Journal of the Royal Statistical Society Series B: Statistical Methodology, 2026, **00**, 1–21
https://doi.org/10.1093/jrsssb/qkag060

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Original Article

Nonparametric inference for censored data using deep neural networks

Wen Su[1,†], Qiang Wu[2,†], Kin-Yat Liu[3], Guosheng Yin[4], Jian Huang[2] and Xingqiu Zhao[2]

[1]Department of Biostatistics, City University of Hong Kong, Hong Kong
[2]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
[3]Department of Statistics, The Chinese University of Hong Kong, Hong Kong
[4]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

Address for correspondence: Xingqiu Zhao, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Email: xingqiu.zhao@polyu.edu.hk

Abstract

We propose a novel deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach enhances model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We establish the nonasymptotic error bound and functional asymptotic normality of the proposed estimator. Subsequently, we develop new one-sample tests for goodness-of-fit evaluation and two-sample tests for treatment comparison. Notably, we design a new test specifically tailored for testing nonparametric Cox models. The consistency of these tests is established by analyzing the power functions. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.

Keywords: asymptotic normality, asymptotic power, censored data, goodness-of-fit, neural networks, nonparametric inference

1 Introduction

Censored data are often encountered in many fields such as econometrics, biology, social sciences, and clinical trials. Right censoring is the most typical mechanism in survival analysis. For example, in the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT) (Knaus et al., 1995), seriously ill patients from five tertiary care academic centers across the United States were followed for a 180-day period during 1989–1991 and 1992–1994. The exact survival time of subjects who died during the follow-up period was recorded, while those who exited the study alive were treated as right-censored. Commonly used semiparametric models for survival data include the Cox proportional hazards (PH) model (Andersen & Gill, 1982; Cox, 1972, 1975), the accelerated failure time (AFT) model (Buckley & James, 1979; Tsiatis, 1990; Wei et al., 1990), and the additive hazards (AH) model (Aalen, 1980; Cox & Oakes, 1984; Lin & Ying, 1994; Mckeague & Sasieni, 1994). Furthmore, there are some studies on the time-varying coefficient Cox model. For example, Cai et al. (2008) proposed a partial linear hazard regression model with a pseudo-partial likelihood estimation method, providing a theoretically robust and empirically validated approach for flexible and efficient multivariate survival data analysis; Yan and Huang (2012) developed an adaptive group lasso method that

† Wen Su and Qiang Wu are co-first authors.

Received: October 1, 2024. **Revised:** March 11, 2026. **Accepted:** March 14, 2026
© The Royal Statistical Society 2026.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

simultaneously selects important variables and determines whether their effects are time-independent or time-varying.

However, most of the existing methods impose structural and functional assumptions on conditional hazards or conditional distributions of survival time, which may lead to unreliable and even spurious inference and prediction results when the underlying model assumptions are not satisfied. For hypothesis testing in survival analysis, Fleming and Harrington (1991) and Klein and Moeschberger (2003) provided comprehensive reviews of one-sample and two-sample tests, including the Kolmogorov–Smirnov test, log-rank test, and a class of linear rank tests. These tests do not consider conditional survival functions and thus cannot be used for testing or comparing conditional survival functions. When covariate effects are involved, most goodness-of-fit tests focus on evaluating the assumptions of the popular Cox model (Fox & Weisberg, 2019; Grambsch & Therneau, 1994; Hosmer et al., 2008; May & Hosmer, 1998).

Rapid growth in computational power and developments of machine learning tools open up a myriad of new possibilities for data analysis, which enable us to build sumptuous extensions on traditional approaches. Examples of deep learning applications include computer vision (Krizhevsky et al., 2012; Russakovsky et al., 2015), natural language processing (Wu et al., 2016), and reinforcement learning (Silver et al., 2016; van Otterlo & Wiering, 2012). Refer to Chakraborty et al. (2017), Murdoch et al. (2019) and Rudin (2019) for details on interpretability of neural networks. There have also been interesting works on applications of deep learning in survival analysis. For example, the random survival forest (RSF) (Ishwaran et al., 2008) uses the log-rank test as a splitting rule to construct decision trees and estimate cumulative hazards by aggregating the cumulative hazard functions of leaf nodes across the ensemble. DeepHit (Lee et al., 2018) leverages deep neural networks (DNNs) to directly estimate the cumulative incidence function (CIF) for time-to-event outcomes. DeepSurv (Katzman et al., 2018) extends the Cox PH model by using deep neural networks to capture nonlinear relationships, while the PH assumption is required. However, none of the these methods have studied the asymptotic properties of the estimators.

Kvamme et al. (2019) extended the Cox model using neural networks and constructed an approximated partial log-likelihood for estimation. Sun et al. (2020) combined the Cox PH model and the deep learning framework to propose a survival prediction method for analyzing high-dimensional data. Survival stacking, introduced by Craig et al. (2021), reformulates survival problems as classification problems by reshaping survival data. However, different from the Cox model, survival stacking does not require the PH assumption, as the flexibility of the model depends on the choice of the classification method. More recently, Zhong et al. (2021, 2022) provided a comprehensive literature review on combining DNN and survival analysis methods, and developed DNN approaches for a partly linear Cox model and a general class of hazards models; and more references on deep survival analysis can be found therein. Wolock et al. (2024) proposed a decomposition-based conditional survival function estimation framework, which mitigates the complexity involved in directly dealing with censoring and truncation. Tarkhan and Simon (2023) introduced a scalable survival analysis framework based on a modified Cox PH model, which uses stochastic gradient descent (SGD) for efficient computation on large datasets. This method provides asymptotically valid confidence intervals in linear Cox PH models.

With the goal to provide a novel framework of DNN-based statistical inference for survival analysis, this paper makes four significant contributions:

(i) We propose a novel model-free approach for estimating the log conditional hazard function and survival function using DNNs without imposing any assumption on its form. As a result, any potential bias resulting from model misspecification can be mitigated. It is worth highlighting that we assume a conditional hazard rate function that does not have a baseline hazard in contrast to the Cox model, which, to the best of our knowledge, is completely new to the literature.

(ii) We establish the nonasymptotic error bound and the functional asymptotic normality of the proposed estimator. These results are crucial theoretical foundations for formulating test statistics and deriving the asymptotic properties of tests required for hypothesis testing.

(iii) We construct test statistics for one-sample and two-sample tests as well as goodness-of-fit tests, and further derive the asymptotic normality of the proposed test statistics. The

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

consistency of the proposed tests is established by analyzing the asymptotic power functions. In particular, we develop a new goodness-of-fit test for testing nonparametric Cox models. These testing methods are breakthroughs in survival analysis, which bridge statistical inferences and deep learning.

(iv) We provide comprehensive examples in simulation studies under the Cox, AFT and AH models in comparisons with existing methods and demonstrate the robustness of the proposed DNN method.

The remainder of this paper is organized as follows. Section 2 introduces the proposed methodology with setups for the DNNs and exhaustive details. Section 3 presents the asymptotic error bound, and in Section 4 we first establish the functional asymptotic normality and then construct the test statistics for hypothesis testing on goodness-of-fit as well as one-sample and two-sample tests for conditional hazard functions. Simulation studies and application to the glioma biopsy imaging dataset are provided in Sections 5 and 6, respectively. Section 7 concludes with some remarks. A test of nonparametric Cox models and all technical proofs as well as additional applications are relegated to the online supplementary material.

2 Methodology

2.1 Deep neural networks

We provide a brief review for the use of a DNN as a function approximator (Goodfellow et al., 2016). Let L be a positive integer and $\boldsymbol{p} = (p_0, p_1, \ldots, p_L, p_{L+1})$ be a sequence of positive integers. An $(L+1)$-layer DNN with width $\boldsymbol{p}$ is a composite function $g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ recursively defined as

$$\begin{aligned} &g(\boldsymbol{u}) = W_L g_L(\boldsymbol{u}) + v_L, \\ &g_L(\boldsymbol{u}) = \sigma(W_{L-1} g_{L-1}(\boldsymbol{u}) + v_{L-1}), \ \ldots, g_1(\boldsymbol{u}) = \sigma(W_0 \boldsymbol{u} + v_0), \end{aligned} \tag{1}$$

where the matrices $W_l \in \mathbb{R}^{p_{l+1} \times p_l}$ and vectors $v_l \in \mathbb{R}^{p_{l+1}}$ (for $l = 0, \ldots, L$) are the parameters of this DNN g, and σ is the component-wise activation function, i.e. $\sigma((u_1, \ldots, u_{p_l})^\top) = (\sigma(u_1), \ldots, \sigma(u_{p_l}))^\top$. This leads to $g_l = (g_{l1}, \ldots, g_{lp_l})^\top : \mathbb{R}^{p_{l-1}} \to \mathbb{R}^{p_l}$ for $l = 0, \ldots, L$. The most popular activation function is called the rectified linear unit (ReLU) defined in Nair and Hinton (2010): $\sigma(u) = \max\{u, 0\}$. For the DNN in (1), L is the depth of the network and vector $\boldsymbol{p}$ is the width of each layer, p_0 is the dimension of the input variables, $p_1, \ldots, p_L$ are the dimensions of the L hidden layers, and p_{L+1} is the dimension of the output layer. The matrix entries $(W_l)_{i,j}$ are the weight linking the jth neuron in layer l to the ith neuron in layer $l+1$, and the vector entries $(v_l)_i$ represent a shift term associated with the ith neuron in layer $l+1$.

Following Schmidt-Hieber (2020), we consider DNNs with all parameters bounded by one. Let $\mathbb{N}_+$ be the set of all positive natural numbers. Given $L \in \mathbb{N}_+$ and $\boldsymbol{p} \in \mathbb{N}_+^{L+2}$, we consider a class of DNNs:

$$\begin{aligned} \mathcal{G}(L, \boldsymbol{p}, B) = \{&g : g \text{ is a DNN with } (L+1) \text{ layers and width vector } \boldsymbol{p} \text{ such that} \\ &\max\{\|W_l\|_\infty, \|v_l\|_\infty\} \le 1, \ \text{for all } l = 0, \ldots, L, \ \|g\|_\infty \le B\}, \end{aligned} \tag{2}$$

where $\|\cdot\|_\infty$ denotes the sup-norm of a matrix, a vector or a function, and $B > 1$.

2.2 Estimation with survival data

Let T and C denote survival and censoring times respectively for each subject, and let X be a p-dimensional covariate vector of interest. In the presence of right censoring, the observed data consist of $\{Z_i = (Y_i, \Delta_i, X_i); \ i = 1, \ldots, n\}$, which are independent copies of $Z = (Y, \Delta, X)$, with $Y = \min(T, C)$ and $\Delta = I(T \le C)$. Assume that T and C are independent conditional on X.

