Statistica Sinica Preprint No: SS-2021-0050	
Title	Unbiased Boosting Estimation for Censored Survival Data
Manuscript ID	SS-2021-0050
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0050
Complete List of Authors	Li-Pang Chen and Grace Yi
Corresponding Authors	Li-Pang Chen
E-mails	lchen723@nccu.edu.tw
Notice: Accepted version subject to English editing.	

Statistica Sinica

Unbiased Boosting Estimation for Censored Survival Data

Li-Pang Chen[1,3] and Grace Y. Yi[1,2]

[1] *Department of Statistical and Actuarial Sciences, University of Western Ontario*

[2] *Department of Computer Science, University of Western Ontario*

[3] *Department of Statistics, National Chengchi University*

Abstract: Boosting methods have been broadly discussed for various settings, and many methods have been developed to handle data with complete observations. While some boosting methods have been available to handle survival data with censored responses, those methods were mainly developed under an assumed model for the survival process, and most of them focused on numerical implementation procedures without rigorous theoretical justifications. In this paper, we develop an unbiased boosting estimation method to handle censored survival data without assuming an explicit model, and explore three strategies to adjust loss functions with censoring effects accommodated. We describe a functional gradient descent algorithm to implement the proposed method. Further, we rigorously establish theoretical results, including consistency and optimization convergence. Numerical studies demonstrate satisfactory performance of the proposed method

Grace Yi is the corresponding author. Email: gyi5@uwo.ca

for finite sample settings.

Key words and phrases: Adjusted loss functions; boosting; consistency; empirical processes; machine learning; right-censoring; survival data.

1. Introduction

Boosting is a popular technique of deriving a strong learner from weak yet simple learners by iteratively updating learning results. Since the publications of Schapire (1990) and Freund (1995), increasing interest in boosting has been stimulated and various boosting algorithms have been developed for different settings. A summary of early developments of boosting methods for regression and classification problems was given by Ridgeway (1999). More details can be found in Bühlmann and Hothorn (2007), Hastie et al. (2008, Chapter 10), and Schapire and Freund (2014).

While considerable attention has been directed to settings with complete responses as considered by many authors (e.g., Bühlmann and Yu 2003; Lugosi and Vayatis 2004; Zhang and Yu 2005), in recent years, boosting algorithms have also been utilized to analyze survival data which typically involve censored responses. Those methods have been mainly developed under parametric or semiparametric survival models. For example, under the Cox proportional hazards model, Li and Luan (2005) de-

scribed a gradient boosting procedure with cubic smoothing splines; Chen et al. (2013) derived a boosting algorithm with the concordance index; He et al. (2016) considered a component-wise gradient boosting procedure; Bühlmann and Hothorn (2007) and Binder and Schumacher (2008) respectively developed the R packages `mboost` and `CoxBoost`. Focusing on the accelerated failure time model, Schmid and Hothorn (2008) examined a boosted estimating procedure. Considering nonlinear transformation models, Lu and Li (2008) proposed a gradient boosting method by employing the negative log marginal likelihood function as the loss function. Other boosting procedures concerning survival data include Benner (2002), Mayr et al. (2016), Lee et al. (2021), and Bellot and van der Schaar (2018a,b).

While boosting methods under assumed survival models can be useful, they are vulnerable to model misspecification and their application scope is model-dependent, and thus, restrictive. More notably, most existing boosting methods for survival data simply focus on implementation procedures with their feasibility assessed through numerical studies. To the best of our knowledge, except Lee et al. (2021), little attention has been directed to establishing theoretical results for boosting methods on survival data.

Motivated by those research gaps, in this paper we develop a boosting method under the general setup which does not impose specific models

for survival data. Our research not only supplies a model-free boosting method for censored survival data but also establishes theoretical results. In particular, our contributions are threefold. First, under the same framework, we examine three strategies to adjust for usual loss functions to address the effects due to right-censored responses, yielding three classes of adjusted loss functions, respectively, called *Buckley-James-type* (BJ) loss functions, *inverse censoring probability weighted* (ICPW) loss functions, and *augmented inverse censoring probability weighted* (AICPW) loss functions. These strategies enable us to handle survival data with flexibility to incorporate different features. The BJ adjustment applies to settings with reasonable information about the survival process, whereas the ICPW scheme works for cases with adequate knowledge of the censoring process. On the contrary, the AICPW method gives us room of misspecifying either the survival or censoring process, thus, enjoying the *double robustness* property. Secondly, based on the functional gradient descent algorithm, we devise a two-stage minimization procedure to derive the prediction function for survival times. Finally and importantly, we rigorously establish theoretical results, including consistency and convergence.

Further, we comment that our development has some commonality with the following work. Hothorn et al. (2006) merely considered the ICPW

scheme to adjust for censoring effects without providing theoretical explorations, yet our development offers additional adjustment methods and provides theoretical justifications. Wang and Wang (2010) used the Buckley-James formulation (Buckley and James 1979) to create a pseudo-response to account for censoring effects where the L_2-norm loss function was employed and no theoretical justification was provided; yet our BJ scheme adjusts for any loss functions but not responses, with rigorous justifications presented. In comparison to Lee et al. (2021), while their work established theoretical results, their goal differs from ours. They aimed at estimating the hazard function nonparametrically by minimizing a scaled negative log likelihood function, whereas we focus on finding a prediction model for survival times by minimizing the risk function that may assume various forms.

The remainder is organized as follows. In Section 2, we introduce the notation and framework. In Section 3, we consider censored survival data and examine three schemes for accommodating censoring effects in loss functions. In Section 4, we devise the implementation procedure for the proposed unbiased boosting estimation method for censored survival data. In Section 5, we establish theoretical results to rigorously justify the validity of the proposed method. Numerical results, including real data analysis and simulation studies, are presented in Section 6 and Section S5 of the Sup-

plementary Material, respectively. Concluding discussions are included in Section 7, and technical details are reported in the Supplementary Material.

2. Notation and Framework

2.1 Survival Data and Objective

Let $T \geq 0$ represent the survival time of an individual and let X denote the vector of associated p-dimensional covariates. To remove the positivity constraint on T, we consider a transformed outcome. In particular, let $\widetilde{T} = \log T$. We are interested in finding a function of X so that its value can well predict $\widetilde{T}$. To this end, we consider a class of useful functions. Specifically, let $\mathcal{F}$ denote the convex set of real-valued functions from $\mathbb{R}^p$ to $\mathbb{R}$ satisfying Condition (C5) in Section S1.1 of the Supplementary Material. For $f \in \mathcal{F}$, let $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ denote the *loss function* of using $f(X)$ to predict $\widetilde{T}$ which is a convex and differentiable function in the second argument as stated in Condition (C3) in Section S1.1 of the Supplementary Material.

Define the *risk function* as

$$R(f) = E\left\{L\left(\widetilde{T}, f(X)\right)\right\}, \tag{2.1}$$

where and hereafter, E represents the expectation with respect to the joint

distribution of random variables appearing in the loss function. By the convexity and differentiability of $L(\cdot,\cdot)$, the risk function (2.1) is convex with respect to $f \in \mathcal{F}$ as discussed by Zhang and Yu (2005) as well as differentiable, provided the interchange ability of the expectation and differentiation.

