Wasserstein-Kaplan-Meier Survival Regression

Yidong Zhou
Department of Statistics, University of California, Davis
and
Hans-Georg Müller
Department of Statistics, University of California, Davis

Abstract

Survival analysis plays a pivotal role in medical research, offering valuable insights into the timing of events such as survival time. One common challenge in survival analysis is the necessity to adjust the survival function to account for additional factors, such as age, gender, and ethnicity. We propose an innovative regression model for right-censored survival data across heterogeneous populations, leveraging the Wasserstein space of probability measures. Our approach models the probability measure of survival time and the corresponding non-parametric Kaplan-Meier estimator for each subgroup as elements of the Wasserstein space. The Wasserstein space provides a flexible framework for modeling heterogeneous populations, allowing us to capture complex relationships between covariates and survival times. We address an underexplored aspect by deriving the non-asymptotic convergence rate of the Kaplan-Meier estimator to the underlying probability measure in terms of the Wasserstein metric. The proposed model is supported with a solid theoretical foundation including pointwise and uniform convergence rates, along with an efficient algorithm for model fitting. The proposed model effectively accommodates random variation that may exist in the probability measures across different subgroups, demonstrating superior performance in both simulations and two case studies compared to the Cox proportional hazards model and other alternative models.

Keywords: Fréchet mean, heterogeneous populations, optimal transport, survival analysis, Wasserstein space

1 Introduction

Survival analysis (Kalbfleisch and Prentice, 2002) is a critical component of medical research, providing insights into the temporal aspects of various events, such as patient survival in healthcare settings. A frequently encountered problem when analyzing survival data is the need to adjust the survival function in order to account for concomitant information (e.g., age, gender and ethnicity). This is particularly relevant when studying heterogeneous populations, which are prevalent in clinical trials, cohort studies, and observational studies. For example, one may want to compare the survival functions for two or more treatments and determine the prognosis of a patient presenting with various characteristics, while controlling for relevant confounders.

In the statistical literature, several approaches for modeling covariate effects on survival are well established, including the Accelerated Failure Time (AFT) model (Pike, 1966; Wei, 1992) and the Cox Proportional Hazards (CPH) model (Cox, 1972; Kalbfleisch and Schaubel, 2023). The AFT model posits that a covariate's effect accelerates or decelerates the progression of a disease by a constant factor. The CPH model provides a semi-parametric specification of the hazard function and has served as the cornerstone for survival analysis, offering a robust framework for examining the effects of covariates on time-to-event outcomes. Nevertheless, it hinges on the proportional hazards assumption, which presumes that covariates exhibit a multiplicative relationship with the hazard function. In practice, this assumption may not hold, potentially leading to inaccurate estimates of covariate effects on survival times (Babińska et al., 2015).

In this study, we introduce an innovative and unified approach for analyzing right-censored survival data for cohorts that share the same (discrete) covariates. We achieve this by treating the survival distribution and the corresponding non-parametric Kaplan-Meier (KM) survival curve (Kaplan and Meier, 1958) as elements of the Wasserstein space, the space of one-dimensional distributions equipped with the Wasserstein metric. The

Wasserstein space is a geodesic metric space related to optimal transport (Villani, 2003) and is increasingly applied in various research domains, including population pyramids (Bigot et al., 2017) and financial returns (Zhang et al., 2022), among others (Petersen et al., 2022). Within the field of survival analysis, the Wasserstein metric was initially employed as a dissimilarity measure for comparing KM survival curves when constructing survival trees (Gordon and Olshen, 1985). However, its potential remained largely unexplored until recent work by Sylvain et al. (2021), which demonstrated the Wasserstein metric's utility for imputing survival times in the presence of censored data.

Our research focuses on studying survival times in heterogeneous populations. We treat the survival function as one of the various manifestations of an underlying probability measure, whereas other representations include the cumulative distribution function, density function, or hazard function. After converting survival functions into probability measures, we treat these measures as responses in a new regression model where the predictors are the covariate levels (or discrete characteristics) associated with each survival function. The proposed regression model is inspired by recently developed regression models that feature metric-space-valued responses and Euclidean predictors (Petersen and Müller, 2019). One option for the space in which the responses are situated that we explore here is the Wasserstein space. While normally the responses in such models are considered to be fully observed, in practical data scenarios this is rarely the case. Instead, one has available samples of independent data generated by the probability measure of interest and one can then use empirical measures as surrogates for the unobservable probability measures when there is no censoring involved (Zhou and Müller, 2023). This strategy offers the unique advantage of accommodating varying sample sizes for different subgroups, some of which may contain as few as one observation. However, in the context of survival data, commonly only partial observations of survival times are available due to the presence of random right censoring. Consequently, the empirical measure is not directly applicable. The proposed

regression model bridges this gap by employing the KM estimator as a surrogate for the unobservable survival distribution.

This paper provides three key innovations: First, we derive the non-asymptotic convergence rate of the KM estimator to the underlying survival distribution in terms of the Wasserstein metric. Second, we obtain pointwise and uniform convergence rates of the estimated regression function when dealing with sparsely populated subgroups, even those with as few as one observation. Third, we develop an efficient algorithm inspired by the least common multiple for model fitting, enhancing the applicability of the proposed model. Further innovations include the introduction of Wasserstein/optimal transport methods for survival analysis, which are particularly suited for the modeling of survival times for heterogeneous populations. The proposed model effectively accommodates random variation in the probability measures underlying the survival functions across different subgroups. In finite sample scenarios, the proposed regression model is shown to outperform the conventional CPH model and other alternatives. We showcase its utility with the analysis of data from the Women's Interagency HIV Study (WIHS) and the Surveillance, Epidemiology, and End Results (SEER) Program.

The remainder of this paper is organized as follows. In Section 2, we introduce the Wasserstein space of probability measures and the Wasserstein metric. Section 3 presents the proposed regression model. Asymptotic properties, including pointwise and uniform convergence rates, are established in Section 4. Section 5 covers computational details and simulation results. The practical application of the proposed model is illustrated in Section 6 using data from the Women's Interagency HIV Study (WIHS) and the Surveillance, Epidemiology, and End Results (SEER) Program, followed by a brief discussion in Section 7.

2 Preliminaries

We denote the set of probability measures on $(\mathcal{T}, \mathcal{B}(\mathcal{T}))$ with finite second moments as $\mathcal{W}$, where $\mathcal{T}$ is a closed interval in $\mathbb{R}$ and $\mathcal{B}(\mathcal{T})$ is the Borel σ-algebra on $\mathcal{T}$. The space $\mathcal{W}$ is a metric space with the 2-Wasserstein, or simply Wasserstein, metric between two measures $\mu_1, \mu_2 \in \mathcal{W}$, defined as

$$d^2_{\mathcal{W}}(\mu_1, \mu_2) = \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{T} \times \mathcal{T}} |s - t|^2 d\pi(s, t),$$

where $\Pi(\mu_1, \mu_2)$ is the set of joint measures on $\mathcal{T} \times \mathcal{T}$ with marginals μ_1 and μ_2 (Kantorovich, 1942). For any measure $\mu \in \mathcal{W}$ with cumulative distribution function F_μ, we consider the quantile function F^{-1}_μ to be the left continuous inverse of F_μ, i.e., $F^{-1}_\mu(p) = \inf\{t \in \mathcal{T} : F_\mu(t) \geq p\}$, for $p \in [0, 1]$. It is well known (Villani, 2003) that the Wasserstein metric can be expressed as the L^2 distance between the corresponding quantile functions,

$$d^2_{\mathcal{W}}(\mu_1, \mu_2) = \int_0^1 \{F^{-1}_{\mu_1}(p) - F^{-1}_{\mu_2}(p)\}^2 dp. \tag{1}$$

It can be shown that $\mathcal{W}$ endowed with $d_{\mathcal{W}}$ is a complete and separable metric space, the Wasserstein space (Villani, 2003).

We assume there is an underlying probability space $(\mathcal{W}, \mathcal{F}, P)$ that induces a probability measure on the space $\mathcal{W}$ with respect to which we can calculate moments for any random element μ in $\mathcal{W}$. Assuming $E\{d^2_{\mathcal{W}}(\mu, \omega)\} < \infty$ for all $\omega \in \mathcal{W}$, the Fréchet mean of μ (Fréchet, 1948), extending the usual notion of mean, is

$$\mu_\oplus = \arg\min_{\omega \in \mathcal{W}} E\{d^2_{\mathcal{W}}(\mu, \omega)\},$$

which is well-defined and unique as the Wasserstein space $\mathcal{W}$ is a Hadamard space (Kloeckner, 2010). It follows from (1) that the quantile function of the Fréchet mean $\mu_\oplus$ is $F^{-1}_{\mu_\oplus}(\cdot) = E\{F^{-1}_\mu(\cdot)\}$. In the following, we use $S_\mu(\cdot) = 1 - F_\mu(\cdot)$ to denote the survival function of the measure μ. Furthermore, $a \lesssim b$ means that there exists a positive constant C such that $a \leq Cb$ and $a \asymp b$ that $a \lesssim b$ and $b \lesssim a$. The Euclidean norm in $\mathbb{R}^p$ is denoted by $\|\cdot\|_E$.

3 Methodology

3.1 Kaplan-Meier estimator

We present the Kaplan-Meier (KM) estimator within the framework of traditional survival analysis and characterize its non-asymptotic convergence rate in terms of the Wasserstein metric. Let $T_1, \ldots, T_N$ be independent and identically distributed (i.i.d.) survival times with probability measure μ, and let $C_1, \ldots, C_N$ be i.i.d. censoring times that are independent of the T_i with probability measure ν. For right-censored survival data, T_j is censored on the right by C_j, so that we only observe the pairs $\{(Y_j, \delta_j)\}_{j=1}^N$, where $Y_j = \min\{T_j, C_j\}$ and $\delta_j = \mathbf{1}_{T_j \leq C_j}$ is the censoring indicator.

Consider the order statistics $Y_{(1)} \leq \cdots \leq Y_{(N)}$ from $\{Y_j\}_{j=1}^N$ and their associated censoring indicators $\delta_{(j)}$. The KM estimator $S_{\hat{\mu}}(\cdot)$ (Kaplan and Meier, 1958) is well-defined up to the largest observation $Y_{(N)}$. However, if $Y_{(N)}$ is censored, the value of the survival function beyond this point is undetermined. Several methods have been proposed to estimate the survival function beyond this point, with Efron (1967) suggesting a value of 0 and Gill (1980) recommending $S_{\hat{\mu}}(Y_{(N)})$.

In practice, the distinction between these methods is typically inconsequential (Andersen et al., 1993). We conducted simulations and real-data applications implementing both methods and the results indicate that they yield comparable results. However, Gill's approach complicates the calculation of the Wasserstein metric, necessitating adjustments to the integral's upper limit as per (1), when the largest observation is censored. Consequently, we opt for Efron's method and define the KM estimator as

$$S_{\hat{\mu}}(t) = \begin{cases} \prod_{j:Y_{(j)} \leq t} \left(\frac{N-j}{N-j+1}\right)^{\delta_{(j)}} & 0 \leq t < Y_{(N)}, \\ 0 & t \geq Y_{(N)}. \end{cases} \tag{2}$$

The following result formalizes the non-asymptotic convergence rate of the KM estimator to the underlying nonrandom probability measure μ in terms of the Wasserstein metric,

where $S_\nu(\cdot)$ denotes the survival function of censoring times.