With no baseline hazard, the conditional hazard rate function of T given $X = x$ is directly modeled as $\lambda(t|x) = \exp\{g(t, x)\}$, where $g : [0, \tau] \times [0, 1]^p \to \mathbb{R}$ is an unknown function. The conditional survival function and density function of T given $X = x$ are respectively given by

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

$$S(t|x) = \exp\left\{-\int_0^t \lambda(s|x)ds\right\} \quad \text{and} \quad f(t|x) = \lambda(t|x)\exp\left\{-\int_0^t \lambda(s|x)ds\right\}.$$

The log-likelihood function of g based on the observed data $Z = \{Y, \Delta, X\}$ is given by

$$l(g; Z) = \log\left\{ f(Y|X)^{\Delta} S(Y|X)^{1-\Delta} \right\} = \Delta g(Y, X) - \int_0^Y \exp\{g(s, X)\}ds.$$

Let $\mathbb{P}$ and $\mathbb{P}_n$ denote the probability measure and empirical measure, respectively. The log-likelihood of g based on the observed data is given by $l_n(g) = \mathbb{P}_n l(g; Z)$. Let $\mathcal{G} = \mathcal{G}(L, \boldsymbol{p}, B)$ with $p_0 = p + 1$ and $p_{L+1} = 1$. Let g_0 denote the log transformation of the true conditional hazard rate function of T given $X = x$. Define an estimator of g_0 by $\hat{g}_n \in \operatorname{argmax}_{g\in\mathcal{G}} l_n(g)$.

2.3 Computation

Our proposed approach is implemented using `PyTorch` as the framework. The maximization of the log-likelihood is performed by the `Adam` optimizer due to its reliability across numerous models and datasets (Kingma & Ba, 2015). The integral in the log-likelihood is evaluated using Gaussian quadrature with 100 nodes in all numerical experiments.

For the initial values of the neural network parameters, we adopt the default random initialization method. For simplicity, we use the same number of neurons in all hidden layers. Table 1 in the online supplementary material lists the candidate ranges of DNN hyperparameters, including the number of hidden layers, the number of neurons per layer, the dropout rate and the learning rate, considered in the simulations and data applications.

The maximum number of training epochs is set to 1,000. To further alleviate overfitting during the training process, we implement an early stopping strategy. Specifically, if the validation loss does not improve to be less than the current minimum for 10 consecutive epochs, training is stopped. The DNN parameters corresponding to the epoch with the lowest validation loss are then retained (Raskutti et al., 2014). Algorithm 1 of the online supplementary material lists the detailed computation procedure of the proposed estimator.

3 Nonasymptotic error bound

To establish the theoretical properties of the proposed DNN estimator $\hat{g}_n$, we assume the true function g_0 belongs to a Hölder class of smooth functions, which is fairly broad as discussed by Schmidt-Hieber (2020), Zhong et al. (2022) and Jiao et al. (2023). Specifically, The ball of α-Hölder functions with radius $B_0 > 0$ is define as

$$\mathcal{C}^{\alpha}_{p+1}(B_0) = \Big\{ g : [0, \tau] \times [0, 1]^p \subset \mathbb{R}^{p+1} \to \mathbb{R} : \sum_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|<\alpha} \|D^{\boldsymbol{\gamma}} g\|_{\infty} + \sum_{\boldsymbol{\gamma}:|\boldsymbol{\gamma}|=\lfloor\alpha\rfloor} \sup_{\boldsymbol{u}\neq\boldsymbol{v}} \frac{|D^{\boldsymbol{\gamma}} g(\boldsymbol{u}) - D^{\boldsymbol{\gamma}} g(\boldsymbol{v})|}{|\boldsymbol{u} - \boldsymbol{v}|_{\infty}^{\alpha - \lfloor\alpha\rfloor}} \le B_0 \Big\},$$

where $\lfloor\alpha\rfloor$ is the largest integer strictly smaller than α, $D^{\boldsymbol{\gamma}} g(\boldsymbol{u}) = \partial^{|\boldsymbol{\gamma}|} g(\boldsymbol{u}) / \partial u_1^{\gamma_1} \cdots \partial u_{p+1}^{\gamma_{p+1}}$ with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p+1}) \in \mathbb{N}^{p+1}$ and $|\boldsymbol{\gamma}| = \sum_{k=1}^{p+1} \gamma_k$.

To establish the theoretical properties of the proposed estimator, we need the following regularity conditions.

(C1) The true nonparametric function g_0 belongs to $\mathcal{F}_0 = \mathcal{C}^{\alpha}_{p+1}(B_0)$.

(C2) The covariate X takes values in a bounded subset of $\mathbb{R}^p$ with a bounded density function. Without loss of generality, we assume $X \in [0, 1]^p$.

(C3) The study stops at a finite time $\tau > 0$, and there exists a small positive constant c_1 such that $P(\Delta = 1|X) \ge c_1$.

Condition (C1) assumes the smoothness of g_0, ensuring the existence of all partial derivatives up to the $\lfloor\alpha\rfloor$th order. This condition is common in smoothing techniques for nonparametric inference. Conditions (C2) and (C3) are standard assumptions in traditional survival analysis.

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

For any function $g \in \mathcal{F}_0$, define $\|g\|_2 = [E\{\Delta g^2(T, X)\}]^{1/2}$ and for any functions $g_1, g_2 \in \mathcal{F}_0$, define $d(g_1, g_2) = \|g_1 - g_2\|_2$.

Theorem 3.1 ((Nonasymptotic error bound)). Suppose that conditions (C1)–(C3) hold. For any $N, M \in \mathbb{N}_+$, let $\mathcal{G} = \mathcal{G}(L, \boldsymbol{p}, B)$ be the function class of ReLU multilayer perceptrons with width $\|\boldsymbol{p}\|_\infty = 38(\lfloor\alpha\rfloor + 1)^2 p_0^{\lfloor\alpha\rfloor+1} N\lfloor\log_2(8N)\rfloor$ and depth $L = 21(\lfloor\alpha\rfloor + 1)^2 M\lfloor\log_2(8M)\rfloor$. Let $\mathcal{S} = \sum_{l=0}^{L} p_{l+1}(p_l + 1)$ denote the size of $\mathcal{G}$, and $U = L\prod_{l=0}^{L}(p_l + 1)$. Then, we have

$$\mathbb{E}d^2(\hat{g}_n, g_0) \le c(B)\frac{\mathcal{S}\log(nU)}{n} + c_0 B_0^2(\lfloor\alpha\rfloor + 1)^4 p_0^{2\lfloor\alpha\rfloor+(\alpha\vee 1)}(NM)^{-4\lfloor\alpha\rfloor/p_0}, \tag{3}$$

where c_0 is a universal constant, and $c(B)$ is a constant depending on B.

In addition to Conditions (C1)–(C3), Theorem 3.1 requires some neural network structures to derive the nonasymptotic error bound. The conditions on the DNN are required for establishing the nonasymptotic DNN approximation error bound (Jiao et al., 2023). By suitably choosing M and N to balance stochastic and approximation errors, one can attain the optimal convergence rate. In particular, we present three different network parameter specifications:

(i) Choose $N = O((\frac{n}{\log n})^{\frac{p_0}{2p_0+4\lfloor\alpha\rfloor}})$ and let M be a fixed positive constant independent of n. Then the depth of $\mathcal{G}$ satisfies $L = O(1)$, the width of $\mathcal{G}$ satisfies $\|\boldsymbol{p}\|_\infty = O\left(n^{\frac{p_0}{2p_0+4\lfloor\alpha\rfloor}}(\log n)^{\frac{p_0+4\lfloor\alpha\rfloor}{2p_0+4\lfloor\alpha\rfloor}}\right)$, $U = O(L\|\boldsymbol{p}\|_\infty^{L+1})$, and the size of $\mathcal{G}$ satisfies

$$\mathcal{S} = O\big(L\|\boldsymbol{p}\|_\infty^2\big) = O\left(n^{\frac{p_0}{p_0+2\lfloor\alpha\rfloor}}(\log n)^{\frac{p_0+4\lfloor\alpha\rfloor}{p_0+2\lfloor\alpha\rfloor}}\right);$$

(ii) Choose $M = O\left((\frac{n}{\log n})^{\frac{p_0}{2p_0+4\lfloor\alpha\rfloor}}\right)$ and let N be a fixed positive constant independent of n. Then the depth of $\mathcal{G}$ satisfies $L = O\left(n^{\frac{p_0}{2p_0+4\lfloor\alpha\rfloor}}(\log n)^{\frac{p_0+4\lfloor\alpha\rfloor}{2p_0+4\lfloor\alpha\rfloor}}\right)$, the width of $\mathcal{G}$ satisfies $\|\boldsymbol{p}\|_\infty = O(1)$, $U = O(L\|\boldsymbol{p}\|_\infty^{L+1})$, and the size of $\mathcal{G}$ satisfies

$$\mathcal{S} = O\big(L\|\boldsymbol{p}\|_\infty^2\big) = O\left(n^{\frac{p_0}{2p_0+4\lfloor\alpha\rfloor}}(\log n)^{\frac{p_0+4\lfloor\alpha\rfloor}{2p_0+4\lfloor\alpha\rfloor}}\right);$$

(iii) Choose $M = O\left((\frac{n}{\log n})^{\frac{p_0}{4(p_0+2\lfloor\alpha\rfloor)}}\right)$ and $N = O\left((\frac{n}{\log n})^{\frac{p_0}{4(p_0+2\lfloor\alpha\rfloor)}}\right)$. Then the depth of $\mathcal{G}$ satisfies $L = O\left(n^{\frac{p_0}{4(p_0+2\lfloor\alpha\rfloor)}}(\log n)^{\frac{3p_0+8\lfloor\alpha\rfloor}{4(p_0+2\lfloor\alpha\rfloor)}}\right)$, the width of $\mathcal{G}$ satisfies $\|\boldsymbol{p}\|_\infty = O\left(n^{\frac{p_0}{4(p_0+2\lfloor\alpha\rfloor)}}(\log n)^{\frac{3p_0+8\lfloor\alpha\rfloor}{4(p_0+2\lfloor\alpha\rfloor)}}\right)$, $U = O(L\|\boldsymbol{p}\|_\infty^{L+1})$, and the size of $\mathcal{G}$ satisfies

$$\mathcal{S} = O\big(L\|\boldsymbol{p}\|_\infty^2\big) = O\left(n^{\frac{3p_0}{4(p_0+2\lfloor\alpha\rfloor)}}(\log n)^{\frac{3(3p_0+8\lfloor\alpha\rfloor)}{4(p_0+2\lfloor\alpha\rfloor)}}\right).$$

Consequently, the above three different specifications of the network parameters achieve the same convergence rate:

$$\mathbb{E}d(\hat{g}_n, g_0) \le c(B)\frac{\mathcal{S}\log(nU)}{n} + c_0 B_0^2(\lfloor\alpha\rfloor + 1)^4 p_0^{2\lfloor\alpha\rfloor+(\alpha\vee 1)}(NM)^{-4\lfloor\alpha\rfloor/p_0} = O\left(\left(\frac{\log n}{n}\right)^{\frac{\lfloor\alpha\rfloor}{2\lfloor\alpha\rfloor+p_0}}\right).$$

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

4 Asymptotic results

We first establish the asymptotic distribution for the estimator $\hat{g}_n$, and then apply the result to hypothesis testing.