To find a function in $\mathcal{F}$ to well predict $\widetilde{T}$, we wish to find the element in $\mathcal{F}$ with

$$f_0 = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, R(f),$$

assuming the existence and uniqueness of $\min_{f\in\mathcal{F}} R(f)$, or equivalently,

$$R(f_0) = \min_{f\in\mathcal{F}} R(f). \tag{2.2}$$

Since the joint distribution of $\widetilde{T}$ and X is unknown, we invoke the sample information and use the empirical average to replace the expectation in (2.1). That is, we aim to estimate f_0 by finding $\widehat{f}_{\text{comp}} \in \mathcal{F}$ such that

$$\widehat{f}_{\text{comp}} = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L(\widetilde{T}_i, f(X_i)) \right\}, \tag{2.3}$$

where we assume the availability of a random sample of independent observations $\mathcal{O}_{\text{comp}} \triangleq \left\{ \{\widetilde{T}_i, X_i\} : i = 1, \cdots, n \right\}$ with n being the sample size and $\{\widetilde{T}_i, X_i\}$ being an independent copy of $\{\widetilde{T}, X\}$.

2.2 Usual Steepest Descent Algorithm

To find the solution of (2.3), one may consider a boosting method such as the steepest descent method (e.g., Hastie et al. 2008, Section 10.10). This method basically employs the gradient of the loss function to enhance iterative estimates of the function $f(\cdot)$ by iteratively using a varying learning rate which can be treated as a weak learner. In doing so, we take the method of Hastie et al. (2008, Section 10.10) by "parameterizing" the function $f(X)$ as $\{f(X_1), \cdots, f(X_n)\}$ for the n observations of the covariates X, and then define the partial derivative of the loss function $L(\widetilde{T}_i, f(X_i))$ as

$$\partial L(\widetilde{T}_i, f(X_i)) \triangleq \frac{\partial L(u,v)}{\partial v}\Big|_{u=\widetilde{T}_i, v=f(X_i)}, \tag{2.4}$$

where $\frac{\partial L(u,v)}{\partial v}$ represents the partial derivative of the loss function $L(u,v)$ with respect to the second argument while keeping the first argument fixed.

With an estimate of $f(\cdot)$ at iteration m, denoted $f^{(m)}(\cdot)$, we employ the steepest descent method to enhance estimation of $f(\cdot)$ at iteration $(m+1)$ by adding an increment term, $-\widehat{\alpha}_{m+1}\partial L(\widetilde{T}_i, f^{(m)}(X_i))$, to $f^{(m)}(\cdot)$, where $\widehat{\alpha}_{m+1}$ is a scalar learning rate determined by the information obtained at the previous iteration:

$$\widehat{\alpha}_{m+1} = \underset{\alpha_{m+1}\in\mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L(\widetilde{T}_i, f^{(m)}(X_i) - \alpha_{m+1}\partial L(\widetilde{T}_i, f^{(m)}(X_i))) \right\}. \tag{2.5}$$

That is, the estimate of $f(\cdot)$ at iteration $(m+1)$ is given by

$$f^{(m+1)}(X_i) = f^{(m)}(X_i) - \widehat{\alpha}_{m+1}\partial L(\widetilde{T}_i, f^{(m)}(X_i)), \tag{2.6}$$

and the final boosting estimator is taken as $\widehat{f}_{\text{comp}}(X_i) \triangleq f^{(m+1)}(X_i)$ if the iteration stops at $(m+1)$.

2.3 Loss Functions

For continuous responses, three loss functions, listed in the middle column of Table 1, are commonly used in applications (Friedman 2001, p.1197), where for the Huber loss function, η can be taken as the αth-quantile of the $\left|\widetilde{T}_i - f(X_i)\right|$ for constant α with $0 < \alpha < 100$ and $i = 1, \cdots, n$ (Hastie et al. 2008, p.349, p.360).

Table 1: Commonly used loss functions

Name	Loss function: $L(\widetilde{T}_i, f(X_i))$	Derivative: $\partial L(\widetilde{T}_i, f(X_i))$
L_2-norm	$\left\{\widetilde{T}_i - f(X_i)\right\}^2$	$-2\left\{\widetilde{T}_i - f(X_i)\right\}$
L_1-norm	$\left\|\widetilde{T}_i - f(X_i)\right\|$	$\text{sign}\left\{\widetilde{T}_i - f(X_i)\right\}$ if $\left\|\widetilde{T}_i - f(X_i)\right\| \neq 0$
Huber	$\begin{cases} \frac{1}{2}\left\{\widetilde{T}_i - f(X_i)\right\}^2, & \text{if } \left\|\widetilde{T}_i - f(X_i)\right\| \leq \eta, \\ \eta\left(\left\|\widetilde{T}_i - f(X_i)\right\| - \frac{\eta}{2}\right), & \text{otherwise} \end{cases}$	$\begin{cases} -\left\{\widetilde{T}_i - f(X_i)\right\}, & \text{if } \left\|\widetilde{T}_i - f(X_i)\right\| < \eta, \\ -\eta\ \text{sign}\left\{\widetilde{T}_i - f(X_i)\right\}, & \text{if } \left\|\widetilde{T}_i - f(X_i)\right\| > \eta \end{cases}$

Differentiation of the loss function is crucial in implementing (2.4).

While this is not an issue when using the L_2-norm, it is a concern when the L_1-norm or Huber loss function is used since they are not differentiable over the entire domain. In such an instance, one may modify $\partial L(\widetilde{T}_i, f(X_i))$ with the *subdifferential*. For ease of exposition, here we take $\partial L(\widetilde{T}_i, f(X_i))$ as defined in (2.4). For the loss functions listed in Table 1, we present the values of $\partial L(\widetilde{T}_i, f(X_i))$ in the last column of Table 1, where the constraint $|\widetilde{T}_i - f(X_i)| \neq a$ is included for the L_1-norm loss with $a = 0$ and the Huber loss with $a = \eta$. This requirement is not as restrictive as it appears, and in fact, it holds in practical settings because that "$|\widetilde{T}_i - f(X_i)| = a$" occurs in zero probability.

3. Adjusting Loss Functions with Censoring Effects Accommodated

The development in Section 2 relies on the availability of complete observations, say $\mathcal{O}_{\text{comp}}$, of a random sample. This condition is, however, often not true for survival data due to the presence of censoring. For $i = 1, \cdots, n$, let C_i denote the censoring time for T_i, let $\Delta_i = \mathbb{I}(T_i \leq C_i)$ denote the censoring indicator, and write $Y_i = \min\{T_i, C_i\}$ and $\widetilde{Y}_i = \log Y_i$, where $\mathbb{I}(\cdot)$ is the indicator function. Let $[0, \tau]$ denote the study period with τ being finite. Consistent with the most discussions about survival analysis, we assume

that T_i and C_i are conditionally independent, given X_i. Following Rubin and van der Laan (2007), Zhu and Kosorok (2012), and Steingrimsson et al. (2016), we further assume that C_i and X_i are independent. These assumptions are listed as Condition (C6) in Section S1.1 of the Supplementary Material.

