DSSR

Discussion Paper No. 157

Partial Identification of Survival Prediction Performance under Dependent Right Censoring

Yuji Komiyama and Yasumasa Matsuda

October, 2026

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Partial Identification of Survival Prediction Performance under Dependent Right Censoring

Yuji Komiyama* Yasumasa Matsuda†

Graduate School of Economics and Management, Tohoku University, Japan

Abstract

Survival prediction models are usually evaluated under right censoring by inverse probability of censoring weighting (IPCW), which assumes independent censoring, an assumption that cannot generally be verified from right-censored data. Leaving the dependence between event and censoring times unrestricted, we derive sharp bounds on the integrated Brier score (IBS) of a fixed model and on IBS differences between models by partial identification. The bounds use the fact that each individual's event time is shared across evaluation times and models, and for more than two models we also derive the sharp upper bound on the difference from the best competitor. A negative upper bound on a difference shows that one model has a smaller IBS than the other under every full-data distribution compatible with the observed data. For a fixed observed-data distribution, we develop estimators, confidence regions, a test of robust strict dominance, and an intersection–union test. In simulations, the confidence regions and pairwise tests performed close to their nominal levels. In three applications, the tests supported the IPCW ranking only for mortgage prepayment data with little censoring, whereas for the other two data sets the observed data alone did not support it.

Keywords: partial identification; dependent censoring; survival prediction; integrated Brier score; model comparison

1 Introduction

Evaluating a survival prediction model requires each individual's survival status at each time, but right censoring leaves part of it unobserved. Such a model uses individual characteristics, called covariates, to predict the probability that an event of interest, such as death or disease recurrence, has not occurred by each time. Its performance is often measured by the Brier score at a given time, the expected squared difference between this predicted probability and the indicator that the event has not occurred by that time (Gerds & Schumacher, 2006; Graf et al., 1999). The integrated Brier score (IBS) averages the Brier score over a prespecified evaluation period. These measures are used to compare candidate models, such as the Cox proportional hazards model (Cox, 1972), random survival forests (Ishwaran et al., 2008), and DeepSurv (Katzman et al., 2018), and to choose the one with the smallest IBS. When follow-up

*Email: komiyama.yuji.r2@dc.tohoku.ac.jp

†Corresponding author. Email: yasumasa.matsuda.a4@tohoku.ac.jp

ends before the event, however, the observation is right censored: the event time is known only to exceed the censoring time, so the survival status after censoring is unknown. Evaluating the Brier score and the IBS therefore requires accounting for the unobserved event times.

Previous studies have estimated prediction performance from right-censored data by inverse probability of censoring weighting (IPCW) under independent censoring. IPCW weights each individual whose survival status is observed by the inverse of the probability of remaining uncensored until that status is observed. For weights based on the marginal censoring distribution, independent censoring means that the censoring time is independent of both the event time and the covariates. Graf et al. (1999) proposed such an estimator of the Brier score, and Gerds and Schumacher (2006) proved it consistent and extended it to conditionally independent censoring, under which the two times are independent given the covariates. Even this condition fails when an unobserved factor, such as a deterioration in health, makes a patient both more likely to drop out and more likely to die soon. IPCW estimates of the IBS are then biased, and the ranking of models based on them can differ from the ranking under the full-data distribution. Moreover, neither assumption can generally be verified from right-censored data, because different dependence structures can generate the same observed-data distribution (Tsiatis, 1975).

Few studies have addressed the evaluation of survival predictions under dependent censoring. Lillelund et al. (2026) propose a dependent Brier score, which they describe as the first such measure. Under an assumed Archimedean copula, which specifies the dependence between event and censoring times, they replace the unknown event time of each censored individual with its conditional mean beyond the censoring time. Their score therefore targets a surrogate defined by these imputed times rather than the Brier score defined by the true event times. It also requires specifying the copula family and its parameter, although the dependence structure is generally not identifiable from right-censored data. A method that evaluates the IBS based on the true event times without specifying this dependence is therefore needed.

In this paper, we evaluate survival predictions by partial identification, leaving the dependence between event and censoring times unrestricted. Partial identification characterizes the set of values of a target quantity that are compatible with the observed data and the maintained conditions (Manski, 2003; Molinari, 2020; Tamer, 2010). Under dependent censoring, it has yielded bounds on survival functions (Peterson, 1976) and on regression parameters and covariate effects (Sakaguchi, 2024; Willems et al., 2026), but not on the prediction performance of a given model. In our setting, many full-data distributions of the event time, censoring time, and covariates generate the same observed-data distribution, yet they can imply different values of the IBS, our measure of prediction performance. The set of these values is called the identified set, and its bounds describe the range of performance supported by the observed data. The identified set thus shows what can be concluded about prediction performance without assuming independent censoring or specifying a copula.

Partial identification has also been used to evaluate prediction performance when outcomes are only partly observed. Rambachan et al. (2026) bound performance measures under selective observation of outcomes, allowing unobserved confounding subject to restrictions on outcome differences between selected and unselected individuals with the same covariates. Maia Polo et al. (2024) bound the accuracy and related measures of a fixed classifier using observed weak labels and a model linking them to the unobserved true labels. Guerdan et al. (2024) bound the difference in performance between two decision policies directly, using the fact that the policies share the same unobserved outcomes, which avoids counting the uncertainty common to both

twice. Liu et al. (2026) characterize the sharp identification region of the fairness–accuracy frontier under selective labels for binary outcomes, measuring accuracy by classification error and fairness by differences in positive decision rates between groups. These studies do not explicitly address right-censored survival outcomes, for which a single event time determines survival status across all evaluation times. We derive sharp bounds on the IBS and on differences between models while respecting this common event time across evaluation times and models.

Our main contribution is to derive sharp bounds on the IBS and on IBS differences between models, using the fact that each individual's event time is shared across evaluation times and models. Sharpness means that the bounds cannot be narrowed using the observed-data distribution alone. Bounding the Brier score separately at each time would allow survival statuses that no single event time produces, and bounding each model separately can widen the bounds on an IBS difference. When more than two models are compared, we also derive the sharp upper bound on the difference between each model's IBS and that of its best competitor. For models trained separately from the evaluation sample and a fixed observed-data distribution, we then develop estimators of these bounds, confidence regions, a test of robust strict dominance, and an intersection–union test. A negative upper bound on one model's IBS minus another's establishes a smaller IBS for the first model under every compatible full-data distribution. If the two bounds have opposite signs, the observed data alone cannot rank the two models.

The remainder of this paper is organized as follows. Section 2 derives the bounds for a single model, two models, and more than two models, and Section 3 develops estimators, confidence regions, and tests. Section 4 examines their finite-sample behavior through simulations, and Section 5 applies them to three real data sets. Section 6 concludes, and Appendices A and B contain the proofs.

2 Partial identification of prediction performance

This section derives sharp bounds on prediction performance without restricting the dependence between event and censoring times. Section 2.1 introduces the setting, the evaluation targets, and the identified set. Sections 2.2 and 2.3 derive sharp lower and upper bounds on the IBS of a single model and on the IBS difference between two models. Section 2.4 derives the sharp upper bound on the difference from the best competitor and uses it to define when a model is robustly strictly best among the K models.

2.1 Framework and evaluation targets

For each individual, we consider the event time $T \geq 0$, the censoring time $C \geq 0$, and the covariates $X \in \mathcal{X}$, but we observe only the observed time, the event indicator, and the covariates:

$$Y = \min(T, C), \qquad \Delta = \mathbf{1}\{T \leq C\}, \qquad X.$$

We write the observed data for one individual as $O = (Y, \Delta, X)$ and the evaluation sample as $O_1, \ldots, O_n$. If $\Delta = 1$, the event occurs during follow-up, and its time is observed as $T = Y$. If $\Delta = 0$, the individual is censored at $C = Y$, and the event time satisfies $T > Y$ but is otherwise unknown. Hence, for a censored individual, the data reveal only that the event time is later than the observed time.

We distinguish two distributions: the full-data distribution Q of (T, C, X) and the observed-data distribution P of (Y, Δ, X). Each full-data distribution determines an observed-data distribution. We say that Q is compatible with P if the distribution of $(\min(T, C), \mathbf{1}\{T \le C\}, X)$ under Q equals P, and we denote by $\mathcal{Q}(P)$ the set of full-data distributions compatible with P. We impose no restriction on the dependence between T and C, so a single observed-data distribution is generally compatible with multiple full-data distributions.

Let $0 < \tau < \infty$ be a prespecified evaluation horizon, defining the evaluation period $[0, \tau]$. We measure prediction performance by the Brier score at each time t and by the IBS, its average over this period. A prediction model outputs a predicted survival probability $\widehat{S}(t \mid x) \in [0, 1]$ for each time t and covariate value x, measurable in (t, x). Like a survival function, $\widehat{S}(t \mid x)$ is nonincreasing in t for every x. The model is trained on data separate from the evaluation sample and is held fixed during evaluation. Under a full-data distribution Q, the Brier score at time t and the IBS are

$$\mathrm{BS}(t; Q) = \mathbb{E}_Q\Big[\big\{\mathbf{1}(T > t) - \widehat{S}(t \mid X)\big\}^2\Big], \qquad \mathrm{IBS}(\tau; Q) = \frac{1}{\tau}\int_0^\tau \mathrm{BS}(t; Q)\, dt. \tag{1}$$

A single event time determines the survival status at every evaluation time: $\mathbf{1}(T > t)$ equals one before the event time and zero afterward. To make this explicit, write $r = \min(T, \tau)$ for an individual's event time truncated at the evaluation horizon. Define the loss for truncated event time r and covariates x by

$$L(r, x; \tau) = \int_0^r \big\{1 - \widehat{S}(t \mid x)\big\}^2\, dt + \int_r^\tau \widehat{S}(t \mid x)^2\, dt, \qquad 0 \le r \le \tau. \tag{2}$$

When several models are compared, we index them by j and write $\widehat{S}_j$, BS_j, IBS_j, and L_j. Exchanging the expectation and the integral in (1) gives $\mathrm{IBS}(\tau; Q) = \tau^{-1}\mathbb{E}_Q L(\min(T, \tau), X; \tau)$.

The Brier scores and IBS values depend on the full-data distribution Q and therefore are generally not uniquely determined by the observed-data distribution P alone. Under conditionally independent censoring, $T \perp C \mid X$, the Brier score is uniquely determined by P given suitable positivity conditions. In that case, IPCW can use the censoring distribution given X (Gerds & Schumacher, 2006).

For a target quantity $\theta(Q)$, such as $\mathrm{IBS}(\tau; Q)$ or $\mathrm{IBS}_j(\tau; Q) - \mathrm{IBS}_k(\tau; Q)$, we define its identified set as the set of values that the target takes under the full-data distributions compatible with P:

$$\Theta_{\mathrm{I}}(P) = \big\{\theta(Q) : Q \in \mathcal{Q}(P)\big\}.$$

The observed data cannot distinguish among the values in $\Theta_{\mathrm{I}}(P)$. A set $B(P)$ is an outer set if $\Theta_{\mathrm{I}}(P) \subseteq B(P)$. Such a set contains $\theta(Q)$ for every $Q \in \mathcal{Q}(P)$, but it may also include values that cannot be attained under any compatible full-data distribution. An outer set is said to be sharp if $B(P) = \Theta_{\mathrm{I}}(P)$.

We define the sharp lower and upper bounds on θ by

$$\underline{\theta}(P) = \inf_{Q \in \mathcal{Q}(P)} \theta(Q), \qquad \overline{\theta}(P) = \sup_{Q \in \mathcal{Q}(P)} \theta(Q).$$

These are, respectively, the largest lower bound and the smallest upper bound that hold for every $Q \in \mathcal{Q}(P)$.

In the following subsections, we express the sharp lower and upper bounds in terms of P for the IBS of a single model and for the IBS difference between two models. For comparisons among $K \geq 3$ models, we derive the sharp upper bound on the difference from the best competitor and use its sign to assess whether a model is robustly strictly best.

2.2 Single model

At a fixed time $t > 0$, the observed data determine the survival status $\mathbf{1}(T > t)$ except for individuals censored before t. Specifically, the survival status equals one if $Y > t$ or $(Y = t, \Delta = 0)$, and zero if $Y \leq t$ and $\Delta = 1$. For individuals with $Y < t$ and $\Delta = 0$, both survival states are compatible with the observed data, so the squared prediction loss can be either $\widehat{S}(t \mid X)^2$ or $\{1 - \widehat{S}(t \mid X)\}^2$. Taking the minimum and maximum of these losses yields the following bounds, which are sharp when $P(Y = t, \Delta = 0) = 0$.

Proposition 2.1. Fix $t > 0$ and define

$$\underline{\mathrm{BS}}(t; P) = \mathbb{E}_P\Big[\mathbf{1}(Y > t)\big\{1 - \widehat{S}(t \mid X)\big\}^2 + \mathbf{1}(Y \leq t, \Delta = 1)\,\widehat{S}(t \mid X)^2 + \mathbf{1}(Y \leq t, \Delta = 0)\min\big\{\widehat{S}(t \mid X)^2, \big\{1 - \widehat{S}(t \mid X)\big\}^2\big\}\Big],$$

$$\overline{\mathrm{BS}}(t; P) = \mathbb{E}_P\Big[\mathbf{1}(Y > t)\big\{1 - \widehat{S}(t \mid X)\big\}^2 + \mathbf{1}(Y \leq t, \Delta = 1)\,\widehat{S}(t \mid X)^2 + \mathbf{1}(Y \leq t, \Delta = 0)\max\big\{\widehat{S}(t \mid X)^2, \big\{1 - \widehat{S}(t \mid X)\big\}^2\big\}\Big].$$

Then the following hold.

(i) *Outer set.* The interval $[\underline{\mathrm{BS}}(t; P), \overline{\mathrm{BS}}(t; P)]$ is an outer set for the identified set of $\mathrm{BS}(t; Q)$; that is,

$$\underline{\mathrm{BS}}(t; P) \leq \mathrm{BS}(t; Q) \leq \overline{\mathrm{BS}}(t; P) \quad \text{for every } Q \in \mathcal{Q}(P).$$

(ii) *Sharpness.* If $P(Y = t, \Delta = 0) = 0$, then $\underline{\mathrm{BS}}(t; P)$ and $\overline{\mathrm{BS}}(t; P)$ are the sharp lower and upper bounds on $\mathrm{BS}(t; Q)$, and both are attained. That is, the identified set of $\mathrm{BS}(t; Q)$ is the closed interval $[\underline{\mathrm{BS}}(t; P), \overline{\mathrm{BS}}(t; P)]$.

(iii) *Width.* The width of the bounds is

$$\overline{\mathrm{BS}}(t; P) - \underline{\mathrm{BS}}(t; P) = \mathbb{E}_P\big[\mathbf{1}(Y \leq t, \Delta = 0)\,\big|1 - 2\widehat{S}(t \mid X)\big|\big].$$

See proof on page 34.

The width of these bounds depends not only on the proportion of individuals censored by t but also on how far their predicted survival probabilities are from $1/2$. By Proposition 2.1(iii), each individual censored by t contributes $|1 - 2\widehat{S}(t \mid X)|$, the absolute difference between its two candidate losses, to the bound width. An individual with $\widehat{S}(t \mid X) = 1/2$ incurs the loss $1/4$ whatever its survival status and therefore does not widen the bounds. In contrast, when the predicted survival probability is close to 0 or 1, the two candidate losses are close to 0 and 1, so their difference is close to 1. Thus, predictions farther from $1/2$ for individuals censored by t contribute more to the width of the Brier score bounds.