Lemma 1. *Let $\tau = F_\mu^{-1}(1)$ be the upper bound of the support for μ. Suppose $F_\mu(\cdot)$ is continuous and $\tau < \infty$. If $S_\nu(\tau) \geq \eta$ for some small constant η, then it holds for the Kaplan-Meier estimator as per (2) that*

$$E\{d_{\mathcal{W}}^2(\hat{\mu}, \mu)\} \leq 2\tau^{3/2}\{\int_0^\tau A(t)dt\}^{1/2} \cdot N^{-1/2},$$

where

$$A(t) = \int_0^t \frac{dF_\mu(u)}{S_\nu(u)}.$$

All proofs are provided in Section S.1 of the Supplementary Material. The non-asymptotic convergence rate in Lemma 1 depends on the quantity $A(t)$, which remains finite for all $t \leq \tau$ since $\tau < \infty$ and $S_\nu(\tau) \geq \eta$. As a result, the KM estimator converges to the underlying probability measure at a rate of $N^{-1/2}$. This non-asymptotic convergence rate is key for working with the Wasserstein metric in survival analysis.

3.2 Wasserstein-Kaplan-Meier survival regression

Let (Z, μ, ν) be a random pair with joint distribution $\mathcal{F}$ on the product space $\mathbb{R}^p \times \mathcal{W} \times \mathcal{W}$ where Z denotes the baseline characteristics and μ, ν denote probability measures of the survival and censoring times, respectively. Suppose that $\{(Z_i, \mu_i, \nu_i)\}_{i=1}^n$ are n independent realizations of (Z, μ, ν), each representing a subgroup defined by baseline characteristics.

For the ith subgroup, let $T_{ij} \sim \mu_i$ be i.i.d. survival times, independent of the censoring times $C_{ij} \sim \nu_i$ for $j = 1, \ldots, N_i$. In practice, we only observe the pairs $\{(Y_{ij}, \delta_{ij})\}_{j=1}^{N_i}$, where $Y_{ij} = \min\{T_{ij}, C_{ij}\}$ and $\delta_{ij} = \mathbf{1}_{T_{ij} \leq C_{ij}}$ is the censoring indicator. Note that there are two independent layers of randomness in the data: The first generates independent pairs of probability measures (μ_i, ν_i) taking values in $\mathcal{W} \times \mathcal{W}$; the second generates independent observations according to each probability measure, $T_{ij} \sim \mu_i$ and $C_{ij} \sim \nu_i$.

Here we aim to investigate survival time within heterogenous populations, with the probability measure of survival time μ which is now a random measure as the response

and the baseline characteristics Z as predictors. The regression function is defined as the conditional Fréchet mean of μ given $Z = z$,

$$m_{\oplus}(z) = \underset{\omega \in \mathcal{W}}{\arg\min}\ E\{d^2_{\mathcal{W}}(\mu, \omega)|Z = z\}.$$

Note that the conditional Fréchet mean extends the concept of the conditional mean to metric-space-valued responses.

Unlike the Euclidean space $\mathbb{R}$, the Wasserstein space $\mathcal{W}$ lacks a vector space structure and does not possess an inner product so that projections and basis expansions are not available. To address this challenge, Petersen and Müller (2019) suggested leveraging the algebraic structure within the predictor space $\mathbb{R}^p$ and modelling the conditional Fréchet mean as a weighted Fréchet mean,

$$m(z) = \underset{\omega \in \mathcal{W}}{\arg\min}\ E\{w(z)d^2_{\mathcal{W}}(\mu, \omega)\}. \tag{3}$$

Here, the weight function is defined as $w(z) = 1 + (Z - \theta)^{\mathrm{T}}\Sigma^{-1}(z - \theta)$, where θ and Σ represent the mean and variance of Z, respectively, with Σ being positive definite. These are the weights given to responses when viewing multiple linear regression as a weighted average and thus this approach generalizes multiple linear regression to metric-space-valued responses, providing a flexible framework for modeling non-Euclidean data.

The empirical version of the weighted Fréchet mean is expressed as

$$\tilde{m}(z) = \underset{\omega \in \mathcal{W}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z)d^2_{\mathcal{W}}(\mu_i, \omega), \tag{4}$$

where $\hat{w}_i(z) = 1 + (Z_i - \bar{Z})^{\mathrm{T}}\hat{\Sigma}^{-1}(z - \bar{Z})$ with

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^{\mathrm{T}}.$$

In practice, the optimization problem in (4) is intractable since the survival distribution μ_i is not directly observed due to random right censoring. To address this issue, we introduce the **W**asserstein-**K**aplan-**M**eier Survival Regression (WKM),

$$\hat{m}(z) = \underset{\omega \in \mathcal{W}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z)d^2_{\mathcal{W}}(\hat{\mu}_i, \omega), \tag{5}$$

where $\hat{\mu}_i$ is the probability measure corresponding to the KM estimator $S_{\hat{\mu}_i}(\cdot)$ for the ith subgroup. In particular,

$$S_{\hat{\mu}_i}(t) = \begin{cases} \prod_{j:Y_{i(j)} \le t} \left(\frac{N_i - j}{N_i - j + 1} \right)^{\delta_{i(j)}} & 0 \le t < Y_{i(N_i)}, \\ 0 & t \ge Y_{i(N_i)}, \end{cases}$$

where $Y_{i(j)}$ is the ordered observation and $\delta_{i(j)}$ the associated censoring indicator. The KM estimator is employed to address the issue of random right censoring, making it possible to estimate the conditional Fréchet mean for survival data.

4 Theory

In this section, we investigate the asymptotic properties of the proposed regression model within the framework of M-estimation, establishing both pointwise and uniform convergence rates. We require the following conditions, where μ and ν denote probability measures of the survival and censoring times, respectively.

(C1) The cumulative distribution function of survival time $F_\mu(\cdot)$ is continuous, with $\tau = F_\mu^{-1}(1) < \infty$. There exists a small constant $\eta > 0$ such that $S_\nu(\tau) \ge \eta$ almost surely.

(C2) The sample size N_i is independently Poisson distributed with parameter λ_n, where $\lambda_n / \log n \to \infty$ as $n \to \infty$.

Note that in Condition (C1), μ and ν are random elements in $\mathcal{W}$, where μ is supported on the closed interval $[0, \tau]$ with τ a constant. The first part of Condition (C1) is a regularity condition for the probability measure of the survival time that is commonly assumed in survival analysis. The assumption that $S_\nu(\tau) \ge \eta$ almost surely is a mild condition to ensure the effect of the censoring remains limited and is also a common assumption (Kulasekera, 1995). Condition (C2) introduces a distributional requirement for sample sizes N_i and allows for a large degree of heterogeneity in sample sizes among subgroups. It specifically

allows for very small or zero-sized subgroups alongside large subgroups and ensures that N_i is positive almost surely for sufficiently large n; compare Lemma 3 in Panaretos and Zemel (2016). The following result provides pointwise and uniform convergence rates for the proposed WKM estimate (5).

Theorem 1. *Under Conditions (C1) and (C2), the WKM estimate, defined in* (5)*, satisfies*

$$d_{\mathcal{W}}\{\hat{m}(z), m(z)\} = O_p(n^{-1/2} + \lambda_n^{-1/4}).$$

Furthermore, for any constant B it holds that for any $\varepsilon > 0$,

$$\sup_{\|z\|_E \leq B} d_{\mathcal{W}}\{\hat{m}(z), m(z)\} = O_p(n^{-1/\{2(1+\varepsilon)\}} + \lambda_n^{-1/4}).$$

Theorem 1 involves two layers of randomness. The first layer generates random elements (μ_i, ν_i) within $\mathcal{W} \times \mathcal{W}$, while the second layer generates random samples based on μ_i and ν_i. To handle the first layer of randomness, empirical process methods, specifically M-estimation, are employed, while the second layer of randomness is addressed through Lemma 1. The pointwise convergence rate is $O_p(n^{-1/2})$ as long as the sample size N_i, on average, grows at the same rate as n^2, i.e., $\lambda_n \asymp n^2$. This rate aligns with the well-established optimal rate for multiple linear regression, showing that there is no loss associated with the fact that the responses are probability measures rather than scalars.

5 Implementation and Simulation

5.1 Implementation details

To implement the proposed model, one needs to solve the optimization problem in (5). By standard properties of the $L^2(0,1)$ inner product, (5) can be simplified to

$$\hat{m}(z) = \arg\min_{\omega \in \mathcal{W}} d^2_{L^2}\{\hat{B}(z), F_\omega^{-1}\}, \tag{6}$$

where $\hat{B}(z) = n^{-1}\sum_{i=1}^{n}\hat{w}_i(z)F_{\hat{\mu}_i}^{-1}$ and $d_{L^2}(\cdot,\cdot)$ denotes the L^2 distance; see the proof of Theorem 1 for details. The minimizer $\hat{m}(z)$, as a projection onto $\mathcal{W}$, exists and is unique for any z due to the convexity and closedness of $\mathcal{W}$ (Bigot et al., 2017). Note that $\hat{B}(z)$ represents a weighted average of the KM quantile functions $F_{\hat{\mu}_i}^{-1}$. For computational convenience, we approximate the varying jump sizes in the KM quantile function $F_{\hat{\mu}_i}^{-1}$, resulting from censoring, by multiples of a small unit jump size, chosen as $1/L$ with a sufficiently large $L = 5000$. Specifically, a jump size of 0.3 at location t will be approximated by allocating $5000 \times 0.3 = 1500$ unit jumps at location t. After this step, each jump location is associated with a specific number of unit jumps. Subsequently, we calculate $\hat{B}(z)$ as an element-wise weighted average of the jump locations. Refer to Algorithm 1 for a more detailed explanation.

In the first step of Algorithm 1, we use the `survival` package (Therneau, 2024) in R to obtain the KM estimator for each subgroup. The fifth step of Algorithm 1 involves an optimization problem to obtain the projection onto $\mathcal{W}$. For this step, we use a Riemann sum approximation of the L^2 distance. This leads to the following convex quadratic optimization problem,

$$\begin{aligned} \text{minimize} \quad & \sum_{l=1}^{L}(q_l - \bar{U}_l)^2 \\ \text{subject to} \quad & q_1 \le q_2 \le \cdots \le q_{L-1} \le q_L; \\ & q_j \in [0,\tau], \quad j = 1,\ldots,L. \end{aligned} \tag{7}$$

The solution of (7) represents a discretized version of the predicted quantile function. We use the `osqp` package (Stellato et al., 2020) in R to solve this optimization problem. The R implementation of Algorithm 1 can be accessed at https://github.com/yidongzhou/Wasserstein-Kaplan-Meier-Survival-Regression.

Algorithm 1: Wasserstein-Kaplan-Meier Survival Regression

Input: data for n heterogeneous subgroups $\{(Z_i, \{(Y_{ij}, \delta_{ij})\}_{j=1}^{N_i})\}_{i=1}^{n}$, and a new predictor level z.

1 Obtain the Kaplan-Meier estimator $S_{\hat{\mu}_i}(t)$ for $i = 1, \ldots, n$, where $S_{\hat{\mu}_i}(t) = 0$ for $t \geq Y_{i(N_i)}$. Denote $\{V_{ij}\}_{j=1}^{N_i^*}$ as the time points where $S_{\hat{\mu}_i}(t)$ jumps, and $\{W_{ij}\}_{j=1}^{N_i^*}$ as the corresponding jump sizes, where N_i^* is the number of jumps;

2 For each $i = 1, \ldots, n$, stretch the vector $(V_{i1}, \ldots, V_{iN_i^*})^{\mathrm{T}}$ to a vector of length L, $\mathbf{U}_i = (U_{i1}, \ldots, U_{iL})^{\mathrm{T}}$, by repeating the jth element LW_{ij} times for $j = 1, \ldots, N_i^*$ and then arranging in ascending order;

3 For each $i = 1, \ldots, n$, compute the weight function $\hat{w}_i(z)$ as $1 + (Z_i - \bar{Z})^{\mathrm{T}}\hat{\Sigma}^{-1}(z - \bar{Z})$ where $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^{\mathrm{T}}$ are the sample mean and variance of $\{Z_i\}_{i=1}^{n}$, respectively;

4 Denote $\bar{\mathbf{U}} = (\bar{U}_1, \ldots, \bar{U}_L)^{\mathrm{T}}$ as the element-wise weighted average of $\{\mathbf{U}_i\}_{i=1}^{n}$ with weights being $\{\hat{w}_i(z)\}_{i=1}^{n}$;

5 Obtain the estimated regression function $\hat{m}(z) = \underset{\omega \in \mathcal{W}}{\arg\min}\, d_{L^2}^2\{\hat{B}(z), F_\omega^{-1}\}$, where $\hat{B}(z) = \sum_{l=1}^{L} \bar{U}_l \mathbf{1}_{q \in (\frac{l-1}{L}, \frac{l}{L}]}$ is a step function defined on $q \in [0, 1]$;

Output: prediction $\hat{m}(z)$.