4.1 Asymptotic functional normality

Theorem 4.1 ((Asymptotic normality)). Suppose that conditions (C1)–(C3) hold and $\sqrt{n}\tau_n^2 = o_p(1)$ where τ_n^2 is the right-hand side of (3). Then, for any $h \in \mathcal{F}_0$ with $\|h\|_\infty \leq 1$, we have

$$\sqrt{n}\mathbb{P}\Big[\int_0^Y h(s, X)\{\hat{g}_n(s, X) - g_0(s, X)\} \exp\left(g_0(s, X)\right) ds\Big] = \sqrt{n}\mathbb{P}_n \psi(g_0; Z)[h] + o_p(1),$$

where

$$\psi(g; Z)[h] = \Delta h(Y, X) - \int_0^Y \exp\{g(s, X)\} h(s, X) ds.$$

Therefore,

$$\sqrt{n}\mathbb{P}\big[\Delta h(Y, X)\{\hat{g}_n(Y, X) - g_0(Y, X)\}\big] = \sqrt{n}\mathbb{P}\Big[\int_0^Y h(s, X)\{\hat{g}_n(s, X) - g_0(s, X)\} e^{g_0(s,X)} ds\Big]$$

converges to a normal distribution with mean zero and variance $\sigma_h^2 = \mathbb{E}\{\psi^2(g_0; Z)[h]\}$.

A key condition in Theorem 4.1 is $\sqrt{n}\tau_n^2 = o_p(1)$, where τ_n^2 denotes the nonasymptotic error bound. That means the convergence rate of the proposed DNN estimator should be $o(n^{-1/4})$, which is a fundamental requirement for establishing the asymptotic distribution theory of estimators. This asymptotic result plays a fundamental role for developing hypothesis testing procedures as discussed in the next section.

4.2 Hypothesis testing

4.2.1 One-sample tests

To test the null hypothesis $H_0 : g = g_0$, we let $\hat{g}_n(t, x)$ be the proposed DNN estimator, and define a class of test statistics as

$$T_w = \sqrt{n}\mathbb{P}_n\big[\Delta W_n(Y, X)\{\hat{g}_n(Y, X) - g_0(Y, X)\}\big],$$

where W_n is a weight function.

Theorem 4.2 ((Asymptotic distribution of the test statistic)). Suppose the conditions in Theorem 4.1 hold, W_n is a bounded weight process and there exists $W \in \mathcal{F}_0$ with $\|W\|_\infty \leq 1$ such that $\{\mathbb{P}(W_n - W)^2\}^{1/2} = O_p(n^{-1/4})$. Then under H_0, T_w converges to a normal distribution with mean zero and variance $\sigma_w^2 = \mathbb{E}\{\psi^2(g_0; Z)[W]\}$, where σ_w^2 can be consistently estimated by

$$\hat{\sigma}_w^2 = \frac{1}{n}\sum_{i=1}^n \Big[\Delta_i W_n(Y_i, X_i) - \int_0^{Y_i} \exp\{\hat{g}_n(s, X_i)\} W_n(s, X_i) ds\Big]^2.$$

For the choice of a weight function, we consider the Fleming Harrington (1991) weight with parameters (ρ, γ),

$$W_n^{(\rho,\gamma)}(t, x) = \frac{1}{n}\sum_{i=1}^n I(Y_i \geq t)\hat{S}_n(t, x)^\rho (1 - \hat{S}_n(t, x))^\gamma, \quad \rho \geq 0, \ \gamma \geq 0,$$

where $\hat{S}_n(t, x) = \exp\{-\int_0^t \exp\left(\hat{g}_n(s, x)\right) ds\}$.

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

For a given significance level α, we reject H_0 if $|T_w/\hat{\sigma}_w| > z_{\alpha/2}$ where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Define the power function of the test by $\Gamma_n(g_1) = P(|T_w/\hat{\sigma}_w| > z_{\alpha/2}|g = g_1)$ under the alternative hypothesis $H_1 : g = g_1$.

Theorem 4.3 ((Asymptotic power function of the test)). Consider the alternative hypothesis $H_1 : g = g_1 \neq g_0$, and let $\mu = \mathbb{E}\{\Delta W(g_1(T, X) - g_0(T, X))\} \neq 0$ and

$$\sigma^2 = \mathrm{Var}(\psi(g_1; Z)[W] + \Delta W(T, X)(g_1(T, X) - g_0(T, X))) > 0.$$

Then, we have

$$\Gamma_n(g_1) = 1 - \Phi\left(\frac{\sigma_w(g_1)z_{\alpha/2} - \sqrt{n}\mu}{\sigma}\right) + \Phi\left(\frac{-\sigma_w(g_1)z_{\alpha/2} - \sqrt{n}\mu}{\sigma}\right) + o(1),$$

where $\sigma_w^2(g_1) = \mathbb{E}\{\psi^2(g_1; Z)[W]\}$.

It is easy to see that $\Gamma_n(g_0) = \alpha + o(1)$, and the above theorem entails the consistency of the test.

4.2.2 Two-sample tests

Consider two samples: $Z_i^{(k)} = (Y_i^{(k)}, \Delta_i^{(k)}, X_i^{(k)})$ with $Y_i^{(k)} = \min(T_i^{(k)}, C_i^{(k)})$ and $\Delta_i^{(k)} = I(T_i^{(k)} \leq C_i^{(k)})$, $i = 1, \ldots, n_k$; $k = 1, 2$. Suppose the conditional hazard rate function of $T^{(k)}$ given $X^{(k)} = x$ is $\lambda_k(t|x) = \exp\{g_k(t, x)\}$, $k = 1, 2$.

Our goal is to test whether the conditional hazard rate functions of two groups are identical. The null hypothesis is $H_0 : g_1 = g_2 = g_0$. Denote $\mathcal{O}_k = \{Z_i^{(k)}, i = 1, \ldots, n_k\}$ and $\mathcal{O} = \{\mathcal{O}_k, k = 1, 2\}$. Let $n = n_1 + n_2$, and rewrite $\mathcal{O} = \{Z_i, i = 1, \ldots, n\}$ with $Z_i = (Y_i, \Delta_i, X_i)$, $i = 1, \ldots, n$ under the null hypothesis. Let $\hat{g}_k$ and $\hat{g}_0$ be the proposed estimators of g_k and g_0 based on the data $\mathcal{O}_k$ from sample k and the pooled data $\mathcal{O}$, respectively. Define a class of test statistics as

$$U_w = \sqrt{n}\mathbb{P}_n\big[\Delta W_n(Y, X)\{\hat{g}_1(Y, X) - \hat{g}_2(Y, X)\}\big],$$

where W_n is a weight function.

Theorem 4.4 ((Asymptotic distribution of the test statistic)). Suppose the conditions in Theorem 4.2 hold. If $\frac{n_1}{n} \to \xi$ as $n \to \infty$ where $0 < \xi < 1$, then under H_0, U_w converges to a normal distribution with mean zero and variance $\tau_w^2 = \frac{1}{\xi}\sigma_{1,w}^2 + \frac{1}{1-\xi}\sigma_{2,w}^2 = (\frac{1}{\xi} + \frac{1}{1-\xi})\sigma_w^2$, where $\sigma_{k,w}^2 = \mathbb{E}\{\psi^2(g_k; Z^{(k)})[W]\}$, $\sigma_w^2 = \mathbb{E}\{\psi^2(g_0; Z)[W]\}$, τ_w^2 can be consistently estimated by $\hat{\tau}_w^2 = \frac{n}{n_1}\hat{\sigma}_{1,w}^2 + \frac{n}{n_2}\hat{\sigma}_{2,w}^2$ or $\tilde{\tau}_w^2 = \frac{n^2}{n_1 n_2}\hat{\sigma}_w^2$ where $\hat{\sigma}_w^2$ and $\hat{\sigma}_{k,w}^2$ $(k = 1, 2)$ are defined as in Theorem 4.2 based on the pooled sample and sample k, respectively.

To derive the asymptotic power function of the test statistic, we use $\hat{\tau}_w^2$ rather than $\tilde{\tau}_w^2$ because the asymptotic property of $\hat{g}_n$ based on the combined sample is unknown under the alternative hypothesis. For a given significance level α, we reject H_0 if $|U_w/\hat{\tau}_w| > z_{\alpha/2}$. Define the power function of the test by $\Gamma_n(g_1, g_2) = P(|U_w/\hat{\tau}_w| > z_{\alpha/2}|g_1 \neq g_2)$.

Theorem 4.5 ((Asymptotic power function of the test)). Consider the alternative hypothesis $H_1 : g_1 \neq g_2$ with $\eta \neq 0$ where

$$\begin{aligned}\eta = {} & \xi E[\Delta^{(1)} W(T^{(1)}, X^{(1)})\{g_1(T^{(1)}, X^{(1)}) - g_2(T^{(1)}, X^{(1)})\}] \\ & + (1-\xi)E[\Delta^{(2)} W(T^{(2)}, X^{(2)})\{g_1(T^{(2)}, X^{(2)}) - g_2(T^{(2)}, X^{(2)})\}].\end{aligned}$$

For simplicity, assume the censoring times and covariates $(C^{(k)}, X^{(k)})$ for the two samples have the same distribution. Let $f_k(t|x)$ be the conditional density function of $T^{(k)}$ given $X^{(k)}$ with $M_1 \leq f_k(t|x) \leq M_2$, $(t, x) \in [0, \tau] \times [0, 1]^p$ for constants M_1 and M_2, $k = 1, 2$. Then, we have

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

$$\Gamma_n(g_1, g_2) = 1 - \Phi\left(\frac{\sqrt{\frac{1}{\xi}\sigma^2_{1,w} + \frac{1}{1-\xi}\sigma^2_{2,w}}z_{\alpha/2} - \sqrt{n}\eta}{\sqrt{\xi\sigma^2_{1,w}(g_1, g_2) + (1-\xi)\sigma^2_{2,w}(g_1, g_2)}}\right) + \Phi\left(\frac{-\sqrt{\frac{1}{\xi}\sigma^2_{1,w} + \frac{1}{1-\xi}\sigma^2_{2,w}}z_{\alpha/2} - \sqrt{n}\eta}{\sqrt{\xi\sigma^2_{1,w}(g_1, g_2) + (1-\xi)\sigma^2_{2,w}(g_1, g_2)}}\right) + o(1),$$

where

$$\sigma^2_{1,w}(g_1, g_2) = \mathrm{Var}\big(\psi(g_1; Z^{(1)})[W + ((1-\xi)/\xi)W_{11}] + \varphi_W(g_1 - g_2; Z^{(1)})\big)$$

and

$$\sigma^2_{2,w}(g_1, g_2) = \mathrm{Var}\big(\psi(g_2; Z^{(2)})[W + (\xi/(1-\xi))W_{22}] - \varphi_W(g_1 - g_2; Z^{(2)})\big)$$

with $W_{11}(t, x) = W(t, x)\{f_2(t|x)/f_1(t|x)\}$, $W_{22}(t, x) = W(t, x)\{f_1(t|x)/f_2(t|x)\}$, and $\varphi_W(g; Z) = \Delta W(Y, X)g(Y, X)$.