In the presence of censoring, the survival time $\widetilde{T}_i$ is not available for every study subject, thus the estimation procedure in Section 2.2 cannot be used directly. Consequently, we consider new loss functions expressed in terms of the observed censored data $\mathcal{O}_{\text{cd}} \triangleq \{\{\widetilde{Y}_i, X_i, \Delta_i\} : i = 1, \cdots, n\}$, with the censoring effects accounted for. The basic idea is to ensure that the expectation of a new loss function, denoted $L^*(\widetilde{Y}_i, f(X_i))$, recovers the expectation of the original loss function $L(\widetilde{T}_i, f(X_i))$ expressed in terms of $\widetilde{T}_i$ and X_i; namely, $E\left\{L^*(\widetilde{Y}_i, f(X_i))\right\} = E\left\{L(\widetilde{T}_i, f(X_i))\right\}$. Thus, the minimizer of the expectation of the workable new loss function $E\left\{L^*(\widetilde{Y}_i, f(X_i))\right\}$ also minimizes the risk function $R(f)$ in (2.1) as if $\widetilde{T}_i$ were always observed.

3.1 Adjustment Strategies

In this subsection we describe three strategies for constructing an adjusted loss functions. Let $F_{T0}(y|X_i) = P(T_i > y|X_i)$ represent the true condi-

tional survivor function of T_i, given X_i, and let $F_T(y|X_i)$ denote a working function used to model $F_{T0}(y|X_i)$ with $f_T(t|X_i)$ denoting the corresponding conditional density of T_i. Let $G_0(c) = P(C_i > c)$ stand for the true survivor function of C_i and let $G(c)$ denote its working function.

The first adjusted loss function is motivated by the *Buckley-James* (BJ) formulation (Buckley and James 1979):

$$L_{\mathrm{BJ}}(\widetilde{Y}_i, f(X_i)) = \Delta_i L(\widetilde{Y}_i, f(X_i)) + (1 - \Delta_i)\Psi(Y_i, X_i), \tag{3.7}$$

where $\Psi(y, X_i) = E\left\{L(\widetilde{T}_i, f(X_i))|T_i > y, X_i\right\}$, determined by

$$\Psi(y, X_i) = \int_y^\infty \frac{L(t, f(X_i))f_T(t|X_i)}{F_T(y|X_i)}dt. \tag{3.8}$$

The conditional expectation $\Psi(Y_i, X_i)$ in (3.7) facilitates the contribution from the censored subjects, yet the paid price is the requirement of having the working model $F_T(y|X_i)$ for the survivor process.

On the other hand, one may be interested in using the information from uncensored subjects only, in the hope of not needing $F_T(y|X_i)$. Hothorn et al. (2006) considered an adjusted loss function using the *inverse censoring probability weight* (ICPW) scheme:

$$L_{\mathrm{ICPW}}(\widetilde{Y}_i, f(X_i)) = \frac{\Delta_i L(\widetilde{Y}_i, f(X_i))}{G(Y_i)}. \tag{3.9}$$

While (3.9) does free us from involving $F_T(y|X_i)$ in the formulation like in (3.7), it calls for the working model $G(c)$ for the censoring process. With the

different involvements of $F_T(y|X_i)$ and $G(c)$ in (3.7) and (3.9), one might be piqued with the question: what happens if using both $F_T(y|X_i)$ and $G(c)$ to adjust the original loss function $L(\widetilde{T}_i, f(X_i))$? Motivated by the idea in Rubin and van der Laan (2007), we further consider the *augmented inverse censoring probability weighted* (AICPW) loss function:

$$L_{\text{AICPW}}(\widetilde{Y}_i, f(X_i)) \;=\; L_{\text{ICPW}}(\widetilde{Y}_i, f(X_i)) + \Gamma(Y_i, X_i, \Delta_i) \qquad (3.10)$$

with

$$\Gamma(Y_i, X_i, \Delta_i) = \frac{(1-\Delta_i)}{G(Y_i)}\Psi(Y_i, X_i) - \int_0^{Y_i} \frac{\Psi(t, X_i)}{G^2(t)} dG(t).$$

While (3.10) might look more restrictive than (3.7) and (3.9) due to its involvement of both $F_T(y|X_i)$ and $G(c)$ but not just one, it has a hidden advantage to be shown in Section 3.2. It is clear that (3.9) merely makes use of the information from those subjects in the sample who are not censored, and thus, ignoring the partial information contributed from those subjects who are censored. Adding the term $\Gamma(Y_i, X_i, \Delta_i)$ to (3.10) enables us to further use the measurements contributed from the subjects who are censored, and hence, possibly enhances the efficiency. As $F_T(y|X_i)$ and $G(c)$ appear in the denominators in the preceding formulations, they are assumed to be greater than zero (almost surely), as stated in Condition (C7) in Section S1.1 of the Supplementary Material. It is noted that an

empirical version similar to (3.10) was employed by Steingrimsson et al. (2016) to construct survival trees.

3.2 Properties of the Proposed Loss Functions

The three adjusted loss functions in Section 3.1 are formulated from different perspectives to accommodate censoring effects. Their validity is justified in the following two propositions.

Proposition 1. *The proposed adjusted loss functions (3.7) and (3.9) have the same expectation as* $L(\widetilde{T}_i, f(X_i))$. *That is,*

(a) $E\left\{L_{ICPW}(\widetilde{Y}_i, f(X_i))\right\} = E\left\{L(\widetilde{T}_i, f(X_i))\right\}$;

(b) $E\left\{L_{BJ}(\widetilde{Y}_i, f(X_i))\right\} = E\left\{L(\widetilde{T}_i, f(X_i))\right\}$,

where the expectations are evaluated with respect to the joint distribution of the associated random variables under the working models.

The proof of this proposition is placed in Section S4.1 of the Supplementary Material. Proposition 1 says that the expectation of the two adjusted loss functions, $L_{\mathrm{ICPW}}(\cdot,\cdot)$ and $L_{\mathrm{BJ}}(\cdot,\cdot)$, recovers the risk function (2.1) based on the failure time $\widetilde{T}_i$. With $L(\cdot,\cdot)$ taken as the L_2-norm loss function, Bühlmann and Hothorn (2007) established the identity in Proposition 1 (a).

The formulation of $L_{\text{AICPW}}(\cdot,\cdot)$ involves the distribution for both survival and censoring processes, which, at first sight, appears more restrictive than either $L_{\text{BJ}}(\cdot,\cdot)$ or $L_{\text{ICPW}}(\cdot,\cdot)$. However, the following proposition ensures that $L_{\text{AICPW}}(\cdot,\cdot)$ is more flexible than $L_{\text{BJ}}(\cdot,\cdot)$ or $L_{\text{ICPW}}(\cdot,\cdot)$: as long as $F_T(y|X_i)$ or $G(c)$ is correctly specified (even if we do not know which one), $L_{\text{AICPW}}(\cdot,\cdot)$ has the expectation identical to that of the initial loss function $L(\cdot,\cdot)$.