5.2 Simulations

In this section, we illustrate the finite sample performance of the proposed model and provide numerical comparisons with the CPH model. The survival distributions serving as responses for the WKM model are chosen as Weibull distributions with shape parameter $k = 2$ and scale parameter λ. The regression function $m(Z)$ relates λ to Z through $E(\lambda|Z)$. The corresponding hazard and quantile functions are

$$h_{m(Z)}(t) = 2t \cdot \{E(\lambda|Z)\}^{-2}, \quad F_{m(Z)}^{-1}(p) = E(\lambda|Z) \cdot \{-\log(1-p)\}^{1/2},$$

respectively. Two different simulation settings are examined as follows.

- Setting I: $\lambda|Z$ follows a Gamma distribution with shape parameter $(Z\beta+0.1)^2/\rho$ and scale parameter $\rho/(Z\beta+0.1)$.

- Setting II: $\lambda|Z$ follows a Gamma distribution with shape parameter $e^{-Z\beta}/\rho$ and scale parameter $\rho/e^{-Z\beta/2}$.

The scale parameter for the Weibull distribution in Setting I and Setting II is $E(\lambda|Z) = Z\beta+0.1$ and $e^{-Z\beta/2}$, respectively. Setting I aligns with the setting of the proposed model, while Setting II satisfies the proportional hazards assumption. We explore three different values of ρ: 0.05, 0.1, and 0.5. Larger values of ρ correspond to increased random variation in the survival distributions across subgroups. To introduce random right censoring, we independently generate censoring times from Weibull distributions with shape parameter $k = 2$ and scale parameters $2E(\lambda|Z)$, $E(\lambda|Z)$, respectively. This results in overall censoring rates of 20% and 50%, respectively.

We choose a five-dimensional predictor Z with $\beta = (0.01, 0.02, 0.03, 0.04, 0.05)^{\mathrm{T}}$. In each simulation setting, we perform $Q = 1000$ simulation runs with sample sizes $n = 100, 200, 500$. For each n, the sample size N_i is independently sampled from the Poisson distribution with parameter $\lambda_n = 0.5n$. In each simulation run, Z_i is generated independently from five Bernoulli distributions with the same parameters $p = 0.5$. For each i, the scale parameter for the Weibull distribution is generated conditionally on Z_i as described in Setting I and II. N_i random survival times $\{T_{ij}\}_{j=1}^{N_i}$ and N_i random censoring times $\{C_{ij}\}_{j=1}^{N_i}$ are then independently sampled from the Weibull distribution with shape parameter $k = 2$ and the generated scale parameter. With $Y_{ij} = \min\{T_{ij}, C_{ij}\}$ and $\delta_{ij} = \mathbf{1}_{T_{ij}\leq C_{ij}}$ being the censoring indicator, the proposed model is evaluated using the observations of the form: $(Z_i, Y_{ij}, \delta_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, N_i$.

For the qth simulation run under a given setting, the quality of the estimation is quan-

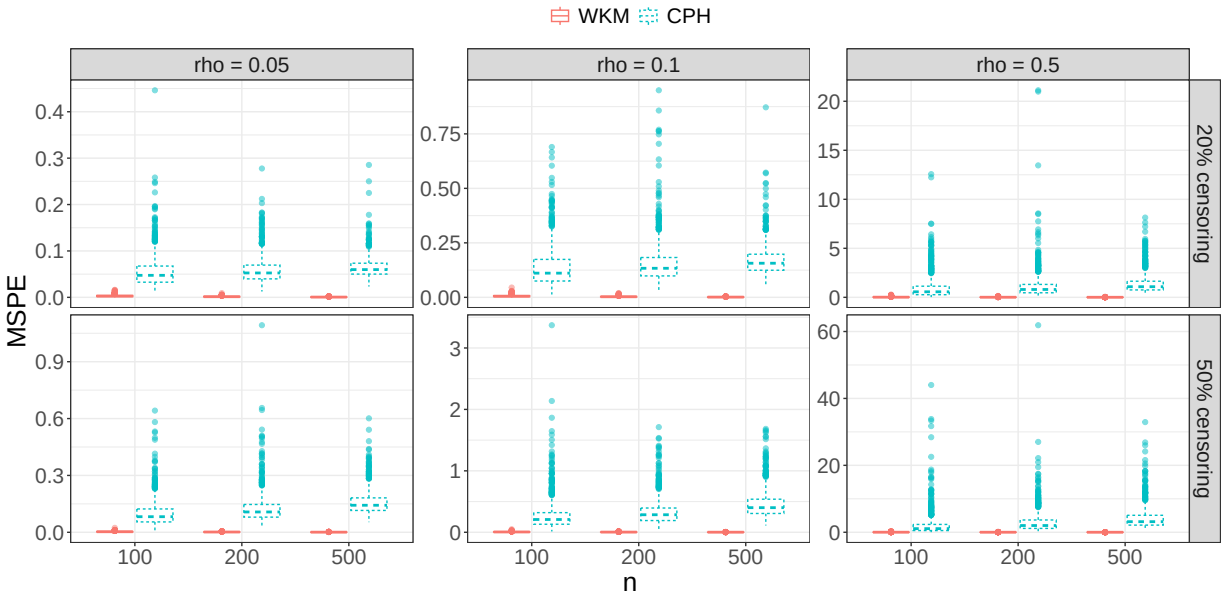
tified by the mean squared prediction error (MSPE) using the Wasserstein metric,

$$\text{MSPE}_q = E_Z[d^2_{\mathcal{W}}\{\hat{m}_q(Z), m(Z)\}] = \frac{1}{32}\sum_{z\in\{0,1\}^5} d^2_{\mathcal{W}}\{\hat{m}_q(z), m(z)\},$$

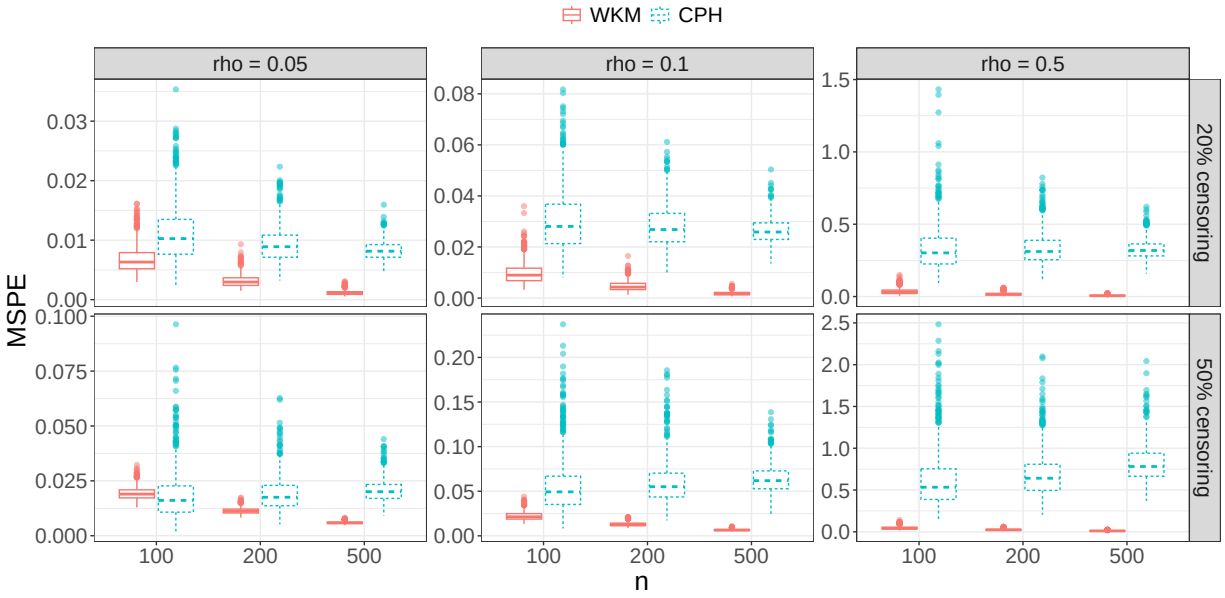
where $m(\cdot)$ is the true regression function and $\hat{m}_q(\cdot)$ is the predicted regression function. The MSPEs for all simulation runs under the two simulation settings using the WKM and CPH models are visualized as boxplots in Figure 1, with their averages summarized in Table 1.

As expected, lower censoring rates yield more precise estimates. Across all simulation scenarios, we observe a reduction in average MSPEs as the sample size increases, demonstrating the convergence of the WKM model to the target. Notably, the WKM model exhibits remarkable robustness, performing exceptionally well even under the proportional hazards assumption that typically favors the CPH model. In fact, the WKM model consistently outperforms the CPH model across all simulation scenarios, particularly in cases involving moderate to large sample sizes. The only exception, as highlighted in Table 1, is when $n = 100$, $\rho = 0.05$ and the censoring rate is 50% under the proportional hazards assumption. A significant advantage of the WKM model is its capacity to effectively accommodate random variation inherent in survival distributions across different subgroups. In contrast, the CPH model is susceptible to such random variation, as evident in Figure 1.

Additional simulations are detailed in Section S.2 of the Supplementary Material, where we assess the performance of the WKM model using varying values of β and different baseline hazard functions. These simulations also include comparisons with nonparametric survival models and evaluations in scenarios involving high-dimensional predictors.



(a)



(b)

Figure 1: Mean squared prediction errors for $Q = 1000$ simulation runs and different simulation settings using the WKM and CPH models. (a) Setting I; (b) Setting II.

Table 1: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM and CPH models.

		$\rho = 0.05$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting I	WKM	0.0033	0.0016	0.0007	0.0036	0.0020	0.0009
		(0.0020)	(0.0009)	(0.0004)	(0.0019)	(0.0008)	(0.0003)
	CPH	0.0550	0.0586	0.0643	0.0993	0.1256	0.1559
		(0.0346)	(0.0287)	(0.0230)	(0.0698)	(0.0791)	(0.0626)
Setting II	WKM	0.0068	0.0031	0.0012	0.0193	0.0114	0.0059
		(0.0022)	(0.0010)	(0.0004)	(0.0029)	(0.0013)	(0.0005)
	CPH	0.0110	0.0092	0.0083	0.0182	0.0190	0.0205
		(0.0048)	(0.0029)	(0.0016)	(0.0107)	(0.0075)	(0.0050)
		$\rho = 0.1$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting I	WKM	0.0063	0.0032	0.0012	0.0062	0.0032	0.0013
		(0.0045)	(0.0022)	(0.0007)	(0.0042)	(0.0019)	(0.0007)
	CPH	0.1384	0.1554	0.1691	0.2716	0.3333	0.4525
		(0.0932)	(0.0952)	(0.0687)	(0.2507)	(0.2176)	(0.2248)
Setting II	WKM	0.0097	0.0047	0.0018	0.0222	0.0130	0.0066
		(0.0039)	(0.0019)	(0.0007)	(0.0047)	(0.0023)	(0.0009)
	CPH	0.0302	0.0279	0.0264	0.0552	0.0590	0.0642
		(0.0121)	(0.0080)	(0.0050)	(0.0296)	(0.0234)	(0.0154)
		$\rho = 0.5$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting I	WKM	0.0268	0.0135	0.0063	0.0253	0.0134	0.0059
		(0.0323)	(0.0131)	(0.0048)	(0.0310)	(0.0113)	(0.0041)
	CPH	0.9359	1.1254	1.3856	2.1137	3.0142	4.0954
		(1.1680)	(1.3855)	(1.0072)	(3.4221)	(3.4628)	(3.2621)
Setting II	WKM	0.0357	0.0173	0.0066	0.0452	0.0245	0.0114
		(0.0188)	(0.0093)	(0.0034)	(0.0179)	(0.0087)	(0.0037)
	CPH	0.3316	0.3286	0.3269	0.6086	0.6808	0.8208
		(0.1532)	(0.1054)	(0.0668)	(0.3252)	(0.2580)	(0.2200)

6 Data Applications

6.1 Risk factors for time to death of women living with HIV

Human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) continues to pose a significant global public health challenge, despite extensive efforts to combat this disease. As of the end of 2018, an estimated 37.9 million individuals were living with HIV worldwide. Understanding the risk factors that impact HIV survival is crucial for designing targeted interventions, especially in the context of diverse geographic regions.

