of t_i and y_i as in (2). (b) If we do not observe t_i but know $t_i \in \Psi$ with probability $\Pr(t_i \in \Psi \,|\, a, \{\lambda_{ij}\}_j) = \int_\Psi f_{\tau_i}(s \,|\, a, \sum_j \lambda_{ij}) ds$, then we draw t_i by the density $f_{\tau_i,\Psi}(t \,|\, a, \sum_j \lambda_{ij})$, resulting in the likelihood

$$p(t_i, t_i \in \Psi \,|\, a, \textstyle\sum_j \lambda_{ij}, \tau_i) = f_{\tau_i,\Psi}(t_i \,|\, a, \sum_j \lambda_{ij}) \Pr(t_i \in \Psi \,|\, a, \sum_j \lambda_{ij}) = f_{\tau_i}(t \,|\, a, \textstyle\sum_j \lambda_{ij}).$$

With y_i that can be drawn by (1) if missing, the likelihood $p(y_i, t_i, t_i \in \Psi \,|\, a, \{\lambda_{ij}\}_j, \tau_i)$ becomes the same as the right hand side of (2). Therefore, under different censoring conditions or with missing event times or types, sampling t_i and/or y_i gives Weibull racing the same factorized likelihood as in (2).

If all the event times are unobserved, that is, if y_i is the only dependent variable, WDR survival analysis will be reduced to a classification model. Specifically, with latent times t_{ijk}'s and $t_{ij} = \min_k t_{ijk}$, the category $y_i = \text{argmin}_j\, t_{ij}$. This provides an alternative view of Weibull (delegate) racing from the perspective of discrete choice models (Hanemann, 1984; Greene, 2003; Train, 2009; Zhang and Zhou, 2017). Specifically, the observed event type (or category) y_i is equal to the one whose latent arrival time is earlier than all the others. Distinct from ordinary discrete choice models where y_i corresponds to the category that brings the maximum latent utility and the utility values are not identifiable or of interest, in Weibull (delegate) racing, the event type y_i is determined to minimize the waiting time for the first arrival, and the minimum waiting time t_i can be either observed and studied in the survival

model, or missing but imputed as an auxiliary variable in both the survival and classification models. This finding unifies Bayesian inference of the WDR classification and survival analysis with missing event times. See Appendix E for more details on WDR classification.

3.3 Hierarchical Model of WDR and MCMC

For the implementation of WDR, we follow Zhou and Carin (2015) to truncate the number of atoms of the gamma processes at K, which is a sufficiently large integer, by choosing a finite and discrete base measure $G_{0j} = \sum_{k=1}^{K} \gamma_{0j}\delta_{\boldsymbol{\beta}_{jk}}/K$, $j = 1, \ldots, J$. In this way, we allow up to K latent sub-events within each competing event j. To obtain a factorized likelihood of WDR, we first follow Section 3.2 to sample y_i and t_i if missing or censored. Next, WDR requires another auxiliary variable $\kappa_{iy_i} \sim \text{Categorical}(\lambda_{iy_i1}/\sum_{k=1}^{K}\lambda_{iy_ik}, \ldots, \lambda_{iy_iK}/\sum_{k=1}^{K}\lambda_{iy_ik})$, which is the label of the winning sub-event within competing event y_i in the first phase of the race. Consequently, the factorized likelihood of subject i becomes

$$p(t_i, y_i, \kappa_{iy_i} \,|\, a, \{\lambda_{ijk}\}_{jk}, \tau_i) = \lambda_{iy_i\kappa_{iy_i}} a t_i^{a-1} \exp(-(t_i^a - \tau_i^a)\textstyle\sum_{j,k}\lambda_{ijk}).$$

We write the hierarchical model of WDR as

$$\begin{aligned}
&t_i = t_{iy_i},\ y_i = \text{argmin}_{j\in\{1,\ldots,J\}}\, t_{ij},\ t_{ij} = t_{ij\kappa_{ij}},\ \kappa_{ij} = \text{argmin}_{k\in\{1,\ldots,K\}}\, t_{ijk},\\
&t_{ijk} \sim \text{Weibull}_{\tau}(a, \lambda_{ijk}),\ \lambda_{ijk} \sim \text{Gamma}(r_{jk}, \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})),\\
&\boldsymbol{\beta}_{jk} \sim \textstyle\prod_{v=1}^{V} \text{N}(0, \alpha_{vjk}^{-1}),\ \alpha_{vjk} \sim \text{Gamma}(a_0, 1/b_0),\ r_{jk} \sim \text{Gamma}(\gamma_{0j}/K, 1/c_{0j}),
\end{aligned}$$

where $i = 1, \cdots, n$, $j = 1, \cdots, J$, and $k = 1, \cdots, K$. We choose non-informative hyperpriors $\gamma_{0j} \sim \text{Gamma}(d_0, 1/e_0)$ and $c_{0j} \sim \text{Gamma}(d_1, 1/e_1)$, with $d_0 = e_0 = d_1 = e_1 = 0.01$.

With the factorized likelihood and additional auxiliary variables from the Pólya gamma (Polson et al., 2013) and Chinese restaurant table (Zhou and Carin, 2015) distributions, our MCMC algorithm updates all the parameters by Gibbs sampling except the Weibull shape parameter a. A possible solution is to update a by Metropolis-Hastings, but the proposal distribution has to be tuned. Considering the fact that the full conditional distribution of a is unimodal when left truncation times are 0 (see Step 4 of the MCMC algorithm in Appendix B), we alternatively use slice sampling (Damien et al., 1999; Neal et al., 2003) that is more efficient and less sensitive to tuning parameters. To remove unnecessary modeling capacity, the gamma processes regulate the weights of the sub-events by pushing some of r_{jk}'s towards zero. In addition, we propose a scheme (Step 9 of the MCMC algorithm) based on latent count allocations to actively prune negligible latent sub-events during MCMC iterations and thus accelerate the convergence of the algorithm. Appendix B shows the complete MCMC algorithm for WDR survival analysis including censoring and missing outcome imputation.

WDR assumes that competing events are conditionally independent given covariates. To relax this assumption, one can incorporate random effects in WDR, such as a random intercept concatenated to $\boldsymbol{\beta}$ for each subject. In this way, flexible dependence among competing events and/or subjects is allowed, even if the covariates are given. Notably, incorporating random effects in WDR does not undermine the Gaussian conjugacy of $\boldsymbol{\beta}_{jk}$'s using our data augmentation scheme and MCMC if Gaussian random effects are used. Moreover, the Gaussian conjugacy admits easy and various regularizations on $\boldsymbol{\beta}_{jk}$'s by

Data 1	Data 2	Data 3
$t_{i1} \sim \text{Weibull}(0.8, \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1))$	$t_{i1} \sim \text{Weibull}(3, \lvert\sinh(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1)\rvert)$	$t_{i1} \sim \text{Weibull}(1, \exp((\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1)^2))$
$t_{i2} \sim \text{Weibull}(0.8, \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2))$	$t_{i2} \sim \text{Weibull}(3, \cosh(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2))$	$t_{i2} \sim \text{Weibull}(1, \exp((\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2)^2))$
$t_i = \min(t_{i1}, t_{i2}, T_{r.c.} = 2)$	$t_i = \min(t_{i1}, t_{i2}, T_{r.c.} = 1.2)$	$t_i = \min(t_{i1}, t_{i2}, T_{r.c.} = 1.2)$
Left truncation time: 0.05	Left truncation time: 0.3	Left truncation time: 0.05
Covariates update time: $0.15, 0.25$	Covariates update time: $0.5, 0.7$	Covariates update time: $0.1, 0.15$
Evaluation time: $.1, .3, .5, .7, .9$	Evaluation time: $.4, .55, .7, .85, 1$	Evaluation time: $0.1, 0.3, 0.5, 0.7, 0.9$
Data 4	**Data 5**	**Data 6**
$t_{i1} \sim \text{logNormal}(\mu_{i1}, 0.2^2)$	$t_{i1} \sim \text{logNormal}(\mu_{i1}, 0.2^2)$	$t_{i1} \sim \text{logNormal}(\mu_{i1}, 0.05^2)$
$\mu_{i1} = \frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_3 \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1) - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_4 \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2)}{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1) + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2)}$	$\text{logit}(\mu_{i1}) = \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_3 - \boldsymbol{b}_2^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_4$	$\text{logit}(\mu_{i1}) = \cosh(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1) - \cosh(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2)$
$t_{i2} \sim \text{logNormal}(\mu_{i2}, 0.2^2)$	$t_{i2} \sim \text{logNormal}(\mu_{i2}, 0.2^2)$	$t_{i2} \sim \text{logNormal}(\mu_{i2}, 0.05^2)$
$\mu_{i2} = \frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_4 \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2) - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_3 \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1)}{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1) + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2)}$	$\text{logit}(\mu_{i2}) = \boldsymbol{b}_2^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_4 - \boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_3$	$\text{logit}(\mu_{i2}) = \cosh(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_2) - \cosh(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}_1)$
$t_i = \min(t_{i1}, t_{i2}, T_{r.c.} = 1)$	$t_i = \min(t_{i1}, t_{i2}, T_{r.c.} = 1.4)$	$t_i = \min(t_{i1}, t_{i2}, T_{r.c.} = 1.6)$
Left truncation time: 0.2	Left truncation time: 0.8	Left truncation time: 0.8
Covariates update time: $0.3, 0.4$	Covariates update time: $0.9, 1$	Covariates update time: $0.9, 1$
Evaluation time: $.3, .45, .6, .75, .9$	Evaluation time: $0.9, 1, 1.1, 1.2, 1.3$	Evaluation time: $1, 1.1, 1.2, 1.3, 1.4$

Table 1: Data synthesis.

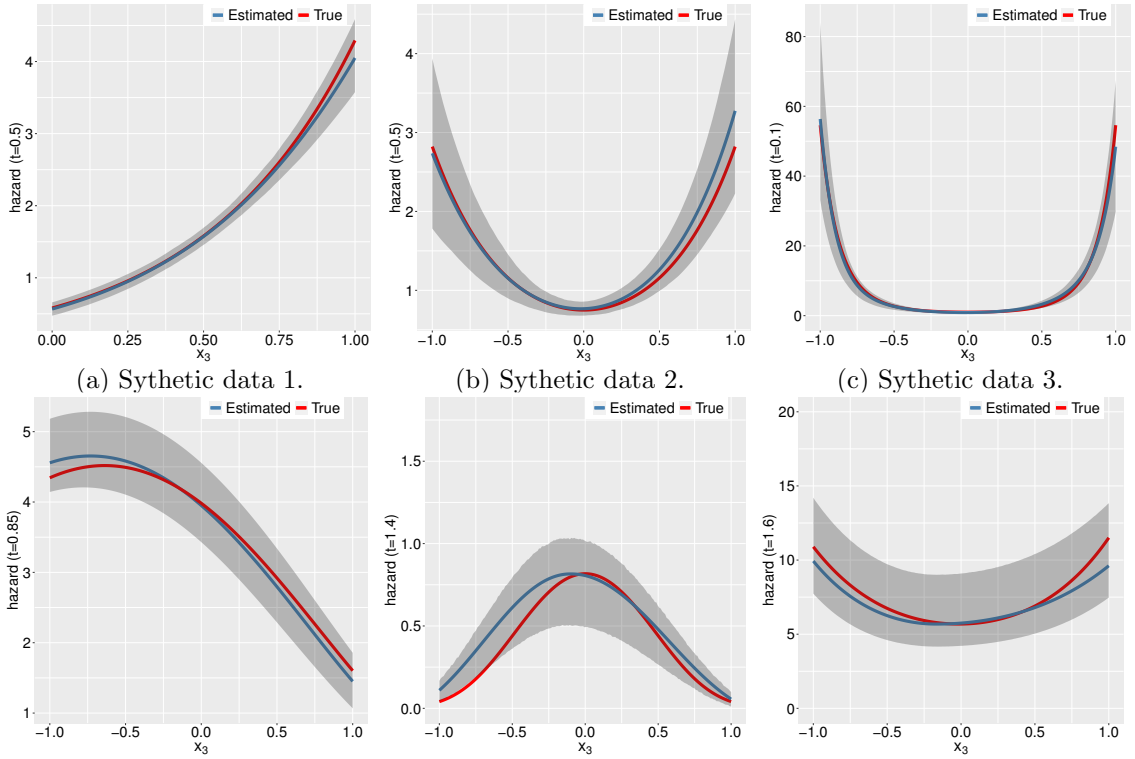
imposing a prior distribution. We use Gaussian-inverse-gamma priors, but other priors, such as the Laplace prior (Park and Casella, 2008) and the horseshoe prior (Carvalho et al., 2010; Johndrow et al., 2020) also apply. We defer mixed-effects WDR and other priors on $\boldsymbol{\beta}_{jk}$'s to future work.

4. Synthetic Data Analysis and Model Comparison

We validate WDR survival analysis on synthetic data by showing its parsimonious nonlinearity and comparing its prediction accuracy with benchmark models. In Section 4.1, we introduce the synthetic data and the quantification of prediction accuracy. In Section 4.2, we illustrate the estimation of nonlinear covariate effects by WDR. In Section 4.3, we introduce the benchmark models and compare the models on the data with competing events, left truncation, and time-varying covariates. Technical details and supplementary results are deferred to Appendix D. We show that WDR is an attractive approach for its interpretability, versatility, and prediction accuracy.

4.1 Data and Model Evaluation

We simulate six data sets with $J = 2$ competing events for each. The data generating process is provided in Table 1, where each subject i has covariates $\boldsymbol{x}_i$ from a Gaussian or uniform distribution. Times to competing events follow Weibull distributions in data 1 to 3 and log-normal distributions in data 4 to 6. For data 1, the covariates have linear effects on survival times. For data 2 to 6, we use nonlinear functions of $\boldsymbol{x}_i$ as the distribution parameters. Specifically, we use the hyperbolic sine and cosine functions in data 2 and 6, quadratic functions in data 3 and 5, and weighted linear functions with the weights being covariate dependent in data 4. Subject i is right censored at time $T_{r.c.}$ if the latent event times t_{i1} and t_{i2} are greater than $T_{r.c.}$. The observed event type $y_i = \text{argmin}_j\, t_{ij}$ if $t_i < T_{r.c.}$, and $y_i = 0$ is used to denote right censoring if $t_i = T_{r.c.}$. In Section 4.2, we simulate time-



(a) Sythetic data 1. (b) Sythetic data 2. (c) Sythetic data 3.

(d) Sythetic data 4. (e) Sythetic data 5. (f) Sythetic data 6.