We now consider bounds on the IBS over $[0, \tau]$. Integrating the pointwise Brier score bounds over this period yields an outer set for the identified set of the IBS. Since the inequalities in Proposition 2.1(i) hold for every t and every $Q \in \mathcal{Q}(P)$, integration gives

$$\frac{1}{\tau}\int_0^\tau \underline{\mathrm{BS}}(t; P)\, dt \leq \mathrm{IBS}(\tau; Q) \leq \frac{1}{\tau}\int_0^\tau \overline{\mathrm{BS}}(t; P)\, dt. \tag{3}$$

Even when the pointwise bounds are sharp, the integrated upper bound need not be sharp. At each time, the endpoints in (3) choose for each censored individual the survival status that gives the smaller or the larger of the two candidate losses. Under a single event time, the survival status $\mathbf{1}(T > t)$ changes from one to zero at most once as t increases and never returns to one. Since $\widehat{S}(t \mid x)$ is nonincreasing in t, the loss of being event-free, $\{1 - \widehat{S}(t \mid x)\}^2$, grows relative to the loss of having experienced the event, $\widehat{S}(t \mid x)^2$, as t increases. The ordering of the two losses therefore changes at most once, and so does each pointwise choice of survival status. For the lower bound, the chosen status changes from one to zero, which a single event time produces or approaches. For the upper bound, it increases from zero to one: the individual has experienced the event at earlier times and is event-free at later times. No event time produces this sequence when both statuses are chosen on periods of positive length after censoring. Therefore, the integrated lower bound is sharp, whereas the sharp upper bound must be obtained by choosing a single event time for each censored individual.

The observed data determine each individual's contribution to the IBS except for individuals censored before τ. For individuals with $Y \geq \tau$ or $\Delta = 1$, this contribution is $L(\min(Y, \tau), X; \tau)/\tau$, with L defined in (2). For individuals with $Y < \tau$ and $\Delta = 0$, it can be $L(r, X; \tau)/\tau$ for any $r \in (Y, \tau]$, where $r = \min(T, \tau)$. The sharp bounds therefore require the smallest and largest values of $L(r, X; \tau)$ over $r \in (Y, \tau]$ for each censored individual.

These smallest and largest values are determined by where the predicted survival probability crosses $1/2$. At time t, the two candidate losses differ by $\{1-\widehat{S}(t \mid x)\}^2 - \widehat{S}(t \mid x)^2 = 1 - 2\widehat{S}(t \mid x)$, so being event-free gives the smaller loss exactly when $\widehat{S}(t \mid x) > 1/2$. For $0 \leq y < \tau$ and covariate value x, the first time after y at which the predicted survival probability falls to $1/2$ is

$$r^*(y, x) = \inf\big\{s \in [y, \tau] : \widehat{S}(s \mid x) \leq 1/2\big\},$$

with $r^*(y, x) = \tau$ if this set is empty. An event at $r^*(Y, X)$ gives the smaller loss at every time after censoring, so the loss of a censored individual is smallest at $r = r^*(Y, X)$. The loss $L(r, X; \tau)$ increases in r at the rate $1 - 2\widehat{S}(r \mid X)$, which is nondecreasing, so it is convex in r and largest as $r \downarrow Y$ or at $r = \tau$. Averaging these smallest and largest losses with the known contributions yields the following sharp bounds.

Proposition 2.2. Define

$$\begin{aligned}\underline{\mathrm{IBS}}(\tau; P) = \frac{1}{\tau}\mathbb{E}_P\Big[&\mathbf{1}(Y \geq \tau \text{ or } \Delta = 1)\, L(\min(Y, \tau), X; \tau) \\ &+ \mathbf{1}(Y < \tau, \Delta = 0)\, L\big(r^*(Y, X), X; \tau\big)\Big],\end{aligned}$$

$$\overline{\mathrm{IBS}}(\tau; P) = \frac{1}{\tau}\mathbb{E}_P\Big[\mathbf{1}(Y \geq \tau \text{ or } \Delta = 1)\, L(\min(Y, \tau), X; \tau) + \mathbf{1}(Y < \tau, \Delta = 0) \max\big\{L(Y, X; \tau), L(\tau, X; \tau)\big\}\Big],$$

and, for $0 \leq y < \tau$ and x,

$$g(y, x) = \min\Big\{\int_y^\tau \big\{2\widehat{S}(t \mid x) - 1\big\}_+ \, dt,\ \int_y^\tau \big\{1 - 2\widehat{S}(t \mid x)\big\}_+ \, dt\Big\},$$

where $a_+ = \max\{a, 0\}$. Then the following hold.

(i) *Outer set.* The interval $[\underline{\mathrm{IBS}}(\tau; P), \overline{\mathrm{IBS}}(\tau; P)]$ is an outer set for the identified set of $\mathrm{IBS}(\tau; Q)$; that is,

$$\underline{\mathrm{IBS}}(\tau; P) \leq \mathrm{IBS}(\tau; Q) \leq \overline{\mathrm{IBS}}(\tau; P) \quad \text{for every } Q \in \mathcal{Q}(P).$$

(ii) *Sharpness.* $\underline{\mathrm{IBS}}(\tau; P)$ and $\overline{\mathrm{IBS}}(\tau; P)$ are the sharp lower and upper bounds on $\mathrm{IBS}(\tau; Q)$, but they need not be attained. That is, the closure of the identified set of $\mathrm{IBS}(\tau; Q)$ is $[\underline{\mathrm{IBS}}(\tau; P), \overline{\mathrm{IBS}}(\tau; P)]$.

(iii) *Width.* The width of the bounds is

$$\overline{\mathrm{IBS}}(\tau; P) - \underline{\mathrm{IBS}}(\tau; P) = \frac{1}{\tau}\mathbb{E}_P\Big[\mathbf{1}(Y < \tau, \Delta = 0) \times \Big\{\max\big\{L(Y, X; \tau), L(\tau, X; \tau)\big\} - L\big(r^*(Y, X), X; \tau\big)\Big\}\Big].$$

(iv) *Comparison with the integrated pointwise bounds.*

$$\frac{1}{\tau}\int_0^\tau \underline{\mathrm{BS}}(t; P)\, dt = \underline{\mathrm{IBS}}(\tau; P) \leq \overline{\mathrm{IBS}}(\tau; P) \leq \frac{1}{\tau}\int_0^\tau \overline{\mathrm{BS}}(t; P)\, dt,$$

and the difference in the last inequality is

$$\frac{1}{\tau}\int_0^\tau \overline{\mathrm{BS}}(t; P)\, dt - \overline{\mathrm{IBS}}(\tau; P) = \frac{1}{\tau}\mathbb{E}_P\big[\mathbf{1}(Y < \tau, \Delta = 0)\, g(Y, X)\big].$$

In particular, the last inequality is strict if and only if $P\big(Y < \tau,\ \Delta = 0,\ g(Y, X) > 0\big) > 0$.

See proof on page 35.

By part (iii), only individuals censored before τ can widen the IBS bounds. Each such individual's contribution to the width is $1/\tau$ times the difference between its largest loss, as $r \downarrow Y$ or at $r = \tau$, and its smallest loss, at $r = r^*(Y, X)$. The lower bound thus corresponds to an event at the time the predicted survival probability falls to $1/2$, when the prediction is most consistent with the realized survival status at every time. The upper bound corresponds to an event immediately after censoring or no event before τ, whichever makes the prediction worse. Individuals with observed events or censoring at or after τ contribute the same known loss to both bounds. Thus the sharp bounds on the IBS place each censored individual's event time at one of these points.

2.3 Comparison of two models

To compare two models j and k, we take as the target quantity the difference in their IBSs,

$$\mathrm{IBS}_{jk}(\tau; Q) = \mathrm{IBS}_j(\tau; Q) - \mathrm{IBS}_k(\tau; Q).$$

Since a smaller IBS indicates a smaller prediction error, $\mathrm{IBS}_{jk}(\tau; Q) < 0$ means that model j predicts better than model k under Q. Exchanging the two models reverses the sign, $\mathrm{IBS}_{kj}(\tau; Q) = -\mathrm{IBS}_{jk}(\tau; Q)$. The identified set of this difference is defined as in Section 2.1, with $\theta(Q) = \mathrm{IBS}_{jk}(\tau; Q)$. The ranking of the two models is therefore determined by where this identified set lies relative to zero.

Combining the bounds on each model's IBS in Proposition 2.2 yields an outer set for the identified set of the difference, but this outer set is generally not sharp. Writing $\underline{\mathrm{IBS}}_j(\tau; P)$ and $\overline{\mathrm{IBS}}_j(\tau; P)$ for the bounds on the IBS of model j, every $Q \in \mathcal{Q}(P)$ satisfies

$$\underline{\mathrm{IBS}}_j(\tau; P) - \overline{\mathrm{IBS}}_k(\tau; P) \le \mathrm{IBS}_{jk}(\tau; Q) \le \overline{\mathrm{IBS}}_j(\tau; P) - \underline{\mathrm{IBS}}_k(\tau; P). \tag{4}$$

The lower end of this interval, however, minimizes the IBS of model j and maximizes that of model k with separate event times, whereas each individual has a single event time shared by both models. In the extreme case in which the two models output the same predictions, $\mathrm{IBS}_{jk}(\tau; Q) = 0$ for every Q, so the identified set is $\{0\}$. The interval in (4) nevertheless has width $2\{\overline{\mathrm{IBS}}_j(\tau; P) - \underline{\mathrm{IBS}}_j(\tau; P)\}$, which is positive whenever the IBS of model j is not point identified. Hence, the sharp bounds must be obtained by minimizing and maximizing the difference in losses itself under a single event time common to both models.

To do so, we assign one value of r to each censored individual and use it for both models. Let

$$D_{jk}(r, x) = L_j(r, x; \tau) - L_k(r, x; \tau), \qquad 0 \le r \le \tau,$$

denote the difference in the losses of the two models for an individual whose event time truncated at τ is r and whose covariates are x. By the representation in Section 2.1, $\mathrm{IBS}_{jk}(\tau; Q) = \tau^{-1}\mathbb{E}_Q D_{jk}(\min(T, \tau), X)$. As in Section 2.2, the data determine $\min(T, \tau) = \min(Y, \tau)$ unless the individual is censored before τ, in which case $\min(T, \tau)$ can take any value in $(Y, \tau]$. The sharp bounds therefore require the smallest and largest values of $D_{jk}(r, X)$ over $r \in (Y, \tau]$ for each censored individual.

The smallest and largest values depend on where the two predicted survival curves cross. As in the proof of Proposition 2.2, the definition of L_j in (2) directly gives the following identity:

$$L_j(r, x; \tau) - L_j(y, x; \tau) = \int_y^r \left\{1 - 2\widehat{S}_j(t \mid x)\right\} dt, \qquad 0 \le y \le r \le \tau.$$

Taking the difference between the two models gives

$$D_{jk}(r, x) = D_{jk}(y, x) + 2\int_y^r \left\{\widehat{S}_k(t \mid x) - \widehat{S}_j(t \mid x)\right\} dt, \qquad 0 \le y \le r \le \tau. \tag{5}$$

Hence $D_{jk}(r, X)$ decreases in r where model j predicts a higher survival probability than model k, and increases where it predicts a lower one. For example, if $\widehat{S}_j(t \mid X) \ge \widehat{S}_k(t \mid X)$ for all $t \in [Y, \tau]$, the infimum over $r \in (Y, \tau]$ is attained at $r = \tau$, and the supremum is approached

as $r \downarrow Y$. In general, the infimum and supremum for a censored individual are attained or approached at $r \downarrow Y$, at $r = \tau$, or at a point where $\widehat{S}_j(t \mid X) - \widehat{S}_k(t \mid X)$ changes sign. Unlike the loss of a single model, $D_{jk}(r, X)$ need not be convex in r, so these extremes do not reduce to a closed form. We therefore define the individual contributions to the lower and upper bounds by

$$\underline{d}_{jk}(O) = \frac{1}{\tau}\Big\{\mathbf{1}(Y \geq \tau \text{ or } \Delta = 1)\, D_{jk}(\min(Y,\tau), X) + \mathbf{1}(Y < \tau, \Delta = 0) \inf_{r \in (Y,\tau]} D_{jk}(r, X)\Big\},$$
$$\overline{d}_{jk}(O) = \frac{1}{\tau}\Big\{\mathbf{1}(Y \geq \tau \text{ or } \Delta = 1)\, D_{jk}(\min(Y,\tau), X) + \mathbf{1}(Y < \tau, \Delta = 0) \sup_{r \in (Y,\tau]} D_{jk}(r, X)\Big\},$$

and set $\underline{\mathrm{IBS}}_{jk}(\tau; P) = \mathbb{E}_P\, \underline{d}_{jk}(O)$ and $\overline{\mathrm{IBS}}_{jk}(\tau; P) = \mathbb{E}_P\, \overline{d}_{jk}(O)$. Since $D_{kj} = -D_{jk}$, exchanging the two models gives $\overline{\mathrm{IBS}}_{kj}(\tau; P) = -\underline{\mathrm{IBS}}_{jk}(\tau; P)$.

Proposition 2.3. For models j and k, the following hold.

(i) *Outer set.* The interval $[\underline{\mathrm{IBS}}_{jk}(\tau; P), \overline{\mathrm{IBS}}_{jk}(\tau; P)]$ is an outer set for the identified set of $\mathrm{IBS}_{jk}(\tau; Q)$; that is,

$$\underline{\mathrm{IBS}}_{jk}(\tau; P) \leq \mathrm{IBS}_{jk}(\tau; Q) \leq \overline{\mathrm{IBS}}_{jk}(\tau; P) \quad \text{for every } Q \in \mathcal{Q}(P).$$

(ii) *Sharpness.* The lower bound $\underline{\mathrm{IBS}}_{jk}(\tau; P)$ and the upper bound $\overline{\mathrm{IBS}}_{jk}(\tau; P)$ on $\mathrm{IBS}_{jk}(\tau; Q)$ are sharp, but they need not be attained. That is, the closure of the identified set of $\mathrm{IBS}_{jk}(\tau; Q)$ is

$$\big[\underline{\mathrm{IBS}}_{jk}(\tau; P), \overline{\mathrm{IBS}}_{jk}(\tau; P)\big].$$

(iii) *Width.* The width of the bounds is

$$\overline{\mathrm{IBS}}_{jk}(\tau; P) - \underline{\mathrm{IBS}}_{jk}(\tau; P) = \frac{1}{\tau}\mathbb{E}_P\Big[\mathbf{1}(Y < \tau, \Delta = 0)\Big\{\sup_{r \in (Y,\tau]} D_{jk}(r, X) - \inf_{r \in (Y,\tau]} D_{jk}(r, X)\Big\}\Big].$$

(iv) *Comparison with the combined bounds.*

$$\underline{\mathrm{IBS}}_{j}(\tau; P) - \overline{\mathrm{IBS}}_{k}(\tau; P) \leq \underline{\mathrm{IBS}}_{jk}(\tau; P) \leq \overline{\mathrm{IBS}}_{jk}(\tau; P) \leq \overline{\mathrm{IBS}}_{j}(\tau; P) - \underline{\mathrm{IBS}}_{k}(\tau; P).$$

See proof on page 38.

If the upper bound on the IBS difference is negative, model j has a smaller IBS than model k under every full-data distribution compatible with the observed data. We give this situation a name.

Definition 2.4 (Robust strict dominance)**.** We say that model j robustly strictly dominates model k if

$$\overline{\mathrm{IBS}}_{jk}(\tau; P) < 0.$$

By Proposition 2.3(i), robust strict dominance implies $\mathrm{IBS}_{jk}(\tau; Q) \leq \overline{\mathrm{IBS}}_{jk}(\tau; P) < 0$ for every $Q \in \mathcal{Q}(P)$, so model j is better than model k by at least $-\overline{\mathrm{IBS}}_{jk}(\tau; P)$ regardless of Q. It is sufficient but not necessary for model j to be strictly better under every Q, because the upper bound may equal zero without being attained. If instead $\underline{\mathrm{IBS}}_{jk}(\tau; P) < 0 < \overline{\mathrm{IBS}}_{jk}(\tau; P)$, Proposition 2.3(ii) implies that some compatible full-data distributions favor model j and others

favor model k. In that case, the observed data alone cannot determine which model predicts better.

2.4 Comparison of multiple models

When $K \geq 3$ models are compared, whether model j is the best is determined by the difference between its IBS and that of its best competitor,

$$M_j(\tau; Q) = \max_{k \neq j} \mathrm{IBS}_{jk}(\tau; Q) = \mathrm{IBS}_j(\tau; Q) - \min_{k \neq j} \mathrm{IBS}_k(\tau; Q).$$

Since a smaller IBS is better, $M_j(\tau; Q) < 0$ means that model j predicts strictly better than every other model under Q. The identified set of $M_j(\tau; Q)$ is defined as in Section 2.1, with $\theta(Q) = M_j(\tau; Q)$. If this identified set lies below zero, model j is the best under every full-data distribution compatible with the observed data.