Proposition 2. *(Double Robustness)*

Let $L_{AICPW,0}(\widetilde{Y}_i, f(X_i))$ be determined by (3.10) with $G(c)$ and $F_T(y|X_i)$ respectively replaced by $G_0(c)$ and $F_{T0}(y|X_i)$. Then

(a) the expectation of the loss function (3.10) is given by

$$\begin{aligned} & E\left\{L_{AICPW,0}(\widetilde{Y}_i, f(X_i))\right\} \\ = \; & E\left\{\frac{G(T_i)}{G_0(T_i)} L(\widetilde{T}_i, f(X_i))\right\} \\ & -E\left\{L(\widetilde{T}_i, f(X_i)) \times \left(\int_0^{T_i} \frac{F_T(t|X_i)}{F_{T0}(t|X_i)} \left[\frac{d}{dt}\left\{\frac{G(t)}{G_0(t)}\right\}\right] dt\right)\right\}; \end{aligned}$$

(b) if either $F_T(y|X_i) = F_{T0}(y|X_i)$ or $G(c) = G_0(c)$, then we have

$$E\left\{L_{AICPW,0}(\widetilde{Y}_i, f(X_i))\right\} = E\left\{L(\widetilde{T}_i, f(X_i))\right\}, \tag{3.11}$$

where the expectations are evaluated with respect to the working models for the associated random variables.

The proof of this proposition is deferred to Section S4.2 of the Supplementary Material. Proposition 2 (b) resembles the property of double robust estimators in regression analysis (e.g., Rubin and van der Laan 2007, Theorem 1), where only one of the two involved models is required to be correctly specified.

For ease of referral, we let $L^*(\cdot,\cdot)$ denote the loss function defined by (3.7), (3.9), or (3.10). By (2.1), Propositions 1 and 2 indicate $R(f) = E\left\{L^*(\widetilde{Y}_i, f(X_i))\right\}$. Consequently, we now modify (2.3) with the complete observations $\mathcal{O}_{\text{comp}}$ of a random sample replaced by the available censored data $\mathcal{O}_{\text{cd}}$. That is, we want to find $\widehat{f}_{\text{cd}} \in \mathcal{F}$ such that

$$\widehat{f}_{\text{cd}} = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L^*(\widetilde{Y}_i, f(X_i)) \right\}. \tag{3.12}$$

However, the minimization problem (3.12) cannot directly proceed due to the involvement of the adjusted loss function $L^*(\cdot,\cdot)$ with unknown (conditional) survivor functions $F_T(y|X_i)$ or/and $G(c)$. To circumvent this difficulty, we work on an approximation of $L^*(\cdot,\cdot)$, denoted $\widehat{L}^*(\cdot,\cdot)$, by replacing $F_T(y|X_i)$ and $G(c)$ with their consistent estimators which are denoted by $\widehat{F}_T(y|X_i)$ and $\widehat{G}(c)$, respectively, where the construction of $\widehat{F}_T(y|X_i)$ and $\widehat{G}(c)$ is deferred to Section 4.2. Our ultimate goal is to apply the observed censored data $\mathcal{O}_{\text{cd}}$ to find the solution of the following minimization prob-

lem:

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \widehat{L}^*(\widetilde{Y}_i, f(X_i)) \right\}. \tag{3.13}$$

4. Unbiased Boosting Estimation with Censored Data

4.1 Boosting Estimation Procedure

Propositions 1 and 2(b) say that an adjusted loss function, constructed from censored data, has the same expectation as that of the initial loss function $L(\widetilde{T}_i, f(X_i))$, indicating that on average, the risk function induced from an adjusted loss function is identical to that derived from the original failure time $\widetilde{T}_i$ as if $\widetilde{T}_i$ were available for all $i \in \{1, \cdots, n\}$. In this sense, we regard the procedure described in this subsection as *unbiased boosting estimation for censored (UBEC) data*.

Now we describe an implementation procedure for the minimization problem (3.13). With a given form of $\widehat{L}^*(\cdot,\cdot)$, it may be tempting to use (2.6) by replacing $\partial L(\widetilde{T}_i, f(X_i))$ with $\partial \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i)) \triangleq \frac{\partial \widehat{L}^*(u,v)}{\partial v}\Big|_{u=\widetilde{Y}_i, v=f^{(m)}(X_i)}$ (assuming its existence). However, as discussed by Schapire and Freund (2014, p.189), directly using the gradient descent update may lead to an entirely unconstrained new update $f^{(m+1)}(\cdot)$. A scheme to overcome this problem is to impose constraints to ensure each updated estimate of $f(\cdot)$ to be contained in a class of functions.

Let $\mathcal{C}$ represent a certain class of continuous functions mapping $\mathbb{R}^p$ to $\mathbb{R}$ which are uniformly bounded over any finite domain. Instead of using the increment $-\widehat{\alpha}_{m+1}\partial L(\widetilde{T}_i, f^{(m)}(X_i))$ in (2.6) with $\partial L(\widetilde{T}_i, f^{(m)}(X_i))$ replaced by $\partial \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i))$ to update estimation of $f(\cdot)$, here we take the increment to be $\widehat{\alpha}_{m+1}h_{m+1}(X_i)$ with $h_{m+1}(\cdot)$ taken from $\mathcal{C}$, and update the estimate of $f(\cdot)$ at iteration $(m+1)$ using a modified version of (2.6):

$$f^{(m+1)}(X_i) = f^{(m)}(X_i) + \widehat{\alpha}_{m+1}\widehat{h}_{m+1}(X_i), \tag{4.14}$$

where $\widehat{\alpha}_{m+1}$ and $\widehat{h}_{m+1}(\cdot)$ are determined by

$$\left(\widehat{\alpha}_{m+1}, \widehat{h}_{m+1}\right) = \underset{\substack{\alpha_{m+1}\in\mathbb{R}\\ h_{m+1}\in\mathcal{C}}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1}h_{m+1}(X_i)) \right\}. \tag{4.15}$$

While using (4.15) involves the determination of two unknown components α_{m+1} and $h_{m+1}(\cdot)$ rather than merely one unknown parameter α_{m+1} as in (2.6), (4.15) ensures the estimates of $f(\cdot)$ at each iteration to be bounded, and thus, assuring the final estimate to fall in the class $\mathcal{F}$.

To find the minimizer of (4.15) at iteration $(m+1)$, we may take two iterative steps to find $\widehat{\alpha}_{m+1}$ and $\widehat{h}_{m+1}(\cdot)$ separately rather than jointly. First, treating $\widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1}h_{m+1}(X_i))$ as a function of the argument $h_{m+1}(X_i)$ with other quantities held fixed, we apply the first order Taylor

series expansion around $h_{m+1}(X_i) = 0$:

$$\begin{aligned} & \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1} h_{m+1}(X_i)) \\ \approx \; & \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i)) + \left\{ \partial \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i)) \times h_{m+1}(X_i) \right\} \alpha_{m+1}. \end{aligned} \tag{4.16}$$

Since the first term in (4.16) is free of $h_{m+1}(\cdot)$ and α_{m+1} is fixed, then the minimizer of $h_{m+1}(\cdot)$, denoted $\widehat{h}_{m+1}(\cdot)$, is determined by

$$\widehat{h}_{m+1} = \underset{h_{m+1} \in \mathcal{C}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \partial \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i)) \times h_{m+1}(X_i) \right\} \right]. \tag{4.17}$$

Next, replacing $h_{m+1}(\cdot)$ in (4.15) with $\widehat{h}_{m+1}(\cdot)$ gives

$$\widehat{\alpha}_{m+1} = \underset{\alpha_{m+1} \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i) + \alpha_{m+1} \widehat{h}_{m+1}(X_i)) \right\}. \tag{4.18}$$

Consequently, $\widehat{\alpha}_{m+1}$ and $\widehat{h}_{m+1}(\cdot)$ can be used to update $f^{(m)}(\cdot)$ and then produce $f^{(m+1)}(\cdot)$ using (4.14). This strategy is also called the *functional gradient descent algorithm* (e.g., Boyd and Vandenberghe 2004, p.475; Schapire and Freund 2014, p.190). A pseudo-code for the implementation is presented in Algorithm 1.