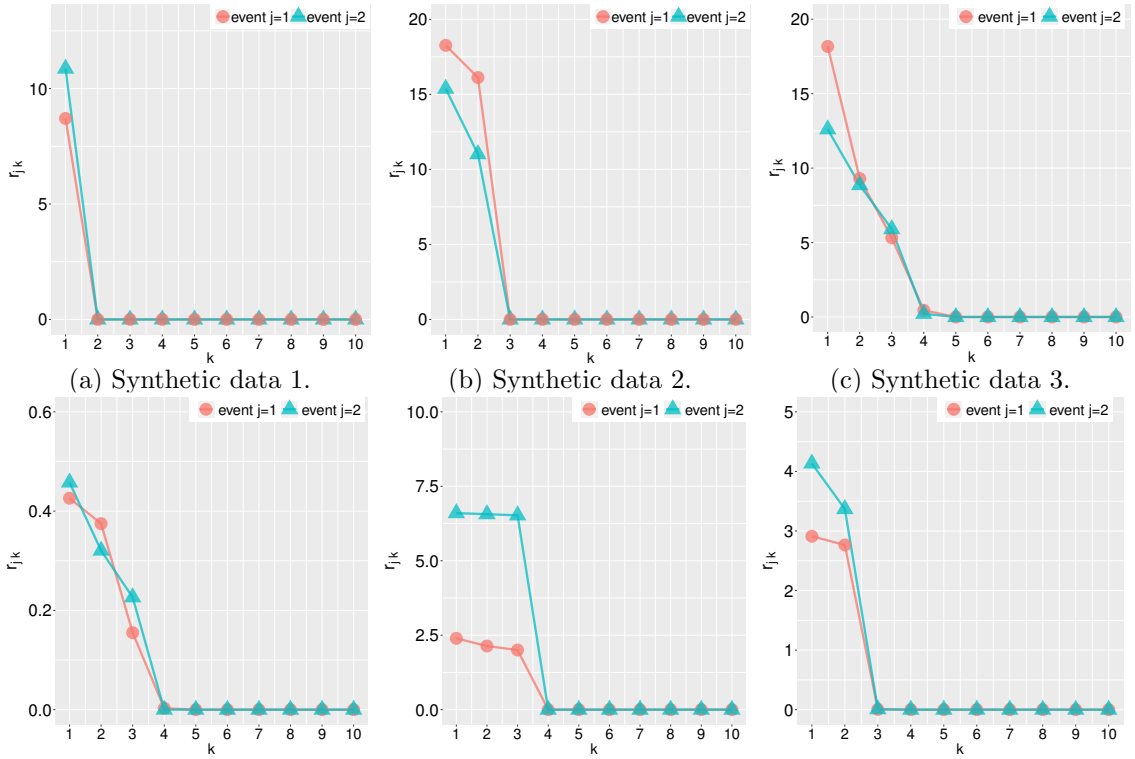
Figure 1: WDR estimations (with 95% credible intervals) of hazards of event 2 against x_3.

invariant covariates for each subject and let the left truncation equal to 0. In Section 4.3, we consider left truncation and time-varying covariates, and the covariate update times and left truncation times are given in Table 1. The distribution of $\boldsymbol{x}_i$, the values of $\boldsymbol{b}$'s, and the simulation of time-varying covariates are deferred to Tables 8 and 9 and Algorithm 1, respectively, in Appendix C.

We use the Brier score (Gerds et al., 2008; Steyerberg et al., 2010) to quantify the prediction accuracy. Specifically, with subjects $i = 1, \ldots, n$, the Brier score (BS) for event j at time t is $\mathrm{BS}_j(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}(t_i \le t, y_i = j) - \Pr(t_i \le t, y_i = j)\right]^2$. The Brier score represents the mean squared distances between the observed survival status and the estimated cumulative incidence function (CIF) that is equal to $\Pr(t_i \le t, y_i = j)$. Brier scores are between 0 and 1, and a smaller value indicates a more accurate prediction. In Section 4.3, we evaluate Brier scores at five time points whose range covers roughly the middle 80% of the survival times in each data. The five evaluation time points are given in Table 1. In all the experiments, we set $K = 10$ in WDR to allow up to 10 latent sub-events for each competing event, run 20,000 MCMC iterations, and collect the last 2,000 for posterior estimations.

4.2 Parsimonious Nonlinearity of WDR

We first illustrate the performance of WDR in estimating nonlinear covariate effects. Specifically, we simulate time-invariant covariates $\boldsymbol{x}_i = (x_{i1}, x_{i1}, x_{i3})^{\mathrm{T}}$ from a uniform distribution (see Table 8 in the Appendix) and follow the data generating process in Table 1 to syn-



(a) Synthetic data 1. (b) Synthetic data 2. (c) Synthetic data 3.

(d) Synthetic data 4. (e) Synthetic data 5. (f) Synthetic data 6.

Figure 2: Sub-event weights $\{r_{jk} \,|\, k = 1, \ldots, 10\}$ in descending order for $j = 1$ and 2.

thesize the six data sets without left truncation. We predict the hazards of event 2 for subjects with $x_1 = 0$, $x_2 = 0.5$ and $x_3 \in [0, 1]$ in data 1 at $t = 0.5$, with $x_1 = 0$, $x_2 = 0$ and $x_3 \in [-1, 1]$ in data 2, 4, 5, and 6 at $t = 0.5$, 0.85, 1.4, and 1.6, respectively, and with $x_1 = 0.5$, $x_2 = 0.5$ and $x_3 \in [-1, 1]$ in data 3 at $t = 0.1$. The true and estimated hazards along with the 95% credible intervals are shown in Figure 1. We see WDR successfully recovers the log-linear (data 1) and non-monotonic (data 2 to 6) covariate effects on the hazards.

Plotted in Figure 2 are r_{jk}'s that reflect the data-adaptive, parsimonious nonlinearity of WDR. For data 1, only one sub-event is discovered for competing events 1 and 2, respectively, suggesting the linear covariate effects in data 1. Racing between two sub-events within each competing event can approximate the generating process of data 2 and 6. Furthermore, WDR uses racing among three latent sub-events to model the quadratic and interacting covariate effects in data 3, 4, and 5. The other prespecified sub-events are redundant as their weights r_{jk}'s are very close to zero.

4.3 Model Comparison

Considering the lack of nonlinear benchmarks accommodating competing events, left truncation, and time-varying covariates, we develop a kernel-based Fine-Gray (KFG) model that is similar to the kernel Cox models for the analysis of a single event (Li and Luan, 2002; Evers and Messow, 2008). Specifically, KFG replaces the covariate matrix of the Fine-Gray

	WDR	FG	KFG	RF	DeepHit	PCH
Nonlinear	✓	✗	✓	✓	✓	✓
Easy to interpret	✓	✓	✗	✗	✗	✗
Continuous time	✓	✓	✓	✓	✗	✓
Competing events	✓	✓	✓	✓	✓	✗
Left truncation	✓	✓	✓	✗	✗	✗
Time-varying covariates	✓	✓	✓	✗	✗	✗
Heuristics	-	-	-	Pseudo subjects	Pseudo subjects	Censoring and pseudo subjects

Table 2: Model capability.

model with a radial basis function kernel matrix and uses the gradient boosting method of Binder et al. (2009) for sparsity. Note that KFG is a combination of existing approaches and is used for the purpose of benchmarking; we do not claim a contribution to this model. We also compare with four other models: the Fine-Gray model (FG) (Fine and Gray, 1999) that is a linear model, random survival forests (RF) (Ishwaran et al., 2014), DeepHit (Lee et al., 2018), and the piecewise constant hazards (PCH) method (Kvamme and Borgan, 2021).

We compare the model capability in Table 2. WDR, FG, and KFG are able to deal with competing events, left truncation, and time-varying covariates. RF and DeepHit accommodate competing events but are incapable of left truncation or time-varying covariates. DeepHit transforms survival analysis into a classification problem by time discretization and models the class probabilities by a neural network. PCH is proposed for single-event survival analysis without left truncation or time-varying covariates, assumes piecewise constant hazards that are modeled by a neural network, and uses interpolation for continuous-time prediction. We use the censoring trick to adapt PCH to competing-event analysis. Concretely, we analyze one competing event at a time and treat subjects having other events as right censored. To adapt RF, DeepHit, and PCH to the scenario with time-varying covariates, we use the left-truncated and right-censored pseudo subjects as described in Section 3.1. Accommodating left truncation using RF, DeepHit, or PCH has not been investigated and is out of the scope of this paper. So for these three models, we pretend any left truncation is at time 0 and will show that overlooking left truncation results in poor predictions. For DeepHit and PCH, we discretize the continuous survival time into 20 intervals of an equal length, in each of which the survival or hazard function is constant and modeled by a neural network. Detailed experiment settings are provided in Appendix C. Overall, WDR is versatile compared to RF, DeepHit, and PCH, and its nonlinearity and interpretability can be more appealing than the FG or KFG model.

4.3.1 Time-Invariant Covariates and No Left Truncation

We first provide the model comparison on the synthetic data without left truncation or time-varying covariates. Specifically, we simulate covariates $\boldsymbol{x}_i \in \mathbb{R}^{10}$ from a Gaussian or uniform distribution (see Table 9 in the Appendix) and follow Table 1 to generate the six data sets but without left truncation or covariate updates. In this scenario, all the models

		Data 1	Data 2	Data 3	Data 4	Data 5	Data 6
Event 1	WDR	**0.190**	0.127	0.170	0.073	**0.132**	0.093
	FG	0.192	0.150	0.209	0.117	0.202	0.193
	KFG	0.192	0.124	**0.166**	**0.066**	0.160	0.130
	RF	0.200	**0.122**	0.168	0.086	0.147	0.130
	DeepHit	0.227	0.143	0.247	0.089	0.154	**0.081**
	PCH	0.205	0.148	0.192	0.076	0.147	0.085
Event 2	WDR	**0.184**	0.181	**0.168**	0.069	**0.131**	0.101
	FG	0.185	0.193	0.213	0.124	0.202	0.208
	KFG	0.186	0.179	**0.168**	0.070	0.162	0.136
	RF	0.193	**0.176**	0.173	0.088	0.152	0.140
	DeepHit	0.218	0.195	0.253	0.082	0.151	0.088
	PCH	0.200	0.193	0.180	**0.066**	0.140	**0.072**

Table 3: Brior scores for synthetic data with *constant* covariates and *no* left truncation.

compared (the censoring heuristics is used for PCH) are able to handle the survival analysis. For each data set, we simulate 2000 subjects and take 20 random partitions into a training set of 1800 and a testing set of 200. We evaluate the Brier scores at the five time points as in Table 1 and report in Table 3 the average score over the partitions and the time evaluated. The Brier scores at each specific time are deferred to Tables 10 and 11 in the Appendix.

On data 1 where the covariates have linear effects, WDR, FG, and KFG have similar performances in survival prediction and are slightly better than RF, DeepHit, and PCH. On data 2 to 6 with nonlinear covariate effects, FG does not work well, and WDR is among the best-performing models. Note that DeepHit has a larger variation in prediction accuracy across data or time points than the other models, likely because it models survival probabilities in discrete time and the performance is sensitive to time discretization (Kvamme and Borgan, 2021). This implies the importance of continuous-time modeling in survival analysis.

4.3.2 Time-Varying Covariates and Left Truncation

We compare the model performance on synthetic data with left truncation and time-varying covariates as described in Table 1. In each data set, we simulate covariates $\boldsymbol{x}_i \in \mathbb{R}^{10}$ from a Gaussian or uniform distribution (see Table 9 in the Appendix) for each subject and allow up to two covariate updates after the first measurement. In this case, WDR, FG, and KFG can handle the survival analysis with competing events, left truncation, and time-varying covariates. We use PCH with the censoring heuristics for competing events and RF, DeepHit, and PCH with the pseudo subjects heuristics for time-varying covariates. For each data set, we simulate 1000 subjects and take 20 random partitions into a training set of 900 and a testing set of 100. We report in Tables 4 and 5 the Brier scores (mean $\pm$ standard error) for events 1 and 2, respectively, by the six models.

On data 1 where the covariate effects are linear, WDR, FG, and KFG have comparable performance. But on all the other data, WDR consistently outperforms the other models, suggesting its excellent prediction accuracy. Notably, RF, DeepHit, and PCH deliver com-

Data 1	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.069**±0.006	**0.176**±0.005	**0.202**±0.004	0.217±0.003	0.223±0.003
FG	0.071±0.006	0.181±0.006	0.203±0.005	**0.215**±0.004	**0.219**±0.003
KFG	0.071±0.006	0.182±0.006	0.204±0.005	0.216±0.004	0.221±0.003
RF	0.071±0.006	0.188±0.006	0.216±0.006	0.231±0.004	0.237±0.003
DeepHit	0.075±0.007	0.184±0.006	0.221±0.007	0.231±0.004	0.236±0.003
PCH	0.072±0.006	0.204±0.008	0.226±0.008	0.237±0.005	0.250±0.006
Data 2	$t = 0.4$	$t = 0.55$	$t = 0.7$	$t = 0.85$	$t = 1$
WDR	**0.030**±0.004	**0.086**±0.004	**0.131**±0.005	**0.166**±0.005	**0.182**±0.004
FG	0.032±0.004	0.090±0.005	0.153±0.006	0.189±0.005	0.210±0.004
KFG	0.032±0.004	0.091±0.005	0.143±0.006	0.177±0.005	0.197±0.004
RF	0.032±0.004	0.089±0.005	0.149±0.006	0.184±0.005	0.203±0.004
DeepHit	0.032±0.004	0.089±0.005	0.154±0.007	0.186±0.005	0.205±0.004
PCH	0.032±0.004	0.091±0.005	0.157±0.006	0.199±0.007	0.232±0.008
Data 3	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.079**±0.004	**0.190**±0.003	**0.202**±0.003	**0.207**±0.003	**0.210**±0.003
FG	0.091±0.005	0.219±0.005	0.236±0.004	0.242±0.003	0.245±0.002
KFG	0.090±0.005	0.193±0.004	0.207±0.003	0.213±0.003	0.216±0.002
RF	0.089±0.005	0.21±0.005	0.227±0.004	0.233±0.003	0.236±0.003
DeepHit	0.096±0.005	0.312±0.01	0.373±0.012	0.398±0.012	0.416±0.013
PCH	0.096±0.005	0.286±0.015	0.333±0.02	0.351±0.023	0.21±0.005
Data 4	$t = 0.3$	$t = 0.45$	$t = 0.6$	$t = 0.75$	$t = 0.9$
WDR	**0.028**±0.003	**0.076**±0.003	**0.091**±0.003	**0.100**±0.003	**0.094**±0.003
FG	0.039±0.005	0.119±0.005	0.153±0.004	0.166±0.003	0.171±0.002
KFG	0.039±0.006	0.082±0.006	0.098±0.004	0.101±0.003	0.105±0.003
RF	0.039±0.006	0.114±0.006	0.145±0.005	0.160±0.004	0.163±0.002
DeepHit	0.036±0.005	0.129±0.007	0.118±0.005	0.128±0.005	0.126±0.004
PCH	0.032±0.005	0.086±0.003	0.091±0.004	0.108±0.006	0.138±0.009
Data 5	$t = 0.9$	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$
WDR	**0.087**±0.004	**0.167**±0.005	0.177±0.004	**0.161**±0.003	**0.153**±0.003
FG	0.090±0.005	0.200±0.005	0.240±0.004	0.247±0.003	0.250±0.002
KFG	**0.087**±0.006	0.168±0.006	0.187±0.003	0.188±0.002	0.188±0.001
RF	**0.087**±0.006	0.176±0.006	0.200±0.004	0.199±0.003	0.199±0.002
DeepHit	0.093±0.007	0.218±0.009	0.210±0.009	0.180±0.013	0.171±0.014
PCH	**0.087**±0.006	0.185±0.007	**0.170**±0.005	0.175±0.008	0.198±0.01
Data 6	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$	$t = 1.4$
WDR	**0.101**±0.005	**0.150**±0.005	**0.130**±0.004	**0.118**±0.003	**0.120**±0.002
FG	0.117±0.005	0.227±0.005	0.241±0.004	0.244±0.003	0.246±0.002
KFG	0.102±0.006	0.174±0.003	0.177±0.002	0.177±0.002	0.178±0.002
RF	0.109±0.006	0.182±0.004	0.187±0.003	0.191±0.003	0.195±0.002
DeepHit	0.116±0.007	0.215±0.005	0.140±0.003	0.138±0.003	0.139±0.003
PCH	0.108±0.004	0.194±0.005	0.186±0.005	0.132±0.008	0.185±0.011

Table 4: Brier scores for event 1 of the synthetic data (mean ± stander error).