The Women's Interagency HIV Study (WIHS) (Barkan et al., 1998; Bacon et al., 2005; Adimora et al., 2018) is a comprehensive, multi-center cohort study initiated in the mid-1990s. It has been continuously observing women living with HIV (WLWH) and women at risk for HIV across the United States. The WIHS has played a vital role in unraveling the influence of various risk factors on HIV survival among WLWH. It enrolled a diverse group of 3,772 women from various sources, including HIV primary care clinics, research programs, community outreach sites, women's support groups, drug rehabilitation programs, and HIV testing sites in major cities including Chicago, Los Angeles, New York City, San Francisco, and Washington, DC.

We focus on a subset of 1,689 participants who were HIV positive, with a censoring rate of 42.92%. Our main interest lies in the time to death following HIV diagnosis, and the corresponding survival distribution is the response in the WKM model. We consider four baseline characteristics: age, educational level, marital status, and employment status. Age is categorized as those aged 30 years or less (0), between 30 and 40 years (1), and over 40 years (2). Educational levels are stratified into low (0), medium (1), and high (2), based on the completion of high school or graduate school. Similarly, marital status is dichotomized as either married (0) or unmarried (1), and employment status as either unemployed (0) or employed (1). These four baseline characteristics result in the division of the entire sample

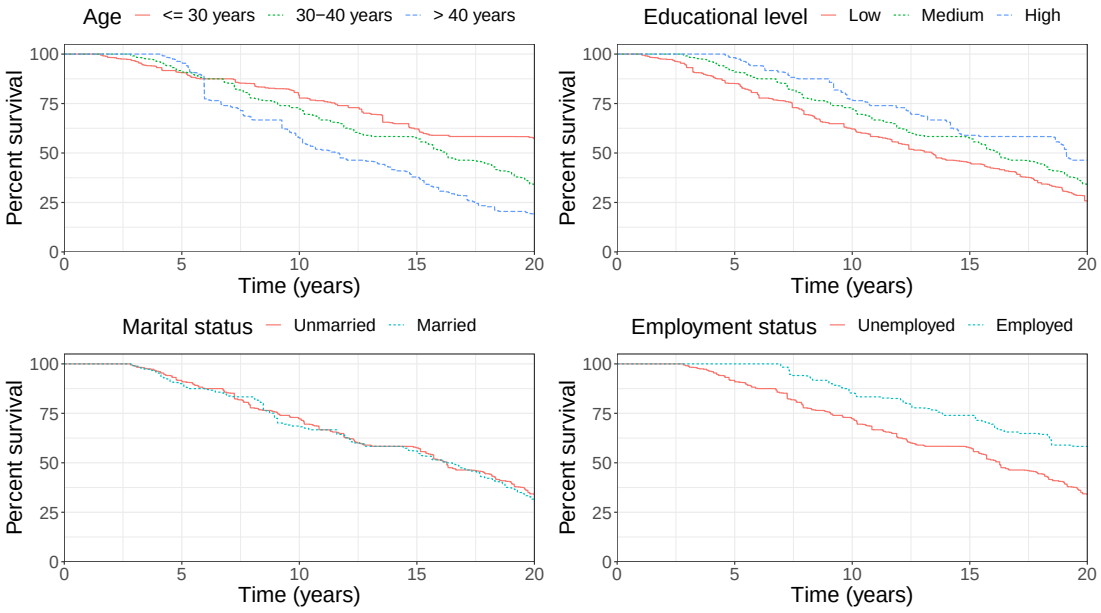
Table 2: Number of deaths and censoring for the Women's Interagency HIV Study (WIHS).

Subgroup		0000	1000	2000	0100	1100	2100	0200	1200	2200	0010	1010
deaths		69	115	58	93	140	76	3	9	9	66	54
censoring		51	63	16	79	78	32	6	10	2	45	32
Total		120	178	74	172	218	108	9	19	11	111	86
	2010	0110	1110	2110	0210	1210	2210	0001	1001	2001	0101	1101
	12	42	70	32	3	4	2	4	6	1	14	23
	3	51	42	21	5	3	1	6	4	2	37	36
	15	93	112	53	8	7	3	10	10	3	51	59
	2101	0201	1201	2201	0011	1011	2011	0111	1111	2111	1211	2211
	11	4	4	2	2	7	2	9	9	5	3	1
	9	8	13	4	6	3	0	25	19	5	4	2
	20	12	17	6	8	10	2	34	28	10	7	3

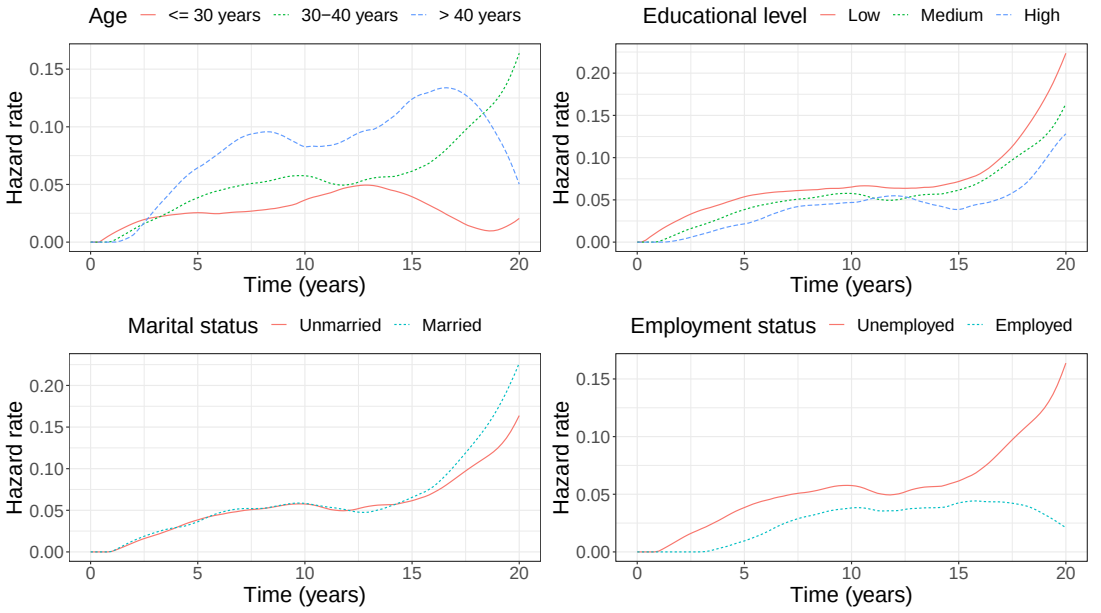
into 32 subgroups, each with distinct characteristics, as shown in Table 2. Note that the subgroup "0211" has been excluded from the analysis as it consists of only two participants, where the lifetimes for both were censored.

As illustrated in Table 2, there is a notable variation in sample sizes across subgroups, with "2011" being the smallest, involving just two participants with uncensored survival times. One of the strengths of the WKM model is its robustness in estimating survival functions for these sparsely populated subgroups. To further study the factors impacting HIV survival among WLWH, we present predicted survival and hazard functions for time to death following HIV diagnosis at various levels of baseline characteristics in Figure 2. To display the effect of varying a selected characteristic in this figure, all other characteristics are fixed at the levels with the highest number of participants. Our observations reveal that advanced age, lower educational attainment, and unemployment are risk factors associated with shorter time to death for WLWH, which is in line with the results from previous studies (Chandran et al., 2020). Conversely, marital status appears to have a negligible impact, which also has been noted before (Tadege, 2018).

Additionally, we compare the WKM model with the CPH model using the Wasserstein metric. The leave-one-out cross-validation MSPE for the WKM model is notably smaller,



(a)



(b)

Figure 2: Predicted (a) survival and (b) hazard functions for time to death following HIV diagnosis with different levels of baseline characteristics.

constituting only 47.65% of the MSPE for the CPH model. This significant reduction in MSPE underscores the superiority of the WKM model.

6.2 Prognostic factors for liver cancer among adolescents and young adults

The National Cancer Institute has recognized adolescents and young adults (AYAs) aged 15-39 as a distinct patient population in cancer research due to unique socio-economic factors and specific biologic characteristics, distinguishing AYA cancers from those in children and older adults (Sender and Zabokrtsky, 2015). Liver cancer, ranking as the sixth most common malignant cancer globally and the fourth leading cause of cancer-related death, is the focus of our study among AYAs in the United States.

The Surveillance, Epidemiology, and End Results (SEER) Program (Surveillance, Epidemiology, and End Results (SEER) Program, www.seer.cancer.gov), operated by the National Cancer Institute, stands as a comprehensive and authoritative resource in the realm of cancer-related data and research. This program offers invaluable insights into cancer incidence and outcomes throughout the United States. SEER 17 registries include 17 regions across the United States, covering approximately 26.5% of the U.S. population These registries collect a wealth of data, including information on cancer incidence, survival, patient demographics such as sex and age at diagnosis, and details regarding the year of diagnosis and death, among other critical factors.

Our analysis includes 2,040 AYA patients diagnosed with liver cancer between 2000 and 2020, with a censoring rate of 36.26%. We are primarily interested in the survival time following liver cancer diagnosis among AYAs and consider the baseline characteristics age, sex, tumor stage and income. Age is categorized as 15–24 years (0) and 25–39 years (1). Tumor stage is classified as localized (0), regional (1), or distant (2). Income levels are categorized as less than 55,000 (0) and 55,000 or more (1). The entire sample is divided

Table 3: Number of deaths and censoring for the Surveillance, Epidemiology, and End Results (SEER) Program.

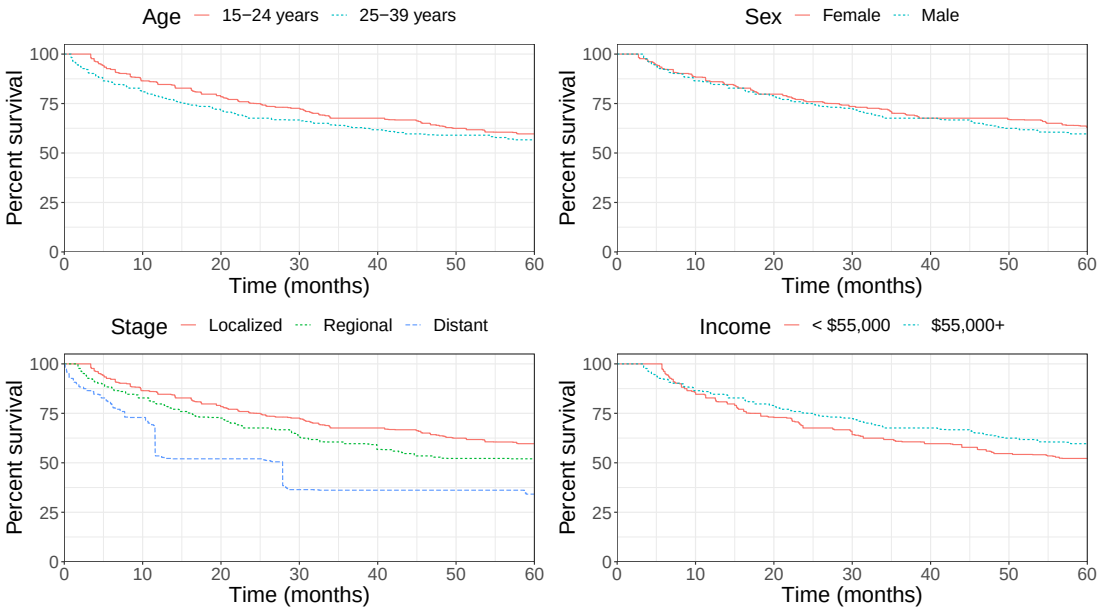
Subgroup	1000	0100	1100	0010	1010	0110	1110	0020	1020	0120	1120
deaths	19	5	30	4	17	7	22	4	13	5	26
censoring	17	3	12	1	1	4	9	0	4	3	3
Total	36	8	42	5	18	11	31	4	17	8	29
0001	1001	0101	1101	0011	1011	0111	1111	0021	1021	0121	1121
18	61	29	158	17	83	29	254	32	76	45	213
34	109	32	184	26	53	27	72	9	13	17	26
52	170	61	342	43	136	56	326	41	89	62	239

into 24 subgroups, each with distinct characteristics, as detailed in Table 3. Subgroup "0000" was excluded as all patients in this category were censored.