Algorithm 1 differs from the *greedy algorithm* of Zhang and Yu (2005) and Schapire and Freund (2014, p.178) which obtains the minimizers of α_{m+1} and h_{m+1} simultaneously. In implementing Algorithm 1, we need to set a *stopping* criterion. To highlight the feature of censored responses, we may examine the difference of $\widehat{L}^*(\cdot, \cdot)$ evaluated at two successive estimates

Algorithm 1: Functional Gradient Descent Algorithm

Let $f^{(0)} \in \mathcal{F}$ denote the initial value and set $\zeta = n^{-\varpi}$ for a given $\varpi \geq 1$;

for *step* m *with* $m = 0, 1, 2, \cdots$ **do**

(a) calculate $\partial \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i))$ for $i = 1, \cdots, n$;

(b) find $\widehat{h}_{m+1}$ by solving (4.17);

(c) solve the minimization problem (4.18) and obtain $\widehat{\alpha}_{m+1}$;

(d) update $f^{(m)}(X_i)$ using (4.14) and denote the resultant estimate as $f^{(m+1)}(X_i)$;

if

$$\left| \frac{1}{n} \sum_{i=1}^{n} \widehat{L}^*(\widetilde{Y}_i, f^{(m)}(X_i)) - \frac{1}{n} \sum_{i=1}^{n} \widehat{L}^*(\widetilde{Y}_i, f^{(m+1)}(X_i)) \right| \leq \zeta \quad (4.19)$$

then

Stop iteration and let

$$\widehat{f}_n(\cdot) \leftarrow f^{(\widetilde{m})}(\cdot) \quad (4.20)$$

be the final estimator, where $\widetilde{m}$ represents the iteration number m at the stopping step such that (4.19) is met for ζ.

end

end

$f^{(m+1)}(\cdot)$ and $f^{(m)}(\cdot)$ using the squared error (L_2-norm), the absolute error (L_1-norm), or Huber error, as a stopping criterion, and compare it to a pre-specified threshold value; an example of using the L_1-norm is shown in (4.19). Alternatively, the stopping step may be determined by examining the values of the empirical risk function against the number of iterations.

For data $\mathcal{O}_{\text{cd}}$ with the size n described in Section 3, let $\widetilde{m}$ represent the iteration number at the stopping step such that (4.19) is met for a pre-specified positive value ζ, and let $\widehat{f}_n(\cdot)$ denote the resultant estimator for the solution (3.13) that is determined by (4.20). The value of ζ determines when to stop iterations. Taking $\zeta = n^{-\varpi}$ with $\varpi \geq 1$ gives us a convenient way to discuss the *asymptotic* behavior of $\widehat{f}_n(\cdot)$, as shown in Section S4.4. In applications with a finite sample, one may often set ζ as a small value, such as 10^{-6}, regardless of the value of n.

4.2 Implementation Remarks

The boosting procedure described in Section 4.1 hinges on the specification of the class $\mathcal{C}$ as well as the use of a consistent estimator of $F_T(y|X_i)$ or $G(c)$.

Similar to Li and Luan (2005) and Bühlmann and Yu (2003), we employ the cubic spline method to characterize functions in $\mathcal{C}$. Specifically, any

function $h(\cdot)$ in $\mathcal{C}$ is assumed to take the additive form

$$h(X_i) = h_1(X_{i1}) + \cdots + h_p(X_{ip}),$$

with $X_i = \big(X_{i1}, \cdots, X_{ip}\big)^\top$ and each $h_j(X_{ij})$ expressed as an M-order spline with J knots, where M and J are positive integers. That is, using the truncated power basis functions $\big\{1, X_{ij}, X_{ij}^2, \cdots, X_{ij}^{M-1}, (X_{ij}-\rho_{j1})_+^{M-1}, \cdots, (X_{ij}-\rho_{jJ})_+^{M-1}\big\}$ with knots $\rho_{j1}, \cdots, \rho_{jJ}$, we write

$$h_j(X_{ij}) = \sum_{r=0}^{M-1} \beta_{jr} X_{ij}^r + \sum_{k=1}^{J} \gamma_{jk} (X_{ij} - \rho_{jk})_+^{M-1},$$

where β_{jr} for $r = 0, 1, \cdots, M-1$ and γ_{jk} for $k = 1, \cdots, J$ are unknown parameters, and $a_+ \triangleq \max(0, a)$ for constant a. In applications, cubic spline is often used with M set as 4, where J may be set as 2 (Hastie et al. 2008, p.143).

Regarding estimation of $F_T(y|X_i)$, one may employ available strategies based on parametric or semiparametric regression models (e.g., Lawless 2003). Such methods are straightforward to implement, but a major drawback is the sensitivity of the results to the model assumptions. Alternatively, one may invoke the kernel conditional Kaplan-Meier estimator (e.g., Dabrowska 1989) to consistently estimate $F_T(y|X_i)$. However, this approach requires the proper specification of the bandwidth and entails the curse of dimensionality when the dimension p of covariates is large (e.g.,

Geenens 2011).

To provide consistent yet robust estimation of $F_T(y|X_i)$, one often employs the random survival forest (RSF) to estimate $F_T(y|X_i)$. The estimation procedure is outlined as follows. First, set a positive integer D (e.g., $D = 1000$), and draw D independent bootstrap samples from the initial sample data $\mathcal{O}_{\text{cd}}$, denoted $\mathcal{S}_1, \cdots, \mathcal{S}_D$. For each bootstrap sample $\mathcal{S}_d$ with $d = 1, \cdots, D$, build a binary survival tree by recursive random splitting rules using the procedures, for instance, described by Cui et al. (2021); and let $\{\mathcal{A}_{ud} : u \in \mathcal{U}_d\}$ denote the collection of the resulting terminal nodes with $\mathcal{U}_d$ denoting a set of indices based on the dth bootstrap sample. Then the Nelson-Aalen estimator for the cumulative baseline hazard function based on a terminal node $\mathcal{A}_{ud}$ is given by

$$\widehat{\Lambda}_{\mathcal{A}_{ud}}(t) = \sum_{u \le t} \left\{ \frac{\sum_{i=1}^{n} \mathbb{I}\big(\Delta_i = 1\big)\mathbb{I}\big(Y_i = u\big)\mathbb{I}\big(X_i \in \mathcal{A}_{ud}\big)}{\sum_{i=1}^{n} \mathbb{I}\big(Y_i \ge u\big)\mathbb{I}\big(X_i \in \mathcal{A}_{ud}\big)} \right\} \quad \text{for } t > 0,$$

and the conditional cumulative baseline hazards function, given X_i, is thus given by

$$\widehat{\Lambda}_d(t|X_i) = \sum_{u \in \mathcal{U}_d} \mathbb{I}\big(X_i \in \mathcal{A}_{ud}\big)\widehat{\Lambda}_{\mathcal{A}_{ud}}(t).$$

Finally, the RSF estimate of $\Lambda(t|X_i)$ is given by $\widehat{\Lambda}(t|X_i) = \frac{1}{D}\sum_{d=1}^{D} \widehat{\Lambda}_d(t|X_i)$, and thus, leading to an estimate for $F_T(t|X_i)$ to be $\widehat{F}_T(t|X_i) = \exp\Big\{-\widehat{\Lambda}(t|X_i)\Big\}$.