The sharp upper bound on $M_j(\tau; Q)$ equals the largest of the upper bounds on the pairwise differences. To make the maximum of these differences large, it suffices to make one of them large. The supremum over compatible full-data distributions can therefore be interchanged with the maximum over the finitely many competitors. The following proposition expresses the resulting upper bound using the pairwise bounds from Section 2.3.

Proposition 2.5. For a model j, define

$$\overline{M}_j(\tau; P) = \max_{k \neq j} \overline{\mathrm{IBS}}_{jk}(\tau; P).$$

Then $M_j(\tau; Q) \leq \overline{M}_j(\tau; P)$ for every $Q \in \mathcal{Q}(P)$, and this upper bound is sharp:

$$\sup_{Q \in \mathcal{Q}(P)} M_j(\tau; Q) = \overline{M}_j(\tau; P).$$

See proof on page 39.

If the upper bound on $M_j(\tau; Q)$ is negative, model j is the best under every full-data distribution compatible with the observed data. We give this situation a name.

Definition 2.6 (Robustly strictly best model)**.** We say that model j is robustly strictly best among the K models if

$$\overline{M}_j(\tau; P) < 0.$$

By Proposition 2.5, model j is robustly strictly best among the K models if and only if it robustly strictly dominates every other model in the sense of Definition 2.4. As in the two-model case, this condition is sufficient but not necessary for model j to be strictly best under every Q, because the upper bound may equal zero without being attained. If $\overline{M}_j(\tau; P) > 0$, sharpness implies that some compatible full-data distribution gives model j a larger IBS than at least one competitor. Thus certification of a robustly strictly best model requires only the upper bounds on the pairwise differences.

3 Estimation and inference

This section develops estimation and inference for the sharp bounds derived in Section 2: both bounds for a single model and for a pairwise difference, and the upper bound for the difference

from the best competitor. Section 3.1 defines estimators of the bounds as sample averages over the evaluation sample, and Section 3.2 states the assumptions and establishes their consistency and asymptotic distributions for a fixed observed-data distribution. Section 3.3 then constructs confidence regions for a single model and a pairwise difference, together with tests of robust strict dominance and of whether a model is robustly strictly best among the K models.

3.1 Estimators

For fitted prediction models, we estimate the sharp bounds in Section 2 using the evaluation sample $O_1, \ldots, O_n$. For a single model and for a pairwise difference, each bound is the expectation of an individual contribution, so we replace that expectation with the corresponding sample average. For the difference from the best competitor, we estimate the upper bound by the largest estimated pairwise upper bound.

For a single model, we estimate the bounds on the IBS by the sample averages of the individual contributions in Proposition 2.2:

$$\widehat{\underline{\mathrm{IBS}}}(\tau) = \frac{1}{n\tau}\sum_{i=1}^{n}\Big\{\mathbf{1}(Y_i \geq \tau \text{ or } \Delta_i = 1)\, L(\min(Y_i,\tau), X_i; \tau) + \mathbf{1}(Y_i < \tau, \Delta_i = 0)\, L\big(r^*(Y_i, X_i), X_i; \tau\big)\Big\},$$

$$\widehat{\overline{\mathrm{IBS}}}(\tau) = \frac{1}{n\tau}\sum_{i=1}^{n}\Big\{\mathbf{1}(Y_i \geq \tau \text{ or } \Delta_i = 1)\, L(\min(Y_i,\tau), X_i; \tau) + \mathbf{1}(Y_i < \tau, \Delta_i = 0)\max\big\{L(Y_i, X_i; \tau), L(\tau, X_i; \tau)\big\}\Big\}.$$

Here $r^*(Y_i, X_i)$ is the infimum of the times $s \in [Y_i, \tau]$ such that $\widehat{S}(s \mid X_i) \leq 1/2$, with value τ if no such time exists, as defined in Section 2.2. By the nonincreasing property in Assumption 2, the lower contribution of a censored individual is expressed using $r^*(Y_i, X_i)$, whereas the upper contribution is the larger of $L(Y_i, X_i; \tau)$ and $L(\tau, X_i; \tau)$. When several models are compared, we write $\widehat{\underline{\mathrm{IBS}}}_j(\tau)$ and $\widehat{\overline{\mathrm{IBS}}}_j(\tau)$ for model j. The estimated interval $[\widehat{\underline{\mathrm{IBS}}}(\tau), \widehat{\overline{\mathrm{IBS}}}(\tau)]$ is an estimate of the closure of the identified set of the IBS.

For two models j and k, we estimate the bounds on the IBS difference by the sample averages of the individual contributions $\underline{d}_{jk}(O_i)$ and $\overline{d}_{jk}(O_i)$ in Section 2.3, rather than by differences between the single-model estimators:

$$\widehat{\underline{\mathrm{IBS}}}_{jk}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\underline{d}_{jk}(O_i), \qquad \widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\overline{d}_{jk}(O_i).$$

Since $\overline{d}_{kj} = -\underline{d}_{jk}$, exchanging the two models gives $\widehat{\overline{\mathrm{IBS}}}_{kj}(\tau) = -\widehat{\underline{\mathrm{IBS}}}_{jk}(\tau)$. The differences $\widehat{\underline{\mathrm{IBS}}}_j(\tau) - \widehat{\overline{\mathrm{IBS}}}_k(\tau)$ and $\widehat{\overline{\mathrm{IBS}}}_j(\tau) - \widehat{\underline{\mathrm{IBS}}}_k(\tau)$ would instead estimate the wider interval in (4). In $\underline{d}_{jk}(O_i)$ and $\overline{d}_{jk}(O_i)$, the same value of r enters the losses of both models. Thus the estimators share the event time of each censored individual between the two models, as the sharp bounds do.

For the difference between model j and its best competitor, we estimate the upper bound by

the largest estimated pairwise upper bound, following Proposition 2.5:

$$\widehat{\overline{M}}_j(\tau) = \max_{k \neq j} \widehat{\overline{\mathrm{IBS}}}_{jk}(\tau).$$

The individual contributions for each pair have already been computed, so this estimator requires only taking their largest sample average. No additional optimization over event times is needed. Since this estimator is a maximum of sample averages rather than a sample average itself, Section 3.2 treats its asymptotic distribution separately.

3.2 Asymptotic properties

We establish the asymptotic properties of the estimators under the following assumptions.

Assumption 1 (Independent evaluation sample)**.** The evaluation sample $O_1, \ldots, O_n$ consists of independent and identically distributed copies of $O = (Y, \Delta, X)$ with a fixed observed-data distribution P that does not depend on n.

Assumption 2 (Predicted survival functions)**.** For each model j, the predicted survival function $\widehat{S}_j(t \mid x)$ is measurable in (t, x) and takes values in $[0, 1]$. For every $x \in \mathcal{X}$, it is nonincreasing in time: $\widehat{S}_j(t_1 \mid x) \geq \widehat{S}_j(t_2 \mid x)$ whenever $0 \leq t_1 \leq t_2 \leq \tau$.

Assumption 3 (Independent model construction)**.** All data used to construct the prediction models, including preprocessing, training, and hyperparameter tuning, are independent of the evaluation sample. The fitted prediction models are held fixed during evaluation.

Assumption 4 (Prespecified comparison)**.** The candidate model j, the set of K compared models, and the evaluation horizon τ are fixed before the evaluation sample is examined.

Assumption 2 formalizes the properties of the predicted survival functions introduced in Section 2.1. Under Assumption 3, inference is conditional on the data used to construct the models, so the fitted prediction functions are treated as fixed. A validation sample may be used for tuning if it is independent of the evaluation sample. The results therefore describe the sampling variability of the evaluation sample, conditional on the fitted models.

For model j, let $\underline{\ell}_j(O)$ and $\overline{\ell}_j(O)$ denote the summands of $\widehat{\underline{\mathrm{IBS}}}_j(\tau)$ and $\widehat{\overline{\mathrm{IBS}}}_j(\tau)$ in Section 3.1, so that, for example, $\widehat{\underline{\mathrm{IBS}}}_j(\tau) = n^{-1} \sum_{i=1}^n \underline{\ell}_j(O_i)$ and $\underline{\mathrm{IBS}}_j(\tau; P) = \mathbb{E}_P\, \underline{\ell}_j(O)$. Since $0 \leq L_j \leq \tau$, the contributions $\underline{\ell}_j$ and $\overline{\ell}_j$ take values in $[0, 1]$, and $\underline{d}_{jk}$ and $\overline{d}_{jk}$ take values in $[-1, 1]$. The law of large numbers and the central limit theorem therefore apply without further moment conditions.

The estimators of the bounds all converge almost surely to the corresponding bounds.

Proposition 3.1 (Strong consistency)**.** Under Assumptions 1, 2, and 3, the following hold as $n \to \infty$.

(i) For each model j, $\widehat{\underline{\mathrm{IBS}}}_j(\tau) \xrightarrow{\text{a.s.}} \underline{\mathrm{IBS}}_j(\tau; P)$ and $\widehat{\overline{\mathrm{IBS}}}_j(\tau) \xrightarrow{\text{a.s.}} \overline{\mathrm{IBS}}_j(\tau; P)$.

(ii) For each pair of models j and k, $\widehat{\underline{\mathrm{IBS}}}_{jk}(\tau) \xrightarrow{\text{a.s.}} \underline{\mathrm{IBS}}_{jk}(\tau; P)$ and $\widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) \xrightarrow{\text{a.s.}} \overline{\mathrm{IBS}}_{jk}(\tau; P)$.

(iii) For each model j, $\widehat{\overline{M}}_j(\tau) \xrightarrow{\text{a.s.}} \overline{M}_j(\tau; P)$.

See proof on page 39. Part (iii) does not require that the competitor attaining the upper bound be unique. The proposition concerns the bounds, not the IBS itself, because the IBS is only

known to lie in its identified set. Increasing the sample size removes the estimation error, but the width of the identified set remains.

The estimators of the bounds for a single model and for two models are sample averages, so the lower and upper estimators are jointly asymptotically normal.

Theorem 3.2 (Joint asymptotic normality)**.** Under Assumptions 1, 2, and 3, the following hold as $n \to \infty$.

(i) For each model j,

$$\sqrt{n}\begin{pmatrix}\widehat{\underline{\mathrm{IBS}}}_j(\tau) - \underline{\mathrm{IBS}}_j(\tau; P)\\ \widehat{\overline{\mathrm{IBS}}}_j(\tau) - \overline{\mathrm{IBS}}_j(\tau; P)\end{pmatrix} \xrightarrow{d} N_2(0, \Sigma_j), \qquad \Sigma_j = \mathrm{Var}_P\begin{pmatrix}\underline{\ell}_j(O)\\ \overline{\ell}_j(O)\end{pmatrix}.$$

(ii) For each pair of models j and k,

$$\sqrt{n}\begin{pmatrix}\widehat{\underline{\mathrm{IBS}}}_{jk}(\tau) - \underline{\mathrm{IBS}}_{jk}(\tau; P)\\ \widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) - \overline{\mathrm{IBS}}_{jk}(\tau; P)\end{pmatrix} \xrightarrow{d} N_2(0, \Sigma_{jk}), \qquad \Sigma_{jk} = \mathrm{Var}_P\begin{pmatrix}\underline{d}_{jk}(O)\\ \overline{d}_{jk}(O)\end{pmatrix}.$$

(iii) The sample covariance matrices of the contributions over $O_1, \ldots, O_n$ converge almost surely to Σ_j and Σ_{jk}.

The covariance matrices Σ_j and Σ_{jk} may be singular, in which case the limiting normal distributions are degenerate.

See proof on page 40. Because the lower and upper estimators are computed from the same individuals, they are generally correlated, and Σ_j and Σ_{jk} keep this correlation. The covariance matrix can be singular; for example, Σ_{jk} is the zero matrix if the two models output the same predictions. The limit is normal regardless of whether the value of r attaining the infimum or supremum for each individual is unique, because the optimization over r takes place inside the expectation, individual by individual. Section 3.3 uses this result to construct confidence regions and tests of robust strict dominance.

The estimator of the upper bound on the difference from the best competitor is a maximum of sample averages, so its limiting distribution is generally not normal. Let

$$\mathcal{K}_j^* = \arg\max_{k \neq j} \overline{\mathrm{IBS}}_{jk}(\tau; P)$$

be the set of competitors attaining the upper bound in Proposition 2.5. This set is nonempty because the number of competitors is finite.

Theorem 3.3 (Asymptotic distribution of the upper bound for the best competitor)**.** Under Assumptions 1, 2, and 3, for each model j, the vector

$$\sqrt{n}\big\{\widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) - \overline{\mathrm{IBS}}_{jk}(\tau; P)\big\}_{k \neq j}$$

converges in distribution to a mean-zero normal vector $Z_j = (Z_{jk})_{k \neq j}$ with

$$\mathrm{Cov}(Z_{jk}, Z_{jk'}) = \mathrm{Cov}_P\{\overline{d}_{jk}(O), \overline{d}_{jk'}(O)\}.$$

Moreover,

$$\sqrt{n}\{\widehat{\overline{M}}_j(\tau) - \overline{M}_j(\tau; P)\} \xrightarrow{d} \max_{k \in \mathcal{K}_j^*} Z_{jk}.$$

See proof on page 40. If $\mathcal{K}_j^*$ contains several competitors, the limit is the maximum of several correlated normal variables and is not normal in general. If $\mathcal{K}_j^* = \{k^*\}$ is a singleton, the limit reduces to the normal variable Z_{jk^*}. Section 3.3 constructs a test and a one-sided upper confidence bound from the pairwise upper bounds without directly approximating the distribution of their maximum.

3.3 Confidence regions and tests

An estimated identified set reflects two kinds of uncertainty: the width of the identified set, which remains because some event times are unobserved, and the sampling variability from estimating its bounds, and a confidence region adds the latter to the former. For the IBS of one model and the difference between two models, we construct confidence regions that contain the entire closure of the identified set with asymptotic probability $1 - \alpha$ (Chernozhukov et al., 2007). A confidence interval designed to cover the true value of the target, rather than the entire identified set, with asymptotic probability at least $1 - \alpha$ is generally narrower (Imbens & Manski, 2004; Romano & Shaikh, 2008; Stoye, 2009). We take the identified set itself as the object of inference, because it summarizes what the observed data reveal about the target without restricting the dependence between event and censoring times. A region that covers the identified set is the sample counterpart of this object, whereas an interval for the true value is calibrated to cover a single unknown point within the set.

Through the normal approximation to the bound estimators, we construct confidence regions for the IBS of a single model and the IBS difference between two models by widening the estimated bounds by a common multiple of their standard errors. Let θ denote either $\mathrm{IBS}_j(\tau; Q)$ or $\mathrm{IBS}_{jk}(\tau; Q)$, and let $\underline{\theta}(P)$ and $\overline{\theta}(P)$ be its sharp bounds, estimated by $\widehat{\underline{\theta}}$ and $\widehat{\overline{\theta}}$ as in Section 3.1. Let $\underline{\sigma}^2$ and $\overline{\sigma}^2$ be the variances of the corresponding individual contributions, ρ their correlation when both are positive, and $\widehat{\underline{\sigma}}$, $\widehat{\overline{\sigma}}$, and $\widehat{\rho}$ the sample counterparts. The region extends $c\,\widehat{\underline{\sigma}}/\sqrt{n}$ below the estimated lower bound and $c\,\widehat{\overline{\sigma}}/\sqrt{n}$ above the estimated upper bound, where the critical value c is common to the two endpoints. The joint asymptotic normality in Theorem 3.2 determines how large c must be.