It is easy to see that $\Gamma_n(g_0, g_0) = \alpha + o(1)$, and the above theorem entails the consistency of the test.

4.2.3 *Goodness-of-fit tests*

Consider a hypothesis testing problem,

H_0: T follows model M_0 vs. H_1: T does not follow model M_0.

To construct a goodness-of-fit test, we randomly split the sample into two sub-samples with equal sample size $n_1 = n_2 = n/2$. Denote $\mathcal{O} = \{\Delta_i, Y_i, X_i, i = 1, \ldots, n\} = \mathcal{O}_1 \cup \mathcal{O}_2$. Let $\hat{g}_n(t, x)$ be the proposed DNN estimator of $g_0(t, x)$ based on sample $\mathcal{O}_1$, and let $\hat{g}_0(t, x)$ be an estimator of $g_0(t, x)$ under model M_0 based on sample $\mathcal{O}_2$. Define a class of test statistics as

$$\tilde{T}_w = \sqrt{n}\mathbb{P}_n\big[\Delta W_n(Y, X)\{\hat{g}_n(Y, X) - \hat{g}_0(Y, X)\}\big]$$

where W_n is a weight function.

Theorem 4.6 ((Asymptotic distribution of the test statistic)). Suppose the conditions in Theorem 4.1 hold, W_n is a bounded weight process and there exists $W \in \mathcal{F}_0$ with $\|W\|_\infty \le 1$ such that $\{\mathbb{P}(W_n - W)^2\}^{1/2} = O_p(n^{-1/4})$. Further, assume that $\hat{g}_0(t, x)$ satisfies

$$\sqrt{n}\mathbb{P}_n\big[\Delta W_n(Y, X)\{\hat{g}_0(Y, X) - g_0(Y, X)\}\big] = \sqrt{n}\mathbb{P}_n\phi_w(g_0; Z) + o_p(1)$$

with $E\phi_w(g_0; Z) = 0$. Then under H_0, $\tilde{T}_w$ converges to a normal distribution with mean zero and variance $\tilde{\sigma}^2_w = 2(\tilde{\sigma}^2_{1,w} + \tilde{\sigma}^2_{2,w})$, where $\tilde{\sigma}^2_{1,w} = \mathbb{E}\psi^2(g_0; Z)[W]$ and $\tilde{\sigma}^2_{2,w} = \mathbb{E}\phi^2_w(g_0; Z)$ can be consistently estimated by

$$\widehat{\tilde{\sigma}^2_{1,w}} = \frac{1}{n_1}\sum_{Z_i \in \mathcal{O}_1}\psi^2(\hat{g}_n; Z_i)[W_n] \quad \text{and} \quad \widehat{\tilde{\sigma}^2_{2,w}} = \frac{1}{n_2}\sum_{Z_i \in \mathcal{O}_2}\phi^2_{w_n}(\hat{g}_0; Z_i).$$

Let $\hat{\tilde{\sigma}}_w$ denote the corresponding estimator of $\tilde{\sigma}^2_w$, and then we reject H_0 if $|\tilde{T}_w/\hat{\tilde{\sigma}}_w| > z_{\alpha/2}$. Define the power function of the test by $\Gamma_n(g_1) = P(|\tilde{T}_w/\hat{\tilde{\sigma}}_w| > z_{\alpha/2}|g = g_1)$.

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Theorem 4.7 ((Asymptotic power function of the test)). Consider the alternative hypothesis H_1: T follows model M_1 with $g_1 \neq g_0$ and $\mu = \mathbb{E}\{\Delta W(T, X)(g_1(T, X) - g_0(T, X))\} \neq 0$. Assume the conditions in Theorem 4.6 hold. Then, we have

$$\Gamma_n(g_1) = 1 - \Phi\left(\frac{\tilde{\sigma}_w z_{\alpha/2} - \sqrt{n}\mu}{\tilde{\sigma}_w(g_1)}\right) + \Phi\left(\frac{-\tilde{\sigma}_w z_{\alpha/2} - \sqrt{n}\mu}{\tilde{\sigma}_w(g_1)}\right) + o(1),$$

where $\tilde{\sigma}_w^2(g_1) = 2\mathbb{E}\psi^2(g_1; Z)[W] + 2\mathbb{E}\phi_W^2(g_0; Z) + \mathrm{Var}(\Delta W(Y, X)\{g_1(Y, X) - g_0(Y, X)\})$.

It is easy to see that $\Gamma_n(g_0) = \alpha + o(1)$, and the above theorem entails the consistency of the test. Furthermore, we provide three examples for Theorem 4.6 as detailed in Section C of the online supplementary material.

5 Simulation studies

We conduct four simulation studies to evaluate finite sample performance of the proposed estimation procedure, the one-sample, two-sample, and goodness-of-fit tests.

5.1 Evaluation of the DNN estimator

Under our simulation settings, we denote the covariates as $X = (X_1, X_2, X_3, X_4, X_5, X_6)^\top$, where $X_1 \sim \text{Bernoulli}(1/2)$, $X_2 \sim \text{Bernoulli}(1/4)$, and $(X_3, X_4, X_5, X_6)^\top$ were generated from a multivariate normal distribution with zero mean and a covariance matrix determined by $\mathrm{cov}(X_j, X_k) = 1$ for $j = k$ and 0.5 for $j \neq k$, with each component truncated by $[-2, 2]$ and transformed as $(X_j + 2)/4$ $(j = 3, 4, 5, 6)$, resulting in values within $[0, 1]$. We generated the survival time T from the following seven scenarios, including the Cox proportional hazards model with linear and nonlinear covariate components, as well as the Cox model with time-varying coefficients:

Cox I: $\lambda(t|X) = \frac{\sqrt{t+0.001}}{6}\exp(X_1 + X_2 + X_3 + X_4/2 + X_5/3 + X_6/4)$;
Cox II: $\lambda(t|X) = \frac{\sqrt{t+0.001}}{14}\exp[(X_1 + X_2)^2 + \log\{\sin(\pi X_3) + 4X_4X_5 + \exp(X_6) + 2\}]$;
Cox III: $\lambda(t|X) = (t + 0.001)^{1/5}\exp[0.5(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)\log(t + 0.001)]$.

For the accelerated failure time (AFT) model, we considered

AFT I: $\log(T) = 0.5(X_1 - X_2 + X_3 - X_4 + X_5 - X_6) + \varepsilon$, with $\varepsilon \sim N(0, 1)$;
AFT II: $\log(T) = X_1 \sin(\pi(X_3 + X_4)) + \exp(X_2X_4 + X_5) - 2\log(X_5^2 + 2X_6 + 1.5) + \varepsilon$, with $\varepsilon \sim N(0, 1)$.

For the additive hazards (AH) model, we considered

AH I: $\lambda(t|X) = 0.1(t + 0.001) + 0.5(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)$;
AH II: $\lambda(t|X) = (t + 0.001) + \exp\{X_1 \sin(X_3) + X_2X_4\} + \log(X_5^2 + X_6 + 1)$.

Censoring time C was generated independently from an exponential distribution with mean μ and then truncated to the interval $[0, \tau]$, with $\tau = 2$. We define the observed time as $Y = \min(T, C)$ and the censoring indicator as $\Delta = I(T \leq C)$. For Scenarios 1–5, we set $\mu = 2$; for Scenario 6 and 7, we set $\mu = 1$ and 0.5, respectively. These parameter values were selected to achieve an approximate censoring rate of 50%.

We considered sample sizes of $n = 400$ and 800, allocating 80% of the data for training and the remaining 20% for tuning the hyperparameters. Additionally, we generated $N = 500$ samples as the test set. All performance results were averaged across 200 simulation replicates.

The prediction performance of the DNN estimator was evaluated using the prediction error (PE) of the conditional survival function, which measures the average squared difference between the predicted and true conditional survival functions,

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

$$\mathrm{PE} = \frac{\sum_{i=1}^{N} \Delta_i |S(Y_i|X_i) - \hat{S}(Y_i|X_i)|}{\sum_{i=1}^{N} \Delta_i}.$$

Smaller PE values reflect more accurate estimates.

Graf et al. (1999) proposed a modification to the Brier score to account for censored observations by incorporating inverse probability of censoring weights. The adjusted Brier score (BS) is defined as

$$\mathrm{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\hat{S}(t \mid X_i)^2 I\{Y_i \le t, \Delta_i = 1\}}{\hat{G}(Y_i)} + \frac{\left(1 - \hat{S}(t \mid X_i)\right)^2 I\{Y_i > t\}}{\hat{G}(t)} \right],$$

where $\hat{G}(t)$ denotes the Kaplan–Meier estimator of the survival function of censoring time, $\Pr(C > t)$. To evaluate the Brier score over a time interval rather than at a single time point, the integrated Brier score (IBS) is defined by

$$\mathrm{IBS} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \mathrm{BS}(s)\, ds,$$

where $t_1 = \min_{i=1,\ldots,N} Y_i$ and $t_2 = \max_{i=1,\ldots,N} Y_i$. In practical applications, this integral can be approximated using numerical methods, with the time range chosen to match the duration of the test data. Based on our experiments, we found that reliable results can be obtained by dividing the interval into 100 evenly spaced grid points.

To comprehensively evaluate our estimation method, we compared it with 11 alternative approaches, including both traditional survival analysis techniques and contemporary machine learning-based methods as follows.

CoxPH-Partial: Cox proportional hazards model with the partial likelihood approach (Cox, 1975).
CoxPH-Full: Cox proportional hazards model with the full likelihood approach; see Example 1 of the online supplementary material.
Cox-Varying: Cox model with time-varying coefficients (Tian et al., 2005).
Lognormal-AFT: Accelerated failure time model with normally distributed errors.
SAFT: Semiparametric accelerated failure time model with the distribution of the error ε unspecified (Ding & Nan, 2011).
AH: Additive hazards model (Lin & Ying, 1994).
Cox-CC: Deep neural network-based proportional Cox model (Kvamme et al., 2019).
Cox-Time: Deep neural network-based non-proportional Cox model (Kvamme et al., 2019).
DeepSurv: Cox proportional hazards deep neural network model (Katzman et al., 2018).
DeepHit: Deep learning-based method specifically designed for survival analysis (Kvamme et al., 2019; Lee et al., 2018).
RSF: random survival forest, a machine learning-based ensemble method for survival analysis that extends random forests to handle censored data (Ishwaran et al., 2008).