Such an estimator of $F_T(t|X_i)$ was shown to be consistent under regularity conditions (e.g., Ishwaran and Kogalur 2010; Cui et al. 2021).

Finally, using the formulation of the Kaplan-Meier estimator (e.g., Lawless 2003), we estimate $G(c)$ by pooling the study subjects with their differences in covariates ignored:

$$\widehat{G}(c) = \prod_{i:Y_i \le c} \left(1 - \frac{1}{\#\{j : Y_j \ge Y_i\}}\right)^{1-\Delta_i},$$

which is shown to be consistent (e.g., Wang 1987), where $\#\{j : Y_j \ge Y_i\}$ represents the count of index j satisfying $Y_j \ge Y_i$.

4.3 Finite Sample Performance

To evaluate the finite sample prediction performance, we use the *integrated Brier score* (IBS) (Graf et al. 1999) as other authors did (e.g., Benner 2002; Zhu and Kosorok 2012). Let $F_T(t) = P(T_i \ge t)$ represent the unconditional survivor function for the survival time T_i. We want to assess how applying the estimator (4.20) to predict $F_T(t)$ may perform by using the censored sample $\mathcal{O}_{\text{cd}}$. To this end, we divide the original observed data $\mathcal{O}_{\text{cd}}$ as training data and validation data so that the censoring proportion in both data is comparable, and let $\mathcal{T}$ and $\mathcal{V}$ denote the set of subject indexes, respectively.

First, we use the training data in $\mathcal{T}$ to find an estimate of $F_T(t)$, where

the procedure described in Section 4.1 is implemented to determine $\widehat{f}_n(\cdot)$ in (4.20), with $F_T(y|X_i)$ and $G(c)$ in $L^*(\cdot,\cdot)$ respectively estimated by the RSF estimator and $\widehat{G}(c)$. Then using $\widehat{f}_n(\cdot)$, we take

$$\widehat{F}_T(t) = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \mathbb{I}\left\{\widehat{f}_n^*(X_i) \geq t\right\} \tag{4.21}$$

as an estimate of $F_T(t)$, where $\widehat{f}_n^*(x) = \exp\left\{\widehat{f}_n(x)\right\}$.

Next, we use the validation data in $\mathcal{V}$ by separating the measurements according to the censoring status, and then calculate the empirical version of the expected value for the squared difference between $\mathbb{I}(T_i > t)$ and $F_T(t)$ with the censoring effects accounted for. That is, for any $t > 0$, the Brier score (Benner 2002) for the validation data in $\mathcal{V}$ is defined as

$$BS(t) = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \left[\left\{0 - \widehat{F}_T(t)\right\}^2 \mathbb{I}\left(Y_i \leq t, \Delta_i = 1\right) \frac{1}{\widehat{G}(Y_i)} + \left\{1 - \widehat{F}_T(t)\right\}^2 \mathbb{I}\left(Y_i > t\right) \frac{1}{\widehat{G}(t)}\right], \tag{4.22}$$

where $\widehat{G}(t)$ and $\widehat{F}_T(t)$ are obtained from the training data. Then the integrated Brier score for the validation data is defined as

$$\text{IBS} = \{y_{\max}\}^{-1} \int_0^{y_{\max}} BS(t)dt, \tag{4.23}$$

where $y_{\max} = \max\{Y_i : i \in \mathcal{V}\}$.

To alleviate the division effects of the original data into the training and validation data, we further employ the K-fold cross-validation procedure

with a positive integer K. Specifically, we split the original data $\mathcal{O}_{\text{cd}}$ into K roughly equal-sized subsets so that the censoring proportion in each subsets is similar. For $k = 1, \cdots, K$, take the kth subset as validation data, and let the remaining $(K-1)$ pooled subsets as training data; let $\mathcal{V}_k$ and $\mathcal{T}_k$ represent the class of the subject indexes for the kth validation and training datasets, respectively.

For $k = 1, \cdots, K$, we apply the preceding steps to the training data in $\mathcal{T}_k$ and the validation data in $\mathcal{V}_k$ to calculate the Brier score at each Y_i with $i \in \mathcal{V}_k$ using (4.22) with $\mathcal{V}$ replaced by $\mathcal{V}_k$. Then we approximate the integrated Brier score (4.23) for the kth validation data by

$$\text{IBS}_k = \{y_{\max,k}\}^{-1} \sum_{i \in \mathcal{V}_k} BS(Y_i),$$

where $y_{\max,k} = \max\{Y_i : i \in \mathcal{V}_k\}$. Consequently, the integrated Brier score for the original sample data $\mathcal{O}_{\text{cd}}$ is approximated by the average of the K-fold cross-validation estimates of IBS:

$$\text{IBS}_{\text{CV}} = \frac{1}{K} \sum_{k=1}^{K} \text{IBS}_k. \tag{4.24}$$

5. Theoretical Results

In this section, we develop theoretical results for the proposed method, including convergence of the proposed iterated algorithm described in Al-

gorithm 1 and the consistency of the estimator $\widehat{f}_n(\cdot)$ defined by (4.20).

5.1 Convergence of the Algorithm

Theorem 1. *Assume regularity conditions (C1)-(C5) in Section S1.1 of the Supplementary Material. Suppose that we are given data of a random sample $\mathcal{O}_{cd}$ with the given size n considered in Section 3. For any initial function $f^{(0)} \in \mathcal{F}$, let $f^{(m+1)}$ denote the updated estimate of the function at step $(m+1)$ of Algorithm 1. Then*

$$\lim_{m\to\infty} R(f^{(m+1)}) = R(f_0). \tag{5.25}$$

This theorem says that with data $\mathcal{O}_{\text{cd}}$ given, repeatedly iterating Algorithm 1 yields convergence as the iteration number approaches infinity.

While Theorem 1 ensures the convergence of iterations in Algorithm 1, in applications, it is not ideal to take a super large number $\widetilde{m}$ to stop iterations and avoid overfitting (e.g., Jiang 2004; Zhang and Yu 2005). In Section S4.3 of the Supplementary Material, we show that for any given nonnegative integer m, there exist positive constants b^* and B^* with $b^* < B^*$ such that

$$R(f^{(m+1)}) - R(f_0) \;\le\; \left(1 - \frac{b^*}{B^*}\right)^m \left\{R(f^{(0)}) - R(f_0)\right\}, \tag{5.26}$$

where f_0 is the true function satisfying (2.2).

Let $\widetilde{m}$ denote a positive integer for stopping iterations in Algorithm 1, then we have

$$\begin{aligned}\left|R(f^{(\widetilde{m})}) - R(f^{(\widetilde{m}+1)})\right| &\leq \left|R(f^{(\widetilde{m})}) - R(f_0)\right| + \left|R(f^{(\widetilde{m}+1)}) - R(f_0)\right| \\ &\leq \left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}-1} \left\{R(f^{(0)}) - R(f_0)\right\} \\ &\quad + \left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}} \left\{R(f^{(0)}) - R(f_0)\right\} \\ &= \left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}-1} \left\{R(f^{(0)}) - R(f_0)\right\} \left(2 - \frac{b^*}{B^*}\right),\end{aligned} \tag{5.27}$$

where the second inequality comes from (5.26).