The critical value is chosen so that the region covers both bounds at once. The region contains the lower bound $\underline{\theta}(P)$ if and only if

$$\widehat{\underline{\theta}} - \frac{c\,\widehat{\underline{\sigma}}}{\sqrt{n}} \le \underline{\theta}(P) \iff \frac{\sqrt{n}(\widehat{\underline{\theta}} - \underline{\theta}(P))}{\widehat{\underline{\sigma}}} \le c,$$

and it contains the upper bound $\overline{\theta}(P)$ if and only if

$$\widehat{\overline{\theta}} + \frac{c\,\widehat{\overline{\sigma}}}{\sqrt{n}} \ge \overline{\theta}(P) \iff -\frac{\sqrt{n}(\widehat{\overline{\theta}} - \overline{\theta}(P))}{\widehat{\overline{\sigma}}} \le c.$$

By Theorem 3.2, its coverage probability is therefore approximately $\Pr(\max\{Z_1, -Z_2\} \le c)$, where Z_1 and Z_2 are the limits of $\sqrt{n}(\widehat{\underline{\theta}} - \underline{\theta}(P))/\underline{\sigma}$ and $\sqrt{n}(\widehat{\overline{\theta}} - \overline{\theta}(P))/\overline{\sigma}$, respectively, and have correlation ρ.

Accordingly, for $\rho \in [-1, 1]$, let (Z_1, Z_2) be a bivariate normal vector with standard normal margins and correlation ρ, and define the critical value

$$c_{1-\alpha}(\rho) = \inf\{c : \Pr(\max\{Z_1, -Z_2\} \leq c) \geq 1 - \alpha\}.$$

When $\widehat{\underline{\sigma}} > 0$ and $\widehat{\overline{\sigma}} > 0$, we define

$$\mathcal{C}_{1-\alpha} = \Big[\widehat{\underline{\theta}} - \frac{c_{1-\alpha}(\widehat{\rho})\,\widehat{\underline{\sigma}}}{\sqrt{n}},\ \widehat{\overline{\theta}} + \frac{c_{1-\alpha}(\widehat{\rho})\,\widehat{\overline{\sigma}}}{\sqrt{n}}\Big],$$

and otherwise we set $\mathcal{C}_{1-\alpha} = \mathbb{R}$.

Proposition 3.4 (Confidence region for one or two models)**.** Suppose that Assumptions 1, 2, and 3 hold and that $\underline{\sigma} > 0$ and $\overline{\sigma} > 0$. Then

$$\Pr\big\{[\underline{\theta}(P), \overline{\theta}(P)] \subseteq \mathcal{C}_{1-\alpha}\big\} \to 1 - \alpha \qquad \text{as } n \to \infty.$$

See proof on page 40. The critical value $c_{1-\alpha}(\widehat{\rho})$ uses the correlation between the lower and upper estimators. A simpler alternative is to replace it with the standard normal quantile $z_{1-\alpha/2}$, which by the Bonferroni inequality gives asymptotic coverage of at least $1 - \alpha$ but can be conservative. If a fixed benchmark value lies outside $\mathcal{C}_{1-\alpha}$, we conclude, at asymptotic level α, that it does not belong to the closure of the identified set.

The tests below concern comparisons fixed before the evaluation sample is examined, as specified in Assumption 4.

Whether model j robustly strictly dominates model k is tested by checking whether a one-sided upper confidence bound for the upper bound on the IBS difference is negative. By Definition 2.4, the hypotheses are

$$H_0 : \overline{\mathrm{IBS}}_{jk}(\tau; P) \geq 0 \qquad \text{versus} \qquad H_1 : \overline{\mathrm{IBS}}_{jk}(\tau; P) < 0.$$

Let $\overline{\sigma}^2_{jk} = \mathrm{Var}_P\{\overline{d}_{jk}(O)\}$, let $\widehat{\overline{\sigma}}_{jk}$ be the sample standard deviation of $\overline{d}_{jk}(O_1), \ldots, \overline{d}_{jk}(O_n)$, and let $z_{1-\alpha}$ be the $(1 - \alpha)$ quantile of the standard normal distribution with $0 < \alpha < 1/2$. Define the one-sided upper confidence bound

$$U_{jk} = \widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) + z_{1-\alpha}\frac{\widehat{\overline{\sigma}}_{jk}}{\sqrt{n}},$$

and reject H_0 if $U_{jk} < 0$.

Proposition 3.5 (Test of robust strict dominance)**.** Under Assumptions 1, 2, 3, and 4, the following hold for each fixed P.

(i) *Level.* If $\overline{\mathrm{IBS}}_{jk}(\tau; P) \geq 0$, then $\limsup_{n\to\infty} \Pr\{U_{jk} < 0\} \leq \alpha$.

(ii) *Consistency.* If $\overline{\mathrm{IBS}}_{jk}(\tau; P) < 0$, then $\Pr\{U_{jk} < 0\} \to 1$.

(iii) *Upper confidence bound.* $\liminf_{n\to\infty} \Pr\{\overline{\mathrm{IBS}}_{jk}(\tau; P) \leq U_{jk}\} \geq 1 - \alpha$.

See proof on page 41. At the boundary $\overline{\mathrm{IBS}}_{jk}(\tau; P) = 0$, the rejection probability converges to α if $\overline{\sigma}_{jk} > 0$ and equals zero if $\overline{\sigma}_{jk} = 0$, which occurs, for example, when the two models output the same predictions. Testing whether model k robustly strictly dominates model j amounts to

rejecting when $\widehat{\underline{\mathrm{IBS}}}_{jk}(\tau) - z_{1-\alpha}\widehat{\underline{\sigma}}_{jk}/\sqrt{n} > 0$, since $\overline{d}_{kj} = -\underline{d}_{jk}$; here $\widehat{\underline{\sigma}}_{jk}$ is the sample standard deviation of $\underline{d}_{jk}(O_i)$. Because $\widehat{\underline{\mathrm{IBS}}}_{jk}(\tau) \leq \widehat{\overline{\mathrm{IBS}}}_{jk}(\tau)$ and $z_{1-\alpha} > 0$, the tests in the two directions never reject simultaneously. If H_0 is rejected, we conclude that model j predicts better than model k under every full-data distribution compatible with the observed data, whereas a failure to reject does not imply that model k is at least as good as model j.

Whether model j is robustly strictly best among the K models is tested by an intersection–union test (IUT), which rejects only if the test in Proposition 3.5 rejects against every competitor. By Definition 2.6 and Proposition 2.5, $\overline{M}_j(\tau; P) = \max_{k \neq j} \overline{\mathrm{IBS}}_{jk}(\tau; P)$, so the hypotheses are

$$H_0 : \bigcup_{k \neq j} \{\overline{\mathrm{IBS}}_{jk}(\tau; P) \geq 0\} \qquad \text{versus} \qquad H_1 : \bigcap_{k \neq j} \{\overline{\mathrm{IBS}}_{jk}(\tau; P) < 0\}.$$

The null hypothesis is a union of the pairwise null hypotheses, and the alternative is their intersection (Berger, 1982; Berger & Hsu, 1996). We reject H_0 if

$$U_j^{\mathrm{IUT}} = \max_{k \neq j} U_{jk} < 0,$$

where each U_{jk} is computed at level α without adjusting for the number of competitors. The comparison is fixed in advance by Assumption 4.

Proposition 3.6 (Intersection–union test for a robustly strictly best model)**.** Under Assumptions 1, 2, 3, and 4, the following hold for each fixed P.

(i) *Level.* If $\overline{M}_j(\tau; P) \geq 0$, then $\limsup_{n \to \infty} \Pr\{U_j^{\mathrm{IUT}} < 0\} \leq \alpha$.

(ii) *Consistency.* If $\overline{M}_j(\tau; P) < 0$, then $\Pr\{U_j^{\mathrm{IUT}} < 0\} \to 1$.

(iii) *Upper confidence bound.* $\liminf_{n \to \infty} \Pr\{\overline{M}_j(\tau; P) \leq U_j^{\mathrm{IUT}}\} \geq 1 - \alpha$.

See proof on page 41. Assumption 4 rules out choosing the candidate model or the evaluation horizon after examining the evaluation sample, which would create a multiplicity problem and could inflate the probability of a false rejection. The IUT does not require estimating $\mathcal{K}_j^*$ or identifying a unique maximizing competitor. It can be conservative, however, because when several competitors lie on the boundary of the null hypothesis, all of their pairwise tests must reject. This test and its one-sided upper confidence bound are thus constructed from the asymptotic normality of the pairwise upper bounds alone.

The coverage and level guarantees above are asymptotic and hold for each fixed observed-data distribution. They are not uniform over classes of observed-data distributions: for example, if a few individuals have large positive contributions and the rest have small negative ones, an evaluation sample that contains none of the former yields a negative estimate with a small sample standard deviation, and the tests can then reject falsely. The phrase "every full-data distribution compatible with the observed data" in Definitions 2.4 and 2.6 addresses the identification problem for a single P. A uniform level, by contrast, would address the sampling problem over many possible P, which is a different issue. This paper does not consider tests with uniform guarantees.

4 Simulation study

This section examines how the sharp bounds in Section 2 and the inference procedures in Section 3 behave in finite samples when the dependence between event and censoring times is controlled. Section 4.1 describes the data-generating process and the conditions shared by all experiments. Sections 4.2–4.4 then evaluate a single model, two models, and three models.

4.1 Simulation design

Because the simulations generate the latent event times, we can compare the full-data IBS numerically with the bounds that delimit the closure of the identified set. As in Sections 2.1 and 3.1, every prediction rule is fixed independently of the evaluation sample, and no model is trained within the replications. This separates the estimation error of the prediction models from the identification problem caused by censoring. We approximate the population bounds and the full-data IBS using an independent sample of 500,000 individuals and use these approximations as reference values. The censoring distribution and the shifts of the competing models in Sections 4.3 and 4.4 are calibrated on separate independent samples of 1,000,000 individuals. On the reference sample, the censoring proportions and the calibrated bounds were within about 0.001 of their targets. Hence the differences in the results below arise from dependent censoring and partial identification, not from how the models are trained.

We use the copula-based construction of Gharari et al. (2023): a pair of uniform random variables is transformed by Weibull inverse survival functions to generate dependent event and censoring times. Our design uses a Gaussian copula in place of their Clayton and Frank copulas, three covariates in place of ten, and a censoring distribution that does not depend on the covariates. Our dependence grid also includes negative values of Kendall's tau. The covariates $X = (X_1, X_2, X_3)$ are mutually independent, with $X_1, X_3 \sim N(0,1)$ and $X_2 \sim \text{Bernoulli}(0.5)$. The covariate X_3 affects neither the event time nor the censoring time; Section 4.4 uses it only to construct competing models that err in opposite directions. Given X, the event time follows the Weibull proportional hazards model

$$S_0(t \mid X) = \exp\big[-\lambda_T t^{\kappa_T} \exp\{\eta(X)\}\big], \qquad \eta(X) = 0.8X_1 + 0.5X_2, \qquad \lambda_T = 0.08, \qquad \kappa_T = 1.4,$$

and the censoring time has the survival function $S_C(c) = \exp(-\lambda_C c^{1.2})$, which does not depend on X. We draw (U_T, U_C) from a Gaussian copula and set $T = S_0^{-1}(U_T \mid X)$ and $C = S_C^{-1}(U_C)$. We specify the strength of dependence by Kendall's tau τ_C of the copula, which determines the copula correlation parameter $\zeta = \sin(\pi\tau_C/2)$ (Nelsen, 2006). The value $\tau_C = 0$ gives $C \perp (T, X)$, since the censoring distribution does not depend on X. Positive and negative values give dependent censoring in opposite directions. Varying τ_C leaves the marginal distribution of the event time unchanged; the censoring scale λ_C is recalibrated for each dependence level, as described below.

The width of the identified set depends directly on the proportion of individuals censored before the end of the evaluation period, $\mu_\tau = P(\Delta = 0, Y < \tau)$, so we specify this proportion separately from the strength of dependence τ_C. The horizon τ is the 75th percentile of the marginal distribution of T. All experiments use the same grid of conditions:

$$\tau_C \in \{-0.6, -0.3, 0, 0.3, 0.6\}, \quad \mu_\tau \in \{0.10, 0.30, 0.50\}, \quad n \in \{250, 500, 1000\}.$$

For each pair (τ_C, μ_τ), we calibrate λ_C numerically to target the censoring proportion μ_τ. Increasing n from 250 to 1000 shows how closely the finite-sample behavior approaches the asymptotic theory for a fixed observed-data distribution in Section 3.2. We call each combination of (τ_C, μ_τ, n) a cell and replicate each cell $R = 1{,}000$ times. With this number of replications, an estimated proportion whose true value is 0.05 or 0.95 has a Monte Carlo standard error of about 0.0069. The confidence level is $1 - \alpha = 0.95$, and the tests use the significance level $\alpha = 0.05$. For each of the 15 combinations of dependence and censoring proportion, we thus compare the behavior of the bounds and the inference procedures across sample sizes.

As a benchmark, we use IPCW estimation of the IBS and an IPCW test of the IBS difference, both of which assume independent censoring in the sense $C \perp (T, X)$. The censoring survival function is estimated by the reverse Kaplan–Meier estimator without covariates (Graf et al., 1999; Kaplan & Meier, 1958). Standard errors are estimated by a nonparametric bootstrap that resamples individuals, with 500 resamples in every replication. IPCW estimates a single value of the full-data IBS, whereas the proposed methods target the identified set. The observed coverage and rejection proportions of IPCW are therefore not comparisons of the two approaches on a common criterion. We report the IPCW results to show what conclusions an analysis that ignores dependent censoring would reach.

4.2 Single model

This subsection evaluates the identified set and the confidence region for the IBS when the prediction rule is the true conditional survival function, $\widehat{S} = S_0$, which removes misspecification of the prediction rule as a source of bias. We examine three questions. First, does the interval between the reference bounds contain the reference full-data IBS across the dependence settings? Second, is the IPCW estimator biased under dependent censoring, and if so, does the bias shrink as the sample size grows? Third, is the proportion of replications in which the confidence region in Proposition 3.4 contains both reference bounds close to the nominal level 0.95? The evaluation measures are the reference width of the identified set, the estimated bias of the IPCW estimator, this empirical coverage proportion, and the length of the confidence region.

In all 15 scenarios, the interval between the reference bounds contained the reference full-data IBS (Table 1). Because the event-time distribution and the prediction rule are the same in all scenarios, the full-data IBS was about 0.153 throughout, and only the observed-data distribution changed with τ_C and μ_τ. The width of the identified set was determined mainly by the censoring proportion: it was 0.022–0.024 for $\mu_\tau = 0.10$, 0.069–0.074 for $\mu_\tau = 0.30$, and 0.119–0.139 for $\mu_\tau = 0.50$, and it varied much less across τ_C.

The IPCW estimator, by contrast, was biased under dependent censoring, and the bias did not shrink as the sample size grew. Under independent censoring, $\tau_C = 0$, the bias was at most 0.0002 in absolute value. Under dependent censoring, its sign and size depended on both τ_C and μ_τ. The bias was largest, -0.0106 at $n = 1000$, for $\tau_C = -0.6$ and $\mu_\tau = 0.30$, which is about 15% of the width of the identified set in that scenario, and it stayed within 0.0022 in absolute value for $\tau_C = 0.3$. In every scenario, the estimated bias was almost the same for $n = 250$, 500, and 1000. This pattern is consistent with bias caused by violation of the independent-censoring assumption.

In every cell, $\mathcal{C}_{1-\alpha}$ contained both reference bounds in close to 95% of the 1,000 replications, and its excess length over the identified set shrank as the sample size grew (Table 2). These

Table 1: Population bounds on the IBS of the true survival function S_0 and bias of the IPCW estimator. Lower, Upper, and Width approximate the sharp bounds and their difference, and Full data approximates the full-data IBS; all four are computed from the reference sample of 500,000 individuals in each scenario. IPCW bias is the mean over 1,000 replications of the IPCW estimate minus the full-data IBS.