For CoxPH-Partial and Lognormal-AFT, the `CoxPHFitter` and `LogNormalAFTFitter` functions in the `lifelines` package were used, respectively (Davidson-Pilon, 2019). As the CoxPH-Full, Cox-Varying, SAFT, and AH methods all involve basis spline curves, the `BSpline` function from the `scipy.interpolate` module was utilized. The initial parameter values for all components were set to 0.1 and the degree of the B-splines was set to 3. We used B-splines with uniformly spaced internal knots and the number of internal knots was determined as $\lfloor n^{1/4} \rfloor$. RSF was implemented using the `RandomSurvivalForest` function from the `sksurv.ensemble` module. The implementation details of four machine learning methods, including Cox-CC, Cox-Time, DeepSurv, and DeepHit, are available at https://github.com/havakv/pycox. As shown in

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Table 1 and Table 3 of the online supplementary material, the prediction errors and the integrated Brier scores decrease in all scenarios as sample size increases. When the model assumptions are correct, the corresponding methods naturally achieve the optimal predictive performance. For example, under Cox I, both CoxPH-Partial and CoxPH-Full demonstrate good predictive performance, with the latter outperforming the former due to the fact that the full likelihood captures more information than the partial likelihood. Under AFT I, the Lognormal-AFT method achieves the best performance, while under AH I, the AH method performed the best.

In scenarios where the model assumptions are violated, such as in Cox II, Cox III, AFT II, and AH II, traditional methods showed poor performance. Among the machine learning-based methods, all methods (including Cox-CC, Cox-Time, DeepSurv, RSF, and our proposed method) except DeepHit perform well in these seven scenarios. In particular, our method consistently outperformed other machine learning-based methods in almost all scenarios. As illustrated in Table 4 of the online supplementary material, our method is slightly slower computationally compared to other machine learning-based methods, requiring approximately tens of seconds per simulation run. Nevertheless, it maintains competitive computational efficiency for practical implementation. We further considered cases with 12 and 20 covariates, and the results are also summarized in Tables 6–11 of the online supplementary material.

We present the biases, standard deviations, and coverage probabilities of the 95% confidence interval of the estimated conditional survival function $S(t|X)$ under Cox II, AFT II and AH II. We focused on the performance of the conditional survival function estimator at a finite set of representative points. Specifically, the time points (0.4, 0.8, 1.2, 1.6) were selected to divide the interval $[0, \tau]$ into five equally spaced segments. We considered three values of covariate vectors for each model, and the specific values are given in the online supplementary material. We chose sample size of 400, where 80% of the data were allocated to the training set and 20% to the validation set, and performed 500 replications. We applied the bootstrap method (with 100 resamples) to compute the 95% confidence intervals by taking the 0.025 and 0.975 percentiles of the bootstrap distribution. Table 5 of the online supplementary material summarizes the results, which demonstrate good performance of the proposed estimators.

Based on our simulation results, we recommend that in practical applications, the proposed method performs effectively with a sample size as small as $n = 400$, provided that the dimension of the covariate vector $p \leq 6$ and the censoring rate is approximately 50% or lower.

5.2 Evaluation of one-sample test

To evaluate the performance of the one-sample test proposed in Section 4.2.1 with finite samples, we considered three data generation scenarios:

Cox: $\lambda(t|X) = \frac{\sqrt{t+0.001}}{6} \exp(X_1 + X_2 + X_3 + X_4/2 + X_5/3 + X_6/4)$;
AFT: $\log(T) = (X_1 - X_2 + X_3 + X_4 + X_5 - X_6)/6 + \varepsilon$, where $\exp(\varepsilon) \sim \text{Weibull}(2, 1)$;
AH: $\lambda(t|X) = 0.1(t + 0.001) + 0.5(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)$.

The covariates $(X_1, \ldots, X_6)$ and the censoring time C were generated in the same way as that in Section 5.1. For the Cox and AFT scenarios, we set $\mu = 2$, and for the AH scenario, $\mu = 1$. These selections were designed to produce an approximate right-censoring rate of 50%, and the sample size was set as 800. To obtain optimal hyperparameters for the estimation of $g(t, x)$, we used 80% of the data in each simulation as the training samples and 20% as the validation samples. The test statistic was calculated using the training data of size $n = 640$.

For each setup, we calculated the proposed test statistic $T_w/\hat{\sigma}_w$ with weights $W_n^{(\rho,\gamma)}$ given in Section 4.2, which asymptotically follows the standard normal distribution under the null hypothesis H_0. Here, we considered Cox, AFT, and AH as true models. The QQ-plots, depicted in Figure 1, serves to confirm both the asymptotic normality of T_w in Theorem 4.2 and the functional normality of the proposed DNN estimator as stated in Theorem 4.1.

Table 12 of the online supplementary material presents the estimated size and power of the proposed one-sample test based on 500 simulation runs, along with the corresponding theoretical powers. The estimated test sizes are close to 0.05, and both the estimated powers and theoretical

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Table 1. Integrated brier scores of 12 methods based on 200 repetitions on the test set with 500 samples across seven different scenarios with sample sizes of $n = 400$ and 800 respectively

		Data Generation Scenarios						
n	Method	Cox I	Cox II	Cox III	AFT I	AFT II	AH I	AH II
400	Proposed	0.1595	0.1509	0.1429	0.1797	0.1715	0.1567	0.1321
	CoxPH-Partial	0.1576	0.1539	0.1437	0.1755	0.1769	0.1558	0.1352
	CoxPH-Full	0.1575	0.1545	0.1436	0.1755	0.1770	0.1556	0.1349
	Cox-Varying	0.1595	0.1520	0.1483	0.1792	0.1799	0.1590	0.1378
	Cox-CC	0.1602	0.1526	0.1433	0.1811	0.1741	0.1565	0.1334
	Cox-Time	0.1600	0.1523	0.1432	0.1805	0.1745	0.1562	0.1334
	DeepSurv	0.1605	0.1531	0.1435	0.1807	0.1717	0.1569	0.1341
	DeepHit	0.1823	0.1799	0.1515	0.2027	0.2087	0.1758	0.1606
	RSF	0.1620	0.1580	0.1440	0.1809	0.1768	0.1572	0.1357
	Lognormal-AFT	0.1580	0.1531	0.1443	0.1747	0.1762	0.1600	0.1417
	SAFT	0.1663	0.1611	0.1535	0.1761	0.1766	0.1562	0.1367
	AH	0.1613	0.1574	0.1572	0.1825	0.1901	0.1550	0.1351
800	Proposed	0.1573	0.1491	0.1416	0.1767	0.1646	0.1548	0.1315
	CoxPH-Partial	0.1563	0.1528	0.1422	0.1735	0.1752	0.1541	0.1334
	CoxPH-Full	0.1562	0.1532	0.1420	0.1736	0.1753	0.1540	0.1334
	Cox-Varying	0.1571	0.1493	0.1442	0.1756	0.1765	0.1557	0.1344
	Cox-CC	0.1587	0.1508	0.1425	0.1786	0.1672	0.1551	0.1324
	Cox-Time	0.1579	0.1504	0.1424	0.1777	0.1672	0.1549	0.1323
	DeepSurv	0.1580	0.1503	0.1424	0.1780	0.1660	0.1548	0.1322
	DeepHit	0.1793	0.1760	0.1475	0.2014	0.2054	0.1651	0.1442
	RSF	0.1591	0.1531	0.1434	0.1783	0.1707	0.1561	0.1350
	Lognormal-AFT	0.1566	0.1519	0.1428	0.1729	0.1748	0.1586	0.1402
	SAFT	0.1642	0.1597	0.1472	0.1735	0.1749	0.1543	0.1346
	AH	0.1595	0.1563	0.1547	0.1824	0.1913	0.1529	0.1330

powers are 1 for all scenarios considered. This result suggests that the asymptotic normality of the proposed test is valid and demonstrates strong power.

5.3 Evaluation of two-sample test

To assess the finite-sample performance of the two-sample test proposed in Section 4.2.2, we investigated two distinct scenarios: one where the conditional survival functions of the two groups do not cross, and the other where they cross (Zeng & Lin, 2007).

First, we present the model specifications and data generation scenarios for the non-crossing cases under the three models:

$$\textbf{Cox: } \lambda(t|X) = \frac{t + 0.001}{3} \exp\left(X_1 + X_2/2 + X_3/3 + X_4/4 + X_5/5 + X_6/6\right);$$
$$\textbf{AFT: } \log(T) = 0.1(X_1 + X_2 + X_3 + X_4 + X_5 + X_6) + \varepsilon, \text{ where } \exp(\varepsilon) \sim \text{Weibull}(2, 1);$$
$$\textbf{AH: } \lambda(t|X) = 0.05(t + 0.001) + 0.4(X_1 + X_2 + X_3 + X_4 + X_5 + X_6).$$

The covariates $(X_1, \ldots, X_6)$ and the censoring time C were generated in the same way as that in Section 5.1. For the Cox and AFT scenarios, we set $\mu = 2$, and for the AH scenario, $\mu = 1$. These selections yielded a right-censoring rate of approximately 50%. Under each setup, we generated two datasets: one based on $g_1(t, x) = \log(\lambda(t|x))$, while the other based on $g_2(t, x) = \log(\lambda(t|x)) + c$