Data 1	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.060**±0.004	**0.172**±0.005	0.201±0.004	0.214±0.004	0.222±0.004
FG	0.062±0.004	0.175±0.006	**0.198**±0.004	**0.210**±0.004	**0.218**±0.004
KFG	0.062±0.004	0.178±0.007	0.201±0.005	0.211±0.004	0.219±0.004
RF	0.062±0.004	0.179±0.006	0.209±0.004	0.221±0.003	0.230±0.003
DeepHit	0.066±0.005	0.180±0.006	0.218±0.005	0.224±0.003	0.232±0.002
PCH	0.063±0.004	0.201±0.008	0.236±0.007	0.256±0.007	0.266±0.008
Data 2	$t = 0.4$	$t = 0.55$	$t = 0.7$	$t = 0.85$	$t = 1$
WDR	**0.042**±0.004	**0.137**±0.005	**0.213**±0.005	**0.238**±0.002	**0.227**±0.002
FG	0.043±0.004	0.144±0.007	0.241±0.007	0.256±0.002	0.247±0.001
KFG	0.043±0.004	0.144±0.007	0.230±0.007	0.243±0.002	0.234±0.001
RF	**0.042**±0.004	0.143±0.007	0.238±0.007	0.255±0.003	0.248±0.001
DeepHit	0.043±0.004	0.144±0.007	0.251±0.007	0.265±0.003	0.253±0.002
PCH	0.043±0.004	0.147±0.007	0.256±0.008	0.276±0.006	0.255±0.002
Data 3	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.074**±0.004	**0.207**±0.002	**0.218**±0.002	**0.218**±0.002	**0.219**±0.003
FG	0.076±0.005	0.237±0.004	0.250±0.002	0.252±0.002	0.253±0.001
KFG	0.076±0.005	0.209±0.004	0.220±0.002	0.222±0.002	0.223±0.001
RF	0.076±0.005	0.230±0.004	0.243±0.002	0.245±0.002	0.246±0.001
DeepHit	0.080±0.005	0.376±0.011	0.445±0.011	0.471±0.011	0.484±0.012
PCH	0.080±0.005	0.340±0.016	0.391±0.021	0.405±0.023	0.228±0.007
Data 4	$t = 0.3$	$t = 0.45$	$t = 0.6$	$t = 0.75$	$t = 0.9$
WDR	**0.020**±0.002	**0.061**±0.004	**0.091**±0.003	**0.088**±0.003	**0.085**±0.003
FG	0.031±0.005	0.114±0.004	0.154±0.002	0.167±0.002	0.166±0.001
KFG	0.029±0.003	0.072±0.005	0.096±0.003	0.099±0.003	0.102±0.003
RF	0.030±0.004	0.105±0.006	0.144±0.004	0.160±0.003	0.161±0.003
DeepHit	0.025±0.003	0.113±0.007	0.124±0.005	0.118±0.005	0.118±0.006
PCH	0.022±0.002	0.079±0.005	0.102±0.003	0.098±0.004	0.121±0.007
Data 5	$t = 0.9$	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$
WDR	**0.091**±0.005	0.156±0.005	**0.172**±0.005	**0.157**±0.003	**0.151**±0.003
FG	0.098±0.005	0.191±0.005	0.232±0.002	0.245±0.002	0.250±0.001
KFG	0.096±0.006	0.163±0.006	0.184±0.004	0.188±0.002	0.190±0.001
RF	0.096±0.006	0.174±0.006	0.195±0.004	0.198±0.002	0.200±0.002
DeepHit	0.102±0.007	0.203±0.008	0.217±0.007	0.161±0.013	0.163±0.013
PCH	0.094±0.007	**0.153**±0.006	0.177±0.006	0.158±0.004	0.163±0.008
Data 6	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$	$t = 1.4$
WDR	0.109±0.005	**0.160**±0.004	**0.131**±0.003	**0.125**±0.003	**0.125**±0.003
FG	0.120±0.005	0.225±0.004	0.234±0.002	0.239±0.002	0.245±0.001
KFG	**0.104**±0.007	0.172±0.004	0.173±0.003	0.175±0.002	0.177±0.002
RF	0.112±0.006	0.183±0.004	0.184±0.003	0.189±0.003	0.195±0.002
DeepHit	0.122±0.008	0.213±0.005	0.136±0.002	0.139±0.003	0.142±0.003
PCH	0.111±0.005	0.182±0.003	0.170±0.004	0.147±0.004	0.169±0.006

Table 5: Brier scores for event 2 of the synthetic data (mean ± stander error).

parable performance to WDR if the covariates are constant and there are no left truncations (see Table 3), but significantly underperform WDR in the presence of left truncation and time-varying covariates. This is likely because of the bias from overlooking left truncation. In addition, using the pseudo-subjects approach to handle time-varying covariates is underpinned by Equations (4) and (5) for WDR, but it serves as a heuristic method for RF, DeepHit, and PCH, and to our knowledge, no theoretical guarantees have been provided. This may be another reason for their underperformance. Overall, the experiments and the model comparison justify the advantage of WDR and its adaptability to various applications.

5. Biomedical Data Analysis

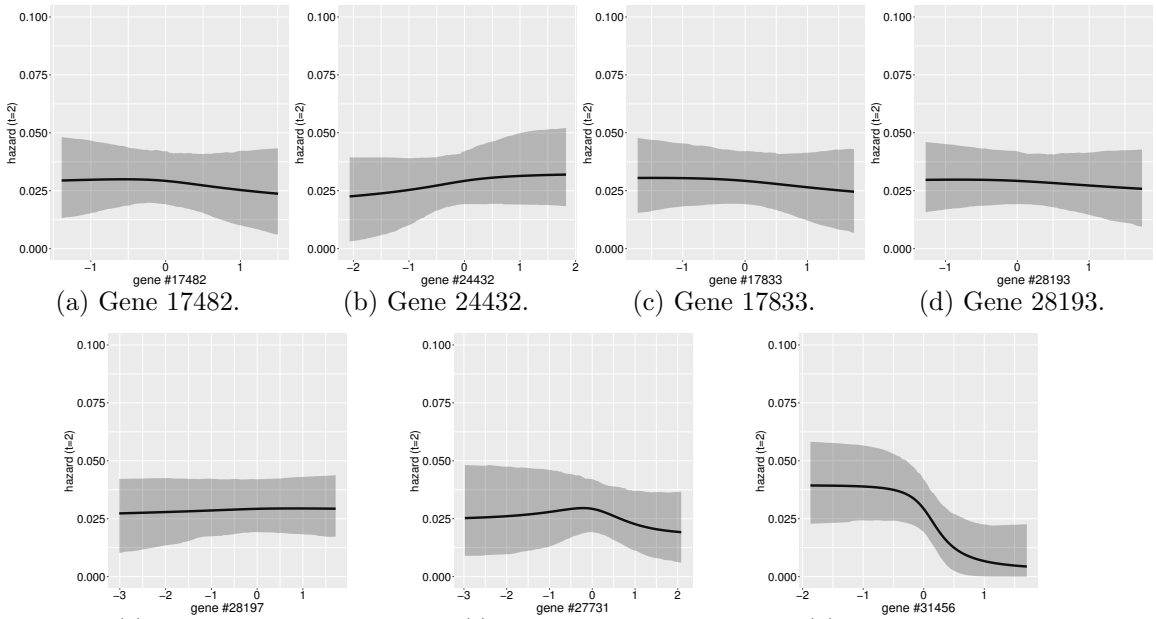
We apply WDR to biomedical data analysis and demonstrate its clinical utility. First, we use WDR on a diffuse large B-cell lymphoma data set where the covariates are time-invariant gene expression and there is no left truncation. In this example, we illustrate the non-monotonic change of hazards with covariates and provide insights into potential new disease subtypes. Second, we study cognitive decline among normal individuals using a data set with left truncation, right censoring, and time-varying covariates. We find that three biomarkers are significantly associated with the progression of Alzheimer's disease.

5.1 Surviving Diffuse Large B-Cell Lymphoma

We apply WDR to the diffuse large B-cell lymphoma (DLBCL) data[3] (Rosenwald et al., 2002) where the covariates are time-invariant. Multiple unsuccessful treatments to increase the survival rate suggest that there may exist several subtypes of DLBCL that differ in responsiveness to chemotherapy. Rosenwald et al. (2002) identify three subgroups of gene expression in the DLBCL data, 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 different mechanisms of malignant transformation. They also conjecture that T3 might be associated with more than one such mechanism. In our analysis, we treat the death of ABC, GCB, or T3 as three competing events and study the time to death due to a specific subtype. Right censoring applies to those who were alive at the end of the study. The data set contains a microarray gene expression profile of 7399 genes from 240 patients, and only 434 genes have no missing values in all the patients. Since missing covariate imputation is beyond our scope and the number of patients is small, we use seven of the 434 genes as covariates, as they have been reported to be related to clinical phenotypes (Li and Luan, 2005). We focus on the WDR estimation of covariate effects and defer the model comparison to Table 13 in the Appendix, where WDR consistently delivers accurate predictions.

We find by WDR one sub-event for ABC and GCB, respectively, implying that the gene expression is linearly associated with the time to death of ABC or GCB. Meanwhile, two sub-events are found under T3, implying nonlinear effects of the genes. Concretely, we show in Figure 3 how the hazard function of T3 changes with the gene expression. In each panel, we plot the hazard function at $t = 2$ against one gene that varies between its minimum and maximum values (the data has been centered) as in the data with the expression of other

3. The data is publicly accessible at `https://llmpp.nih.gov/DLBCL/`. Last access in July 2022.



(a) Gene 17482. (b) Gene 24432. (c) Gene 17833. (d) Gene 28193.

(e) Gene 28197. (f) Gene 27731. (g) Gene 31456.

Figure 3: Hazards of T3 at time $t = 2$ against gene expressions.

genes being zero. The figures only show the relative change of hazards whose absolute values are arbitrary. The WDR parameter estimations are provided in Table 12 in Appendix D.

The decreasing hazard functions in panels (a), (c), (d), and (g) imply that high expression of genes 17482, 17833, 28193, and 31456 is associated with late death of T3 (at $t = 2$), whereas high expression of genes 24432 and 28197 (panels (b) and (e)) is associated with early death. Interestingly noted that panel (g) suggests an inverse S-shaped effect of gene 31456 on the hazard of T3, which is apparently not log-linear (as shown in Figure 1 (a)). Specifically, a plateau is observed between gene expression values -2 and -0.5. Then the hazard sharply goes down until the expression of 0.5 and mildly decreases afterward. Moreover, gene 24432 seems to deliver an S-shaped effect. In contrast to either increasing or decreasing effects, panel (f) reveals a non-monotonic effect of gene 27731 such that the hazard is first increasing and then decreasing as the gene expression grows. The two sub-events found potentially imply two different mechanisms of malignant transformation of type-3 DLBCL whose progression is linearly associated with the genes. One can use WDR to find a patient's sub-event type of T3 (κ_{iy_i} for $y_i = 3$), study the T3 DLBCL progression by $\boldsymbol{\beta}_{3k}$'s, and investigate individualized treatment.

5.2 Alzheimer's Disease Data Analysis

To demonstrate the applicability of WDR to left-truncated and right-censored data with intermittently updated covariates, we apply WDR to a data set from the Biomarkers of Cognitive Decline Among Normal Individuals (BIOCARD) cohort study. It was administrated by the National Institute of Health from 1995 to 2005 and re-established at Johns Hopkins School of Medicine after being stopped for four years. The BIOCARD study recruited participants who were MCI-free at baseline and collected their cognitive performance testing

Age, mean years (SD)	56.9 (9.3)
Sex, % females	58.8%
Education, mean years (SD)	17.1 (2.4)
% ApoE-4 carriers	33.7%

Table 6: Baseline characteristics of the participants in the analysis of the BIOCARD study.

scores along with other biomarkers that are potentially related to Alzheimer's disease (AD) annually or semiannually during the study, aiming to identify biomarkers associated with the development of AD progression.

In the BIOCARD study, we study the failure time defined as the age at the onset of symptoms of mild cognitive impairment (MCI). The data are left-truncated and right-censored, where the truncation time is an individual's age at the time when she or he entered the study. Death is a competing event since subjects can die due to cancer or other diseases without symptoms of MCI. We consider three time-invariant biomarkers: sex, education, and the $\epsilon 4$ allele of the apolipoprotein E (ApoE-4) gene, and 10 time-varying continuous biomarkers: 5 cerebrospinal fluid (CSF) variables and 5 cognitive measures. These biomarkers are selected since they are potentially important measures related to the time to onset of AD clinical symptoms based on the findings from previous literature (Albert et al., 2014). The CSF biomarkers include Abeta 42, Abeta 40, total tau (t-tau), phosphorylated tau 181 (p-tau181), and ptau_amyloid. Cognitive biomarkers were measured by the annual, comprehensive neuropsychological battery test for participants in the BIOCARD to examine whether they are significantly associated with the time to onset of clinical symptoms, which were a harbinger of a diagnosis of MCI. We include logmem (Logical Memory IIA - Delayed), paired (Paired Associates I - New Learning), DSST (WAIS-R Digit Symbol), CVLTTOTL (Number correct Trials 1-5), and MMscore (Total Mini-Mental State Examination score) in the analyses. Of the 291 subjects included in our analyses, 35 subjects were observed death during the study, 82 subjects were diagnosed with MCI or dementia due to AD, and 209 subjects remained cognitively normal at their last visits. Table 6 gives a brief summary of the participants in our analysis.