		Population reference				IPCW bias		
τ_C	μ_τ	Lower	Upper	Width	Full data	$n = 250$	$n = 500$	$n = 1000$
−0.6	0.10	0.1428	0.1646	0.0218	0.1531	−0.0052	−0.0054	−0.0052
−0.3	0.10	0.1459	0.1681	0.0222	0.1533	−0.0022	−0.0024	−0.0026
0.0	0.10	0.1473	0.1702	0.0229	0.1531	−0.0001	0.0000	−0.0001
0.3	0.10	0.1467	0.1699	0.0232	0.1530	−0.0004	−0.0009	−0.0007
0.6	0.10	0.1423	0.1660	0.0237	0.1528	−0.0044	−0.0048	−0.0047
−0.6	0.30	0.1296	0.2010	0.0714	0.1531	−0.0102	−0.0107	−0.0106
−0.3	0.30	0.1336	0.2068	0.0732	0.1531	−0.0036	−0.0039	−0.0040
0.0	0.30	0.1352	0.2092	0.0740	0.1529	0.0001	0.0002	0.0001
0.3	0.30	0.1344	0.2067	0.0724	0.1528	0.0002	0.0001	0.0001
0.6	0.30	0.1297	0.1985	0.0688	0.1531	−0.0063	−0.0058	−0.0057
−0.6	0.50	0.1214	0.2607	0.1393	0.1530	−0.0070	−0.0075	−0.0072
−0.3	0.50	0.1223	0.2604	0.1381	0.1533	−0.0045	−0.0045	−0.0043
0.0	0.50	0.1223	0.2577	0.1354	0.1532	0.0001	0.0000	−0.0001
0.3	0.50	0.1213	0.2503	0.1290	0.1529	0.0022	0.0015	0.0021
0.6	0.50	0.1180	0.2371	0.1192	0.1530	−0.0015	−0.0015	−0.0015

empirical coverage proportions ranged from 0.935 to 0.967 over the 45 cells, and their averages over the 15 scenarios were 0.951, 0.952, and 0.949 for $n = 250$, 500, and 1000. The mean length of the region exceeded the width of the identified set by 0.027–0.032 for $n = 250$ and by 0.013–0.016 for $n = 1000$. As the sample size grows, the part of the length due to sampling variability shrinks, while the width of the identified set remains.

For reference, Table 2 also reports how often the 95% IPCW interval contained the full-data IBS. Under independent censoring, this proportion was 0.931–0.962, close to the nominal level. Under dependent censoring, it fell to as low as 0.117, for $\tau_C = -0.6$, $\mu_\tau = 0.30$, and $n = 1000$. For $\tau_C = -0.6$, -0.3, and 0.6, it decreased from $n = 250$ to $n = 1000$ in every scenario, alongside the estimated bias that changed little over these sample sizes. These proportions concern a single value rather than the identified set, so they do not compare the two approaches on a common criterion. For these three values of τ_C, increasing the sample size over the range examined did not improve the IPCW interval's coverage of the full-data IBS.

4.3 Comparison of two models

We compare two prediction models: model 1 is the true conditional survival function, $\widehat{S}_1 = S_0$, and model 2 adds a positive shift δ to the linear predictor,

$$\widehat{S}_2(t \mid X) = \exp\big[-\lambda_T t^{\kappa_T} \exp\{\eta(X) + \delta\}\big], \qquad \delta > 0.$$

Model 2 predicts a lower survival probability than model 1 for every individual at every time. Because model 1 is the true survival function, it has the smaller IBS under the true full-data distribution. When $\overline{\mathrm{IBS}}_{12}(\tau; P) > 0$, that is, in the interior of the null hypothesis $\overline{\mathrm{IBS}}_{12}(\tau; P) \geq 0$ of the test in Proposition 3.5, some full-data distribution compatible with P gives model 2 the

Table 2: Coverage and length of the 95% confidence regions for the IBS of the true survival function S_0. Coverage is the proportion of 1,000 replications in which $\mathcal{C}_{1-\alpha}$ contained both reference bounds in Table 1, and Mean length is the mean length of $\mathcal{C}_{1-\alpha}$. IPCW coverage is the proportion in which the 95% IPCW interval contained the full-data IBS. Column headings give the sample size n.

		Coverage			Mean length			IPCW coverage		
τ_C	μ_τ	250	500	1000	250	500	1000	250	500	1000
−0.6	0.10	0.938	0.935	0.951	0.0489	0.0407	0.0353	0.837	0.756	0.644
−0.3	0.10	0.942	0.936	0.947	0.0497	0.0418	0.0360	0.923	0.901	0.867
0.0	0.10	0.945	0.947	0.941	0.0511	0.0425	0.0368	0.931	0.946	0.946
0.3	0.10	0.947	0.946	0.953	0.0515	0.0431	0.0374	0.947	0.941	0.950
0.6	0.10	0.942	0.949	0.945	0.0504	0.0427	0.0373	0.880	0.814	0.692
−0.6	0.30	0.959	0.955	0.955	0.1012	0.0926	0.0865	0.644	0.382	0.117
−0.3	0.30	0.955	0.956	0.949	0.1038	0.0946	0.0885	0.909	0.864	0.783
0.0	0.30	0.941	0.947	0.943	0.1046	0.0956	0.0892	0.938	0.945	0.956
0.3	0.30	0.961	0.953	0.947	0.1029	0.0938	0.0876	0.951	0.954	0.936
0.6	0.30	0.947	0.949	0.944	0.0974	0.0888	0.0830	0.847	0.754	0.647
−0.6	0.50	0.962	0.967	0.958	0.1716	0.1626	0.1557	0.843	0.726	0.582
−0.3	0.50	0.959	0.949	0.952	0.1704	0.1607	0.1544	0.898	0.858	0.798
0.0	0.50	0.953	0.963	0.962	0.1672	0.1580	0.1509	0.939	0.953	0.962
0.3	0.50	0.953	0.959	0.944	0.1613	0.1516	0.1445	0.936	0.946	0.926
0.6	0.50	0.962	0.964	0.948	0.1486	0.1407	0.1342	0.932	0.941	0.915

smaller IBS, so the observed data alone cannot establish that model 1 is better.

We choose δ so that the upper bound $\overline{\mathrm{IBS}}_{12}(\tau; P)$, computed on a separate sample of 1,000,000 individuals (Section 4.1), takes each of eight prescribed values, called targets, which place the setting in the interior of the null hypothesis, on its boundary, or in the alternative. As δ increases from zero, the upper bound rises from zero to a positive maximum and then decreases, becoming negative for large δ. The targets are half of this maximum for the interior, zero for the boundary, and -0.0025, -0.005, -0.01, -0.02, -0.04, and -0.06 for the alternative. The eight targets with the same (τ_C, μ_τ, n) share the same evaluation samples, because changing δ changes model 2 but not the data. This design varies the target position relative to the null hypothesis while keeping the observed-data distribution fixed.

For the IBS difference, the proportion of 1,000 replications in which the confidence region contained both reference bounds was also close to the nominal level 0.95 (Table 3). Over the 360 combinations of cell and target level, these empirical coverage proportions ranged from 0.936 to 0.969, and their averages over the 120 combinations for each sample size were 0.951, 0.952, and 0.953. The mean length of $\mathcal{C}_{1-\alpha}$ decreased with n toward the width of the identified set; for $\mu_\tau = 0.50$, it fell from 0.360 at $n = 250$ to 0.325 at $n = 1000$, while the width averaged 0.290. With heavy censoring, most of the length of the region therefore reflects the width of the identified set rather than sampling variability.

At the zero target, the test of robust strict dominance rejected in close to 5% of the replications (Table 3). The rejection proportions ranged from 0.035 to 0.061 over the 45 cells, and their averages over the 15 scenarios were 0.051, 0.049, and 0.048 for $n = 250$, 500, and 1000. These proportions are consistent with the statement after Proposition 3.5 that, at the boundary, the rejection probability converges to α when $\overline{\sigma}_{12} > 0$.

At the positive targets in the interior of the null hypothesis, the test rejected less often than at

Table 3: Empirical coverage for the IBS difference and rejection proportions at the intended null targets. Coverage is the proportion of 1,000 replications in which $\mathcal{C}_{1-\alpha}$ contained both reference bounds on $\mathrm{IBS}_{12}(\tau; Q)$, averaged over all eight target levels. Boundary rejection and Interior rejection are the proportions of 1,000 replications in which the test rejected at $\alpha = 0.05$ when the upper bound was calibrated toward zero and toward half of its largest attainable value, respectively. These labels denote calibration targets rather than verified population null conditions. Column headings give the sample size n.

		Coverage			Boundary rejection			Interior rejection		
τ_C	μ_τ	250	500	1000	250	500	1000	250	500	1000
−0.6	0.10	0.951	0.955	0.952	0.042	0.061	0.054	0.017	0.017	0.007
−0.3	0.10	0.945	0.956	0.955	0.061	0.048	0.040	0.028	0.012	0.006
0.0	0.10	0.953	0.946	0.953	0.054	0.056	0.045	0.025	0.028	0.014
0.3	0.10	0.950	0.943	0.951	0.056	0.044	0.045	0.034	0.024	0.018
0.6	0.10	0.940	0.964	0.938	0.061	0.044	0.057	0.050	0.027	0.031
−0.6	0.30	0.946	0.949	0.962	0.045	0.048	0.040	0.001	0.000	0.000
−0.3	0.30	0.957	0.956	0.967	0.049	0.041	0.035	0.002	0.001	0.000
0.0	0.30	0.948	0.950	0.946	0.049	0.049	0.057	0.009	0.002	0.000
0.3	0.30	0.963	0.944	0.962	0.037	0.051	0.036	0.010	0.002	0.000
0.6	0.30	0.951	0.952	0.941	0.050	0.052	0.047	0.025	0.010	0.003
−0.6	0.50	0.951	0.948	0.956	0.057	0.053	0.055	0.000	0.000	0.000
−0.3	0.50	0.952	0.950	0.955	0.051	0.049	0.053	0.000	0.000	0.000
0.0	0.50	0.948	0.960	0.949	0.046	0.050	0.057	0.001	0.000	0.000
0.3	0.50	0.959	0.951	0.945	0.045	0.044	0.045	0.004	0.000	0.001
0.6	0.50	0.954	0.950	0.959	0.056	0.048	0.052	0.009	0.001	0.000

the zero target, and its average rejection proportion decreased as the sample size grew (Table 3). The rejection proportions were at most 0.050, and their averages over the 15 scenarios were 0.014, 0.008, and 0.005 for $n = 250$, 500, and 1000. The highest proportions occurred for $\mu_\tau = 0.10$, where the interior targets were only 0.0001 to 0.0013 and thus close to zero. These results agree with Proposition 3.5(i), which bounds the asymptotic rejection probability by α whenever the upper bound is nonnegative.

The power of the test increased with the absolute value of the upper bound and with the sample size, and decreased with the censoring proportion (Figure 1). When the absolute value of the upper bound was 0.01 and $n = 1000$, the power was 0.95 to 0.99 for $\mu_\tau = 0.10$ but 0.30 to 0.46 for $\mu_\tau = 0.50$ across the three dependence levels shown. The power was also higher for $\tau_C = 0.6$ than for $\tau_C = -0.6$ and 0, whose curves were similar; at the same absolute value and sample size with $\mu_\tau = 0.30$, it was 0.46, 0.52, and 0.76 for $\tau_C = -0.6$, 0, and 0.6. Heavier censoring before the horizon thus requires a larger difference in performance, or a larger sample, to establish robust strict dominance.

At the null targets, the IPCW test of the IBS difference rejected far more often than the test of robust strict dominance, but it addresses a different question. This test rejects when the one-sided upper 95% confidence bound for the IPCW estimate of the difference is negative. Its rejection proportions averaged 0.83, 0.91, and 0.97 at the interior targets and 0.86, 0.95, and 0.99 at the zero target for $n = 250$, 500, and 1000. In this design, the full-data IBS difference was negative in every scenario and at every target level, so these rejections agree with the ranking under the true full-data distribution. In the interior, however, some full-data distributions compatible with P give model 2 the smaller IBS, and the IPCW test excludes them only through

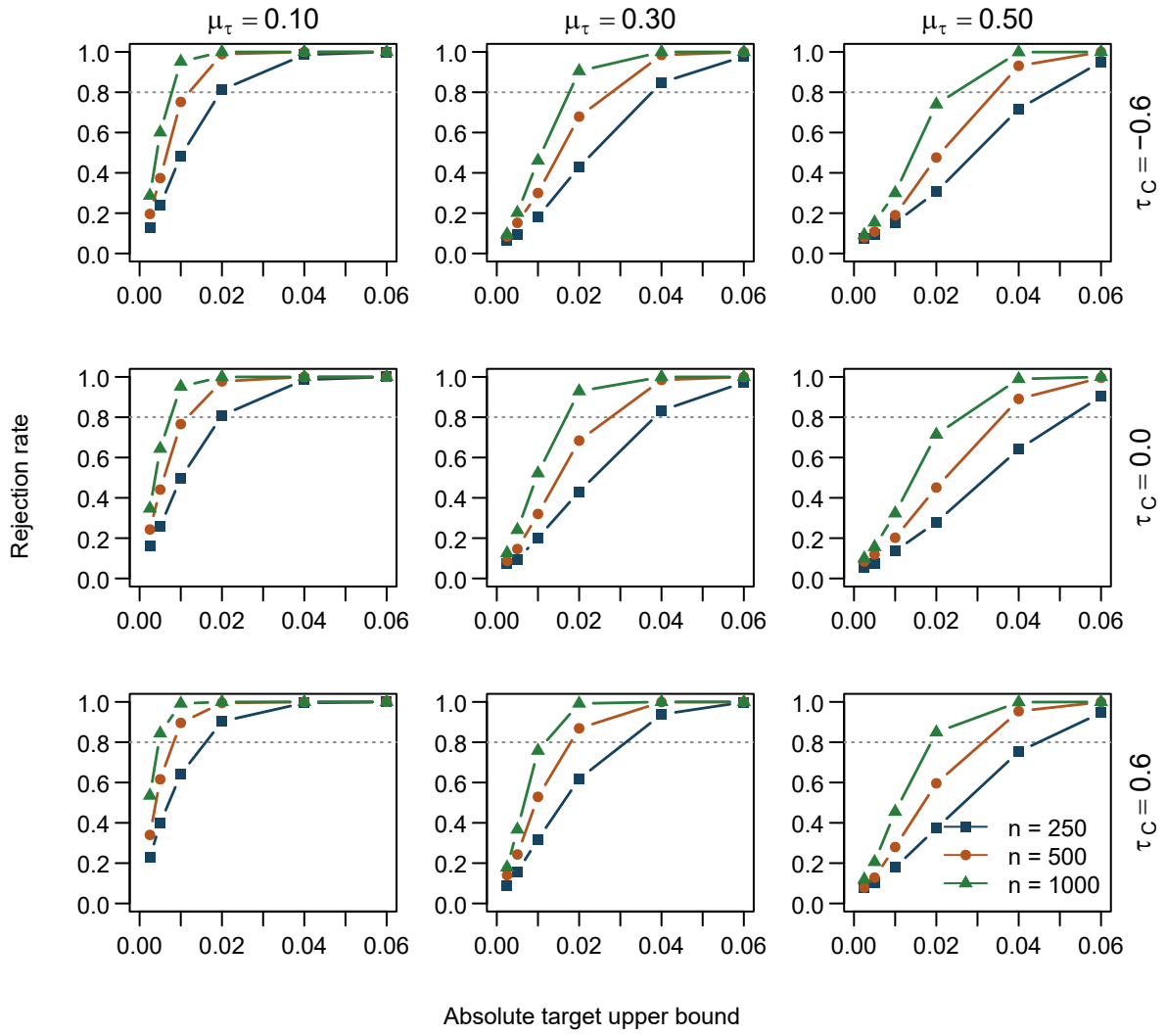


Figure 1: Power of the test of robust strict dominance at $\alpha = 0.05$. The horizontal axis is the absolute value of the target upper bound $\overline{\mathrm{IBS}}_{12}(\tau; P)$, and the vertical axis is the proportion of 1,000 replications in which the test rejected. The rows correspond to the dependence levels $\tau_C = -0.6$, 0, and 0.6, the columns to the censoring proportion μ_τ, and the lines to the sample size n; the dotted line marks 0.8.

the untestable assumption of independent censoring. Moreover, under dependent censoring the IPCW interval often failed to contain the full-data difference: averaged over the scenarios and target levels, its coverage at $n = 1000$ was 0.95 for $\tau_C = 0$ but 0.21 for $\tau_C = -0.6$ and 0.07 for $\tau_C = 0.6$. IPCW therefore ranks the models without reflecting the uncertainty created by dependent censoring, whereas the identified set and the test of robust strict dominance retain it.