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

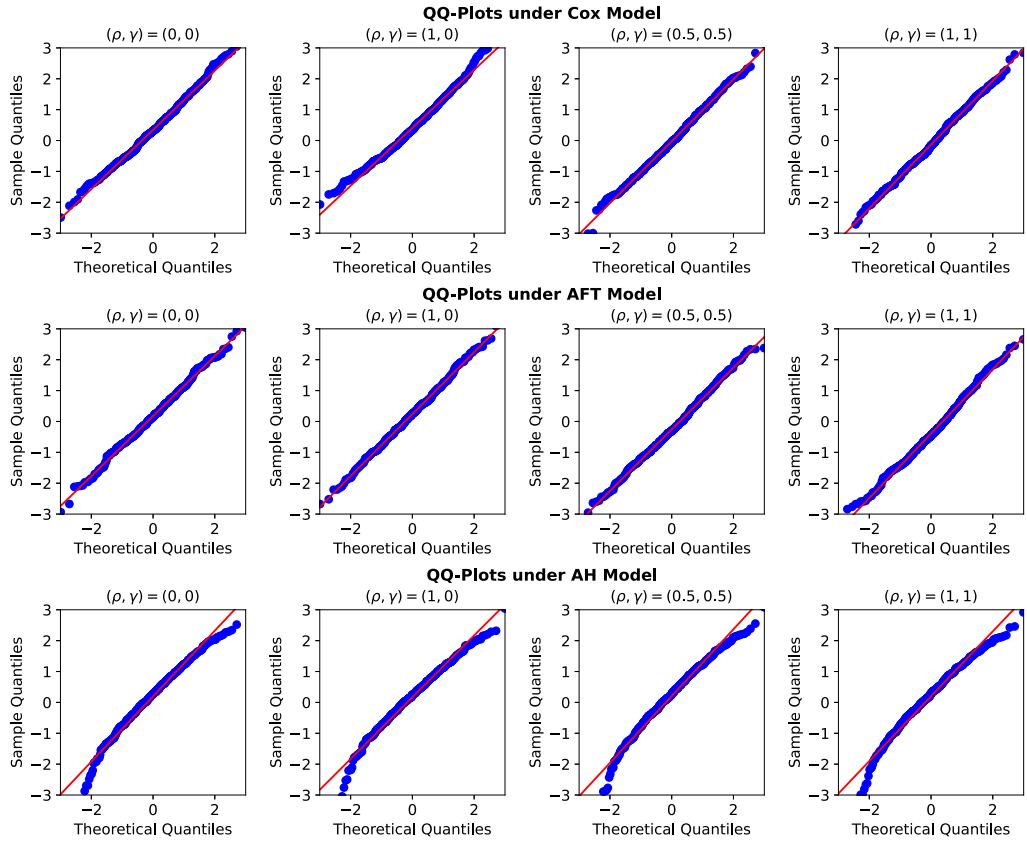

Figure 1. QQ-Plots under three models (Cox, AFT and AH) with four weight configurations.

for the three models, where the constant c represents the deviation from the null. We obtained the size of the test by setting $c = 0$ and the power of the test by setting $c = 0.125, 0.25, 0.5$, respectively.

Subsequently, we present a simulation study to investigate the performance of the proposed two-sample test when two conditional survival functions cross. We considered three models and generated two datasets from two cases of each model:

$$\begin{aligned}
\textbf{Cox}: \lambda_1(t|X) &= \frac{t+0.001}{3}\exp\left(X_1 + X_2/2 + X_3/3 + X_4/4 + X_5/5 + X_6/6\right),\\
\lambda_2(t|X) &= \frac{(t+0.001)^2}{3}\exp\left(X_1 + X_2/2 + X_3/3 + X_4/4 + X_5/5 + X_6/6\right);\\
\textbf{AFT}: \log(T) &= 0.1(X_1 - 2X_2 + X_3 - X_4 + X_5 + 2X_6) + \varepsilon_1,\ \exp(\varepsilon_1) \sim \text{Weibull}(2, 1),\\
\log(T) &= 0.1(X_1 - 2X_2 + X_3 - X_4 + X_5 + 2X_6) + \varepsilon_2,\ \exp(\varepsilon_2) \sim \text{Weibull}(1.5, 1);\\
\textbf{AH}: \lambda_1(t|X) &= 0.3 + 0.2(X_1 + X_2 + X_3 + X_4 + X_5 + X_6),\\
\lambda_2(t|X) &= 0.5(t+0.001) + 0.2(X_1 + X_2 + X_3 + X_4 + X_5 + X_6).
\end{aligned}$$

The covariates $(X_1, \ldots, X_6)$ and the censoring time C were generated in the same way as that in Section 5.1. For the Cox scenario, we set $\mu = 2$, and for the AFT and AH scenarios, $\mu = 1.5$. These choices produced a right-censoring rate of approximately 50%. For each model, we randomly generated two covariate vectors and plotted the corresponding conditional survival functions, which cross as shown in Figure 4 of the online supplementary material.

For each replication, we generated two independent datasets of sample size $n = 800$. To obtain the optimal hyperparameters for the estimation of $g_1(t, x)$ and $g_2(t, x)$, we allocated 80% of the

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

simulated data for training and 20% for validation in each simulation run. The test statistic was only computed using the training data, for which the sample sizes n_1 and n_2 were both set to 640.

According to the theoretical result in Theorem 4.3, we calculated $U_w/\hat{\tau}_w$ and the proportion of rejecting $H_0 : g_1 = g_2$ over 500 simulation runs. Table 2 summarizes the estimated sizes and powers of the two-sample test for the non-crossing cases, the estimated power for the crossing cases, and the theoretical powers across all scenarios. In all cases, our estimated powers were close to the theoretical ones, indicating that our two-sample test method is effective. In the non-crossing cases, when the data were generated according to H_0, the rejection rate was around 0.05; on the contrary, the rejection rate gradually increased when $g_1(t, x)$ and $g_2(t, x)$ differed further away and the powers corresponding to the four weights were very close. In the case of crossing conditional survival functions, the powers corresponding to the four different weights varied across the settings due to the effect of the crossing. Among these, the weight $W_n^{(1,0)}(t, x)$ consistently yielded the highest power, which is recommended for use in practice.

5.4 Evaluation of goodness-of-fit test

To evaluate the performance of the goodness-of-fit (GoF) test proposed in Sections 4.2.3 with finite samples, we consider data generated from two Cox models and four non-Cox models:

Cox I: $\lambda(T|X) = \frac{\sqrt{T+0.001}}{6} \exp(X_1 + X_2 + X_3 + X_4/2 + X_5/3 + X_6/4)$;
Cox II: $\lambda(T|X) = \frac{\sqrt{T+0.001}}{14} \exp[X_1 + 2X_2 \sin(\pi X_3) + \log\{4X_4X_5 + \exp(X_6) + 3\}]$;
AFT I: $\log(T) = 0.25(X_1 - X_2 + X_3 - X_4 + X_5 - X_6) + \varepsilon$, $\varepsilon \sim N(0, 1)$;
AFT II: $\log(T) = 2X_1 \sin(\pi X_3) + \exp(X_2X_4 + X_5) - 2\log(X_5 + X_6 + 2) + \varepsilon$, $\varepsilon \sim N(0, 1)$;
AH I: $\lambda(T|X) = 2(T + 0.001) + 0.25(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)$;
AH II: $\lambda(T|X) = 0.5(T + 0.001) + 0.5(X_1 + X_2) + \cos(\pi(X_3 + X_4)) + \sin(\pi(X_5 + X_6)/2)$.

Under each setup, we test 'H_0: The Cox model is true' against 'H_1: The Cox model is not true'. The covariates $(X_1, \ldots, X_6)$ and the censoring time C were generated in the same way as in Section 5.1. For the first four models, $\mu = 2$ was selected; for AH I and AH II, $\mu = 0.5$ was chosen. These selections yielded an approximate right-censoring rate of 50%, and the sample size was set as 1600. We considered four GoF tests:

GoF test I: The Grambsch–Therneau test that evaluates the proportional hazards assumption; see Grambsch and Therneau (1994).
GoF test II: The overall goodness-of-fit test for the Cox proportional hazards model; see May and Hosmer (1998).
GoF test III: The DNN-based test given by Theorem 4.6 and Example C.1 (online supplementary material).
GoF test IV: The DNN-based test given by Theorem 4.6 and Theorem 5.2, online supplementary material.

In each simulation run, we randomly partitioned the sample into two subsets of equal sample size ($n = 800$). For GoF test III, subset 1 was used to estimate $g(t, x)$, while subset 2 was used to fit the linear Cox model. For GoF test IV, subset 1 was used to estimate $\hat{g}_n(t, x)$, and subset 2 was used to fit the nonparametric Cox model. We adopted an 80%–20% training–validation split for DNN-based models.

The GoF tests III and IV involve estimation of the baseline hazard function. We used a B-spline function with uniformly spaced internal knots to estimate the baseline hazard function. The number of internal knots was chosen as $\lfloor n^{1/4} \rfloor$, initial parameter values of all components were set to 0.1, and the degree of the B-spline function was set to 3.

Table 3 reports the estimated size and power of each GoF test over 500 simulation runs. While GoF tests I and II could produce reasonable test sizes close to 5%, the power values of these methods were small when data were generated from other survival models, suggesting that traditional methods might not be ideal in identifying suitable models. GoF test III also yielded similar test size and power except for the case with data generated under Cox II. This indicates that GoF test III can

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Table 2. Estimated and theoretical sizes and powers of the two-sample test for non-crossing and crossing survival cases (theoretical values in parentheses)

		Non-crossing survival cases			
Data Model	c	$W_n^{(0,0)}$	$W_n^{(1,0)}$	$W_n^{(0.5,0.5)}$	$W_n^{(1,1)}$
Cox	0	0.046 (0.050)	0.040 (0.050)	0.044 (0.050)	0.052 (0.050)
	0.125	0.284 (0.278)	0.278 (0.257)	0.256 (0.268)	0.232 (0.242)
	0.25	0.848 (0.820)	0.798 (0.781)	0.818 (0.801)	0.754 (0.753)
	0.5	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
AFT	0	0.056 (0.050)	0.066 (0.050)	0.060 (0.050)	0.064 (0.050)
	0.125	0.290 (0.281)	0.270 (0.262)	0.258 (0.277)	0.234 (0.253)
	0.25	0.828 (0.825)	0.758 (0.791)	0.832 (0.817)	0.804 (0.778)
	0.5	1.000 (1.000)	0.998 (1.000)	1.000 (1.000)	1.000 (1.000)
AH	0	0.054 (0.050)	0.042 (0.050)	0.064 (0.050)	0.066 (0.050)
	0.125	0.296 (0.284)	0.286 (0.253)	0.290 (0.280)	0.264 (0.266)
	0.25	0.816 (0.815)	0.784 (0.765)	0.806 (0.801)	0.770 (0.789)
	0.5	1.000 (1.000)	0.998 (1.000)	0.998 (1.000)	0.998 (1.000)
		Crossing survival cases			
Data Model		$W_n^{(0,0)}$	$W_n^{(1,0)}$	$W_n^{(0.5,0.5)}$	$W_n^{(1,1)}$
Cox		0.990 (0.996)	0.998 (0.997)	0.958 (0.973)	0.890 (0.879)
AFT		0.614 (0.699)	0.802 (0.833)	0.176 (0.154)	0.042 (0.057)
AH		0.444 (0.478)	0.508 (0.568)	0.324 (0.291)	0.228 (0.176)

differentiate linear Cox models from other survival models, including nonlinear Cox models. The test sizes of our proposed GoF test IV were close to 5% when data were generated from Cox models I or II, while the tests possessed good powers when data were not generated from Cox models. This indicates that GoF test IV is able to differentiate Cox models from other survival models.