We wish to stop the iteration at $\widetilde{m}$ if the difference (in absolute value) of its following two iterations is smaller than a given threshold, say, $\zeta > 0$. By (5.27), we suggest to stop the iteration if

$$\left(1 - \frac{b^*}{B^*}\right)^{\widetilde{m}-1} \left\{R(f^{(0)}) - R(f_0)\right\} \left(2 - \frac{b^*}{B^*}\right) < \zeta.$$

That is,

$$\widetilde{m} < 1 + \frac{\log\left\{\frac{\zeta}{\left|R(f^{(0)}) - R(f_0)\right|\left(2 - \frac{b^*}{B^*}\right)}\right\}}{\log\left(1 - \frac{b^*}{B^*}\right)}, \tag{5.28}$$

suggesting that the iteration number $\widetilde{m}$ for stopping Algorithm 1 is upper bounded by a value depending on b^*, B^*, and ζ, as well as the initial value $f^{(0)}$.

In addition, we observe that the upper bound of (5.28) involves the initial value $f^{(0)}$. While (5.25) holds regardless of choices of the initial value, a better choice of $f^{(0)}$ makes $\left|R(f^{(0)}) - R(f_0)\right|$ and right-hand side of (5.28) become small, and then a smaller number $\widetilde{m}$ of iterations can achieve the required accuracy.

Finally, another concern is whether the sample size n affects the convergence (5.25). As discussed in Section 2.2, based on n observations, the function f is "parametrized" and characterized as $\{f(X_1), \cdots, f(X_n)\}$. From Algorithm 1, the updated value at X_i in step $(m+1)$, $f^{(m+1)}(X_i)$, is determined by $\widehat{h}_{m+1}(X_i)$ and $\widehat{\alpha}_{m+1}$ whose optimization depends on the sample size n, a larger sample size n may enable the updated value $f^{(m+1)}(X_i)$ to be more precise, yet the inequality (5.26) is free of the sample size n but depends on the initial value $f^{(0)}$.

5.2 Consistency and Boundness

In this subsection, we examine the asymptotic behaviour of $\widehat{f}_n$. For $f \in \mathcal{F}$ and $\widehat{L}^*(\cdot,\cdot)$ defined as in (3.13), define

$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \widehat{L}^*(\widetilde{Y}_i, f(X_i)). \tag{5.29}$$

Theorem 2. *Assume regularity conditions (C1)-(C5) in Section S1.1 of the Supplementary Material. Suppose that Algorithm 1 is run to a sequence*

of random samples $\mathcal{O}_{cd}$ with varying size n, and let $\widehat{f}_n$ denote the resultant estimator defined as in (4.20) at the stopping time $\widetilde{m}$ satisfying (4.19). Then for any $\epsilon > 0$,

$$P(\|\widehat{f}_n - f_0\|_\infty \leq \epsilon) \to 1 \text{ as } n \to \infty,$$

where $\|\widehat{f}_n - f_0\|_\infty = \sup_{X_i:\ i=1,\cdots,n} \left|\widehat{f}_n(X_i) - f_0(X_i)\right|$ is the L_∞ norm of $\widehat{f}_n - f_0$ evaluated over $\{X_i : i = 1, \cdots, n\}$.

Theorem 2 shows the limiting behavior of a sequence of the proposed estimator $\widehat{f}_n$ that is obtained by applying the same method (i.e., Algorithm 1) to a sequence of data $\mathcal{O}_{\text{cd}}$ with varying sample sizes n. The result says that the difference between the estimator $\widehat{f}_n$ and its target f_0 expressed in the L_∞ norm converges in probability to zero, suggesting the consistency of $\widehat{f}_n$ in this sense. The proof of Theorem 2 is placed in Section S4.4 of the Supplementary Material.

Next, we examine a lower bound of $\widehat{f}_n - f_0$ in the infinity norm.

Theorem 3. *Under regularity conditions and the setup in Theorem 2, there exist positive constants s and $\alpha > \frac{1}{2}$ such that*

$$\left\|\widehat{f}_n - f_0\right\|_\infty \geq s n^{-\frac{\alpha}{2\alpha+1}}$$

for any sample size n.

This lower bound of $\left\|\widehat{f}_n - f_0\right\|_\infty$ is characterized in terms of the sample size n, which sheds light into the finite sample performance of $\widehat{f}_n$. It implies that with a small sample size n, the difference between $\widehat{f}_n$ and f_0 cannot be arbitrarily small and must be lower bounded by a positive constant related to the size n.

6. Analysis of NKI Breast Cancer Data

To illustrate its utility, we apply the proposed method to analyze the breast cancer data collected by the Netherlands Cancer Institute (NKI) (van de Vijver et al. 2002). Tumors from 295 women with breast cancer were collected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. Tumors for those patients were primarily invasive breast cancer carcinoma that were about 5 centimeter in diameter. Patients at diagnosis were 52 years or younger and the diagnosis was done between 1984 and 1995. Of all those patients, only 79 patients died before the study ended, yielding approximately 73.2% censoring.

For those patients, about 25000 gene expressions were also collected, in which 70 genes with previously determined average profiles are useful for tumor diagnosis (van de Vijver et al. 2002, p.2002); these gene expression values are recorded as the log intensity. In our analysis here, we study the

relationship between survival times and those 70 genes by implementing the proposed method, where to specify $\mathcal{C}$ discussed in Section 4.2, we set $M = 4$ and $J = 2$ together with ρ_{j1} and ρ_{j2} respectively set to be the 25th and 75th percentiles of the jth variable in X_i in the sample. We assess the prediction performance using the measure discussed in Section 4.3 with $K = 5$.

To gain a better understanding how the IBS measure may perform over a number of datasets instead of a single dataset, here we implement the procedure described in Section 4.3 repeatedly to N_{boot} bootstrap samples which are randomly generated from the initial sample $\mathcal{O}_{\text{cd}}$ with replacement, where N_{boot} is taken as 500. Let $\text{IBS}^{(d)}$ denote the value (4.24) yielded from the dth bootstrap sample for $d = 1, \cdots, N_{\text{boot}}$, where the three adjusted loss functions (3.7), (3.9), and (3.10) are respectively used in combination with the three loss functions in Table 1. The boxplots for $\left\{\text{IBS}^{(d)} : d = 1, \cdots, N_{\text{boot}}\right\}$ are displayed in Figure 1, where L1-BJ, L1-ICPW, and L1-AICPW denote the adjusted loss functions (3.7), (3.9), and (3.10), respectively, with the loss function $L(\cdot, \cdot)$ taken as the L_1-norm; L2-BJ, L2-ICPW, and L2-AICPW denote the adjusted loss functions with the loss function $L(\cdot, \cdot)$ set as the L_2-norm; and H-BJ, H-ICPW, and H-AICPW denote the adjusted loss functions with the loss function $L(\cdot, \cdot)$ set as the

Huber loss function. In comparison, we also apply the "coxph" method of Chen et al. (2013) to analyze the data and report the results with the boxplot labelled "coxph".