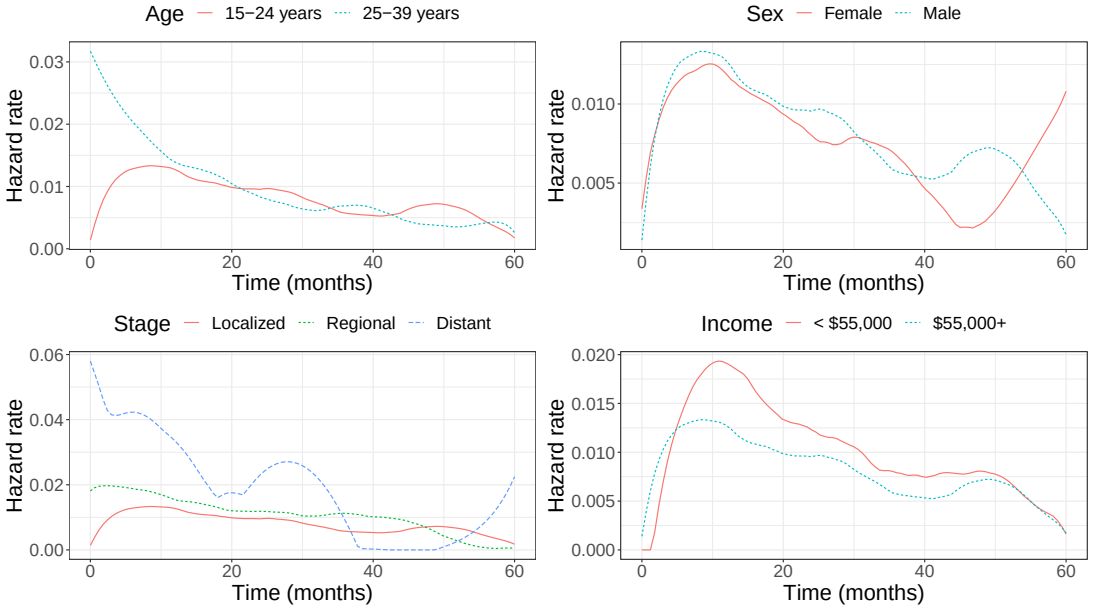
In Figure 3, we present predicted survival and hazard functions for the time to death following liver cancer diagnosis, using the WKM model. We focus on the 5-year survival and seek to identify prognostic factors for liver cancer among AYAs. For each characteristic, we maintain the remaining characteristics at the levels with the highest number of patients. Our findings indicate that older age, advanced tumor stage, and lower income are associated with shorter time to death for AYAs with liver cancer, as noted previously (Ren et al., 2020), while sex seems to have a negligible impact (Liou et al., 2023). Similar to the WIHS, we compared the WKM model with the CPH model and found that the leave-one-out cross-validation MSPE for the WKM model is significantly smaller, with a value of only 53.74% of the MSPE for the CPH model. This substantial reduction in MSPE again underscores the superior performance of the WKM model in this context.

7 Discussion

In this article, we propose a flexible and robust regression model for right-censored survival data. The proposed WKM model effectively accommodates random variation in the probability measures corresponding to survival functions across different subgroups. In real data



(a)



(b)

Figure 3: Predicted (a) survival and (b) hazard functions for time to death following liver cancer diagnosis with different levels of baseline characteristics.

applications, such random variation can account for errors caused by estimating survival distributions from discrete observations, for example, when using the Kaplan-Meier estimator. The proposed algorithm is scalable to high-dimensional settings as demonstrated in subsection S.2.5 of the Supplementary Material. Both theoretical analysis and numerical simulations support the effectiveness and utility of the WKM model.

The implementation of the WKM model involves the conversion of the Kaplan-Meier survival curve to the corresponding quantile function over a common grid. We adopt a gird of length $L = 5000$ to strike a balance between memory efficiency and computational accuracy for most typical application scenarios. Future adaptations of the algorithm could include more accurate approximations of quantile functions or the implementation of more memory-efficient data structures.

Supplementary Materials

Online supplementary material contains proofs of all stated results and additional numerical experiments. R code implementing the proposed regression model is available at https://github.com/yidongzhou/Wasserstein-Kaplan-Meier-Survival-Regression.

Acknowledgments

This research was supported in part by NSF grant DMS-2310450. We deeply appreciate the insightful and constructive comments from the editor, the associate editor, and the two referees.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability

Regarding the datasets utilized in our real data applications, they are not publicly available due to privacy concerns. For researchers interested in accessing these datasets:

- The WIHS dataset can be requested through the portal at https://statepi.jhsph.edu/mwccs/work-with-us/
- The SEER database can be accessed via https://seer.cancer.gov/data/access.html

The application process for both datasets is designed to be straightforward. Approval is typically granted generously, provided all required fields in the application are completed accurately.

References

Adimora, A. A., Ramirez, C., Benning, L., Greenblatt, R. M., Kempf, M.-C., Tien, P. C., Kassaye, S. G., Anastos, K., Cohen, M., Minkoff, H., et al. (2018), "Cohort profile: the Women's Interagency HIV Study (WIHS)," *International Journal of Epidemiology*, 47, 393–394i.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer New York.

Babińska, M., Chudek, J., Chełmecka, E., Janik, M., Klimek, K., and Owczarek, A. (2015), "Limitations of Cox proportional hazards analysis in mortality prediction of patients with acute coronary syndrome," *Studies in Logic, Grammar and Rhetoric*, 43, 33–48.

Bacon, M. C., Von Wyl, V., Alden, C., Sharp, G., Robison, E., Hessol, N., Gange, S., Barranday, Y., Holman, S., Weber, K., et al. (2005), "The Women's Interagency HIV

Study: an observational cohort brings clinical sciences to the bench," *Clinical and Vaccine Immunology*, 12, 1013–1019.

Barkan, S. E., Melnick, S. L., Preston-Martin, S., Weber, K., Kalish, L. A., Miotti, P., Young, M., Greenblatt, R., Sacks, H., and Feldman, J. (1998), "The Women's Interagency HIV Study," *Epidemiology*, 9, 117–125.

Bigot, J., Gouet, R., Klein, T., and López, A. (2017), "Geodesic PCA in the Wasserstein space by convex PCA," *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 53, 1–26.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Chandran, A., Edmonds, A., Benning, L., Wentz, E., Adedimeji, A., Wilson, T. E., Blair-Spence, A., Palar, K., Cohen, M., and Adimora, A. (2020), "Longitudinal associations between neighborhood factors and HIV care outcomes in the WIHS," *AIDS and Behavior*, 24, 2811–2818.

Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B*, 34, 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Deutsch, F. (2001), *Best Approximation in Inner Product Spaces*, volume 7, Springer New York.

Efron, B. (1967), "The two sample problem with censored data," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4.

Fréchet, M. (1948), "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'Institut Henri Poincaré*, 10, 215–310.

Gill, R. D. (1980), "Censoring and stochastic integrals," *Statistica Neerlandica*, 34, 124–124.

Gordon, L. and Olshen, R. A. (1985), "Tree-structured survival analysis." *Cancer Treatment Reports*, 69, 1065–1069.

Han, X., Goldstein, M., and Ranganath, R. (2022), "Survival Mixture Density Networks," in *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, URL https://proceedings.mlr.press/v182/han22a.html.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random survival forests," *Annals of Applied Statistics*, 2, 841–860.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, John Wiley & Sons, 2nd edition.

Kalbfleisch, J. D. and Schaubel, D. E. (2023), "Fifty Years of the Cox Model," *Annual Review of Statistics and its Application*, 10, 1–23.

Kantorovich, L. V. (1942), "On the translocation of masses," *Dokl. Akad. Nauk SSSR (translated version in Journal of Mathematical Sciences, 133, 1381-1382, 2006)*, 37, 227–229.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Kloeckner, B. (2010), "A geometric study of Wasserstein spaces: Euclidean spaces," *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9, 297–323.

Kulasekera, K. (1995), "A bound on the L_1-error of a nonparametric density estimator with censored data," *Statistics & Probability Letters*, 23, 233–238.

Kulasekera, K., Williams, C. L., Coffin, M., and Manatunga, A. (2001), "Smooth estimation of the reliability function," *Lifetime Data Analysis*, 7, 415–433.

Liou, W.-L., Tan, T. J.-Y., Chen, K., Goh, G. B.-B., Chang, J. P.-E., and Tan, C.-K. (2023), "Gender survival differences in hepatocellular carcinoma: Is it all due to adherence to surveillance? A study of 1716 patients over three decades," *JGH Open*, 7, 377–386.

Panaretos, V. M. and Zemel, Y. (2016), "Amplitude and phase variation of point processes," *Annals of Statistics*, 44, 771–812.

Petersen, A. and Müller, H.-G. (2019), "Fréchet regression for random objects with Euclidean predictors," *Annals of Statistics*, 47, 691–719.

Petersen, A., Zhang, C., and Kokoszka, P. (2022), "Modeling probability density functions as data objects," *Econometrics and Statistics*, 21, 159–178.

Pike, M. (1966), "A method of analysis of a certain class of experiments in carcinogenesis," *Biometrics*, 22, 142–161.

Ren, J., Tong, Y.-M., Cui, R.-X., Wang, Z., Li, Q.-L., Liu, W., Qu, K., Zhang, J.-Y., Liu, C., and Wan, Y. (2020), "Comparison of survival between adolescent and young adult vs older patients with hepatocellular carcinoma," *World Journal of Gastrointestinal Oncology*, 12, 1394–1406.

Santambrogio, F. (2015), *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87, Birkhäuser Cham.

Sender, L. and Zabokrtsky, K. B. (2015), "Adolescent and young adult patients with cancer: a milieu of unique features," *Nature Reviews Clinical Oncology*, 12, 465–480.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020), "OSQP: An Operator Splitting Solver for Quadratic Programs," *Mathematical Programming Computation*, 12, 637–672.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov), "SEER*Stat Database: Incidence - SEER Research Data, 17 Registries, Nov 2022 Sub (2000-2020) - Linked To County Attributes - Time Dependent (1990-2021) Income/Rurality, 1969-2021 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2023, based on the November 2022 submission," .

Sylvain, T., Luck, M., Cohen, J., Cardinal, H., Lodi, A., and Bengio, Y. (2021), "Exploring the Wasserstein metric for time-to-event analysis," in *Survival Prediction-Algorithms, Challenges and Applications*, Proceedings of Machine Learning Research.

Tadege, M. (2018), "Time to death predictors of HIV/AIDS infected patients on antiretroviral therapy in Ethiopia," *BMC Research Notes*, 11, 1–6.