6 Application to medical imaging data

We applied the proposed method to a medical imaging dataset from The Cancer Genome Atlas (TCGA) project. It was established to create a collaborative research network dedicated to the comprehensive characterization of molecular events in primary cancers, with the mission of making this information publicly accessible to the global research community. Launched in 2006, the initiative began with a three-year pilot phase targeting glioblastoma (GBM), lung squamous cell carcinoma, and severe ovarian cystadenocarcinoma. Following this, TCGA expanded into full-scale operations from 2009 to 2015. Over the course of this decade-long project, the TCGA network successfully mapped the molecular atlases of tumors from 11,160 patients encompassing 33 cancer types, while also defining their distinct molecular subtypes (Liu et al., 2018). As documented in the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR), tumor specimens were acquired from 161 tissue source sites worldwide.

The TCGA-CDR analysis reports an overall median follow-up of 22.1 months, with significant variation observed across different cancer types. Specifically, the median follow-up times were recorded as 22.1 months for lower-grade glioma (LGG) and 12.0 months for GBM (Liu et al., 2018, Table 2). A comprehensive overview of the demographics, grades, survival, and molecular subtypes is detailed in Table S1 of Mobadersany et al. (2018).

For the analysis, we considered a database of magnified tissue-biopsy slides from 769 patients with brain tumors from TCGA-LGG and TCGA-GBM projects, along with their observed overall survival times (in days). The dataset was retrieved from the TCGA database and can be accessed

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Table 3. Test sizes and powers of the four goodness-of-fit (GoF) tests

	Tests			
Data Model	GoF I	GoF II	GoF III	GoF IV
Cox I	0.066	0.058	0.060	0.060
Cox II	0.436	0.222	0.480	0.064
AFT I	0.254	0.044	0.288	0.368
AFT II	0.654	0.626	0.674	0.872
AH I	0.294	0.084	0.342	0.330
AH II	0.098	0.432	0.868	0.828

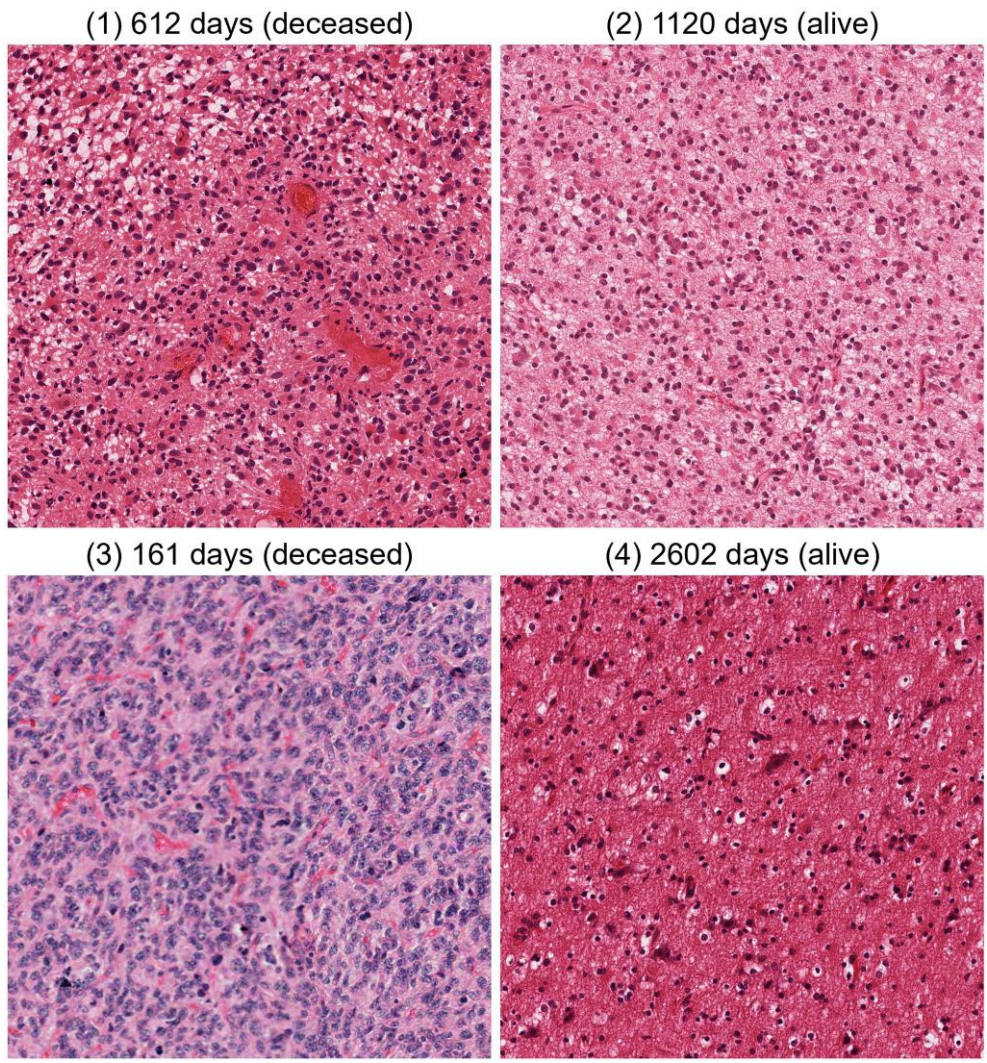

Figure 2. Examples of histological ROIs with survival times and vital status of four subjects: (1) 612 days (deceased); (2) 1,120 days (alive); (3) 1,161 days (deceased); (4) 2,602 days (alive).

via https://github.com/PathologyDataScience/SCNN. One patient had a survival time of 0 days and was therefore excluded from the analysis. In our analysis, we randomly selected one region of interest (ROI) for each patient ($n = 768$ ROIs). The survival time (T) of interest refers to the

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Table 4. Integrated brier score (IBS) and computation time (in seconds) for 13 methods applied to the TCGA dataset. Proposed (GQ) denotes the proposed method with Gaussian quadrature, while proposed (AGQ) refers to the proposed method with adaptive Gaussian quadrature

	Traditional survival methods					
	CoxPH-Partial	**CoxPH-Full**	**Cox-Varying**	**Lognormal-AFT**	**SAFT**	**AH**
IBS	0.1323	0.1332	0.1239	0.1248	0.1368	0.1354
Time	0.05	39.48	6.00	0.30	1.89	4.65
	Machine learning-based methods					
	Proposed (GQ)	**Proposed (AGQ)**	**Cox-CC**	**Cox-Time**	**DeepSurv**	**DeepHit**	**RSF**
IBS	0.1137	0.1151	0.1168	0.1195	0.1153	0.1232	0.1190
Time	8.23	37.50	0.34	0.77	0.25	0.39	0.86

time elapsed from the enrollment of a patient with glioma to death. In contrast, censoring time (C) measures the period from enrollment to the last follow-up of patients who are still alive. The censoring rate was 49.5%. The summary statistics of censoring times for alive patients and survival times for deceased patients (in days) are presented in Table 13 of the online supplementary material. Figure 2 presents four ROIs from different patients with survival times and vital status as follows: 612 days (deceased), 1024 days (alive), 1161 days (deceased), and 2602 days (alive).

We conducted image preprocessing and feature screening, as our model operates on pre-extracted features presented in a tabular format, rather than employing end-to-end processing of raw, unstructured data. This approach aligns with methodologies reported in previous studies (e.g. Mobadersany et al., 2018; Morrison et al., 2024). A commonly used pre-trained network is VGG16 Simonyan & Zisserman, 2014, which a convolutional neural network trained on extensively augmented natural images that improves accuracy by increasing depth through multiple small (3×3) convolutional layers. We used VGG16 with pre-trained weights to extract features from histological ROI images. We removed the last three fully connected layers and passed images through the convolutional stack with fixed pre-trained weights. Given a 1024×1024 RGB ROI image, VGG16 produced a $32 \times 32 \times 512$ feature tensor, which, after flattening, yielded 524,288 features per participant. To handle the high-dimensional input (the feature dimension far exceeds the sample size), we applied the quantile adaptive sure independence screening (QaSIS, He et al., 2013) at the 0.5 quantile to perform nonlinear variable screening, and retain the top 20 features to balance statistical stability and computational tractability in a modest sample with substantial censoring. In addition to the 20 variables extracted from the image features, gender and age at diagnosis were collected for 736 out of 768 patients. Below, we analyze the subset of 736 patients with available data—treating the above 22 variables as covariates—with a censoring rate of 47.6%.

First, we compared the proposed DNN with Gaussian quadrature (Proposed (GQ)) and with adaptive Gaussian quadrature (Proposed (AGQ)) against the 11 methods in Section 6.1 using the IBS metric. We randomly divided the data into training/validation/testing sets at ratios of 60/15/25. As shown in Table 4, machine learning methods generally yield lower IBS than traditional survival models. Within the machine learning methods, our proposed approach performs best, with Proposed (GQ) delivering the strongest predictive performance. The subsequent data analysis thus focuses on the Proposed (GQ) method.

We used the proposed goodness-of-fit test to assess the nonparametric Cox, traditional Cox, AFT, and AH models. All tests returned p-values <0.001, strongly suggesting that none of these models is suitable for this dataset. Furthermore, to test whether there are differences in the conditional hazard function between the male and female groups, we applied the two-sample test from Section 6.3 with four weighting schemes. The corresponding p-values were 0.39, 0.40, 0.37, and 0.38, indicating no significant difference in the conditional hazard function between the male and female groups.

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

To assess sensitivity, we further evaluated the performance of the proposed method using the top 10, 30, and 50 features, respectively. Age and sex are appended, yielding the dimension of covariates as 12, 32, and 52, as shown in Table 14 of the online supplementary material. The results demonstrate that our method is stable and consistently outperforms competing methods in terms of IBS across different feature dimensions.

We emphasize that VGG16 feature extraction and QaSIS screening are preprocessing steps external to the proposed likelihood-based hazard estimator. The proposed DNN approach is applied to the resulting covariate vector. In other words, this application demonstrates survival modeling using image-derived tabular features, rather than an end-to-end pixel-to-hazard model. End-to-end training would require jointly learning the image encoder and the hazard model and developing corresponding theory for the composed estimator, which warrants further development.

7 Conclusion

Motivated by recent developments in deep learning, we proposed a novel nonparametric inference approach for the conditional hazard function using DNNs. In particular, we considered a hazard function without a baseline and hence do not require any structural or functional assumptions on the conditional hazard or distribution function of survival time. Such a versatile setup renders enormous flexibility to the model and mitigates spurious results arising from model misspecification. We have rigorously investigated the asymptotic properties of the proposed estimator and established the nonasymptotic error bound, consistency, convergence rate and functional asymptotic normality. Subsequently, we have developed nonparametric tests to make inference on conditional hazard functions for evaluating goodness-of-fit and conducting treatment comparison. The proposed methods show good performances consistently across all simulated scenarios with various underlying models such as the Cox, AFT, and AH models. Specifically, in the presence of nonlinearity in the hazard functions, the proposed DNN approach yields superior performance to the traditional modeling approaches. To improve the power of the proposed one-sample and two-sample tests, motivated by Yang and Prentice (2010), Lin et al. (2020) and Roychoudhury et al. (2023), we may consider the robust maxCombo test based on the Harrington and Fleming (1982) weight.