4.4 Comparison of multiple models

We compare $K = 3$ models in which model 1 is the true conditional survival function and models 2 and 3 are its competitors. To show that model 1 is robustly strictly best, we must show that it robustly strictly dominates both model 2 and model 3, which requires two tests of the form in Proposition 3.5. If the results of the two tests are reported as separate claims, the probability that at least one of them is false can exceed α, as in multiple testing. The IUT of Proposition 3.6 instead combines the two tests into the single claim that model 1 is best, which it makes only if both tests reject, and its asymptotic level is at most α without any adjustment. We examine its level and power with two kinds of competitors: (i) ordered competitors, for which one competitor alone determines the upper bound $\overline{M}_1(\tau; P)$, and (ii) symmetric competitors, for which the two pairwise upper bounds are equal.

With the ordered competitors (i), the IUT rejected at the zero target in close to 5% of the replications. Both competitors add a positive shift to the linear predictor as in Section 4.3, δ_2 for model 2 and $\delta_3 > \delta_2$ for model 3. The shifts are chosen so that $\overline{\mathrm{IBS}}_{12}(\tau; P)$ takes the eight targets of Section 4.3 and $\overline{\mathrm{IBS}}_{13}(\tau; P)$ takes values 0.005 smaller, so that $\overline{M}_1(\tau; P) = \overline{\mathrm{IBS}}_{12}(\tau; P)$ and model 3 is the weaker competitor. At the zero target, the IUT rejected in proportions of 0.051, 0.049, and 0.048 on average for $n = 250$, 500, and 1000, and from 0.035 to 0.060 in individual scenarios (Table 4). At least one of the two pairwise tests rejected in 0.16, 0.25, and 0.37 on average, and in up to 0.84 in individual scenarios. These rejections come from the test against model 3, whose null hypothesis is false at this target, but model 1 is not robustly strictly best, so they cannot be read as evidence that it is. The IUT is thus not always conservative: when one competitor alone determines the upper bound, its rejection proportion on the null boundary is close to α.

The symmetric competitors (ii) shift the coefficient on X_3 by $\delta > 0$ in opposite directions:

$$\widehat{S}_2(t \mid X) = \exp\bigl[-\lambda_T t^{\kappa_T} \exp\{\eta(X) + \delta X_3\}\bigr], \qquad \widehat{S}_3(t \mid X) = \exp\bigl[-\lambda_T t^{\kappa_T} \exp\{\eta(X) - \delta X_3\}\bigr].$$

Model 2 predicts a lower survival probability than model 1 for individuals with $X_3 > 0$ and a higher one for those with $X_3 < 0$, and model 3 does the opposite. Because X_3 is symmetric about zero and independent of (X_1, X_2, T, C), the two pairwise upper bounds are equal, and we choose δ so that their common value $\overline{M}_1(\tau; P)$ takes the eight targets of Section 4.3. At the zero target, both pairwise null hypotheses therefore lie on their boundaries, and the IUT rejects only if the two pairwise tests reject in the same sample.

With the symmetric competitors at the zero target, each pairwise test rejected in close to 5% of the replications and at least one of them in close to 10%, whereas the IUT almost never rejected (Table 4). Averaged over the 15 scenarios, the tests against models 2 and 3 rejected in proportions from 0.047 to 0.050, at least one of them rejected in 0.094, 0.096, and 0.097 for $n = 250$, 500, and 1000, and the IUT rejected in 0.0006, 0.0004, and 0.0005. In individual

Table 4: Rejection proportions at the zero target, where the shifts were chosen so that $\overline{M}_1(\tau; P) = 0$. Each entry is the proportion of 1,000 replications in which the test rejected at $\alpha = 0.05$. For each kind of competitors, Either is the rejection of at least one of the tests of robust strict dominance of model 1 over models 2 and 3. Column headings give the sample size n.

		(i) Ordered						(ii) Symmetric					
		Either			IUT			Either			IUT		
τ_C	μ_τ	250	500	1000	250	500	1000	250	500	1000	250	500	1000
−0.6	0.10	0.238	0.374	0.601	0.042	0.061	0.054	0.105	0.116	0.080	0.000	0.000	0.000
−0.3	0.10	0.242	0.368	0.591	0.061	0.048	0.040	0.103	0.125	0.108	0.000	0.000	0.000
0.0	0.10	0.258	0.441	0.644	0.054	0.056	0.045	0.096	0.085	0.091	0.000	0.000	0.000
0.3	0.10	0.299	0.486	0.717	0.056	0.044	0.045	0.091	0.074	0.109	0.000	0.000	0.000
0.6	0.10	0.398	0.616	0.844	0.061	0.044	0.057	0.120	0.097	0.093	0.000	0.000	0.000
−0.6	0.30	0.093	0.152	0.202	0.045	0.048	0.040	0.099	0.103	0.110	0.000	0.000	0.000
−0.3	0.30	0.109	0.129	0.215	0.049	0.041	0.035	0.094	0.084	0.097	0.000	0.000	0.001
0.0	0.30	0.095	0.146	0.241	0.049	0.049	0.057	0.073	0.095	0.097	0.001	0.000	0.000
0.3	0.30	0.112	0.199	0.272	0.037	0.051	0.036	0.096	0.088	0.106	0.001	0.001	0.000
0.6	0.30	0.157	0.243	0.367	0.050	0.052	0.047	0.085	0.102	0.089	0.000	0.000	0.001
−0.6	0.50	0.094	0.108	0.154	0.057	0.053	0.055	0.085	0.097	0.092	0.001	0.002	0.003
−0.3	0.50	0.088	0.115	0.144	0.051	0.049	0.053	0.089	0.102	0.103	0.003	0.000	0.002
0.0	0.50	0.072	0.119	0.156	0.046	0.050	0.057	0.087	0.101	0.104	0.001	0.001	0.000
0.3	0.50	0.074	0.107	0.138	0.045	0.044	0.045	0.084	0.080	0.081	0.002	0.001	0.001
0.6	0.50	0.102	0.128	0.206	0.056	0.048	0.052	0.099	0.089	0.091	0.000	0.001	0.000

scenarios, the separate tests rejected in proportions from 0.032 to 0.066, at least one of them in 0.073 to 0.125, and the IUT in at most 0.003. Because both pairwise null hypotheses hold, reporting the two results separately would thus make at least one false claim with probability about 2α. At the positive targets in the interior of the null hypothesis, the IUT never rejected. These proportions of the IUT are consistent with Proposition 3.6(i), which only bounds its asymptotic rejection probability by α from above.

The low rejection proportions of the IUT can be explained by a negative correlation between the upper-bound contributions against the two competitors. An individual with $X_3 > 0$ receives a lower prediction from model 2 and a higher one from model 3, so its contributions $\overline{d}_{12}(O)$ and $\overline{d}_{13}(O)$ tend to move in opposite directions. The test against model k rejects when $U_{1k} < 0$, that is, when $\widehat{\overline{\mathrm{IBS}}}_{1k}(\tau) < -z_{1-\alpha}\widehat{\overline{\sigma}}_{1k}/\sqrt{n}$. Both estimates are sample averages over the same individuals, so they inherit the correlation of $\overline{d}_{12}(O)$ and $\overline{d}_{13}(O)$: a sample whose individuals happen to favor model 1 over model 2 tends to favor model 3 over model 1, and vice versa. The IUT, however, rejects only if the inequality above holds for both $k = 2$ and $k = 3$ in the same sample, that is, only if the sample favors model 1 over both competitors at once. The negative correlation makes such samples rare, so the IUT rarely rejects.

The observed rejection proportions of the IUT agreed with those implied by the correlations in the reference samples. The correlation between the two contributions averaged −0.78, −0.36, and −0.10 for $\mu_\tau = 0.10$, 0.30, and 0.50, and the IUT rejection proportion at the zero target rose from 0.000 to 0.0003 and 0.0012. To compute the implied proportions, we approximate the two standardized estimates $\sqrt{n}\,\widehat{\overline{\mathrm{IBS}}}_{1k}(\tau)/\widehat{\overline{\sigma}}_{1k}$, $k = 2, 3$, by standard normal variables with these correlations, which is their limit when both upper bounds are zero. The IUT rejects when

both are below $-z_{1-\alpha}$, and this probability is about 2×10^{-8}, 0.0003, and 0.0016 for $\mu_\tau = 0.10$, 0.30, and 0.50, close to the observed proportions. This probability depends on the correlation: it would be $\alpha = 0.05$ if the two estimates were perfectly positively correlated and $\alpha^2 = 0.0025$ if they were independent, and it is smaller still when they are negatively correlated. With these competitors, the rejection proportions of the IUT on the null boundary were therefore far below the nominal level α.

At the same value of the upper bound, the IUT had lower power with the symmetric competitors (ii) than with the ordered competitors (i), and the difference was largest when the absolute value of the upper bound was small (Figure 2). Averaged over the 15 scenarios at $n = 1000$, the power with the ordered and the symmetric competitors was 0.20 and 0.03 when the absolute value of the upper bound was 0.0025, 0.62 and 0.51 when it was 0.01, and 0.90 and 0.86 when it was 0.02, and both were at least 0.997 when it was 0.04. The difference shrank as n and the absolute value grew, because every pairwise test then rejects with probability tending to one, as Proposition 3.6(ii) states. Heavier censoring lowered both curves, as in Section 4.3. To show that a model is robustly strictly best when several competitors are equally close to it, a larger difference in performance or a larger sample is thus needed than when one competitor stands out.

5 Real data applications

This section applies the bounds in Section 2 and the inference procedures in Section 3 to three real data sets and shows how the conclusions differ from those of IPCW-based evaluation, which assumes independent censoring. Section 5.1 introduces data on unemployment duration, churn from a music streaming service, and mortgage prepayment, which differ widely in the proportion censored within the evaluation period, and Section 5.2 describes the procedure common to the three data sets. Section 5.3 reports the results for each data set and compares them across the three data sets.

5.1 Data sets

We use three data sets that differ widely in the proportion censored before the horizon and in sample size (Table 5). They record unemployment duration (UnempDur; McCall, 1996), churn from a music streaming service (KKBOX; Kaggle & KKBOX, 2017; Kvamme, 2024), and mortgage prepayment in the Single Family Loan-Level Dataset of Freddie Mac (SFLLD; Freddie Mac, 2026). They contain 3,343, 50,000, and 36,115 individuals, of whom 56.3%, 22.4%, and 0.4% are censored before the horizon, respectively. These differences allow us to compare how the conclusions supported without restricting the dependence between event and censoring times change with the amount of censoring and the sample size.

UnempDur consists of the unemployment spells of 3,343 individuals, recorded in two-week units. The event is reemployment in a full-time job, and spells ending in a part-time job, in a job of unknown type, or in continued unemployment are censored. With the horizon $\tau = 13$, or 26 weeks, 56.3% of the individuals are censored before the horizon, the largest proportion among the three data sets. The censoring may depend on the time to full-time reemployment even given the covariates, because individuals who accepted a part-time job may have had worse prospects of finding a full-time job.

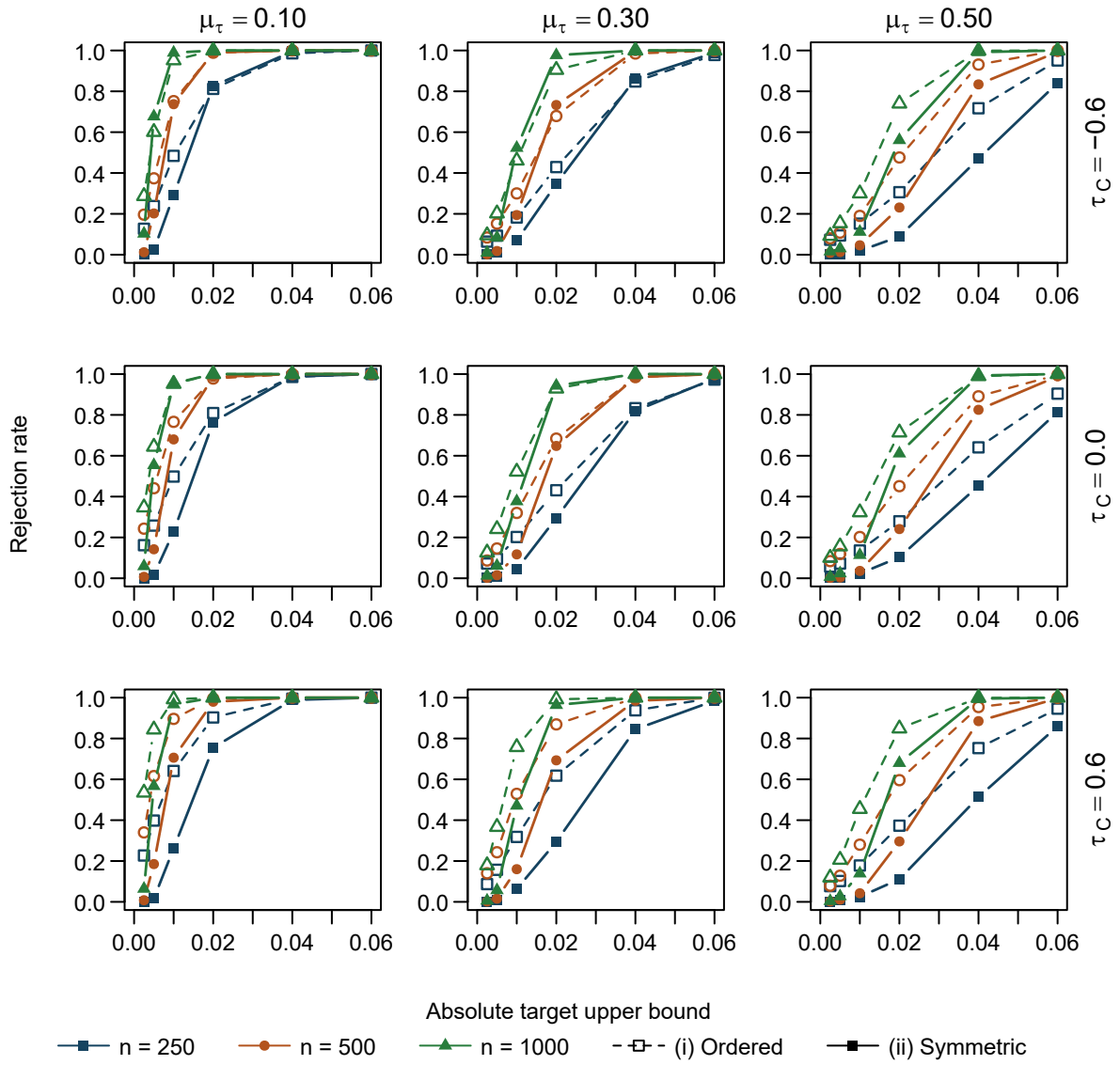


Figure 2: Power of the IUT at $\alpha = 0.05$ with (i) the ordered and (ii) the symmetric competitors. The horizontal axis is the absolute value of the target upper bound $\overline{M}_1(\tau; P)$, and the vertical axis is the proportion of 1,000 replications in which the IUT rejected. The rows correspond to the dependence levels $\tau_C = -0.6$, 0, and 0.6, the columns to the censoring proportion μ_τ, and the colors to the sample size n. Dashed lines with open markers show the ordered competitors, and solid lines with filled markers show the symmetric competitors.

Table 5: The three data sets. The horizon τ is in the recorded time unit of each data set, two-week units for UnempDur. Censored is the percentage of individuals censored before the horizon, with $\Delta = 0$ and $Y < \tau$, and Columns is the number of covariate columns after encoding.

Data	Event	Horizon τ	n	Censored (%)	Columns
UnempDur	Full-time job	13 (26 weeks)	3,343	56.3	6
KKBOX	Churn	365 days	50,000	22.4	44
SFLLD	Prepayment	60 months	36,115	0.4	28

KKBOX consists of the transaction records of a music streaming service released for a Kaggle competition, and the event is churn, defined as in the Python library pycox. We draw 50,000 members from the first subscription spells of members registered on or after 1 January 2015, a restriction intended to avoid left truncation. All spells are administratively censored on 29 January 2017 as in pycox, and with $\tau = 365$ days, 22.4% of the members are censored before the horizon. Because the censoring time is thus determined by the start date of the subscription, independent censoring is more plausible here than in the other two data sets, but the bounds do not assume it.