Tang, W., Ma, J., Mei, Q., and Zhu, J. (2022), "SODEN: A scalable continuous-time survival model through ordinary differential equation networks," *Journal of Machine Learning Research*, 23, 1–29.

Therneau, T. M. (2024), *A Package for Survival Analysis in R*, URL https://CRAN.R-project.org/package=survival. R package version 3.5-8.

Villani, C. (2003), *Topics in Optimal Transportation*, American Mathematical Society.

Wei, L.-J. (1992), "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis," *Statistics in Medicine*, 11, 1871–1879.

Yang, S. (1991), "Minimum Hellinger Distance Estimation of Parameter in the Random Censorship Model," *Annals of Statistics*, 19, 579 – 602.

Zhang, C., Kokoszka, P., and Petersen, A. (2022), "Wasserstein autoregressive models for density time series," *Journal of Time Series Analysis*, 43, 30–52.

Zhou, Y. and Müller, H.-G. (2023), "Wasserstein Regression with Empirical Measures and Density Estimation for Sparse Data," *arXiv preprint arXiv:2308.12540*.

SUPPLEMENTARY MATERIAL

S.1 Proofs

S.1.1 Proof of Lemma 1

Proof. Note that in Lemma 1, the probability measures of survival and censoring times are considered fixed. Consequently, we are only concerned with one layer of randomness, specifically the random sample generated according to μ and ν. Using Lemma 2.2 of Yang (1991), it holds that

$$\int_0^\tau E\{|S_{\hat{\mu}}(t) - S_\mu(t)|^2\}dt \leq \frac{4}{N}\int_0^\tau A(t)dt,$$

where

$$A(t) = \int_0^t \frac{dF_\mu(u)}{S_\nu(u)}.$$

Similar arguments can be found in the proof of Proposition 1 of Kulasekera et al. (2001).

For any two probability measures μ_1, μ_2 with a common support $[0, \tau]$, according to Equation (5.1) and Proposition 2.17 of Santambrogio (2015) we have

$$\begin{aligned} d^2_{\mathcal{W}}(\mu_1, \mu_2) &\leq \tau \int_0^1 |F^{-1}_{\mu_1}(p) - F^{-1}_{\mu_2}(p)|dp \\ &= \tau \int_0^\tau |F_{\mu_1}(t) - F_{\mu_2}(t)|dt. \end{aligned}$$

Now, we can proceed with the following derivations,

$$
\begin{aligned}
E\{d^2_{\mathcal{W}}(\hat{\mu}, \mu)\} &\leq E\{\tau \int_0^\tau |F_{\hat{\mu}}(t) - F_\mu(t)|dt\} \\
&\overset{\text{(i)}}{\leq} \tau(E[\{\int_0^\tau |F_{\hat{\mu}}(t) - F_\mu(t)|dt\}^2])^{1/2} \\
&\overset{\text{(ii)}}{\leq} \tau[E\{\tau \int_0^\tau |F_{\hat{\mu}}(t) - F_\mu(t)|^2 dt\}]^{1/2} \\
&\overset{\text{(iii)}}{=} \tau^{3/2}[\int_0^\tau E\{|F_{\hat{\mu}}(t) - F_\mu(t)|^2\}dt]^{1/2} \\
&= \tau^{3/2}[\int_0^\tau E\{|S_{\hat{\mu}}(t) - S_\mu(t)|^2\}dt]^{1/2} \\
&\leq \tau^{3/2}\{\frac{4}{N}\int_0^\tau A(t)dt\}^{1/2} \\
&= 2\tau^{3/2}\{\int_0^\tau A(t)dt\}^{1/2} \cdot N^{-1/2},
\end{aligned}
$$

where for (i), (ii), and (ii) we use the fact that $E(X) \leq \{E(X^2)\}^{1/2}$, Hölder's inequality, and Fubini's theorem, respectively. The result follows. □

S.1.2 Proof of Theorem 1

Proof. We first establish the pointwise convergence rate for the WKM estimate. By the triangle inequality, we have

$$
d_{\mathcal{W}}\{\hat{m}(z), m(z)\} \leq d_{\mathcal{W}}\{\hat{m}(z), \tilde{m}(z)\} + d_{\mathcal{W}}\{\tilde{m}(z), m(z)\}. \tag{8}
$$

The second term on the right-hand side, corresponding to the pointwise convergence rate for global Fréchet regression with fully observed probability measures, is $O_p(n^{-1/2})$ by Proposition 1 and Theorem 2 of Petersen and Müller (2019).

To analyze the first term, we will utilize simplified expressions of $\tilde{m}(z)$ and $\hat{m}(z)$. Let $\langle \cdot, \cdot \rangle_{L^2}$, $\| \cdot \|_{L^2}$, and $d_{L^2}(\cdot, \cdot)$ be the inner product, norm and distance on the Hilbert space $L^2(0,1)$. For any $\mu \in \mathcal{W}$, the map $Q : \mu \mapsto F_\mu^{-1}$ is an isometry from $\mathcal{W}$ to the subset of $L^2(0,1)$ formed by equivalence classes of left-continuous nondecreasing functions on $(0,1)$. The Wasserstein space $\mathcal{W}$ can thus be viewed as a subset of $L^2(0,1)$, which has been shown to be convex and closed (Bigot et al., 2017).

Define

$$\tilde{B}(z) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) F_{\mu_i}^{-1}, \quad \hat{B}(z) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) F_{\hat{\mu}_i}^{-1},$$

where $F_{\mu_i}^{-1}$ and $F_{\hat{\mu}_i}^{-1}$ are the quantile functions of μ_i and $\hat{\mu}_i$, respectively. Since $n^{-1}\sum_{i=1}^{n} \hat{w}_i(z) = 1$, it follows that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) d_{\mathcal{W}}^2(\mu_i, \omega) &= \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z)[d_{L^2}^2\{F_{\mu_i}^{-1}, \tilde{B}(z)\} + d_{L^2}^2\{\tilde{B}(z), F_{\omega}^{-1}\} \\
&\quad + 2\langle F_{\mu_i}^{-1} - \tilde{B}(z), \tilde{B}(z) - F_{\omega}^{-1}\rangle_{L^2}] \\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) d_{L^2}^2\{F_{\mu_i}^{-1}, \tilde{B}(z)\} + d_{L^2}^2\{\tilde{B}(z), F_{\omega}^{-1}\} \\
&\quad + 2 \cdot \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z)\langle F_{\mu_i}^{-1} - \tilde{B}(z), \tilde{B}(z) - F_{\omega}^{-1}\rangle_{L^2},
\end{aligned}$$

where the last term

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z)\langle F_{\mu_i}^{-1} - \tilde{B}(z), \tilde{B}(z) - F_{\omega}^{-1}\rangle_{L^2} &= \langle \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) F_{\mu_i}^{-1} - \tilde{B}(z), \tilde{B}(z) - F_{\omega}^{-1}\rangle_{L^2} \\
&= \langle \tilde{B}(z) - \tilde{B}(z), \tilde{B}(z) - F_{\omega}^{-1}\rangle_{L^2} \\
&= 0.
\end{aligned}$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) d_{\mathcal{W}}^2(\mu_i, \omega) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) d_{L^2}^2\{F_{\mu_i}^{-1}, \tilde{B}(z)\} + d_{L^2}^2\{\tilde{B}(z), F_{\omega}^{-1}\},$$

whence

$$\begin{aligned}
\tilde{m}(z) &= \operatorname*{arg\,min}_{\omega \in \mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(z) d_{\mathcal{W}}^2(\mu_i, \omega) \\
&= \operatorname*{arg\,min}_{\omega \in \mathcal{W}} d_{L^2}^2\{\tilde{B}(z), F_{\omega}^{-1}\}.
\end{aligned}$$

One can similarly show that

$$\hat{m}(z) = \operatorname*{arg\,min}_{\omega \in \mathcal{W}} d_{L^2}^2\{\hat{B}(z), F_{\omega}^{-1}\}.$$

By the convexity and closedness of $\mathcal{W}$, the minimizers $\tilde{m}(z)$ and $\hat{m}(z)$, viewed as projections onto $\mathcal{W}$, exist and are unique for any $z \in \mathbb{R}^p$ (Deutsch, 2001, chap. 3).

Now consider the first term in (8). The contractive property of the projection onto a closed and convex subset in the Hilbert space $L^2(0,1)$ implies that

$$\begin{aligned} d_{\mathcal{W}}\{\hat{m}(z), \tilde{m}(z)\} &= d_{\mathcal{W}}[\arg\min_{\omega\in\mathcal{W}} d^2_{L^2}\{\hat{B}(z), F_\omega^{-1}\}, \arg\min_{\omega\in\mathcal{W}} d^2_{L^2}\{\tilde{B}(z), F_\omega^{-1}\}] \\ &\le d_{L^2}\{\hat{B}(z), \tilde{B}(z)\} \\ &= d_{L^2}\{\frac{1}{n}\sum_{i=1}^n \hat{w}_i(z) F^{-1}_{\hat{\mu}_i}, \frac{1}{n}\sum_{i=1}^n \hat{w}_i(z) F^{-1}_{\mu_i}\}. \end{aligned}$$

By the triangle inequality, we have that

$$\begin{aligned} d_{\mathcal{W}}\{\hat{m}(z), \tilde{m}(z)\} &\le \frac{1}{n}\sum_{i=1}^n |\hat{w}_i(z)| d_{L^2}(F^{-1}_{\hat{\mu}_i}, F^{-1}_{\mu_i}) \\ &= \frac{1}{n}\sum_{i=1}^n |\hat{w}_i(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i). \end{aligned}$$

Note that $\hat{w}_i(z) = w_i(z) + W_{0n}(z) + Z_i^{\mathrm{T}} W_{1n}(z)$, where

$$\begin{aligned} w_i(z) &= 1 + (Z_i - \theta)^{\mathrm{T}} \Sigma^{-1}(z - \theta), \\ W_{0n}(z) &= \theta^{\mathrm{T}} \Sigma^{-1}(z - \theta) - \bar{Z}\hat{\Sigma}^{-1}(z - \bar{Z}), \\ W_{1n}(z) &= \hat{\Sigma}^{-1}(z - \bar{Z}) - \Sigma^{-1}(z - \theta). \end{aligned}$$

Both $W_{0n}(z)$ and $\|W_{1n}(z)\|_E$ are $O_p(n^{-1/2})$ by the central limit theorem. It follows that

$$\begin{aligned} d_{\mathcal{W}}\{\hat{m}(z), \tilde{m}(z)\} &\le \frac{1}{n}\sum_{i=1}^n |w_i(z) + W_{0n}(z) + Z_i^{\mathrm{T}} W_{1n}(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i) \\ &\le \frac{1}{n}\sum_{i=1}^n |w_i(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^n |W_{0n}(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i) + \frac{1}{n}\sum_{i=1}^n |Z_i^{\mathrm{T}} W_{1n}(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i). \end{aligned} \tag{9}$$

By the Cauchy-Schwarz inequality, for the first term of (9) we have

$$\begin{aligned} E\{\frac{1}{n}\sum_{i=1}^n |w_i(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\} &= \frac{1}{n}\sum_{i=1}^n E\{|w_i(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\} \\ &\le \frac{1}{n}\sum_{i=1}^n [E\{|w_i(z)|^2\}]^{1/2} [E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\}]^{1/2}. \end{aligned} \tag{10}$$

According to Lemma 3 of Panaretos and Zemel (2016), under Condition (C2) one has

$$\liminf_{n\to\infty} \frac{\min_{1\le i\le n} N_i}{\lambda_n} \ge \frac{c(1-e^{-1})}{2} \quad \text{a.s.},$$

which implies that for any $i = 1, \ldots, n$, one has $N_i > 0$ almost surely for large enough n. Note that

$$d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i) = d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i\ge 1} + d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i=0}.$$

The second term $d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i=0}$ hence equals 0 almost surely for large enough n. Since $d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i=0}$ is dominated by $d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i) < \infty$, it follows that $E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i=0}\} = 0$ for large enough n.