We find one sub-event under MCI and death of other causes, respectively, by WDR, implying that the covariates are linearly related to the time to either competing event. Specifically, we run 10,000 MCMC iterations with a burn-in of 8,000 iterations, collect 2,000 post-burn-in MCMC samples, and calculate the posterior means and the 95% credible intervals of the coefficients $\boldsymbol{\beta}_{jk}$ for $j = 1, 2$ and $k = 1$ as reported in Table 7. We find a decelerated progression of MCI significantly associated with higher values of *education*, *paired*, and *DSST*. These results are corroborated by existing literature. A high level of education has been reported to reduce the risk of MCI and AD (Sattler et al., 2012). It has also been reported that the increased risk of progressing from normal cognition to the onset of clinical symptoms is associated with lower scores of paired and DSST (Albert et al., 2014). On the other hand, a higher education level is associated with lower death hazards of other causes, while the other ten AD-related biomarkers do not significantly contribute to explaining the time to death. In addition, we randomly split the subjects into 80% of training data and 20% of testing data to assess the performance of WDR and defer the model comparison to Table 14 in the Appendix.

	MCI ($j = 1$)	Death ($j = 2$)
sex	-0.010 (-0.225, 0.225)	-0.110 (-0.755, 0.463)
education	**-0.102** (-0.124, -0.057)	**-0.124** (-0.203, -0.064)
ApoE-4	0.041 (-0.148, 0.319)	0.265 (-0.523, 0.896)
p-tau181	0.114 (-0.256, 0.562)	-0.009 (-0.451, 0.471)
t-tau	0.080 (-0.090, 0.287)	-0.011 (-0.391, 0.220)
Abeta 42	-0.097 (-0.447, 0.188)	0.037 (-0.494, 0.598)
Abeta 40	0.007 (-0.222, 0.236)	0.098 (-0.377, 0.622)
ptau_amyloid	0.154 (-0.237, 0.574)	0.029 (-0.470, 0.431)
paired	**-0.402** (-0.646, -0.228)	-0.057 (-0.373, 0.151)
logmem	-0.023 (-0.227, 0.248)	-0.012 (-0.287, 0.227)
DSST	**-0.454** (-0.674, -0.207)	-0.139 (-0.466, 0.185)
CVLTTOTL	-0.142 (-0.417, 0.024)	-0.211 (-0.811, 0.047)
MMSCORE	-0.083 (-0.278, 0.05)	-0.105 (-0.363, 0.089)

Table 7: Posterior means and 95% credible intervals of the coefficients $\boldsymbol{\beta}_{jk}$, $k = 1$.

6. Conclusion

Assuming a two-phase racing among latent sub-events, the proposed Weibull delegate racing (WDR) survival model not only accommodates non-monotonic covariate effects but also preserves interpretability. We use a gamma process to support a potentially infinite number of sub-events, rely on its inherent shrinkage property to remove unneeded model capacity, and thus enable WDR to explore mechanisms of surviving competing events. Moreover, WDR can handle left truncation, time-varying covariates, missing event times or types, and different censoring. To the best of our knowledge, WDR is the first nonlinear survival model for competing events, left truncation, and time-varying covariates without using covariate transformations. Simulation studies have shown excellent performance and favorable properties of WDR compared to the Fine-Gray models, the random survival forests, the DeepHit, and the piecewise constant hazards model. The analysis of the lymphoma and Alzheimer's disease data shows intriguing findings that help researchers discover new disease types or interpret covariate effects related to disease progression. Overall, WDR is an attractive alternative to existing models for various applications that require interpretable nonlinearity.

Acknowledgments

We thank the editor and the reviewers for their dedicated time and constructive comments. Yanxun Xu is supported in part by NSF 1918854. Mei-Cheng Wang is supported in part by NIH Grant U19 AG033655. Mingyuan Zhou is supported in part by NSF 1812699 and NSF 2212418.

Appendix A. Theorem and Proof

A.1 Proof of the Weibull Racing Property

We prove Property 1 as follows.

Proof Since $\Pr(t > t_0) = \prod_j \Pr(t_j > t_0) = \prod_j S_\tau(t_0 \,|\, a, \lambda_j) = \frac{\exp(-t_0^a \sum_j \lambda_j)}{\exp(-\tau^a \sum_j \lambda_j)}$ for $t_0 \geq \tau$, then $t = \min_j t_j \sim \text{Weibull}\tau(a, \sum_j \lambda_j)$. Assuming $t_h = \min_j t_j$ and for arbitrary $t_0 \geq \tau$, we have

$$\begin{aligned}
\Pr(\min_j t_j > t_0,\ t_h = \min_j t_j) &= \prod_{j \neq h} \Pr(t_0 < t_h < t_j) \\
&= \int_{t_0}^{\infty} f_\tau(t_h \,|\, a, \lambda_h) \prod_{j \neq h} \Pr(t_j > t_h) dt_h \\
&= \int_{t_0}^{\infty} \frac{f(t_h \,|\, a, \lambda_h)}{S(\tau \,|\, a, \lambda_h)} \prod_{j \neq h} \frac{S(t_h \,|\, a, \lambda_j)}{S(\tau \,|\, a, \lambda_j)} dt_h \\
&= \frac{\lambda_h}{\sum_j \lambda_j} \frac{\exp(-t_0^a \sum_j \lambda_j)}{\exp(-\tau^a \sum_j \lambda_j)}.
\end{aligned}$$

Let $t_0 = \tau$. We have $\Pr(t_h = \min_j t_j) = \frac{\lambda_h}{\sum_j \lambda_j}$. This proves the categorical distribution of $y = \operatorname{argmin}_j t_j$. Consequently,

$$\Pr(\min t_j > t_0,\ t_h = \min_j t_j) = \Pr(t > t_0,\ y = h) = \Pr(t > t_0)\Pr(y = h).$$

This proves the independence of t and y. ■

A.2 WDR Survival Function with Time-Varying Covariates

We derive equations (4) and (5). Omitting the subject index i for brevity and given the piecewise Weibull hazards

$$h^{(l)}(t) = h_{\tau^{(l)}}(t \,|\, a, \sum_{j,k} \lambda_{jk}^{(l)}) = a \sum_{j,k} \lambda_{jk}^{(l)} t^{a-1}$$

for $t \in [\tau^{(l)}, \tau^{(l+1)})$, $l = 0, 1, \ldots, L-1$ and $[\tau^{(l)}, \infty)$ for $l = L$, the cumulative hazard function for $t > \tau^{(L)}$ is

$$\begin{aligned}
H(t \,|\, \{\tau^{(l)}\}_{l=0}^{L}) &= \int_0^t \Big[\sum_{l=0}^{L-1} h^{(l)}(s) \mathbf{1}(s \in (\tau^{(l)}, \tau^{(l+1)}]) + h^{(L)}(s) \mathbf{1}(s \in (\tau^{(L)}, \infty)) \Big] ds \\
&= \sum_{l=0}^{L-1} \int_{\tau^{(l)}}^{\tau^{(l+1)}} a \sum_{j,k} \lambda_{jk}^{(l)} s^{a-1} ds + \int_{\tau^{(L)}}^{t} a \sum_{j,k} \lambda_{jk}^{(L)} s^{a-1} ds \\
&= \sum_{l=0}^{L-1} \Big[\sum_{j,k} \lambda_{jk}^{(l)} (\tau^{(l+1)})^a - \sum_{j,k} \lambda_{jk}^{(l)} (\tau^{(l)})^a \Big] + \sum_{j,k} \lambda_{jk}^{(L)} t^a - \sum_{j,k} \lambda_{jk}^{(L)} (\tau^{(L)})^a.
\end{aligned}$$

Consequently, the survival function at time $t > \tau^{(L)}$ is

$$S(t \mid \{\tau^{(l)}, \lambda_{jk}^{(l)}\}_{l=0}^{L}) = \exp(H(t \mid \{\tau^{(l)}\}_{l=0}^{L})) = S_{\tau^{(L)}}(t \mid a, \sum_{j,k} \lambda_{jk}^{(L)}) \prod_{l=0}^{L-1} S_{\tau^{(l)}}(\tau^{(l+1)} \mid a, \sum_{j,k} \lambda_{j,k}^{(l)}).$$

The density function is obtained by taking the derivative of $1 - S(t \mid \{\tau^{(l)}, \lambda_{jk}^{(l)}\}_{l=0}^{L})$.

A.3 Marginal Distribution of the WDR Event Time

Theorem 4 *If $t_i \sim Weibull_\tau(a, \lambda_{i\bullet\bullet})$ with $\lambda_{i\bullet\bullet} = \sum_{j,k} \lambda_{ijk}$ and $\lambda_{ijk} \sim Gamma(r_{jk}, 1/b_{ijk})$, the density function of t_i given a, $\{r_{jk}\}$, and $\{b_{ijk}\}$ and $t_i > \tau$ is*

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

and the cumulative density function is

$$Pr(t_i < q \mid \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}, t_i > \tau) = 1 - c_i \sum_{m=0}^{\infty} \frac{\delta_{im} b_{i(1)}^{\rho_i+m}}{(q^a - \tau^a + b_{i(1)})^{\rho_i+m}}, \tag{6}$$

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 $m \geq 0$, and $\gamma_{ih} = \sum_{j,k} \frac{r_{jk}}{h} \left(1 - \frac{b_{ijk}}{b_{i(1)}}\right)^h$.

It is difficult to use the density or cumulative distribution functions of t_i in the form of series, but we can use a finite truncation as an approximate. Concretely, as $\Pr(t_i < \infty \mid a, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) = c_i \sum_{m=0}^{\infty} \delta_{im} = 1$, 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 $1 - c_i \sum_{m=0}^{M} \frac{\delta_{im} b_{i(1)}^{\rho_i+m}}{(q^a - \tau^a + b_{i(1)})^{\rho_i+m}}$ as an approximation. Consequently, sampling t_i is feasible by inverting the approximated cumulative distribution function for general cases. We have tried prediction by finite truncation on some synthetic data where $\tau = 0$ and $a = 1$ and found that M is mostly between 10 and 30, which is computationally feasible.

Proof We first study the distribution of gamma convolution. Specifically, if $\lambda_s \overset{ind}{\sim} \text{Gamma}(r_s, 1/b_s)$ with $r_s, b_s \in \mathbb{R}_+$, then the density function of $\lambda = \sum_{s=1}^{S} \lambda_s$ can be written in a form of series (Moschopoulos, 1985) as

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

where $c = \prod_{s=1}^{S} \left(\frac{b_s}{b_{(1)}}\right)^{r_s}$, $b_{(1)} = \max_s b_s$, $\rho = \sum_{s=1}^{S} r_s$, $\delta_0 = 1$, $\delta_{m+1} = \frac{1}{m+1} \sum_{h=1}^{m+1} h\gamma_h \delta_{m+1-h}$ and $\gamma_h = \sum_{t=1}^{T} r_t \left(1 - \frac{b_t}{b_{(1)}}\right)^h / h$. Moschopoulos (1985) has 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 $b_{i0} = max_{j,k}(1 - \frac{b_{ijk}}{b_{i(1)}})$. We want to show the PDF of t_i,

$$
\begin{aligned}
& f(t_i \,|\, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) \\
=& \int_0^\infty f(t_i \,|\, \lambda_{i\bullet\bullet}) f(\lambda_{i\bullet\bullet} \,|\, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) d\lambda_{i\bullet\bullet} \\
=& \int_0^\infty \sum_{m=0}^\infty \frac{c_i \delta_{im} a t_i^{a-1} \lambda_{i\bullet\bullet}^{\rho_i+m} \exp(-(t_i^a - \tau^a)\lambda_{i\bullet\bullet} - b_{i(1)}\lambda_{i\bullet\bullet})}{\Gamma(\rho_i+m)/b_{i(1)}^{\rho_i+m}} d\lambda_{i\bullet\bullet} \\
=& \sum_{m=0}^\infty \int_0^\infty \frac{c_i \delta_{im} a t_i^{a-1} \lambda_{i\bullet\bullet}^{\rho_i+m} \exp(-(t_i^a - \tau^a)\lambda_{i\bullet\bullet} - b_{i(1)}\lambda_{i\bullet\bullet})}{\Gamma(\rho_i+m)/b_{i(1)}^{\rho_i+m}} d\lambda_{i\bullet\bullet} \qquad (7) \\
=& a t_i^{a-1} c_i \sum_{m=0}^\infty \frac{(\rho_i+m)\delta_{im} b_{i(1)}^{\rho_i+m}}{(t_i^a - \tau^a + b_{i(1)})^{1+\rho_i+m}},
\end{aligned}
$$

which suffices to prove the equality in (7). Specifically,

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

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 (7) holds. Next, we want to calculate the CDF of t_i,

$$
\begin{aligned}
\Pr(t_i < q \,|\, n_i, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k}) &= \int_0^q a t_i^{a-1} c_i \sum_{m=0}^\infty \frac{(\rho_i+m)\delta_{im} b_{i(1)}^{\rho_i+m}}{(t_i^a - \tau^a + b_{i(1)})^{1+\rho_i+m}} dt_i \\
&= \sum_{m=0}^\infty \int_0^q a t_i^{a-1} c_i \frac{(\rho_i+m)\delta_{im} b_{i(1)}^{\rho_i+m}}{(t_i^a - \tau^a + b_{i(1)})^{1+\rho_i+m}} dt_i. \qquad (8)
\end{aligned}
$$

It suffices to show (8). Specifically,

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

The last equation holds because the summation of a negative binomial probability mass function is 1. Therefore, $f(t_i \,|\, a, \{r_{jk}\}_{j,k}, \{b_{ijk}\}_{j,k})$ is uniformly convergent and (8) holds. Calculating the integration, we obtain the CDF of t_i. ■

Appendix B. WDR Inference

B.1 MCMC

Let τ_i, T_i and T_{ic} denote the left truncation time, the observed failure time, and the right censoring time, respectively, for subject $i = 1, \ldots, n$, with τ_i less than T_i or T_{ic}. Since left censoring is uncommon and not shown in our real data, we only consider right censoring in our inference and leave to readers other types of censoring which can be analogously done. The inference by MCMC accommodating missing event time or missing event types proceeds by iterating the following steps.