Acknowledgments

The authors are very grateful to the Editor, Professor Xuming He, the Associate Editor and the three referees for their valuable comments and suggestions that greatly improved the paper.

Conflicts of interest: None declared.

Funding

This research was supported in part by the National Natural Science Foundation of China (12401366, 12501387, 12271459), the Research Grant Council of Hong Kong (15307824, 15306521), the research grants from City University of Hong Kong (7200765, 7020159), and the research grants from The Hong Kong Polytechnic University (P0045385, CMFAI-P0046811, JLab-P0052683).

Data availability

The Cancer Genome Atlas (TCGA) data are publicly available at https://github.com/PathologyDataScience/SCNN.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

References

Aalen O. O. (1980) *A model for nonparametric regression analysis of counting processes*. In *Lecture notes in statistics* (Vol. 2, pp. 1–25). Springer-Verlag.

Andersen P. K., & Gill R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, *10*, 1100–1120. https://doi.org/10.1214/aos/1176345976

Buckley J., & James I. (1979). Linear regression with censored data. *Biometrika*, *66*, 429–436. https://doi.org/10.1093/biomet/66.3.429

Cai J., Fan J., Jiang J., & Zhou H. (2008). Partially linear hazard regression with varying coefficients for multivariate survival data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*, 141–158. https://doi.org/10.1111/j.1467-9868.2007.00630.x

Chakraborty S., Tomsett R., Raghavendra R., Harborne D., Alzantot M., Cerutti F., Srivastava M., Preece A., Julier S., Rao R. M., Kelley T. D., Braines D., Sensoy M., Willis C. J., & Gurram P. (2017). *Interpretability of deep learning models: A survey of results*. In *2017 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 1–6). San Francisco, CA. https://doi.org/10.1109/UIC-ATC.2017.8397411

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

Cox D. R. (1975). Partial likelihood. *Biometrika*, *62*, 269–276. https://doi.org/10.1093/biomet/62.2.269

Cox D. R., & Oakes D. (1984). *Analysis of survival data*, Monographs on statistics and applied probability. Chapman and Hall.

Craig E., Zhong C., & Tibshirani R. (2021). 'Survival stacking: Casting survival analysis as a classification problem', arXiv, arXiv:2107.13480, preprint: not peer reviewed.

Davidson-Pilon C. (2019). Lifelines: Survival analysis in python. *Journal of Open Source Software*, *4*, 1317. https://doi.org/10.21105/joss

Ding Y., & Nan B. (2011). A sieve M-theorem for bundled parameters in semi- parametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics*, *39*, 3032–3061. https://doi.org/10.1214/11-AOS934

Fleming T. R., & Harrington D. P. (1991). *Counting processes and survival analysis*. John Wiley & Sons, Inc.

Fox J., & Weisberg S. (2019). *An R companion to applied regression* (3rd ed.). Sage.

Goodfellow I., Bengio Y., & Courville A. (2016). *Deep learning*. MIT Press.

Graf E., Schmoor C., Sauerbrei W., & Schumacher M. (1999). Assessment and com- parison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*, 2529–2545. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18¡2529::aid-sim274¿3.0.co;2-5

Grambsch P., & Therneau T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*, 515–526. https://doi.org/10.1093/biomet/81.3.515

Harrington D. P., & Fleming T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, *69*, 553–566. https://doi.org/10.1093/biomet/69.3.553

He X., Wang L., & Hong H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, *41*, 342–369. https://doi.org/10.1214/13-AOS1087

Hosmer Jr. D. W., Lemeshow S., & May S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (2nd ed.). Wiley.

Ishwaran H., Kogalur U. B., Blackstone E. H., & Lauer M. S. (2008). Random survival forests. *Annals of Applied Statistics*, *2*, 841–860. https://doi.org/10.1214/08-AOAS169

Jiao Y., Shen G., Lin Y., & Huang J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *Annals of Statistics*, *51*, 691–716. https://doi.org/10.1214/23-AOS2266

Katzman J., Shaham U., Bates J., Cloninger A., Jiang T., & Kluger Y. (2018). DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*, 24. https://doi.org/10.1186/s12874-018-0482-1

Kingma D. P., & Ba J. L. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. http://arxiv.org/abs/1412.6980

Klein J. P., & Moeschberger M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). Springer-Verlag.

Knaus W. A., Harrell Jr F. E., Lynn J., Goldman L., Phillips R. S., Connors Jr A. F., Dawson N. V., Fulkerson Jr W. J., Califf R. M., Desbiens N., Layde P., Oye R. K., Bellamy P. E., Hakim R. B., & Wagner D. P. (1995). The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, *122*, 191–203. https://doi.org/10.7326/0003-4819-122-3-199502010-00007

Krizhevsky A., Sutskever I., & Hinton G. E. (2012). *Imagenet classification with deep convolutional neural networks*. In *Advances in neural information processing systems* (pp. 1097–1105).

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Kvamme H., Borgan Ø., & Scheel I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, *20*, 1–30. https://doi.org/10.48550/arXiv.1907.00825
Lee C., Zame W., Yoon J., & van der Schaar M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32(1)).
Lin D. Y., & Ying Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, *81*, 61–71. https://doi.org/10.1093/biomet/81.1.61
Lin R. S., Lin J., Roychoudhury S., Anderson K., Hu T., & Huang B. (2020). Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. *Statistics in Biopharmaceutical Research*, *12*, 187–198. https://doi.org/10.1080/19466315.2019.1697738
Liu J., Lichtenberg T., Hoadley K. A., Poisson L. M., Lazar A. J., Cherniack A. D., Kovatich A. J., Benz C. C., Levine D. A., Lee A. V., Omberg L., Wolf D. M., Shriver C. D., Thorsson V., & Hu H. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, *173*, 400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052
May S., & Hosmer D. W. (1998). A simplified method of calculating an overall goodness-of-fit test for the cox proportional hazards model. *Lifetime Data Analysis*, *4*, 109–120. https://doi.org/10.1023/a:1009612305785
Mckeague I. W., & Sasieni P. D. (1994). A partly parametric additive risk model. *Biometrika*, *81*, 501–514. https://doi.org/10.1093/biomet/81.3.501
Mobadersany P., Yousefi S., Amgad M., Gutman D. A., Barnholtz-Sloan J. S., Velázquez Vega J. E., Brate D. J., & Cooper L. A. D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, *115*, E2970–E2979. https://doi.org/10.1073/pnas.1717139115
Morrison S., Gatsonis C., Eloyan A., & Steingrimsson J. A. (2024). Survival analysis using deep learning with medical imaging. *The International Journal of Biostatistics*, *20*, 1–12. https://doi.org/10.1515/ijb-2022-0113
Murdoch W. J., Singh C., Kumbier K., Abbasi-Asl R., & Yu B. (2019). Interpretable machine learning: Definitions, methods, and applications. The Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1900654116
Nair V., & Hinton G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning* (pp. 807–814).
Raskutti G., Wainwright M. J., & Yu B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, *15*, 335–366.
Roychoudhury S., Anderson K., Ye J., & Mukhopadhyay P. (2023). Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group. *Statistics in Biopharmaceutical Research*, *15*, 280–294. https://doi.org/10.1080/19466315.2021.1874507
Rudin C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg C. A., & Li F. F. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y
Schmidt-Hieber J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, *48*, 1875–1897. https://doi.org/10.1214/19-AOS1931
Silver D., Huang A., Maddison C. J., Guez A., Sifre L., van den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., Dieleman S., Grewe D., Nham J., Kalchbrenner N., Sutskever I., Lillicrap T., Leach M., Kavukcuoglu K., Graepel T., & Hassabis D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*, 484–489. https://doi.org/10.1038/nature16961
Simonyan K., & Zisserman A. (2014). 'Very deep convolutional networks for large-scale image recognition', arXiv, arXiv:1409.1556, preprint: not peer reviewed.
Sun T., Wei Y., Chen W., & Ding Y. (2020). Genome-wide association study-based deep learning for survival prediction. *Statistics in Medicine*, *39*, 4605–4620. https://doi.org/10.1002/sim.8743
Tarkhan A., & Simon N. (2023). An online framework for survival analysis: Reframing cox proportional hazards model for large data sets and neural networks. *Biostatistics*, *25*, 134–153. https://doi.org/10.1093/biostatistics/kxac039
Tian L., Zucker D., & Wei L. J. (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, *100*, 172–183. https://doi.org/10.1198/016214504000000845
Tsiatis A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*, *18*, 354–372. https://doi.org/10.1214/aos/1176347504
van Otterlo M., & Wiering M. (2012). *Reinforcement learning and Markov decision processes*. In *Reinforcement learning*. Springer.
Wei L. J., Ying Z., & Lin D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, *77*, 845–851. https://doi.org/10.1093/biomet/77.4.845

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026

Wolock C. J., Gilbert P. B., Simon N., & Carone M. (2024). A framework for leveraging machine learning tools to estimate personalized survival curves. *Journal of Computational and Graphical Statistics*, *33*, 1098–1108. https://doi.org/10.1080/10618600.2024.2304070

Wu Y., Schuster M., Chen Z., Le Q. V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., & Macherey K. (2016). 'Google's neural machine translation system: Bridging the gap between human and machine translation', arXiv, arXiv:1609.08144, preprint: not peer reviewed.

Yan J., & Huang J. (2012). Model selection for cox models with time-varying coefficients. *Biometrics*, *68*, 419–428. https://doi.org/10.1111/j.1541-0420.2011.01692.x

Yang S., & Prentice R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, *66*, 30–38. https://doi.org/10.1111/j.1541-0420.2009.01243.x

Zeng D., & Lin D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society B*, *69*, 507–564. https://doi.org/10.1111/j.1369-7412.2007.00606.x

Zhong Q., Mueller J. W., & Wang J.-L. (2021). *Deep extended hazard models for survival analysis*. In *Advances in neural information processing systems 34 pre-proceedings*.

Zhong Q., Mueller J. W., & Wang J.-L. (2022). Deep learning for the partially linear cox model. *Annals of Statistics*, *50*, 1348–1375. https://doi.org/10.1214/21-AOS2153

Downloaded from academic.oup.com/jrsssb/advance-article/doi/10.1093/jrsssb/qkag060/8572589 by guest on 15 October 2026