For the first term, according to Lemma 1 we have

$$\begin{aligned} E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i\ge 1}|\mu_i, \nu_i, N_i\} &\le 2\tau^{3/2}\{\int_0^\tau A_i(t)dt\}^{1/2} \cdot N_i^{-1/2} \\ &\le 2\tau^{3/2}\{\int_0^\tau A_i(t)dt\}^{1/2} \cdot (\frac{2}{N_i+1})^{1/2} \\ &= (2\tau)^{3/2}\{\int_0^\tau A_i(t)dt\}^{1/2} \cdot (\frac{1}{N_i+1})^{1/2}, \end{aligned}$$

where

$$A_i(t) = \int_0^t \frac{dF_{\mu_i}(u)}{S_{\nu_i}(u)}.$$

Taking expectation with respect to μ_i and ν_i yields

$$\begin{aligned} E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i\ge 1}|N_i\} &\le (2\tau)^{3/2}[\int_0^\tau E\{A_i(t)\}dt]^{1/2} \cdot (\frac{1}{N_i+1})^{1/2} \\ &= (2\tau)^{3/2}[\int_0^\tau E\{A(t)\}dt]^{1/2} \cdot (\frac{1}{N_i+1})^{1/2}, \end{aligned}$$

where

$$A(t) = \int_0^t \frac{dF_{\mu}(u)}{S_{\nu}(u)}$$

and the equality holds since (μ_i, ν_i) are independent realizations of (μ, ν). Let

$$D = (2\tau)^{3/2}[\int_0^\tau E\{A(t)\}dt]^{1/2}.$$

It follows from Condition (C1) that $E\{A(t)\} < \infty$ for all $t \leq \tau$ and thus D is finite. Upon taking the expectation with respect to N_i, we obtain

$$E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i \geq 1}\} \lesssim E\{(\frac{1}{N_i+1})^{1/2}\}.$$

Using Jensen's inequality, it holds that

$$E\{(\frac{1}{N_i+1})^{1/2}\} \leq \{E(\frac{1}{N_i+1})\}^{1/2}.$$

Now by Condition (C2) N_i follows a Poisson distribution with parameter λ_n, then

$$\begin{aligned}
E(\frac{1}{N_i+1}) &= \sum_{k=0}^{\infty} \frac{1}{k+1} e^{-\lambda_n} \frac{(\lambda_n)^k}{k!} \\
&= \sum_{k=0}^{\infty} e^{-\lambda_n} \frac{(\lambda_n)^k}{(k+1)!} \\
&= \frac{1}{\lambda_n} \sum_{k=0}^{\infty} e^{-\lambda_n} \frac{(\lambda_n)^{k+1}}{(k+1)!} \\
&= \frac{1}{\lambda_n} \sum_{k=1}^{\infty} e^{-\lambda_n} \frac{(\lambda_n)^k}{k!} \\
&= \frac{1}{\lambda_n}(1 - e^{-\lambda_n}) \\
&\leq \frac{1}{\lambda_n}.
\end{aligned}$$

We conclude that

$$E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i \geq 1}\} \lesssim \lambda_n^{-1/2}$$

and hence for large enough n,

$$E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\} = E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i \geq 1}\} + E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\mathbf{1}_{N_i = 0}\} = O(\lambda_n^{-1/2}).$$

This shows that

$$[E\{d^2_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\}]^{1/2} = O(\lambda_n^{-1/4})$$

for large enough n.

Since $[E\{|w_i(z)|^2\}]^{1/2}$ are finite, it follows from (10) that

$$E\{\frac{1}{n}\sum_{i=1}^{n} |w_i(z)| d_{\mathcal{W}}(\hat{\mu}_i, \mu_i)\} = O(\lambda_n^{-1/4}),$$

which implies

$$\frac{1}{n}\sum_{i=1}^{n}|w_i(z)|d_{\mathcal{W}}(\hat{\mu}_i,\mu_i)=O_p(\lambda_n^{-1/4}).$$

By (9) and the fact that $W_{0n}(z)$ and $\|W_{1n}(z)\|_E$ are $O_p(n^{-1/2})$, one has

$$d_{\mathcal{W}}\{\hat{m}(z),\tilde{m}(z)\}=O_p(\lambda_n^{-1/4}).$$

Combined with (8), we conclude that

$$d_{\mathcal{W}}\{\hat{m}(z),m(z)\}=O_p(n^{-1/2}+\lambda_n^{-1/4}).$$

For the uniform result over $\|z\|_E\le B$, use the fact that $W_{0n}(z)$, $\|W_{1n}(z)\|_E$ are both $O_p(n^{-1/2})$, and $[E\{|w_i(z)|^2\}]^{1/2}$ is $O(1)$, uniformly over $\|z\|_E\le B$. Applying similar arguments leads to

$$\sup_{\|z\|_E\le B}d_{\mathcal{W}}\{\hat{m}(z),\tilde{m}(z)\}=O_p(\lambda_n^{-1/4}).$$

By Theorem 2 of Petersen and Müller (2019), one has

$$\sup_{\|z\|_E\le B}d_{\mathcal{W}}\{\tilde{m}(z),m(z)\}=O_p(n^{-1/\{2(1+\varepsilon)\}})$$

for any $\varepsilon>0$. Again by the triangle inequality, we conclude that

$$\sup_{\|z\|_E\le B}d_{\mathcal{W}}\{\hat{m}(z),m(z)\}=O_p(n^{-1/\{2(1+\varepsilon)\}}+\lambda_n^{-1/4})$$

for any $\varepsilon>0$. □

S.2 Additional Simulations

S.2.1 Different values of β

To assess the robustness of the WKM model against violations of model assumptions, we conducted the same simulation procedure under Setting II as described in Section 5, which favors the CPH model, with larger values of β. Specifically, we considered $\beta=(0.1,0.2,0.3,0.4,0.5)^{\mathrm{T}}$ and $(1,2,3,4,5)^{\mathrm{T}}$, representing moderate and significant deviations

from the WKM model. The results, reported in Table 4, show a reduction in average MSPEs as the sample size increases, indicating the convergence of the WKM model to the target. The WKM model consistently outperforms the CPH model across different β values. This superior performance is attributed to the ability of the WKM model to accommodate random variation in survival distributions across different subgroups. This is further illustrated in Table 5, which summarizes simulation results with no random variation in survival distributions, where $\lambda|Z$ is a constant equal to $e^{-Z\beta/2}$. In this scenario, the CPH model demonstrates optimal performance, which aligns with expectations given the perfect match between the setting and assumptions of the CPH model. However, when random variation exists, as shown in Table 4, the WKM model outperforms the CPH model.

S.2.2 Effect of baseline hazard function

We also explored a different baseline hazard function from that in Section 5, namely Weibull distributions with a shape parameter $k = 4$ and a scale parameter λ. The corresponding hazard and quantile functions are

$$h_{m(Z)}(t) = 4t^3 \cdot \{E(\lambda|Z)\}^{-4}, \quad F^{-1}_{m(Z)}(p) = E(\lambda|Z) \cdot \{-\log(1-p)\}^{1/4},$$

respectively. The simulation setting analyzed is as follows.

- Setting III: $\lambda|Z$ follows a Gamma distribution with shape parameter $e^{-Z\beta/2}/\rho$ and scale parameter $\rho/e^{-Z\beta/4}$.

In this setting, the scale parameter for the Weibull distribution is $E(\lambda|Z) = e^{-Z\beta/4}$. We carried out $Q = 1000$ simulation runs with sample sizes of $n = 100, 200, 500$, and $\beta = (0.1, 0.2, 0.3, 0.4, 0.5)^{\mathrm{T}}$. The results, presented in Table 6, indicate a decrease in average MSPEs as the sample size increases, demonstrating the convergence of the WKM model towards the target. In Setting III, where a nonlinear baseline hazard function is used, the

Table 4: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM and CPH models with different values of β in Setting II.

		$\rho = 0.05$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Moderate β	WKM	0.0052	0.0026	0.0011	0.0120	0.0071	0.0037
		(0.0019)	(0.0010)	(0.0004)	(0.0023)	(0.0011)	(0.0004)
	CPH	0.0151	0.0135	0.0129	0.0236	0.0260	0.0278
		(0.0065)	(0.0041)	(0.0026)	(0.0131)	(0.0095)	(0.0072)
Large β	WKM	0.0223	0.0196	0.0185	0.0224	0.0202	0.0186
		(0.0284)	(0.0076)	(0.0026)	(0.0155)	(0.0149)	(0.0013)
	CPH	0.1620	0.1924	0.2313	0.2982	0.3761	0.3502
		(1.6707)	(1.7359)	(1.3734)	(2.3006)	(4.7434)	(1.1562)
		$\rho = 0.1$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Moderate β	WKM	0.0083	0.0042	0.0017	0.0149	0.0085	0.0043
		(0.0038)	(0.0019)	(0.0007)	(0.0039)	(0.0019)	(0.0008)
	CPH	0.0426	0.0409	0.0410	0.0729	0.0810	0.0907
		(0.0185)	(0.0126)	(0.0083)	(0.0382)	(0.0322)	(0.0240)
Large β	WKM	0.0225	0.0248	0.0199	0.0232	0.0212	0.0193
		(0.0122)	(0.1393)	(0.0282)	(0.0186)	(0.0187)	(0.0071)
	CPH	0.1294	1.6680	1.2283	0.2853	0.5348	1.1196
		(0.5494)	(44.810)	(20.628)	(2.3528)	(4.4282)	(9.6569)
		$\rho = 0.5$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Moderate β	WKM	0.0325	0.0162	0.0063	0.0379	0.0200	0.0093
		(0.0204)	(0.0095)	(0.0033)	(0.0180)	(0.0081)	(0.0036)
	CPH	0.4422	0.4579	0.4843	0.8242	0.9679	1.1754
		(0.2370)	(0.1995)	(0.1265)	(0.5255)	(0.4678)	(0.4460)
Large β	WKM	0.0433	0.0292	0.0216	0.1124	0.0383	0.0218
		(0.2610)	(0.0978)	(0.0169)	(2.4568)	(0.3967)	(0.0128)
	CPH	0.8479	1.4236	1.4953	10.7234	9.4663	3.7271
		(10.6051)	(25.1739)	(10.6187)	(306.767)	(249.699)	(19.9637)

Table 5: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM and CPH models across different values of β in Setting II, with no random variation in the survival distributions.

		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Moderate β	WKM	0.0021 (0.0002)	0.0011 (0.0001)	0.0005 (0.0000)	0.0091 (0.0007)	0.0056 (0.0003)	0.0031 (0.0001)
	CPH	0.0003 (0.0002)	0.0001 (0.0000)	0.0000 (0.0000)	0.0007 (0.0004)	0.0003 (0.0002)	0.0001 (0.0001)
Large β	WKM	0.0190 (0.0019)	0.0183 (0.0011)	0.0180 (0.0007)	0.0195 (0.0019)	0.0187 (0.0013)	0.0183 (0.0007)
	CPH	0.0002 (0.0002)	0.0000 (0.0000)	0.0000 (0.0000)	0.0003 (0.0003)	0.0001 (0.0001)	0.0000 (0.0000)

proportional hazards assumption is maintained, paralleling the semi-parametric nature of the CPH model and presenting a greater challenge for the WKM model. When random variation is minimal, with $\rho = 0.05$, the WKM model tends to underperform relative to the CPH model. However, as the degree of random variation increases, the performance of the WKM model enhances significantly, soon outstripping that of the CPH model.

S.2.3 Comparison with survival mixture density networks

Recent research has explored the use of neural networks to improve feature representation in survival analysis. Tang et al. (2022) proposed modeling the distribution of event time through an ordinary differential equation (ODE) to numerically evaluate the likelihood and gradients. More recently, to address the high computational complexity of neural ODE solvers, Han et al. (2022) introduced survival mixture density networks (MDN) to model survival distributions of right-censored data.