Step 1: If y_i is observed, we first sample κ_{iy_i} by

$$\Pr(\kappa_{iy_i} = k \,|\, y_i, \cdots) = \frac{\lambda_{iy_ik}}{\sum_{k'=1}^{K}\lambda_{iy_ik'}}.$$

If y_i is unobserved which means a missing event type, we sample (y_i, κ_{iy_i}) by

$$\Pr(y_i = j, \kappa_{iy_i} = k \,|\, \cdots) = \frac{\lambda_{ijk}}{\sum_{j'=1}^{S}\sum_{k'=1}^{K}\lambda_{ij'k'}}.$$

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-event that has the minimum event time.

Step 2: Update t_i for $i = 1, \cdots, n$, $j = 1, \cdots, J$ and $k = 1, \cdots, K$.

(a) If the event time T_i is observed, we set $t_i = T_i$.

(b) Otherwise, we sample $t_i \sim \text{Weibull}_{(T_{ic},\infty)}(a, \sum_{j=1}^{S}\sum_{k=1}^{K}\lambda_{ijk})$ where $\text{Weibull}_{(T_{ic},\infty)}(\cdot,\cdot)$ is a truncated Weibull distribution so that $t_i \in (T_{ic}, \infty)$. Note $T_{ic} = 0$ if both event time and censoring time are missing for observation i.

Step 3: Sample $(\lambda_{ijk} \,|\, -) \sim \text{Gamma}\left(r_{jk} + n_{ijk}, \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})}{1+(t_i^a - \tau_i^a)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})}\right)$, for $i = 1, \cdots, n$, $j = 1, \cdots, J$, and $k = 1, \cdots, K$.

Step 4: Sample a by slice sampling. a determines how the hazard varies over time, and we assume an improper prior of $a \in \mathbb{R}_+$, $p(a) \propto \mathbf{1}(a > 0)$ to reduce the impact of the prior on the posterior. The full conditional distribution of a is

$$p(a \,|\, \ldots) \propto a^n \prod_i \left[t_i^{a-1} \prod_{j,k} \left(1 + (t_i^a - \tau_i^a)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})\right)^{-n_{ijk}-r_{jk}} \right].$$

If $\tau_i = 0$, the unimodality of $p(a \,|\, \ldots)$ can be shown so that slice sampling can be implemented without tuning parameters. Concretely, when $\tau_i = 0$,

$$\frac{d \log p(a|\ldots)}{da} = \frac{n}{a} + \sum_i \log t_i - \sum_{i,j,k}(n_{ijk} + r_{jk})\frac{\exp(a\log t_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})\log t_i}{1 + \exp(a\log t_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})}.$$

Since $\frac{n}{a}$ is decreasing in a while $\sum_{i,j,k}(n_{ijk} + r_{jk})\frac{\exp(a\log t_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})\log t_i}{1+\exp(a\log t_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})}$ is increasing in a, there must be at most one $a \in \mathbb{R}_+$ satisfying $\frac{d\log p(a|\ldots)}{da} = 0$. So, $p(a|\ldots)$ is unimodal. We use the `mcmc` function in the `R` package `diversitree` (FitzJohn, 2012).

Step 5: Sample $\boldsymbol{\beta}_{jk}$, for $j = 1, \cdots, J$ and $k = 1, \cdots, K$, by Pólya Gamma (PG) data augmentation. First sample $(\omega_{ijk} \,|\, -) \sim \text{PG}(r_{jk} + n_{ijk}, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk} + \log(t_i^a - \tau_i^a))$. Then sample $(\boldsymbol{\beta}_{jk} \,|\, -) \sim \text{MVN}(\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ where $\boldsymbol{\Sigma}_{jk} = (V_{jk} + \boldsymbol{X}^{\mathrm{T}}\Omega_{jk}\boldsymbol{X})^{-1}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^{\mathrm{T}}$, $\Omega_{jk} = \text{diag}(\omega_{1jk}, \cdots, \omega_{njk})$ and $\boldsymbol{\mu}_{jk} = \boldsymbol{\Sigma}_{jk}\left[-\sum_{i=1}^{N}\left(\omega_{ijk}\log(t_i^a - \tau_i^a) + \frac{r_{jk}-n_{ijk}}{2}\right)\boldsymbol{x}_i\right]$. To sample from the Pólya-Gamma distribution, we use a fast and accurate approximate sampler (Zhou, 2016) that matches the first two moments of the original distribution; we set the truncation level of that sampler as five.

Step 6: 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 $k = 1, \cdots, K$.

Step 7: Sample r_{jk} and γ_{0j}, for $j = 1, \cdots, J$ and $k = 1, \cdots, K$, by Chinese restaurant table (CRT) data augmentation (Zhou and Carin, 2015).

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 $(r_{jk} \,|\, -) \sim \text{Gamma}\left(\sum_{i=1}^{N} n_{ijk}^{(2)} + \gamma_{0j}/K, \frac{1}{c_{0j} + \sum_{i=1}^{N}\log(1+(t_i^a - \tau_i^a)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk}))}\right)$,

and $(\gamma_{0j} \mid -) \sim \text{Gamma}\left(d_0 + \sum_{k=1}^{K} l_{jk}, \frac{1}{e_0 - \frac{1}{K}\sum_{k=1}^{K} \log(1-p_{jk})}\right)$,
where $p_{jk} = \frac{\sum_{i=1}^{N} \log(1+(t_i^a - \tau_i^a)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk}))}{c_{0j} + \sum_{i=1}^{N} \log(1+(t_i^a - \tau_i^a)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk}))}$.

Step 8: Sample $(c_{0j} \mid -) \sim \text{Gamma}\left(d_1 + \gamma_{0j}, \frac{1}{e_1 + \sum_{k=1}^{K} r_{jk}}\right)$ for $j = 1, \cdots, J$.

Step 9: (Optional.) Prune unneeded sub-events. For $j = 1, \cdots, J$ and $k = 1, \cdots, K$, if m_{jk} is small enough, say $m_{jk} = 0$ or $m_{jk} \leq 1\% \times n$, prune sub-event k of competing event j by setting $\lambda_{ijk} = 0$ and $t_{ijk} = \infty$ for $i = 1, \ldots, n$.

Steps 1 through 8 are MCMC updates of parameters. Although the item weights $\{r_{jk}\}$ in the gamma processes are almost surely positive, the latent count allocation of n_{ijk} and $n_{ijk}^{(2)}$ in Steps 1 and 7 makes it possible to prune a sub-event that is not an important component in a phase-one race within a competing event. Step 9 is used to explicitly prune the unneeded nonlinear modeling capacity. Specifically, m_{jk} counts the number of observations that have sub-event k winning the race within competing event j. The sub-event k is redundant if m_{jk} is small compared to the total number of observations n. For small data sets, we prune it if $m_{jk} = 0$. For larger data sets with $n \geq$10,000, we suggest pruning it if $m_{jk} \leq (n \times 1\%)$.

Step 1 tries to find to which subevent the event of type y_i belongs by sampling κ_{iy_i} from a multinomial distribution. This latent counts allocation is similar to those in Zhou (2016) and Zhou et al. (2016), where a deep gamma hierarchy is used. Using a similar deep hierarchical model does not remarkably increase the capacity of WDR, possibly because of the following reason. In Zhou (2016) and Zhou et al. (2016), a Bernoulli-Poisson link is used, and the latent counts can be greater than one and are allocated to different layers of the deep gamma hierarchy. But in our context, the latent count is always 1 or 0, indicating whether a subject suffers from a latent subevent k under competing event j or not. Effectively, a deep gamma hierarchy is analogous to WDR where we have one latent layer with a potentially infinite number of nodes (subevents). We refer the readers to the two aforementioned papers for details. In addition, another reason we do not consider such a deep hierarchy is that it may sabotage the interpretability of WDR.

B.2 Maximum a Posteriori Estimation

We propose a Maximum a posteriori (MAP) estimation of the WDR parameters for big data applications. Since the MAP estimations do not involve latent count allocations as in Step 1 of the MCMC algorithm, and thus the sub-events cannot be explicitly pruned, we can determine the number of latent sub-events K by cross-validation and model selection rules, like AIC or BIC. We show the estimations for data with right-censored subjects. Other types of censoring can be done analogously.

With the reparameterization that $\lambda_{ijk} = \tilde{\lambda}_{ijk}\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{jk})$ 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.

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

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

$$p_{t_i} = \begin{cases} a t_i^{a-1} (\sum_{j,k} \tilde{\lambda}_{ijk} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{jk})) \exp\big(- (t_i^a - \tau_i^a) \sum_{jk} \tilde{\lambda}_{ijk} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{jk})\big) & \text{if } t_i \text{ is uncensored,} \\ \exp\big(- T_{ic}^a \sum_{jk} \tilde{\lambda}_{ijk} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{jk})\big) & \text{if } t_i \text{ is right-censored at } T_{ic} \\ 1 \quad \text{if } t_i \text{ is missing, but } y_i \text{ is not,} & \end{cases}$$

$$p_{y_i} = \begin{cases} \frac{\sum_k \tilde{\lambda}_{iy_ik} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{y_ik})}{\sum_{j,k} \tilde{\lambda}_{ijk} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{jk})} & \text{if } y_i \text{ is not missing,} \\ 1 & \text{if } y_i \text{ is missing, but } t_i \text{ is not.} \end{cases}$$

We do not define the likelihood of subject i if both t_i and y_i are missing and remove such observations from the data. For brevity, we write $p_t(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r})$ as p_{t_i} and $p_y(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r})$ as p_{y_i}.

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

$$\log P = \sum_i \log p_i + \log p(a) + \sum_{j,k} \log p(\boldsymbol{\beta}_{jk}) + \sum_{j,k} \log p(r_{jk}) + C \tag{9}$$

where C is a constant function of a, $\{\boldsymbol{\beta}_{jk}\}$ and $\{r_{jk}\}$. In practice we assume an improper prior $p(a) \propto 1$ on a, a Student's t distribution with degrees of freedom α on each element of $\boldsymbol{\beta}_{jk}$ and a $\text{Gamma}(0.01/K, 1/0.01)$ prior on r_{jk}. We also find that a $\text{Gamma}(1/K, 1)$ prior on r_{jk} or an L_2-regularizer, $0.001||\boldsymbol{r}||_2$, is more numerically stable. Subsequently, we have

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

where c is also a constant function of a, $\{\boldsymbol{\beta}_{jk}\}$ 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, score function gradients, and self-normalization. Specifically,

$$\nabla_{a,\boldsymbol{\beta}} \log p_i = \frac{\int \left[\nabla_{a,\boldsymbol{\beta}} \left(p_{t_i} \times p_{y_i}\right)\right] p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i}{\int \left(p_{t_i} \times p_{y_i}\right) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i} \approx \frac{\frac{1}{M} \sum_{m=1}^M \nabla_{a,\boldsymbol{\beta}} \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})\right]}{\frac{1}{M} \sum_{m=1}^M \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})\right]} \tag{10}$$

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

$$
\begin{aligned}
\nabla_{\boldsymbol{r}} \log p_i &= \frac{\int \nabla_{\boldsymbol{r}} \left[(p_{t_i} \times p_{y_i})\, p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) \right] d\tilde{\boldsymbol{\lambda}}_i}{\int (p_{t_i} \times p_{y_i})\, p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i} \\
&= \frac{\int (p_{t_i} \times p_{y_i})\, \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i}{\int (p_{t_i} \times p_{y_i})\, p(\tilde{\boldsymbol{\lambda}}_i \,|\, \boldsymbol{r}) d\tilde{\boldsymbol{\lambda}}_i} \\
&\approx \frac{\frac{1}{M} \sum_{m=1}^{M} p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})}{\frac{1}{M} \sum_{m=1}^{M} \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \right]} \\
&= \sum_{m=1}^{M} \frac{p_t(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r})}{\sum_{m'=1}^{M} \left[p_t(\tilde{\boldsymbol{\lambda}}_i^{(m')} \,|\, \boldsymbol{r}) \times p_y(\tilde{\boldsymbol{\lambda}}_i^{(m')} \,|\, \boldsymbol{r}) \right]} \nabla_{\boldsymbol{r}} \log p(\tilde{\boldsymbol{\lambda}}_i^{(m)} \,|\, \boldsymbol{r}). \qquad (11)
\end{aligned}
$$

Therefore, we can approximate the derivatives of $\log P$ with respect to a, $\boldsymbol{\beta}$ and $\boldsymbol{r}$ by plugging in (10) and (11), respectively, and maximize $\log P$ by (stochastic) gradient ascent.

B.3 Cumulative Incidence Function for WDR

The cumulative incidence function (CIF) of subject i for event j at time t is $\text{CIF}_j(i,t) = \Pr(t_i \le t, y_i = j)$ (Fine and Gray, 1999; Kalbfleisch and Prentice, 2011; Crowder, 2001), indicating the probability of event j by time t. Given time-invariant $\boldsymbol{x}_i$, a left truncation time τ, a, $\{r_{jk}\}_{j,k}$ and $\{\boldsymbol{\beta}_{jk}\}_{j,k}$, WDR has

$$
\text{CIF}_j(i,t) = \Pr(t_i \le t, y_i = j) = E\left[\frac{\sum_k \lambda_{ijk}}{\sum_{j',k} \lambda_{ij'k}} \Big(1 - \exp\big(-(t^a - \tau^a) \sum_{j',k} \lambda_{ij'k} \big) \Big) \right],
$$

where the expectation is taken over $\lambda_{ijk} \sim \text{Gamma}(r_{jk}, \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{jk}))$. The CIF for WDR can be evaluated using Monte Carlo estimation if we have point estimates or a collection of post-burn-in MCMC samples of r_{jk} and $\boldsymbol{\beta}_{jk}$.