Compared with the "coxph" method, the proposed UBEC method generally performs well with small values of IBS regardless of the choice of the loss function $L(\cdot, \cdot)$ or its adjusted version. With a given loss function form, the AICPW adjustment tends to perform the best; with a given adjustment strategy, the Huber loss function outperforms other two loss functions, and the L_2-norm loss function seems to incur the most variability.

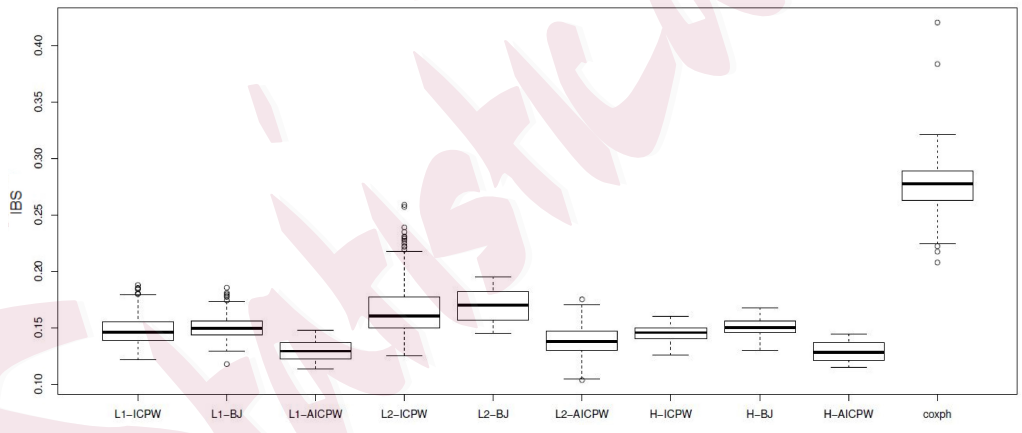

Figure 1: Boxplots of integrated Brier score with 500 repeated bootstrapping.

On the other hand, the "coxph" method produces noticeable larger IBS values than our proposed estimators, suggesting unsatisfactory prediction performance. To see why the "coxph" method fails to work, we apply

the Schoenfeld residuals (e.g., Lawless 2003, p.364) and implement the R function `cox.zph` to test the proportional hazards assumption, giving the p-value 7.3e-06, which suggests inapplicability of the proportional hazards assumption.

7. Discussion

In Section 6, we use an example to illustrate the usage of the proposed method. To assess the finite sample performance of our method, we conduct simulation studies under different settings and report the details in Section S5 of the Supplementary Material to save space, where the numerical results demonstrate that the proposed UBEC method produces satisfactory results.

As discussed in Section 3.2, the proposed method hinges on consistent estimation of $F_T(t|X_i)$ and $G(c)$, which is often handled with existing methods, as remarked in Section 4.2. Different from many available parametric or semiparametric methods which focus on inference about the model parameters or estimating the conditional survivor function $F_T(t|X_i)$, our goal here concentrates on finding an optimal function of covariates to predict the transformed failure time T_i. Further, the development enables us to sensibly estimate the unconditional survivor function of T_i without know-

ing the distribution of the covariates X_i nor specifying any model forms. In contrast to most existing boosting methods for censored data which focus on developing implementation steps without theoretical justifications, the validity of the proposed method is asserted discreetly.

In the development here we assume that the censoring time C_i is independent of the covariates X_i. This consideration is basically driven by its common use in the literature, which enables one to estimate the survivor function of the censoring process consistently using the Kaplan-Meier estimator. This assumption, however, is not essential in the development. One may relax it by replacing the unconditional survivor function $G(c)$ of C_i with the conditional survivor function, $G(c|X_i) \triangleq P(C_i > c|X_i)$, of C_i given X_i. Then the development can be modified accordingly, where $G(c|X_i)$ is required to be consistently estimated.

Supplementary Material

The Supplementary Material contains a full set of regularity conditions with discussion, detailed proofs for the results in Sections 3.2 and 5, and simulation studies.

Acknowledgements

The authors thank the editor and the review team for their comments on the initial version of the manuscript. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs program.

References

Bellot, A. and van der Schaar, M. (2018a). Boosted trees for risk prognosis. *Proceedings of Machine Learning Research*, 85, 1-15.

Bellot, A. and van der Schaar, M. (2018b). Multitask boosting for survival analysis with competing risks. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*, 1-10.

Benner, A. (2002). Application of "aggregated classifiers" in survival time studies. In Härdle, W. and Rönz, B. (eds), *Proceedings in Computational Statistics: COMPSTAT 2002*. Heidelberg: Physica-Verlag, 171-176.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:14, 1-10.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge, New York.

Buckley, J. and James, I. (1979) Linear regression with censored data. *Biometrika*, 66, 429-436.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98, 324-339.

Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine*, Article ID 873595, 1-8.

Cui, Y., Zhu, R., Zhou, M., and Kosorok, M. (2021) Consistency of survival tree and forest models: splitting bias and correction. *Statistica Sinica*, 1-40. DOI: 10.5705/ss.202020.0263.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics*, 17, 1157-1167.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 121, 256-285.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 1189-1232.

Geenens, G. (2011). Curse of dimensionality and related issues in nonparametric functional

regression. *Statistics Surveys*, 5, 30–43.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18, 2529-2545.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, New York.

He, K., Li, Y., Zhu, J., Liu, H., Lee, J. E., Amos, C. I., Hyslop, T., Jin, J., Lin, H, Wei, Q., and Li, Y. (2016). Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. *Bioinformatics*, 32, 50-57.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7, 355–373.

Ishwaran, H. and Kogalur, U. B. (2010) Consistency of random survival forests. *Statistics and Probability Letters*, 80, 1056-1064.

Jiang, W. (2004). Process consistency for AdaBoost. *The Annals of Statistics*, 32, 13-29.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lee, D. K. K., Chen, N., and Ishwaran, H. (2021). Boosted nonparametric hazards with time-dependent covariates. *The Annals of Statistics*, 49, 2101-2128.

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

Lu, W. and Li, L. (2008). Boosting method for nonlinear transformation models with censored survival data. *Biostatistics*, 9, 658–667.

Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32, 30-55.

Mayr, A., Hofner, B., and Schmid, M. (2016). Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC Bioinformatics*, 17:288, 1-12.

Ridgeway, G. (1999). The state of boosting. In Computing Science and Statistics. Models, Predictions, and Computing. *Proceedings of the 31st Symposium on the Interface*, 172–181.

Rubin, D. and van der Laan, M. J. (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3, 1-21.

Schapire, R. E. and Freund, Y. (2014). *Boosting: Foundations and Algorithms*. The MIT Press, Cambridge.

Schapire, R. E. (1990). The strength of Weak learnability. *Machine Learning*, 5, 197-227.

Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. *BMC Bioinformatics*, 9:269, 1:13.

Steingrimsson, J. A., Diao, L., Molinaro, A. M., Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in Medicine*, 35, 3595-3612.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A.M., Voskuil, D. W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347, 1999 - 2009.

Wang, J. G. (1987) A note on the uniform consistency of the Kaplan-Meier estimator. *The Annals of Statistics*, 15, 1313-1316.

Wang, Z. and Wang, C.Y. (2010). Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*, 9:1, 1-31.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: convergence and consistency. *The Annals of Statistics*, 33, 1538-1579.

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, 107, 331-340.