To compare the performance of the WKM model and survival MDN, we follow the simulation setting described in Han et al. (2022), where the one-dimensional predictor Z follows a Bernoulli distribution with parameter $p = 0.5$. The survival distribution is exponential with a scale parameter 0.5 for $Z = 0$, and Rayleigh with a scale parameter 0.5

Table 6: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM and CPH models with a different baseline hazard function.

		$\rho = 0.05$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting III	WKM	0.0003	0.0001	0.0000	0.0019	0.0011	0.0005
		(0.0000)	(0.0000)	(0.0000)	(0.0002)	(0.0001)	(0.0000)
	CPH	0.0001	0.0000	0.0000	0.0002	0.0001	0.0000
		(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)
		$\rho = 0.1$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting III	WKM	0.0031	0.0015	0.0006	0.0047	0.0024	0.0011
		(0.0017)	(0.0008)	(0.0003)	(0.0018)	(0.0008)	(0.0003)
	CPH	0.0135	0.0125	0.0120	0.0299	0.0324	0.0344
		(0.0041)	(0.0028)	(0.0017)	(0.0125)	(0.0097)	(0.0069)
		$\rho = 0.5$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting III	WKM	0.0059	0.0028	0.0011	0.0075	0.0038	0.0016
		(0.0033)	(0.0016)	(0.0006)	(0.0032)	(0.0016)	(0.0006)
	CPH	0.0373	0.0349	0.0340	0.0835	0.0924	0.1011
		(0.0114)	(0.0080)	(0.0049)	(0.0345)	(0.0289)	(0.0210)

for $Z = 1$. The censoring time follows a uniform distribution on $[0, 2]$.

We performed $Q = 100$ simulation runs with sample sizes $n = 100, 200, 500$. Due to the training of neural networks, survival MDN is time-consuming with a runtime of 271.14 minutes, while the WKM model completes in 39.85 seconds. For the qth simulation run, the quality of the estimation is quantified by the mean squared prediction error (MSPE) using the Wasserstein metric:

$$\text{MSPE}_q = E_Z[d^2_{\mathcal{W}}\{\hat{m}_q(Z), m(Z)\}] = \frac{1}{2}[d^2_{\mathcal{W}}\{\hat{m}_q(0), m(0)\} + d^2_{\mathcal{W}}\{\hat{m}_q(1), m(1)\}],$$

where $m(\cdot)$ is the true regression function and $\hat{m}_q(\cdot)$ is the predicted regression function. The average MSPEs and the corresponding standard deviations for $Q = 100$ simulation runs using the WKM model and survival MDN are summarized in Table 7. We observe that the WKM model outperforms survival MDN across all sample sizes, suggesting the

Table 7: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM and survival MDN under the simulation setting in Han et al. (2022).

n	100	200	500
WKM	0.0276 (0.0020)	0.0196 (0.0008)	0.0138 (0.0003)
Survival MDN	0.1673 (0.0058)	0.1535 (0.0074)	0.1606 (0.0073)

superior performance of the WKM model under this simulation setting.

S.2.4 Comparison with random survival forests

Random survival forests (RSF) (Ishwaran et al., 2008) is an extension of the random forests method (Breiman, 2001) designed for analyzing right-censored survival data. To evaluate the performance of the WKM model against RSF, we employed the simulation procedure outlined in Section 5. Due to the computational intensity of RSF, we conducted $Q = 100$ simulation runs with sample sizes $n = 50, 100, 200$. Table 8 summarizes the comparison results. The WKM model consistently outperformed RSF across all simulation scenarios. Due to the existence of random variation in survival distributions across different subgroups, the average MSPE of RSF is substantial even with a sample size of 200.

To further investigate this performance difference, we conducted an additional simulation without random variation in survival distributions, mirroring the approach in subsection S.2.1. In this scenario, $\lambda|Z$ was set as a constant equal to $Z\beta + 0.1$ for Setting I and $e^{-Z\beta/2}$ for Setting II. As shown in Table 9, the WKM model still outperforms RSF in most cases, with exceptions only occurring in Setting II with a 50% censoring rate. These findings suggest that the WKM model offers robust performance across various survival data scenarios, particularly in the presence of random variations in survival distributions.

Table 8: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM model and RSF.

		$\rho = 0.05$					
		20% censoring			50% censoring		
	n	50	100	200	50	100	200
Setting I	WKM	0.0067	0.0033	0.0016	0.0073	0.0033	0.0017
		(0.0058)	(0.0019)	(0.0009)	(0.0043)	(0.0015)	(0.0009)
	RSF	0.0559	0.0665	0.0855	0.0552	0.0757	0.1383
		(0.0436)	(0.0454)	(0.0484)	(0.0370)	(0.0417)	(0.0716)
Setting II	WKM	0.0155	0.0064	0.0030	0.0342	0.0192	0.0112
		(0.0040)	(0.0021)	(0.0009)	(0.0067)	(0.0025)	(0.0013)
	RSF	0.0520	0.0424	0.0319	0.0575	0.0492	0.0561
		(0.0145)	(0.0155)	(0.0151)	(0.0205)	(0.0199)	(0.0323)
		$\rho = 0.1$					
		20% censoring			50% censoring		
	n	50	100	200	50	100	200
Setting I	WKM	0.0115	0.0066	0.0034	0.0126	0.0059	0.0031
		(0.0120)	(0.0054)	(0.0023)	(0.0113)	(0.0032)	(0.0013)
	RSF	0.1152	0.1603	0.2415	0.1290	0.1704	0.3530
		(0.1427)	(0.1422)	(0.2614)	(0.2048)	(0.1061)	(0.2041)
Setting II	WKM	0.0224	0.0097	0.0047	0.0401	0.0223	0.0127
		(0.0097)	(0.0047)	(0.0018)	(0.0096)	(0.0043)	(0.0021)
	RSF	0.0879	0.0798	0.0716	0.0925	0.0936	0.1348
		(0.0296)	(0.0341)	(0.0249)	(0.0319)	(0.0419)	(0.0936)
		$\rho = 0.5$					
		20% censoring			50% censoring		
	n	50	100	200	50	100	200
Setting I	WKM	0.0428	0.0216	0.0134	0.0431	0.0258	0.0150
		(0.0579)	(0.0251)	(0.0165)	(0.0544)	(0.0253)	(0.0175)
	RSF	0.6865	0.8618	1.6046	0.7288	1.2612	2.8392
		(0.8780)	(1.3000)	(2.8725)	(1.0304)	(1.3942)	(3.2091)
Setting II	WKM	0.0746	0.0364	0.0173	0.0893	0.0448	0.0242
		(0.0373)	(0.0186)	(0.0090)	(0.0406)	(0.0184)	(0.0082)
	RSF	0.4519	0.4907	0.6297	0.4259	0.5349	1.0132
		(0.2168)	(0.2151)	(0.3140)	(0.2413)	(0.2713)	(0.4937)

Table 9: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM model and RSF with no random variation in the survival distributions.

		20% censoring			50% censoring		
	n	50	100	200	50	100	200
Setting I	WKM	0.0003	0.0001	0.0001	0.0010	0.0006	0.0003
		(0.0001)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)
	RSF	0.0009	0.0005	0.0002	0.0011	0.0006	0.0004
		(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0002)	(0.0002)
Setting II	WKM	0.0087	0.0036	0.0016	0.0288	0.0163	0.0100
		(0.0016)	(0.0004)	(0.0001)	(0.0036)	(0.0013)	(0.0005)
	RSF	0.0235	0.0114	0.0043	0.0277	0.0166	0.0087
		(0.0050)	(0.0025)	(0.0013)	(0.0055)	(0.0047)	(0.0028)

S.2.5 High-dimensional predictors

To assess the performance of the WKM model with high-dimensional predictors, we extended the simulation procedure outlined in Section 5 to incorporate a 50-dimensional predictor Z. Each dimension of Z independently follows a Bernoulli distribution with parameter $p = 0.5$. In both Setting I and Setting II, the parameter β is a 50-dimensional vector with all entries equal to 0.01. We conducted $Q = 100$ simulation runs with sample sizes $n = 100, 200, 500$. For the qth simulation run, we quantified the estimation quality using the mean squared prediction error (MSPE) based on the Wasserstein metric:

$$\text{MSPE}_q = E_Z[d^2_{\mathcal{W}}\{\hat{m}_q(Z), m(Z)\}] = \frac{1}{n}\sum_{i=1}^{n} d^2_{\mathcal{W}}\{\hat{m}_q(Z_i), m(Z_i)\},$$

where $m(\cdot)$ represents the true regression function and $\hat{m}_q(\cdot)$ is the predicted regression function.

Table 10 summarizes the average MSPEs and corresponding standard deviations for the $Q = 100$ simulation runs, comparing the WKM and CPH models. The results demonstrate that, similar to the findings in Section 5, the WKM model consistently outperforms the CPH model across most simulation scenarios. The only exception occurs under the proportional hazards assumption with a 50% censoring rate. These findings suggest that the WKM model maintains its superior performance even when dealing with high-dimensional pre-

Table 10: Average mean squared prediction errors and standard deviations (in parentheses) for the WKM and CPH models with 50-dimensional predictors.

		$\rho = 0.05$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting I	WKM	0.0255	0.0135	0.0053	0.0257	0.0137	0.0058
		(0.0065)	(0.0028)	(0.0012)	(0.0060)	(0.0027)	(0.0011)
	CPH	0.0464	0.0375	0.0332	0.0551	0.0524	0.0632
		(0.0149)	(0.0102)	(0.0070)	(0.0218)	(0.0168)	(0.0159)
Setting II	WKM	0.0327	0.0153	0.0060	0.0427	0.0222	0.0096
		(0.0062)	(0.0023)	(0.0012)	(0.0049)	(0.0025)	(0.0009)
	CPH	0.0331	0.0205	0.0136	0.0363	0.0281	0.0244
		(0.0082)	(0.0040)	(0.0021)	(0.0085)	(0.0068)	(0.0051)
		$\rho = 0.1$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting I	WKM	0.0477	0.0248	0.0104	0.0471	0.0254	0.0106
		(0.0131)	(0.0053)	(0.0020)	(0.0118)	(0.0062)	(0.0023)
	CPH	0.1199	0.0944	0.0878	0.1322	0.1487	0.1858
		(0.0497)	(0.0257)	(0.0190)	(0.0565)	(0.0613)	(0.0571)
Setting II	WKM	0.0612	0.0272	0.0108	0.0692	0.0344	0.0150
		(0.0130)	(0.0054)	(0.0019)	(0.0109)	(0.0055)	(0.0019)
	CPH	0.0755	0.0486	0.0360	0.0872	0.0727	0.0744
		(0.0193)	(0.0097)	(0.0058)	(0.0291)	(0.0237)	(0.0168)
		$\rho = 0.5$					
		20% censoring			50% censoring		
	n	100	200	500	100	200	500
Setting I	WKM	0.2183	0.1078	0.0469	0.1835	0.1025	0.0447
		(0.1523)	(0.0482)	(0.0150)	(0.1136)	(0.0374)	(0.0102)
	CPH	0.9134	0.8173	0.7579	1.0486	1.4640	1.8646
		(0.8610)	(0.3994)	(0.2764)	(1.0695)	(0.8945)	(0.9308)
Setting II	WKM	0.2393	0.1325	0.0527	0.2542	0.1301	0.0534
		(0.0684)	(0.0289)	(0.0117)	(0.0710)	(0.0291)	(0.0090)
	CPH	0.5281	0.4339	0.3753	0.6583	0.6900	0.7775
		(0.2462)	(0.1176)	(0.0668)	(0.2453)	(0.2559)	(0.1827)

dictors, further highlighting its robustness and versatility in survival analysis tasks. More importantly, the WKM model proves to be scalable to high-dimensional settings, requiring only 18% more time to process a 50-dimensional predictor compared to a 5-dimensional predictor.