Given time-varying covariates $\boldsymbol{X}_i$ whose values are updated to $\boldsymbol{x}_i^{(0)}, \ldots, \boldsymbol{x}_i^{(L)}$ at times $\tau^{(0)}, \ldots, \tau^{(L)}$ before t, the cumulative incidence function of WDR is

$$
\begin{aligned}
&\text{CIF}_j(i,t) \\
=&\Pr(t_i \le \tau^{(1)}, y_i = j) + \Pr(\tau^{(1)} < t_i \le \tau^{(2)}, y_i = j) + \cdots + \Pr(\tau^{(L)} < t_i \le t, y_i = j) \\
=&E\left[\frac{\sum_k \lambda_{ijk}^{(0)}}{\sum_{j',k} \lambda_{ij'k}^{(0)}} \left(1 - \exp(-[(\tau^{(1)})^a - (\tau^{(0)})^a] \sum_{j',k} \lambda_{ij'k}^{(0)} \right) \right] + \\
&E\left[\frac{\sum_k \lambda_{ijk}^{(1)}}{\sum_{j',k} \lambda_{ij'k}^{(1)}} \left(1 - \exp(-[(\tau^{(2)})^a - (\tau^{(1)})^a] \sum_{j',k} \lambda_{ij'k}^{(1)} \right) \exp(-[(\tau^{(1)})^a - (\tau^{(0)})^a] \sum_{j',k} \lambda_{ij'k}^{(0)} \right] \\
&+ \cdots + \\
&E\left[\frac{\sum_k \lambda_{ijk}^{(L)}}{\sum_{j',k} \lambda_{ij'k}^{(L)}} \left(1 - \exp(-[t^a - (\tau^{(L)})^a] \sum_{j',k} \lambda_{ij'k}^{(L)} \right) \prod_{l=0}^{L-1} \exp(-[(\tau^{(l+1)})^a - (\tau^{(l)})^a] \sum_{j',k} \lambda_{ij'k}^{(l)} \right]
\end{aligned}
$$

Data 1	Data 2	Data 3
$x_{i1} \sim \mathrm{U}(-1,1)$	$x_{i1} \sim \mathrm{U}(-1,1)$, $i = 1,\ldots,2000$	$x_{i1} \sim \mathrm{U}(-1,1)$, $i = 1,\ldots,2000$
$x_{i2} \sim \mathrm{U}(0,1)$	$x_{i2} \sim \mathrm{U}(0,1)$, $x_{i3} \sim \mathrm{U}(-1,0)$, $i \le 1000$	$x_{i2} \sim \mathrm{U}(0,1)$, $x_{i3} \sim \mathrm{U}(-1,0)$, $i \le 1000$
$x_{i3} \sim \mathrm{U}(0,1)$	$x_{i2} \sim \mathrm{U}(-1,0)$, $x_{i3} \sim \mathrm{U}(0,1)$, $i > 1000$	$x_{i2} \sim \mathrm{U}(-1,0)$, $x_{i3} \sim \mathrm{U}(0,1)$, $i > 1000$
$\boldsymbol{b}_1 = (1,2,-1)^{\mathrm{T}}$, $\boldsymbol{b}_2 = (1,-1,2)^{\mathrm{T}}$	$\boldsymbol{b}_1 = (1,2,4)^{\mathrm{T}}$, $\boldsymbol{b}_2 = (1,4,2)^{\mathrm{T}}$	$\boldsymbol{b}_1 = (1,-2,1)^{\mathrm{T}}$, $\boldsymbol{b}_2 = (1,-1,2)^{\mathrm{T}}$
Right censoring: $T_{r.c.} = 1.6$	Right censoring: $T_{r.c.} = 1.3$	Right censoring: $T_{r.c.} = 0.6$
Data 4	**Data 5**	**Data 6**
$x_{i1}, x_{i2}, x_{i3} \sim \mathrm{U}(-1,1)$	$x_{i1}, x_{i2}, x_{i3} \sim \mathrm{U}(-1,1)$	$x_{i1}, x_{i2}, x_{i3} \sim \mathrm{U}(-1,1)$
$\boldsymbol{b}_1 = (-0.6, 0.2, -0.8)^{\mathrm{T}}$	$\boldsymbol{b}_1 = (-0.9, 0.2, 1.6)^{\mathrm{T}}$	$\boldsymbol{b}_1 = (-0.6, 0.2, -0.8)^{\mathrm{T}}$
$\boldsymbol{b}_2 = (-0.8, 0.2, -0.6)^{\mathrm{T}}$	$\boldsymbol{b}_2 = (1.6, 0.2, -0.9)^{\mathrm{T}}$	$\boldsymbol{b}_2 = (-0.8, 0.2, -0.6)^{\mathrm{T}}$
$\boldsymbol{b}_3 = (0.9, 0.7, 0.6)^{\mathrm{T}}$	$\boldsymbol{b}_3 = (-1.1, -0.1, 0.1)^{\mathrm{T}}$	
$\boldsymbol{b}_4 = (0.6, 0.7, 0.9)^{\mathrm{T}}$	$\boldsymbol{b}_4 = (0.1, -0.1, -1.1)^{\mathrm{T}}$	
Right censoring: $T_{r.c.} = 1.1$	Right censoring: $T_{r.c.} = 1.7$	Right censoring: $T_{r.c.} = 1.7$

Table 8: $\boldsymbol{x}_i$ and $\boldsymbol{b}$ for synthetic data with $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}) \in \mathbb{R}^3$.

Data 1	Data 2	Data 3
$\boldsymbol{x}_i \in \mathbb{R}^{10},\ \boldsymbol{x}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$	$\boldsymbol{x}_i \in \mathbb{R}^{10},\ \boldsymbol{x}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$	$\boldsymbol{x}_i \in \mathbb{R}^{10},\ \boldsymbol{x}_i \sim \mathrm{U}(-\mathbf{1}, \mathbf{1})$
$\boldsymbol{b}_1 = (-.1, -.1, 0, .2, -.2, .2, .3, .1, .1, -.3)$	$\boldsymbol{b}_1 = (-.2, -.1, .1, .4, -.3, .4, .4, .2, .1, -.4)$	$\boldsymbol{b}_1 = (.5, .7, 1.1, 1.8, .4, 1.8, 1.9, 1.3, 1.3, .1)$
$\boldsymbol{b}_2 = (-.2, -.2, .1, -.1, .2, 0, .1, .3, -.1, .2)$	$\boldsymbol{b}_2 = (-.3, -.3, .2, -.1, .3, 0, .2, .5, -.1, .3)$	$\boldsymbol{b}_2 = (.1, 1.3, 1.3, 1.9, 1.8, .4, 1.8, 1.1, .7, .5)$
Data 4	**Data 5**	**Data 6**
$\boldsymbol{x}_i \in \mathbb{R}^{10},\ \boldsymbol{x}_i \sim \mathrm{U}(-\mathbf{1}, \mathbf{1})$	$\boldsymbol{x}_i \in \mathbb{R}^{10},\ \boldsymbol{x}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$	$\boldsymbol{x}_i \in \mathbb{R}^{10},\ \boldsymbol{x}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$
$\boldsymbol{b}_1 = (-.6, .2, -.8, 1.6, .3, -.8, .5, .7, .6, -.3)$	$\boldsymbol{b}_1 = (-.6, .2, -.8, 1.6, .3, -.8, .5, .7, .6, -.3)$	$\boldsymbol{b}_1 = (-.6, .2, -.8, 1.6, .3, -.8, .5, .7, .6, -.3)$
$\boldsymbol{b}_2 = (-.3, .6, .7, .5, -.8, .3, 1.6, -.8, .2, -.6)$	$\boldsymbol{b}_2 = (-.3, .6, .7, .5, -.8, .3, 1.6, -.8, .2, -.6)$	$\boldsymbol{b}_2 = (-.3, .6, .7, .5, -.8, .3, 1.6, -.8, .2, -.6)$
$\boldsymbol{b}_3 = (.9, .2, .7, .1, .3, .4, 0, .4, .9, .3)$	$\boldsymbol{b}_3 = (1.5, .4, -.6, -2.2, 1.1, 0, 0, .9, .8, .6)$	
$\boldsymbol{b}_4 = (.3, .9, .4, 0, .4, .3, .1, .7, .2, .9)$	$\boldsymbol{b}_4 = (.6, .8, .9, 0, 0, 1.1, -2.2, -.6, .4, 1.5)$	

Table 9: $\boldsymbol{x}_i$ and $\boldsymbol{b}$ for synthetic data with $\boldsymbol{x} \in \mathbb{R}^{10}$.

where the expectations are taken over $\lambda_{ijk}^{(l)}$'s with $\lambda_{ijk}^{(l)} \sim \mathrm{Gamma}(r_{jk}, \exp(\boldsymbol{x}_i^{(l)\mathrm{T}} \boldsymbol{\beta}_{jk}))$.

Appendix C. Data Synthesis and Experiment Settings

The distribution of $\boldsymbol{x}_i$ and the values of $\boldsymbol{b}$'s for data synthesis are provided in Tables 8 and 9. We simulate synthetic data 1 to 6 with time-varying covariates using Algorithm 1. Concretely, we allow up to L updates of covariates and L can differ among subjects. With updated covariates at time $\tau^{(l)}$, we simulate the event time from a Weibull or log-normal distribution that is left truncated at $\tau^{(l)}$. If the event time is greater than $\tau^{(l+1)}$, we update the covariates at $\tau^{(l+1)}$ and repeat this procedure. Otherwise, we stop. Right censoring is also allowed.

We describe the experiment settings as follows. For the kernel Fine-Gray (KFG) model, we use the radial basis function kernel and select the kernel width from $2^{-5}, 2^{-4}, \ldots, 2^5$ by maximizing the partial likelihood of validation data, which are randomly sampled from and accounts for 20% of the training data. For the random survival forests (RF), we set the number of trees equal to 1000 and the number of splits equal to 2 if $\boldsymbol{x}_i \in \mathbb{R}^3$ and equal to 4 if $\boldsymbol{x}_i \in \mathbb{R}^{10}$, which are roughly equal to the square root of the covariate dimensions. For the DeepHit and the piecewise constant hazards (PCH) models, we discretize the continuous time into 20 intervals of an equal length, in each of which the survival or hazard function is constant, and use a feedforward neural network with the ReLU activation functions and two hidden layers, each of which has 20 nodes. Early stopping is implemented by incorporating

Algorithm 1 Simulation of the survival data with time-varying covariates.

Input: Number of subjects n, number of competing events J, right censoring time $T_{r.c.}$, covariate distribution $P_{\boldsymbol{x}}$, maximum number of covariate updates $L \in \mathbb{Z}_+$, potential covariate update times $\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(L)}$ with $\tau^{(L)} < T_{r.c.}$, Weibull parameters a and $\{\lambda_j(\boldsymbol{x})\}_{j=1}^J$

Output: $\{t_i, y_i, \boldsymbol{x}_i^{(0)}, \ldots, \boldsymbol{x}_i^{(L_i)}, \tau_i^{(0)}, \ldots, \tau_i^{(L_i)}\}_{i=1}^n$

for $i = 1, \ldots, n$ **do**
  $l \leftarrow -1$
  while $l < L$ **do**
    $l \leftarrow l + 1$
    Draw $\boldsymbol{x}_i^{(l)} \sim P_{\boldsymbol{x}}$
    $t_{ij} \sim \text{Weibull}_{\tau^{(l)}}(a, \lambda_j(\boldsymbol{x}_i^{(l)}))$ or $\text{logNormal}_{\tau^{(l)}}(\mu_j(\boldsymbol{x}_i^{(l)})), \sigma^2),\ j = 1, \ldots, J$
    if $\min_j t_{ij} < \tau^{(l+1)}$ **then**
      $t_i \leftarrow \min_j t_{ij}$, $y_i \leftarrow \text{argmin}_j\, t_{ij}$
      break
    end if
  end while
  if $t_i > T_{r.c.}$ **then**
    $t_i \leftarrow T_{r.c.}$, $y_i \leftarrow 0$ `# 0 indicates right censoring`
  end if
  $L_i \leftarrow l$
end for

a validation set, which is the same as those for the KFG model. We use `R` for the MCMC algorithm of WDR and the package `riskRegression` (Gerds and Scheike, 2015) for FG, the package `CoxBoost` (Binder, 2013) for the gradient boosting method in KFG, and the package `randomForestSRC` (Ishwaran and Kogalur, 2018) for RF. For DeepHit and PCH, we use the `Python` package `pycox`[4].

Appendix D. Supplementary Results

D.1 Brier Scores on Synthetic Data with Constant Covariates and No Left Truncation

We take 20 random training-testing partitions of each synthetic data in Section 4.3.1, and report in Tables 10 and 11 the Brier scores (mean $\pm$ standard deviation) on the testing data. A smaller Brier score indicates a better model fit. On data 1 with linear covariate effects, WDR, FG, KFG, and RF are comparable. On other data with nonlinear covariate effects, FG does not perform well, and WDR, KFG, and RF are comparable. Note that DeepHit and PCH have a large variation in prediction accuracy over different time and data, possibly because their piecewise constant survival/hazard functions can be sensitive to time discretization.

4. `https://github.com/havakv/pycox`. Last access in February 2023.

Data 1	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.123**±0.006	**0.19**±0.004	**0.207**±0.004	**0.213**±0.004	**0.216**±0.004
FG	0.124±0.007	0.192±0.004	0.209±0.004	0.216±0.004	0.217±0.004
KFG	0.124±0.007	0.193±0.004	0.21±0.004	0.216±0.004	0.218±0.004
RF	0.125±0.007	0.198±0.003	0.22±0.003	0.226±0.003	0.231±0.002
DeepHit	0.148±0.009	0.288±0.009	0.224±0.003	0.234±0.003	0.243±0.004
PCH	0.128±0.007	0.205±0.005	0.225±0.006	0.232±0.004	0.236±0.005
Data 2	$t = 0.4$	$t = 0.55$	$t = 0.7$	$t = 0.85$	$t = 1$
WDR	**0.052**±0.004	**0.097**±0.005	**0.142**±0.005	**0.164**±0.005	**0.18**±0.005
FG	0.054±0.005	0.113±0.006	0.169±0.005	0.199±0.005	0.216±0.004
KFG	0.054±0.005	0.112±0.006	0.146±0.005	0.175±0.005	0.192±0.004
RF	0.054±0.005	0.109±0.006	0.144±0.005	0.173±0.005	0.19±0.004
DeepHit	0.054±0.005	0.108±0.005	0.161±0.005	0.188±0.004	0.206±0.004
PCH	0.053±0.005	0.108±0.006	0.161±0.006	0.198±0.007	0.221±0.007
Data 3	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.095**±0.004	**0.175**±0.005	**0.186**±0.004	**0.194**±0.005	**0.199**±0.004
FG	0.107±0.004	0.208±0.004	0.235±0.003	0.245±0.002	0.249±0.002
KFG	0.105±0.004	0.18±0.004	0.201±0.002	0.209±0.002	0.213±0.002
RF	0.103±0.004	0.179±0.004	0.203±0.002	0.214±0.002	0.219±0.002
DeepHit	0.121±0.006	0.289±0.008	0.364±0.009	0.225±0.004	0.235±0.004
PCH	0.101±0.004	0.188±0.006	0.2±0.006	0.231±0.006	0.24±0.006
Data 4	$t = 0.3$	$t = 0.45$	$t = 0.6$	$t = 0.75$	$t = 0.9$
WDR	0.02±0.002	0.046±0.003	0.076±0.003	**0.107**±0.004	**0.115**±0.005
FG	0.026±0.004	0.075±0.005	0.125±0.005	0.17±0.004	0.189±0.005
KFG	0.024±0.003	**0.045**±0.005	0.086±0.004	0.122±0.003	0.134±0.003
RF	0.026±0.003	0.055±0.005	0.107±0.004	0.151±0.003	0.17±0.002
DeepHit	0.022±0.003	0.055±0.004	0.095±0.003	0.129±0.004	0.142±0.003
PCH	**0.019**±0.002	0.046±0.003	**0.072**±0.004	0.113±0.005	0.129±0.005
Data 5	$t = 0.9$	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$
WDR	**0.097**±0.004	0.145±0.004	**0.147**±0.003	**0.139**±0.003	**0.133**±0.003
FG	0.105±0.005	0.192±0.005	0.227±0.003	0.241±0.002	0.246±0.001
KFG	0.1±0.005	0.163±0.005	0.187±0.003	0.194±0.001	0.197±0.001
RF	0.098±0.005	0.15±0.004	0.168±0.003	0.179±0.002	0.181±0.002
DeepHit	0.099±0.006	0.167±0.006	0.173±0.006	0.168±0.006	0.163±0.008
PCH	**0.097**±0.006	**0.144**±0.005	0.148±0.004	0.154±0.005	0.19±0.007
Data 6	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$	$t = 1.4$
WDR	0.077±0.004	0.098±0.005	0.098±0.005	0.092±0.003	0.098±0.003
FG	0.103±0.004	0.19±0.004	0.213±0.004	0.223±0.004	0.234±0.003
KFG	0.094±0.004	0.14±0.004	0.158±0.004	0.164±0.003	0.173±0.003
RF	0.094±0.004	0.138±0.004	0.158±0.004	0.164±0.003	0.175±0.003
DeepHit	0.099±0.005	0.099±0.004	0.068±0.003	**0.067**±0.003	**0.072**±0.003
PCH	**0.074**±0.004	**0.061**±0.004	**0.067**±0.004	0.096±0.006	0.129±0.006

Table 10: BS for event 1 of synthetic data with constant covariates and no left truncation.