SFLLD records the monthly performance of mortgages, and we use the 36,115 loans in its 2014 sample file whose first payment was due in 2014 and whose records start at loan age zero. The event is voluntary prepayment, and loans that end in a credit-related disposition, a whole loan sale, a reperforming loan securitization, or a confirmed underwriting or major servicing defect before a credit event are censored. With $\tau = 60$ months, which ends before March 2020, only 0.4% of the loans are censored before the horizon, the smallest proportion among the three data sets. The censoring may depend on the time to prepayment even given the covariates, because delinquency may restrict access to refinancing, so loans ending in a credit-related disposition may have been prepaid later than other loans had they continued; this is our interpretation.

5.2 Analysis procedure

We split each data set once into training, validation, and test samples in proportions of 70%, 10%, and 20%, stratified by the event indicator, and use the test sample only for evaluation. The test sample is not used for fitting the models, selecting their hyperparameters, or choosing the model to be evaluated. Continuous covariates are standardized, categorical covariates are encoded as indicator variables, and missing values are imputed, all using quantities computed from the training sample only. The test sample serves as the evaluation sample of Section 3, and the fitted models are independent of it, as Assumption 3 requires.

The three candidate models are fitted with hyperparameters selected on the validation sample. They are the Cox proportional hazards model (CoxPH) (Cox, 1972), which has no hyperparameter to select, the random survival forest (RSF) (Ishwaran et al., 2008), and DeepSurv (Katzman et al., 2018). RSF uses 100 trees with a maximum depth of 3, and its grid varies the fraction of covariate columns tried at each split over $\{1/3, 2/3\}$ and the minimum node size over $\{3, 10\}$. DeepSurv has two hidden layers of equal width with ReLU activations and dropout probability 0.1 and is trained by Adam with learning rate 0.01 and weight decay 10^{-4}, and its grid varies the width over $\{16, 64\}$ and the number of epochs over $\{25, 50, 100\}$. For each model, we select the setting with the smallest IPCW estimate of the IBS on the validation sample and refit the model on the combined training and validation samples. Before looking at the test sample, we

fix the model with the smallest such estimate as the evaluated model of Assumption 4, with the other two models as its competitors. This rule separates model selection from evaluation, so the asymptotic level of the intersection–union test in Proposition 3.6 is not invalidated by reselecting the model on the test sample.

On the test sample, we compute the estimators, confidence regions, and tests of Section 3 for single models, pairs of models, and the three models together. For each model, we compute the estimated bounds on its IBS and the confidence region for the identified set in Proposition 3.4. For each pair of models, we compute the estimated bounds on the IBS difference, the confidence region in the same proposition, and the tests of robust strict dominance in both directions in Proposition 3.5. For the three models together, we estimate the upper bound $\overline{M}_j(\tau; P)$ on the difference of each model from its best competitor and, for the prespecified model, compute the intersection–union test in Proposition 3.6. All tests use the significance level $\alpha = 0.05$, and the confidence regions and tests rely on the asymptotic guarantees for a fixed observed-data distribution in Section 3.3.

We compare these results with IPCW, which estimates a single value of the full-data IBS under independent censoring. The IPCW estimates are weighted by the reverse Kaplan–Meier estimator without covariates, and their 95% confidence intervals are computed from 1,000 bootstrap resamples of individuals. As in Section 4.1, this benchmark assumes $C \perp (T, X)$. The IPCW confidence intervals concern a single value of the full-data IBS, whereas the confidence regions concern the identified set, and we keep this distinction when interpreting the results.

5.3 Results

The estimated identified set for the IBS of each model was narrower in data sets with a smaller proportion censored before the horizon (panel (a) of Tables 6–8). The widths were 0.138 to 0.150 for UnempDur, 0.059 to 0.071 for KKBOX, and 0.0011 to 0.0014 for SFLLD. For SFLLD, the estimated range of IBS values compatible with the observed data thus had width about 0.001 for each model. Only individuals censored before the horizon contribute to this width, but their contributions also depend on the predicted survival curves (Section 2.2).

In all three data sets, the IPCW estimates lay within the estimated identified sets and particularly close to their lower bounds, at most 10% of the width above them for UnempDur and at most 20% for the other two (panel (a) of Tables 6–8). For UnempDur and KKBOX, the estimated upper bounds exceeded the IPCW estimates by 0.05 to 0.14, and for UnempDur they were nearly twice the IPCW estimates. According to these estimates, the IBS of each model in these two data sets could be considerably larger than the IPCW estimate under some full-data distributions compatible with the observed data. The IPCW estimates alone therefore do not show how large the IBS could be without the assumption of independent censoring.

The estimated identified sets for the IBS difference between two models excluded zero for all three pairs in SFLLD but contained zero for all three pairs in UnempDur and KKBOX (panel (b) of Tables 6–8). That the estimated identified sets contain zero does not establish that the population identified sets do. In all data sets, the bounds on the differences were 8% to 20% as wide as those obtained by combining the separate bounds for the two models, because each individual's event time is shared by both models (Proposition 2.3). Only in SFLLD, therefore, did the estimates indicate the sign of every difference without restricting the dependence between event and censoring times.

Table 6: Results for UnempDur on the test sample ($n = 669$). (a) Estimated bounds on the IBS, the 95% confidence region (CR) for the identified set in Proposition 3.4, and the IPCW estimate with its 95% bootstrap confidence interval (CI). (b) For each pair (j, k), estimated bounds on $\mathrm{IBS}_{jk}(\tau; Q) = \mathrm{IBS}_j(\tau; Q) - \mathrm{IBS}_k(\tau; Q)$, the 95% CR, the one-sided upper confidence bounds of the tests of whether model j robustly strictly dominates model k (U_{jk}) and the reverse (U_{kj}) in Proposition 3.5, and the IPCW estimate of the difference with its 95% CI. (c) For each model j, the estimate of the upper bound $\overline{M}_j(\tau; P)$ on the difference $M_j(\tau; Q)$ from the best competitor in Proposition 2.5, and U_j^{IUT} of the intersection–union test in Proposition 3.6 for the prespecified model, marked with an asterisk. Each test rejects at $\alpha = 0.05$ if its upper confidence bound is negative.

(a) IBS of each model

Model	Lower	Upper	95% CR	IPCW	95% CI
CoxPH	0.145	0.287	[0.135, 0.299]	0.158	[0.145, 0.171]
RSF	0.145	0.283	[0.135, 0.295]	0.154	[0.141, 0.167]
DeepSurv	0.139	0.290	[0.129, 0.302]	0.154	[0.141, 0.168]

(b) IBS difference between two models

Pair	Lower	Upper	95% CR	U_{jk}	U_{kj}	IPCW	95% CI
CoxPH − RSF	−0.0144	0.0204	[−0.0174, 0.0234]	0.0229	0.0169	0.0036	[0.0000, 0.0072]
CoxPH − DeepSurv	−0.0161	0.0200	[−0.0196, 0.0228]	0.0224	0.0190	0.0033	[−0.0003, 0.0070]
RSF − DeepSurv	−0.0158	0.0141	[−0.0187, 0.0163]	0.0159	0.0182	−0.0003	[−0.0033, 0.0028]

(c) Difference from the best competitor

Model	Upper	U_j^{IUT}
CoxPH	0.0204	
RSF	0.0144	
DeepSurv*	0.0161	0.0190

The tests of robust strict dominance supported the IPCW ranking in SFLLD but rejected for no pair in UnempDur and KKBOX (panel (b) of Tables 6–8). In SFLLD, the tests, each at the 5% significance level without adjustment for multiple testing, supported DeepSurv over both CoxPH and RSF, and CoxPH over RSF. In UnempDur, IPCW distinguished only CoxPH and RSF, and only barely, with a lower confidence limit of about 5×10^{-6} that is rounded to 0.0000 in Table 6. In KKBOX, by contrast, the IPCW confidence intervals excluded zero for all three pairs. The comparisons supported by these tests without restricting the dependence between event and censoring times therefore differed across the data sets.

The IUT at the 5% significance level certified the prespecified model, DeepSurv, as robustly strictly best only in SFLLD (panel (c) of Tables 6–8). In all three data sets, DeepSurv had the smallest IPCW estimate on the validation sample and was therefore prespecified as the evaluated model. For SFLLD, the estimated upper bound on its difference from the best competitor was -0.0012, and the one-sided upper confidence bound $U_j^{\mathrm{IUT}} = -0.0006$ remained below zero. For UnempDur and KKBOX, the IUT did not reject because the estimated upper bounds themselves were positive, 0.0161 and 0.0027, before any allowance for sampling uncertainty. These nonrejections do not establish that a competitor has a smaller population IBS than DeepSurv.

Table 7: Results for KKBOX on the test sample ($n = 10{,}002$). The panels and columns are as in Table 6.

(a) IBS of each model

Model	Lower	Upper	95% CR	IPCW	95% CI
CoxPH	0.098	0.169	[0.095, 0.173]	0.110	[0.106, 0.113]
RSF	0.107	0.166	[0.104, 0.170]	0.118	[0.115, 0.121]
DeepSurv	0.095	0.165	[0.092, 0.169]	0.107	[0.103, 0.110]

(b) IBS difference between two models

Pair	Lower	Upper	95% CR	U_{jk}	U_{kj}	IPCW	95% CI
CoxPH − RSF	−0.0126	0.0053	[−0.0140, 0.0069]	0.0066	0.0137	−0.0081	[−0.0098, −0.0065]
CoxPH − DeepSurv	−0.0027	0.0087	[−0.0036, 0.0097]	0.0095	0.0034	0.0029	[0.0018, 0.0040]
RSF − DeepSurv	−0.0012	0.0144	[−0.0027, 0.0158]	0.0156	0.0024	0.0111	[0.0094, 0.0127]

(c) Difference from the best competitor

Model	Upper	U_j^{IUT}
CoxPH	0.0087	
RSF	0.0144	
DeepSurv*	0.0027	0.0034

Table 8: Results for SFLLD on the test sample ($n = 7{,}225$). The panels and columns are as in Table 6, with two additional columns in panel (a). Width is Upper minus Lower, and Position is 100(IPCW − Lower)/Width, the IPCW estimate's percentage distance from the lower end. Both columns are calculated from the saved results before rounding the displayed endpoints and IPCW estimates.

(a) IBS of each model

Model	Lower	Upper	95% CR	Width	Position (%)	IPCW	95% CI
CoxPH	0.152	0.153	[0.149, 0.156]	0.0013	17.1	0.152	[0.149, 0.155]
RSF	0.155	0.156	[0.152, 0.158]	0.0011	14.7	0.155	[0.152, 0.157]
DeepSurv	0.151	0.152	[0.148, 0.155]	0.0014	17.5	0.151	[0.148, 0.153]

(b) IBS difference between two models

Pair	Lower	Upper	95% CR	U_{jk}	U_{kj}	IPCW	95% CI
CoxPH − RSF	−0.0027	−0.0023	[−0.0036, −0.0014]	−0.0016	0.0035	−0.0026	[−0.0035, −0.0017]
CoxPH − DeepSurv	0.0012	0.0015	[0.0005, 0.0021]	0.0020	−0.0006	0.0013	[0.0007, 0.0020]
RSF − DeepSurv	0.0036	0.0041	[0.0026, 0.0051]	0.0050	−0.0027	0.0039	[0.0029, 0.0049]

(c) Difference from the best competitor

Model	Upper	U_j^{IUT}
CoxPH	0.0015	
RSF	0.0041	
DeepSurv*	−0.0012	−0.0006

6 Conclusion

This paper has derived sharp bounds on the IBS of survival prediction models and on IBS differences between them without restricting the dependence between event and censoring times. Such bounds are needed because the independent censoring assumed by IPCW cannot generally be verified from right-censored data. Using the fact that each individual's event time is shared across evaluation times and models, the bounds cover a single model, a pair of models, and, through its upper bound, the difference from the best competitor (Propositions 2.2, 2.3, and 2.5). For a fixed observed-data distribution, we constructed estimators of these bounds, confidence regions, a test of robust strict dominance, and an intersection–union test. With these tools, a model can be shown to be better under every full-data distribution compatible with the observed data, without specifying the distribution of the unobserved event times.

The simulations and the applications showed that the bounds separate rankings supported by the observed data from rankings that rely on independent censoring. In the simulations, the confidence regions and tests performed close to their nominal levels, whereas the IPCW intervals often missed the full-data IBS under dependent censoring, although the IUT was far more conservative when two competitors were equally close. In the applications, the tests at the 5% level, without adjustment for multiple testing, supported the IPCW ranking and certified DeepSurv as robustly strictly best only for SFLLD, in which 0.4% of the individuals were censored before the horizon. For UnempDur and KKBOX, with 56.3% and 22.4% censored, no test rejected, so their IPCW rankings were not supported without assuming independent censoring. Estimated bounds should therefore be reported alongside IPCW estimates.

The method has limitations that point to directions for future work. Its inference is asymptotic for a fixed observed-data distribution, so its level need not hold uniformly near distributions at which an upper bound is close to zero or several competitors nearly attain the maximum. It also requires models trained separately from the evaluation sample and comparisons specified in advance (Assumptions 3 and 4). Moreover, the identified sets can be wide when many individuals are censored before the horizon. Future work includes narrowing them with weak assumptions credible in the application, such as administrative censoring times known for all individuals. It also includes extending the bounds to measures other than the IBS, such as the concordance index (Harrell et al., 1982; Uno et al., 2011).

References

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, *24*(4), 295–300. https://doi.org/10.1080/00401706.1982.10487790

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283–319. https://doi.org/10.1214/ss/1032280304

Chernozhukov, V., Hong, H., & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, *75*(5), 1243–1284. https://doi.org/10.1111/j.1468-0262.2007.00794.x

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Freddie Mac. (2026). *Single family loan-level dataset* [Origination and performance data through March 31, 2026] (Release 47). Retrieved October 8, 2026, from https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, *48*(6), 1029–1040. https://doi.org/10.1002/bimj.200610301

Gharari, A. H. F., Cooper, M., Greiner, R., & Krishnan, R. G. (2023). Copula-based deep survival models for dependent censoring. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, *216*, 669–680. Retrieved September 4, 2026, from https://proceedings.mlr.press/v216/gharari23a.html

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, *18*(17–18), 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

Guerdan, L., Coston, A. L., Holstein, K., & Wu, S. (2024). Predictive Performance Comparison of Decision Policies Under Confounding. *Proceedings of the 41st International Conference on Machine Learning*, 16673–16705. Retrieved September 4, 2026, from https://proceedings.mlr.press/v235/guerdan24a.html

Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*(18), 2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, *72*(6), 1845–1857. https://doi.org/10.1111/j.1468-0262.2004.00555.x

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

Kaggle & KKBOX. (2017). *WSDM - KKBox's churn prediction challenge*. Retrieved October 9, 2026, from https://www.kaggle.com/c/kkbox-churn-prediction-challenge

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

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

Kvamme, H. (2024, September 4). *pycox: Survival analysis with PyTorch* (0.3.0). Retrieved October 9, 2026, from https://github.com/havakv/pycox

Lillelund, C. M., Qi, S.-a., & Greiner, R. (2026). Overcoming dependent censoring in the evaluation of survival models. *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, *337*, 3841–3866. Retrieved September 4, 2026, from https://proceedings.mlr.press/v337/lillelund26a.html

Liu, Y., Molinari, F., & Velez, A. (2026, June 12). *Identification and Inference for Algorithmic Frontiers with Selective Labels*. arXiv.org. Retrieved September 4, 2026, from https://arxiv.org/abs/2606.14977v1

Maia Polo, F., Maity, S., Yurochkin, M., Banerjee, M., & Sun, Y. (2024). Weak Supervision Performance Evaluation via Partial Identification. *Advances in Neural Information Processing Systems*, *37*, 135704–135740. https://doi.org/10.52202/079017-4309

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer. https://doi.org/10.1007/b97478

McCall, B. P. (1996). Unemployment insurance rules, joblessness, and part-time work. *Econometrica*, *64*(3), 647–682. https://doi.org/10.2307/2171865

Molinari, F. (2020). Microeconometrics with partial identification. In S. Durlauf, L. P. Hansen, J. J. Heckman, & R. L. Matzkin (Eds.), *Handbook of econometrics* (pp. 355–486, Vol. 7A). Elsevier. https://doi.org/10.1016/bs.hoe.2020.05.002

Nelsen, R. B. (2006). *An introduction to copulas* (2nd ed.). Springer. https://doi.org/10.1007/0-387-28678-0

Peterson, A. V., Jr. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, *73*(1), 11–13. https://doi.org/10.1073/pnas.73.1.11

Rambachan, A., Coston, A., & Kennedy, E. H. (2026). Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding. *Review of Economics and Statistics*, 1–40. https://doi.org/10.1162/REST.a.1815

Romano, J. P., & Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, *138*(9), 2786–2807. https://doi.org/10.1016/j.jspi.2008.03.015

Sakaguchi, S. (2024). Partial identification and inference in duration models with endogenous censoring. *Journal of Applied Econometrics*, *39*(2), 308–326. https://doi.org/10.1002/jae.3024

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, *77*(4), 1299–1315. https://doi.org/10.3982/ECTA7347

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, *2*(1), 167–195. https://doi.org/10.1146/annurev.economics.050708.143401

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, *72*(1), 20–22. https://doi.org/10.1073/pnas.72.1.20

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, *30*(10), 1105–1117. https://doi.org/10.1002/sim.4154

Willems, I., Beyhum, J., & Van Keilegom, I. (2026). Bounds for the regression parameters in dependently censored survival models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Article qkag119. https://doi.org/10.1093/jrsssb/qkag119

Appendix

A Proofs for Section 2

Proof of Proposition 2.1. Part (i). Let $Q \in \mathcal{Q}(P)$. By the definition of compatibility in Section 2.1, the observed data $O = (Y, \Delta, X)$ have distribution P under Q. Write $\underline{b}_{\mathrm{BS}}(t, O)$ and $\overline{b}_{\mathrm{BS}}(t, O)$ for the expressions inside the expectations defining $\underline{\mathrm{BS}}(t; P)$ and $\overline{\mathrm{BS}}(t; P)$, respectively.