	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.117**±0.006	**0.177**±0.004	**0.203**±0.004	**0.21**±0.004	**0.213**±0.004
FG	0.118±0.006	0.179±0.004	0.204±0.004	0.211±0.004	0.215±0.004
KFG	0.118±0.006	0.179±0.004	0.205±0.004	0.212±0.004	0.216±0.004
RF	0.121±0.006	0.184±0.005	0.213±0.004	0.221±0.003	0.225±0.003
DeepHit	0.143±0.009	0.26±0.01	0.22±0.004	0.229±0.004	0.236±0.004
PCH	0.125±0.007	0.191±0.006	0.22±0.007	0.23±0.007	0.235±0.006
Data 2	$t = 0.4$	$t = 0.55$	$t = 0.7$	$t = 0.85$	$t = 1$
WDR	**0.078**±0.005	**0.15**±0.006	0.213±0.003	0.235±0.003	**0.229**±0.004
FG	0.081±0.006	0.155±0.006	0.226±0.003	0.251±0.001	0.25±0.002
KFG	0.081±0.006	0.154±0.006	0.214±0.003	0.238±0.001	0.238±0.001
RF	0.08±0.006	0.154±0.006	**0.209**±0.003	**0.234**±0.001	0.234±0.002
DeepHit	0.081±0.006	0.156±0.007	0.231±0.005	0.255±0.003	0.253±0.002
PCH	0.082±0.006	0.158±0.008	0.229±0.005	0.248±0.002	0.249±0.002
Data 3	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
WDR	**0.104**±0.005	**0.166**±0.005	**0.184**±0.005	**0.19**±0.004	**0.195**±0.004
FG	0.119±0.006	0.212±0.004	0.237±0.002	0.246±0.002	0.25±0.002
KFG	0.115±0.006	0.178±0.004	0.202±0.003	0.209±0.002	0.214±0.002
RF	0.114±0.006	0.183±0.004	0.208±0.002	0.217±0.001	0.221±0.001
DeepHit	0.137±0.008	0.3±0.009	0.371±0.009	0.224±0.003	0.234±0.004
PCH	0.11±0.007	0.169±0.005	0.192±0.005	0.211±0.005	0.219±0.006
Data 4	$t = 0.3$	$t = 0.45$	$t = 0.6$	$t = 0.75$	$t = 0.9$
WDR	0.024±0.002	0.042±0.003	0.074±0.003	0.102±0.004	**0.102**±0.004
FG	0.041±0.003	0.084±0.003	0.138±0.005	0.175±0.005	0.182±0.005
KFG	0.037±0.002	0.052±0.003	0.096±0.004	0.119±0.003	0.126±0.003
RF	0.037±0.003	0.056±0.003	0.111±0.004	0.15±0.003	0.165±0.002
DeepHit	0.028±0.002	0.053±0.003	0.093±0.004	0.114±0.003	0.124±0.003
PCH	**0.023**±0.002	**0.037**±0.002	**0.068**±0.004	**0.095**±0.004	0.106±0.006
Data 5	$t = 0.9$	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$
WDR	0.091±0.005	0.147±0.005	0.151±0.005	**0.135**±0.003	**0.129**±0.002
FG	0.092±0.006	0.188±0.006	0.234±0.004	0.246±0.003	0.251±0.002
KFG	0.089±0.006	0.164±0.006	0.195±0.004	0.2±0.003	0.201±0.002
RF	0.089±0.006	0.156±0.006	0.181±0.003	0.185±0.002	0.187±0.001
DeepHit	0.089±0.006	0.165±0.007	0.178±0.008	0.164±0.008	0.16±0.009
PCH	**0.084**±0.006	**0.144**±0.007	**0.149**±0.007	0.142±0.008	0.183±0.01
Data 6	$t = 1$	$t = 1.1$	$t = 1.2$	$t = 1.3$	$t = 1.4$
WDR	0.09±0.004	0.118±0.004	0.103±0.003	0.094±0.003	0.1±0.003
FG	0.106±0.006	0.218±0.005	0.232±0.004	0.238±0.004	0.247±0.003
KFG	0.1±0.005	0.156±0.004	0.164±0.004	0.168±0.003	0.174±0.002
RF	0.098±0.005	0.154±0.004	0.166±0.004	0.174±0.003	0.186±0.003
DeepHit	0.104±0.006	0.126±0.004	0.07±0.003	**0.067**±0.002	**0.075**±0.003
PCH	**0.082**±0.004	**0.061**±0.004	**0.055**±0.004	0.068±0.005	0.094±0.005

Table 11: BS for event 2 of synthetic data with constant covariates and no left truncation.

$\boldsymbol{\beta}_{jk}$	ABC ($j=1$)	GCB ($j=2$)	T3 ($j=3$)	
Gene #	$k=1$	$k=1$	$k=1$	$k=2$
17482	1.943	0.060	-0.256	-0.497
24432	0.715	0.025	0.248	0.414
17833	0.795	0.128	-0.229	-0.258
28193	0.096	0.090	-0.200	-0.077
28197	0.031	-0.719	-0.021	0.109
27731	-0.015	-0.081	-1.481	1.059
31456	0.246	0.654	-2.871	-5.568
r_{jk}	0.122	0.131	0.030	0.031

Table 12: $\boldsymbol{\beta}_{jk}$ and r_{jk} of WDR for the analysis of DLBCL.

D.2 WDR Estimation on DLBCL and Model Comparison

We analyze the DLBCL data by WDR and estimate the model by MCMC. The posterior mean of the Weibull shape parameter a is 1.379 with the posterior standard deviation 0.151. Provided in Table 12 are the posterior means of $\boldsymbol{\beta}_{jk}$'s and r_{jk}'s. We see two sub-events under T3 ($j=3$), indicating two potential disease subtypes of T3, and each is linearly accelerated by the expression of the seven genes. We compare the performance of WDR, FG, KFG, RF, DeepHit, and PCH on the DLBCL data. To reduce the impact on the training-testing partition, we randomly take 20 partitions, in each of which 200 subjects are used for training and the other 40 for testing. We report the Brier scores in Table 13. As a result, no model consistently outperforms the others, but WDR has a good performance that is comparable to RF. Notably, DeepHit and PCH do not work well and seem to overfit this smaller data, though validation set and early stopping are used.

		t=0.5	t=1	t=1.5	t=2	t=2.5	t=3
ABC	WDR	0.11±0.009	0.172±0.01	0.195±0.011	0.202±0.01	0.223±0.01	**0.223**±0.009
	FG	**0.108**±0.009	0.172±0.01	0.195±0.011	0.203±0.01	0.224±0.01	0.226±0.009
	KFG	0.126±0.01	0.187±0.013	0.201±0.014	0.215±0.013	0.225±0.014	0.257±0.012
	RF	0.111±0.009	**0.168**±0.01	**0.191**±0.011	**0.199**±0.01	**0.217**±0.01	**0.223**±0.008
	DeepHit	0.122±0.01	0.211±0.012	0.221±0.013	0.231±0.012	0.25±0.011	0.268±0.01
	PCH	0.119±0.01	0.197±0.012	0.233±0.013	0.242±0.013	0.265±0.013	0.277±0.011
GCB	WDR	**0.071**±0.007	**0.134**±0.008	0.196±0.012	0.205±0.011	0.221±0.011	0.247±0.01
	FG	0.072±0.007	0.138±0.008	0.2±0.011	0.213±0.012	0.231±0.011	0.255±0.01
	KFG	0.073±0.006	0.135±0.006	**0.19**±0.009	**0.196**±0.009	**0.212**±0.008	**0.232**±0.007
	RF	**0.071**±0.007	0.135±0.008	0.196±0.012	0.207±0.012	0.224±0.012	0.244±0.01
	DeepHit	0.074±0.008	0.148±0.009	0.214±0.013	0.23±0.013	0.243±0.013	0.276±0.013
	PCH	0.075±0.008	0.146±0.008	0.219±0.014	0.232±0.014	0.255±0.014	0.285±0.014
T3	WDR	**0.081**±0.008	**0.11**±0.009	0.143±0.012	**0.145**±0.012	**0.158**±0.011	**0.157**±0.011
	FG	0.087±0.008	0.129±0.009	0.163±0.012	0.165±0.011	0.168±0.011	0.168±0.011
	KFG	0.083±0.007	0.118±0.006	0.163±0.008	0.177±0.008	0.189±0.007	0.205±0.008
	RF	0.085±0.008	0.113±0.009	**0.141**±0.012	0.151±0.012	0.161±0.011	0.161±0.011
	DeepHit	0.087±0.009	0.119±0.01	0.153±0.013	0.172±0.013	0.172±0.011	0.172±0.011
	PCH	0.088±0.009	0.122±0.009	0.157±0.013	0.174±0.012	0.185±0.012	0.185±0.012

Table 13: Model comparison on DLBCL in Brier scores (mean ± standard error).

	MCI			Death		
	Age=80	Age=85	Age=90	Age=80	Age=85	Age=90
WDR	0.157	0.175	0.175	0.020	0.021	0.022
FG	0.179	0.157	0.151	0.017	0.020	0.028
KFG	0.242	0.265	0.264	0.020	0.021	0.025

Table 14: Brier scores for MCI and death of other causes.

D.3 Model Comparison on the AD Data

We compare the performance of WDR, FG, and KFG on the testing set of the Alzheimer's disease data. Since RF, DeepHit, and PCH are not designed for left truncation and time-varying covariates, we do not run these models. We report the Brier scores in Table 14, which are evaluated at ages 80, 85, and 90. As a result, WDR is comparable to FG and outperforms KFG in predicting MCI.

Appendix E. WDR for Classification and Additional Experiments

We show WDR as a nonlinear discrete choice model for classification problems. Considering the existence of many sophisticated classification models, our goal is not the state-of-the-art prediction accuracy, but to provide an alternative approach for interpretable classification as a supplement to the linear ones, like probit and logistic regressions. Weibull (delegate) racing can be regarded as a discrete choice model where the decision of categorization is made to minimize the waiting time for the arrival of the first candidate choice, which is $y = \operatorname{argmin}_j t_j$. Therefore, WDR classification inherits all the advantages of the WDR survival model such as data-adaptive nonlinearity and interpretability as shown in Section 4. In practice, we adopt finite truncation of the gamma processes of WDR classification by allowing each category $j \in 1, \ldots, J$ to consist of up to K subtypes. Consequently, the probability of y given $\{\lambda_{jk}\}_{j,k}$ is

$$\Pr(y = j \mid \{\lambda_{jk}\}_{j,k}) = \frac{\sum_{k=1}^{K} \lambda_{jk}}{\sum_{j'=1}^{J} \sum_{k'=1}^{K} \lambda_{j'k'}} \tag{12}$$

where $\lambda_{jk} \sim \text{Gamma}(r_{jk}, \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_{jk}))$. This probability can be estimated by Monte Carlo methods if we have point or posterior estimates of r_{jk}'s and $\boldsymbol{\beta}_{jk}$'s. Note that (12) does not depend on the Weibull shape parameter a. So we fix a at an arbitrary constant value without sacrificing modeling capacity (we set $a = 1$ for all the experiments of WDR classification). The inference by MCMC or MAP follows the same algorithm as for the WDR survival model as if all the event times are missing, except that the step of estimating a is skipped. In this way, the MCMC for WDR classification turns out to be a Gibbs sampler. For identifiability and good mixing of MCMC, we should fix one of the $\boldsymbol{\beta}_{jk}$'s, say $\boldsymbol{\beta}_{11}$, equal to a zero vector.

E.1 WDR Classification on Toy Data

We first illustrate the data-adaptive nonlinearity of WDR for the classification of *square*, which is synthesized with two-dimensional covariates and $J = 3$ categories. Figure 4 shows the classification of the three categories, two rectangles (denoted by black and grey dots) in

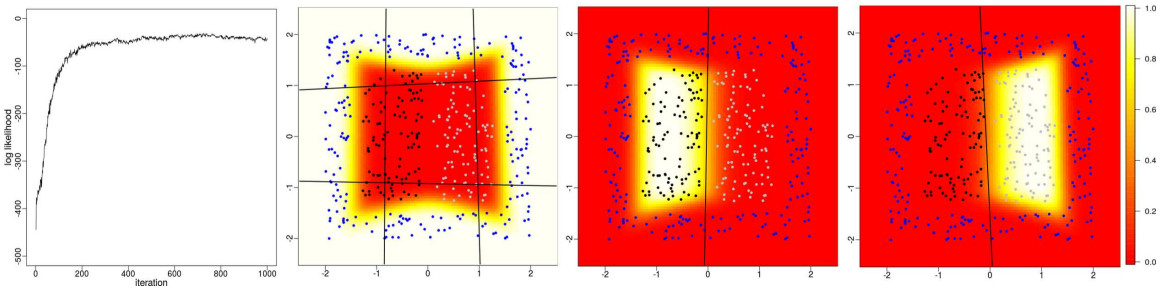


Figure 4: Trajectory plots of log likelihood (column 1) and predictive probability heat maps and hyperplanes for category 1 to 3 (column 2 to 4) of *square* data. Blue points are labeled as category 1, black 2 and gray 3.

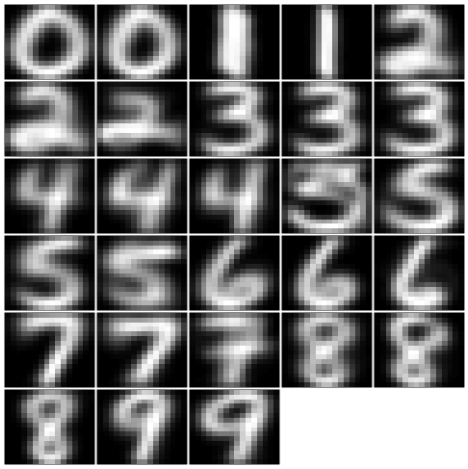


Figure 5: Subtypes of *usps* handwritten numbers by WDR.

one square frame (blue dots), by MCMC with K set to be 10. The panel on the left gives the trajectory plot of the log likelihood by MCMC iterations, and the other three panels show the heat maps of predictive probabilities of category 1 (blue dots), 2 (black dots), and 3 (gray dots), respectively. The solid lines on these three panels denote the hyperplanes $\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{jk} = 0$ where $\hat{\boldsymbol{\beta}}_{jk}$ is the estimated posterior means of $\boldsymbol{\beta}_{jk}$. We see the three categories have been perfectly separated within 1,000 MCMC iterations and WDR have found four subtypes of category 1 and one subtype for category 2 or 3, respectively. For illustration purpose, we do not care for identifiability and have not fixed any $\boldsymbol{\beta}_{jk}$ equal to zero in Figure 4.

We further visualize WDR's capability of finding subtypes on data *usps* that have handwritten numbers zero to nine. Figure 5 shows the subtypes of each number found by WDR. Specifically, we visualize $\sum_i \frac{\hat{\lambda}_{ijk}}{\hat{\lambda}_{ijk}+\sum_{j'\neq j}\sum_{k'}\hat{\lambda}_{ij'k'}}\boldsymbol{x}_i$, a weighted average of all the training samples $\boldsymbol{x}_i$'s, for subtype k of category j where $\hat{\lambda}_{ijk}$ is the estimated posterior mean of λ_{ijk}. We see that the number five has four subtypes while the number zero, one, or nine has only two subtypes, indicating the different amount of nonlinear capacity is required when depicting the classification boundaries.

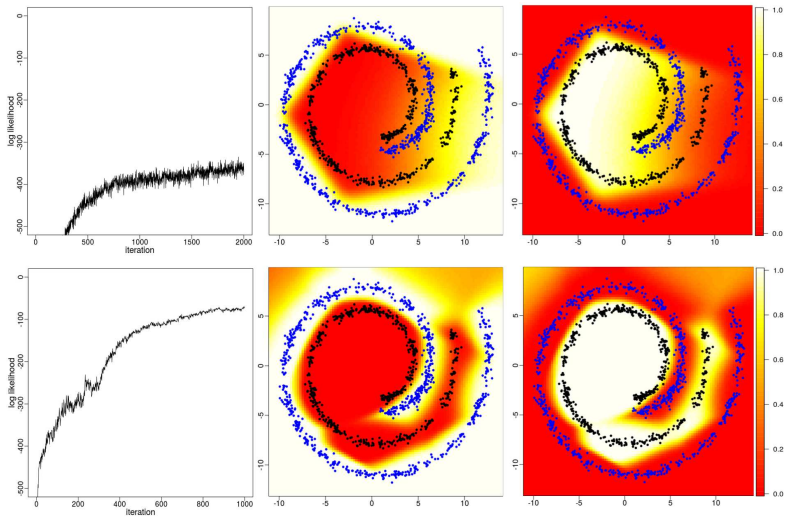


Figure 6: Trajectory plot of log likelihood and predictive probability heat map for *swiss roll* data. Blue points are labeled as category 1 and black 2. Row 1 results from WDR with the original covariates and row 2 with data transformation as in (13).

E.2 Improving Nonlinear Capacity by Data Transformation

If used in classification settings, WDR is not primarily focused on highly accurate prediction for big and complex data. Instead, it is advantageous in interpretation and data-adaptive nonlinearity; WDR can be ideal if one wants to discover subtypes of a category or evaluate how much nonlinearity is relatively needed for each category. In case WDR is applied to reasonably complex classification problems, we propose a data transformation scheme to enhance the nonlinear capacity and make its classification accuracy comparable to kernel support vector machines on moderate-sized data sets.

Unlike kernel methods such as support vector machines (Boser et al., 1992; Cristianini and Shawe-Taylor, 2000) and the relevance vector machine (Tipping, 2001) which make categories more linearly separable in a higher-dimensional transformed covariate space, or neural networks using complex data transformations, WDR uses interactions of linear hyperplanes to construct nonlinear decision boundaries, and hence may have insufficient capacity if the class boundaries are highly complex. To tackle such a problem we can stack another WDR on a previously trained one to enhance its capacity, which is a similar strategy of Zhang and Zhou (2017). Concretely, we first run a WDR to obtain a finite set of hyperplanes denoted as $\tilde{\boldsymbol{\beta}}_{jk}$, and then augment the original covariates $\boldsymbol{x}_i$ into

$$\tilde{\boldsymbol{x}}_i := \left[\boldsymbol{x}_i^{\mathrm{T}}, \log\left(1+\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{11}\right)\right), \log\left(1+\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{12}\right)\right), \cdots, \log\left(1+\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{SK}\right)\right)\right]^{\mathrm{T}}, \tag{13}$$

and then run another WDR with the transformed covariate $\tilde{\boldsymbol{x}}_i$.

We show the increased capacity of WDR using the data transformation on a synthetic 2-D *swiss roll* data in Figure 6. Row 1 shows the results using the data with original covariates and row 2 using the transformed data by (13) where $\tilde{\boldsymbol{\beta}}_{jk}$ are from the estimation

	iris	wine	glass	vehicle	waveform	segment	dna	satimage	usps	mnist
Train size	120	142	171	592	500	231	2000	4435	7291	60000
Test size	30	36	43	254	4500	2079	1186	2000	2007	10000
Covariate number	4	13	9	18	21	19	180	36	256	784
Category number	3	3	6	4	3	7	3	6	10	10

Table 15: Multiclass datasets used in experiments for model comparison.

	WDR	WDR w. data trnsf	L_2-MLR	SVM
iris	4.00±2.79	**2.67**±2.79	3.33±3.33	4.00±1.83
wine	3.89±3.17	3.33±3.04	3.89±3.17	**2.78**±1.47
glass	29.77±8.76	**26.98**±8.32	33.02±3.82	28.84±5.74
vehicle	17.24±1.16	**14.44**±0.23	22.83	18.50
waveform	**14.39**±0.29	14.95±0.26	15.60	15.22
segment	7.09±0.65	6.62±0.16	8.56	**6.20**
dna	**4.06**±0.28	4.69±0.22	5.98	4.97
satimage	11.42±0.20	9.37±0.26	17.80	**8.50**
usps	5.98±0.17	5.64±0.14	8.47	**4.78**
mnist	2.71±0.17	1.78±0.22	7.40	**1.48**

Table 16: Comparison of prediction error rate (%).

in row 1. The improvement is remarkable in not only in-sample fits reflected by log likelihood trajectory plots but also in the out-of-sample prediction shown in the heat maps.

E.3 Model Comparison

We compare the classification performance of WDR on real data sets. Table 15 summarizes the data sizes and the numbers of covariates and categories. The training and testing sets are predefined for *vehicle*, *dna*, *satimage*, *usps* and *mnist* where the validation sets are merged into training. We divide the other data sets into training and testing as follows. For *iris*, *wine*, and *glass*, five random partitions are taken such that for each partition the training set accounts for 80% of the whole data while the testing set accounts for 20%. The classification error rate is calculated by averaging the error rates of all the five partitions. For *waveform* and *segment*, one random partition is taken and 10% of the data points are used for training and the other 90% for testing.

We compare the WDR classification model with an L_2-regularized multinomial logistic regression (L_2-MLR) and support vector machines (SVMs) with radial basis function (RBF) kernels. An observation in testing sets is classified into the category associated with the largest predictive probability if provided by the model. For WDR models we use the Monte-Carlo average for predictive probabilities and report the mean and standard deviation. For the WDR with data transformation, we first run WDR with $K = 10$, using the original training data to obtain $\tilde{\boldsymbol{\beta}}_{jk}$'s, and then we run another WDR with $K = 10$ with the transformation of (13).

We use the `R` package `LiblineaR` (Helleputte, 2015) for L_2-MLR where a bias term is included and the regularization parameter is selected by a five-fold cross-validation on

the training set from $(2^{-10}, 2^{-9}, \ldots, 2^{15})$. For SVMs, we use the LIBSVM (Chang and Lin, 2011) provided by the `R` package `e1071` (Meyer et al., 2015). A Gaussian RBF kernel is used and a three-fold cross validation is adopted to tune both the regularization parameter and kernel width from $(2^{-10}, 2^{-9}, \ldots, 2^{10})$ on the training set. We choose the default settings for all the other parameters.

We report the classification error rates on the testing sets in Table 16 together with standard errors by all the models on *iris*, *wine*, and *glass* (recall that we have five training-testing partitions), as well as the standard errors by WDR on other data sets. We can see the classification accuracy by WDR is comparable to fine-tuned SVMs with RBF kernels.

References

Ahmed M. Alaa and Mihaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Advances in Neural Information Processing Systems*, pages 2326–2334, 2017.

Marilyn Albert, Anja Soldan, Rebecca Gottesman, Guy McKhann, Ned Sacktor, Leonie Farrington, Maura Grega, Raymond Turner, Yi Lu, Shanshan Li, et al. Cognitive changes preceding clinical symptom onset of mild cognitive impairment and relationship to ApoE genotype. *Current Alzheimer Research*, 11(8):773–784, 2014.

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.

Peter C Austin, Aurélien Latouche, and Jason P Fine. A review of the use of time-varying covariates in the Fine-Gray subdistribution hazard competing risk regression model. *Statistics in medicine*, 39(2):103–113, 2020.

James E Barrett and Anthony CC Coolen. Gaussian process regression for survival data with competing risks. *arXiv preprint arXiv:1312.1591*, 2013.

Harald Binder. *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks*, 2013. URL `https://CRAN.R-project.org/package=CoxBoost`. R package version 1.4.

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, 2009.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Daniel Commenges, Luc Letenneur, Pierre Joly, Ahmadou Alioum, and Jean-François Dartigues. Modelling age-specific risk: application to dementia. *Statistics in medicine*, 17 (17):1973–1988, 1998.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

Martin J Crowder. *Classical competing risks*. CRC Press, 2001.

Paul Damien, John Wakefield, and Stephen Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

Coraline Danieli and Michal Abrahamowicz. Competing risks modeling of cumulative effects of time-varying drug exposures. *Statistical methods in medical research*, 28(1):248–262, 2019.

Ludger Evers and Claudia-Martina Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, 2008.

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.

R. G. FitzJohn. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, in press, 2012. doi: 10.1111/j.2041-210X.2012.00234.x.

Thomas A Gerds, Tianxi Cai, and Martin Schumacher. The performance of risk prediction models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(4): 457–479, 2008.

Thomas Alexander Gerds and Thomas Harder Scheike. *riskRegression: Risk Regression Models for Survival Analysis with Competing Risks*, 2015. URL `https://CRAN.R-project.org/package=riskRegression`. R package version 1.1.7.

William H Greene. *Econometric analysis*. Pearson Education India, 2003.

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

W. M. Hanemann. Discrete/continuous models of consumer demand. *Econometrica: Journal of the Econometric Society*, pages 541–561, 1984.

Thibault Helleputte. *LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library*, 2015. R package version 1.94-2.

H. Ishwaran and U.B. Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2018. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.6.0.

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, 2014.

James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73), 2020.

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

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.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

Havard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20:1–30, 2019.

Bryan Lau, Stephen R Cole, and Stephen J Gange. Parametric mixture models to evaluate and summarize hazard ratios in the presence of competing risks with time-dependent hazards and delayed entry. *Statistics in medicine*, 30(6):654–665, 2011.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.

Linda E Lévesque, James A Hanley, Abbas Kezouh, and Samy Suissa. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *Bmj*, 340, 2010.

Hongzhe Li and Yihui Luan. Kernel cox regression models for linking gene expression profiles to censored survival data. In *Biocomputing 2003*, pages 65–76. World Scientific, 2002.

Hongzhe Li and Yihui Luan. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10):2403–2409, 2005.

Mary Lunn and Don McNeil. Applying Cox regression to competing risks. *Biometrics*, pages 524–532, 1995.

Sarah F McGough, Devin Incerti, Svetlana Lyalina, Ryan Copping, Balasubramanian Narasimhan, and Robert Tibshirani. Penalized regression for left-truncated and right-censored survival data. *Statistics in Medicine*, 40(25):5487–5500, 2021.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015. URL `https://CRAN.R-project.org/package=e1071`. R package version 1.6-7.

Peter G Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(1):541–544, 1985.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021.

Radford M Neal et al. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

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.

MA Nicolaie, Hans C van Houwelingen, and Hein Putter. Vertical modeling: a pattern mixture approach for competing risks modeling. *Statistics in medicine*, 29(11):1190–1205, 2010.

Mioara Alina Nicolaie, Jeremy MG Taylor, and Catherine Legrand. Vertical modeling: analysis of competing risks data with a cure fraction. *Lifetime data analysis*, 25(1):1–25, 2019.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American statistical Association*, 108 (504):1339–1349, 2013.

Ross L Prentice, John D Kalbfleisch, Arthur V Peterson Jr, Nancy Flournoy, Vern T Farewell, and Norman E Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554, 1978.

Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

Md Mahmudur Rahman, Koji Matsuo, Shinya Matsuzaki, and Sanjay Purushotham. Deeppseudo: Pseudo value based deep learning models for competing risk analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 479–487, 2021.

Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena 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*, 346(25):1937–1947, 2002.

Christine Sattler, Pablo Toro, Peter Schönknecht, and Johannes Schröder. Cognitive activity, education and socioeconomic status as preventive factors for mild cognitive impairment and alzheimer's disease. *Psychiatry research*, 196(1):90–95, 2012.

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.

Yvonne H Sparling, Naji Younes, John M Lachin, and Oliver M Bautista. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics*, 7(4):599–614, 2006.

Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

Michael E Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001.

Donna Tjandra, Yifei He, and Jenna Wiens. A hierarchical approach to multi-event survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 591–599, 2021.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge University Press, 2009.

Quan Zhang and Mingyuan Zhou. Permuted and augmented stick-breaking Bayesian multinomial regression. *The Journal of Machine Learning Research*, 18(1):7479–7511, 2017.

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.

Mingyuan Zhou. Softplus regressions and convex polytopes. *arXiv:1608.06383*, 2016.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.

Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *Journal of Machine Learning Research*, 17(163):1–44, 2016.