If $Y > t$, then $T \geq Y > t$, so the loss equals $\{1 - \widehat{S}(t \mid X)\}^2$. If $Y \leq t$ and $\Delta = 1$, then $T = Y \leq t$, so the loss equals $\widehat{S}(t \mid X)^2$. In these two cases, both $\underline{b}_{\mathrm{BS}}(t, O)$ and $\overline{b}_{\mathrm{BS}}(t, O)$ equal the loss. If $Y \leq t$ and $\Delta = 0$, then $\mathbf{1}(T > t) \in \{0, 1\}$, so the loss $\{\mathbf{1}(T > t) - \widehat{S}(t \mid X)\}^2$ lies between the minimum and the maximum of $\widehat{S}(t \mid X)^2$ and $\{1 - \widehat{S}(t \mid X)\}^2$. Thus, in all three cases,

$$\underline{b}_{\mathrm{BS}}(t, O) \leq \{\mathbf{1}(T > t) - \widehat{S}(t \mid X)\}^2 \leq \overline{b}_{\mathrm{BS}}(t, O) \qquad Q\text{-almost surely.}$$

Since $\underline{b}_{\mathrm{BS}}(t, O)$ and $\overline{b}_{\mathrm{BS}}(t, O)$ are functions of O alone, compatibility implies that their expectations under Q equal their expectations under P. Taking expectations of the loss inequalities therefore gives

$$\begin{aligned} \underline{\mathrm{BS}}(t; P) &= \mathbb{E}_P[\underline{b}_{\mathrm{BS}}(t, O)] = \mathbb{E}_Q[\underline{b}_{\mathrm{BS}}(t, O)] \\ &\leq \mathrm{BS}(t; Q) \leq \mathbb{E}_Q[\overline{b}_{\mathrm{BS}}(t, O)] \\ &= \mathbb{E}_P[\overline{b}_{\mathrm{BS}}(t, O)] = \overline{\mathrm{BS}}(t; P). \end{aligned}$$

Part (ii). Suppose that $P(Y = t, \Delta = 0) = 0$. We first construct a compatible full-data distribution by choosing event times to minimize the squared prediction loss. For individuals censored before t, the loss is $\widehat{S}(t \mid X)^2$ if $T \leq t$ and $\{1 - \widehat{S}(t \mid X)\}^2$ if $T > t$. Since $\widehat{S}(t \mid X)^2 \leq \{1 - \widehat{S}(t \mid X)\}^2$ if and only if $\widehat{S}(t \mid X) \leq 1/2$, we choose the event times accordingly. Given $O = (Y, \Delta, X)$, define $C = Y$ and

$$T = \begin{cases} Y, & \Delta = 1, \\ Y + 1, & \Delta = 0,\ Y \geq t, \\ t, & \Delta = 0,\ Y < t,\ \widehat{S}(t \mid X) \leq 1/2, \\ t + 1, & \Delta = 0,\ Y < t,\ \widehat{S}(t \mid X) > 1/2, \end{cases}$$

and let $\underline{Q}$ be the distribution of (T, C, X) when O follows P. This construction is measurable because $\widehat{S}(t \mid X)$ is measurable. In every case, $T \leq C$ if and only if $\Delta = 1$, and $\min(T, C) = Y$; hence $\underline{Q} \in \mathcal{Q}(P)$.

We now verify that $\underline{Q}$ attains the lower bound. For individuals with $Y < t$ and $\Delta = 0$, the construction sets $T = t$ when $\widehat{S}(t \mid X) \leq 1/2$ and $T = t + 1$ otherwise. Their loss is therefore $\min\{\widehat{S}(t \mid X)^2, \{1 - \widehat{S}(t \mid X)\}^2\}$. For individuals with $Y > t$ or $Y \leq t$ and $\Delta = 1$, the survival status is determined by the observed data, and the corresponding loss agrees with the definition of $\underline{\mathrm{BS}}(t; P)$. Since $P(Y = t, \Delta = 0) = 0$, taking expectations gives $\mathrm{BS}(t; \underline{Q}) = \underline{\mathrm{BS}}(t; P)$.

The upper bound is attained by reversing the loss assignments for individuals censored before t. Interchanging the values t and $t + 1$ in the last two cases of the definition of T gives $\overline{Q} \in \mathcal{Q}(P)$ under which each such individual has the larger loss. The same probability-zero case and expectation argument give $\mathrm{BS}(t; \overline{Q}) = \overline{\mathrm{BS}}(t; P)$.

Every value between the two attained bounds can also be attained by mixing $\underline{Q}$ and $\overline{Q}$. For $\lambda \in [0,1]$, define $Q_\lambda = (1-\lambda)\underline{Q} + \lambda\overline{Q}$. For every measurable set A of observed-data values, compatibility gives

$$\begin{aligned} Q_\lambda(O \in A) &= (1-\lambda)\underline{Q}(O \in A) + \lambda\overline{Q}(O \in A) \\ &= (1-\lambda)P(A) + \lambda P(A) = P(A). \end{aligned}$$

Thus Q_λ also induces P and belongs to $\mathcal{Q}(P)$. Since the Brier score is the expectation of a fixed loss, its value under Q_λ is $(1-\lambda)\underline{\mathrm{BS}}(t;P) + \lambda\overline{\mathrm{BS}}(t;P)$. By part (i), no value outside this interval is attainable, so the identified set equals $[\underline{\mathrm{BS}}(t;P), \overline{\mathrm{BS}}(t;P)]$. Its endpoints are therefore its infimum and supremum, the sharp bounds defined in Section 2.1.

Part (iii). The definitions of $\underline{\mathrm{BS}}(t;P)$ and $\overline{\mathrm{BS}}(t;P)$ differ only in the term for $Y \le t$ and $\Delta = 0$. For real numbers u, v, $\max\{u,v\} - \min\{u,v\} = |u-v|$, and $\{1-\widehat{S}(t \mid X)\}^2 - \widehat{S}(t \mid X)^2 = 1 - 2\widehat{S}(t \mid X)$. Subtracting the two definitions therefore gives the stated width. □

Proof of Proposition 2.2. Preliminaries. We first establish three properties of the loss L for a censored individual. By (2), for every x and $0 \le s \le r \le \tau$,

$$L(r,x;\tau) - L(s,x;\tau) = \int_s^r \{1 - 2\widehat{S}(t \mid x)\}\, dt,$$

since $\{1-\widehat{S}(t \mid x)\}^2 - \widehat{S}(t \mid x)^2 = 1 - 2\widehat{S}(t \mid x)$. This identity uses only the definition of L. The integrand is bounded in absolute value by one, so $|L(r,x;\tau) - L(s,x;\tau)| \le |r-s|$. Since $\widehat{S}(t \mid x)$ is nonincreasing in t, the integrand is nondecreasing, so its integral, and therefore $r \mapsto L(r,x;\tau)$, is convex on $[0,\tau]$.

Fix $0 \le y < \tau$ and x, and write $r^* = r^*(y,x)$. By (2), for $r \in [y,\tau]$,

$$\begin{aligned} L(r,x;\tau) &= \int_0^y \{1-\widehat{S}(t \mid x)\}^2\, dt + \int_y^r \{1-\widehat{S}(t \mid x)\}^2\, dt + \int_r^\tau \widehat{S}(t \mid x)^2\, dt \\ &\ge \int_0^y \{1-\widehat{S}(t \mid x)\}^2\, dt + \int_y^\tau \min\{\widehat{S}(t \mid x)^2, \{1-\widehat{S}(t \mid x)\}^2\}\, dt. \end{aligned}$$

Since $\widehat{S}(t \mid x)$ is nonincreasing in t, $\widehat{S}(t \mid x) > 1/2$ for $t \in [y, r^*)$ and $\widehat{S}(t \mid x) \le 1/2$ for $t \in (r^*, \tau]$. Hence the minimum in the display equals $\{1-\widehat{S}(t \mid x)\}^2$ for $t \in [y, r^*)$ and $\widehat{S}(t \mid x)^2$ for $t \in (r^*, \tau]$. The single point r^* does not affect the integrals, so taking $r = r^*$ attains the lower bound in the display:

$$\begin{aligned} L(r^*,x;\tau) &= \int_0^y \{1-\widehat{S}(t \mid x)\}^2\, dt + \int_y^\tau \min\{\widehat{S}(t \mid x)^2, \{1-\widehat{S}(t \mid x)\}^2\}\, dt \\ &\le L(r,x;\tau), \qquad r \in [y,\tau]. \end{aligned} \tag{6}$$

By convexity, $L(r,x;\tau) \le \max\{L(y,x;\tau), L(\tau,x;\tau)\}$ for every $r \in [y,\tau]$.

These quantities are measurable functions of the observed data. By Tonelli's theorem and the joint measurability of $\widehat{S}$, the functions $L(Y,X;\tau)$, $L(\tau,X;\tau)$, and the integral expression in (6) with $y = Y$ are measurable. For $r^*(Y,X)$, fix $c \le \tau$. By the definition of r^*, $r^*(Y,X) < c$ holds exactly when $\widehat{S}(s \mid X) \le 1/2$ for some $s \in [Y,c)$. If such an s exceeds Y, every rational $q \in (s,c)$

also satisfies $\widehat{S}(q \mid X) \le \widehat{S}(s \mid X) \le 1/2$ because $\widehat{S}(t \mid X)$ is nonincreasing in t. Hence

$$\{r^*(Y,X) < c\} = \big\{Y < c,\ \widehat{S}(Y \mid X) \le 1/2\big\} \cup \bigcup_{q \in \mathbb{Q}} \big\{Y < q < c,\ \widehat{S}(q \mid X) \le 1/2\big\}.$$

Each event on the right-hand side is measurable by the joint measurability of $\widehat{S}$, so this countable union is measurable. For $c > \tau$, the event $\{r^*(Y,X) < c\}$ contains every point with $Y < \tau$, so $r^*(Y,X)$ is measurable on $Y < \tau$. Write $\underline{\ell}(O)$ and $\overline{\ell}(O)$ for τ^{-1} times the expressions inside the expectations defining $\underline{\mathrm{IBS}}(\tau;P)$ and $\overline{\mathrm{IBS}}(\tau;P)$, so that $\underline{\mathrm{IBS}}(\tau;P) = \mathbb{E}_P\,\underline{\ell}(O)$ and $\overline{\mathrm{IBS}}(\tau;P) = \mathbb{E}_P\,\overline{\ell}(O)$.

Part (i). Let $Q \in \mathcal{Q}(P)$. By compatibility, the observed data $O = (Y, \Delta, X)$ have distribution P under Q, and $\mathrm{IBS}(\tau;Q) = \tau^{-1}\mathbb{E}_Q L(\min(T,\tau), X; \tau)$ by Section 2.1. If $\Delta = 1$, then $T = Y$; if $\Delta = 0$ and $Y \ge \tau$, then $T > Y \ge \tau$. In both cases, $\min(T,\tau) = \min(Y,\tau)$, so both $\underline{\ell}(O)$ and $\overline{\ell}(O)$ equal $\tau^{-1}L(\min(T,\tau), X;\tau)$. If $Y < \tau$ and $\Delta = 0$, then $T > Y$, so $\min(T,\tau) \in (Y,\tau]$, and the preliminaries give $\underline{\ell}(O) \le \tau^{-1}L(\min(T,\tau), X;\tau) \le \overline{\ell}(O)$. Since $\underline{\ell}$ and $\overline{\ell}$ are functions of O alone, their expectations under Q equal those under P. Taking expectations gives $\underline{\mathrm{IBS}}(\tau;P) \le \mathrm{IBS}(\tau;Q) \le \overline{\mathrm{IBS}}(\tau;P)$.

Part (ii). Fix $\varepsilon > 0$. On $Y < \tau$, define $T_\varepsilon(O) = \max\{r^*(Y,X), \min(Y+\varepsilon, \tau)\}$, a measurable function of O. Since $r^*(Y,X) \le \tau$ and $\min(Y+\varepsilon,\tau) > Y$, $T_\varepsilon(O) \in (Y,\tau]$. If $r^*(Y,X) \ge \min(Y+\varepsilon,\tau)$, then $T_\varepsilon(O) = r^*(Y,X)$; otherwise $r^*(Y,X) < T_\varepsilon(O) \le Y + \varepsilon \le r^*(Y,X) + \varepsilon$. Thus $|T_\varepsilon(O) - r^*(Y,X)| \le \varepsilon$, and the Lipschitz bound gives $L(T_\varepsilon(O), X;\tau) \le L(r^*(Y,X), X;\tau) + \varepsilon$.

Given O, define $C = Y$ and

$$T = \begin{cases} Y, & \Delta = 1, \\ Y + 1, & \Delta = 0,\ Y \ge \tau, \\ T_\varepsilon(O), & \Delta = 0,\ Y < \tau, \end{cases}$$

and let Q_ε be the distribution of (T, C, X) when O follows P. In every case, $T \le C$ if and only if $\Delta = 1$, and $\min(T,C) = Y$; hence $Q_\varepsilon \in \mathcal{Q}(P)$. Under Q_ε, $\min(T,\tau) = \min(Y,\tau)$ in the first two cases and $\min(T,\tau) = T_\varepsilon(O)$ in the third. Taking expectations under P therefore gives $\mathrm{IBS}(\tau;Q_\varepsilon) \le \underline{\mathrm{IBS}}(\tau;P) + \varepsilon/\tau$. Together with part (i), the arbitrariness of ε identifies $\underline{\mathrm{IBS}}(\tau;P)$ as the infimum of the identified set.

For the upper bound, on $Y < \tau$ set $\overline{T}_\varepsilon(O) = \tau$ if $L(\tau, X;\tau) \ge L(Y, X;\tau)$ and $\overline{T}_\varepsilon(O) = \min(Y+\varepsilon,\tau)$ otherwise. This choice is measurable, lies in $(Y,\tau]$, and by the Lipschitz bound satisfies $L(\overline{T}_\varepsilon(O), X;\tau) \ge \max\{L(Y,X;\tau), L(\tau,X;\tau)\} - \varepsilon$. Replacing $T_\varepsilon(O)$ by $\overline{T}_\varepsilon(O)$ in the construction gives $\overline{Q}_\varepsilon \in \mathcal{Q}(P)$ with $\mathrm{IBS}(\tau;\overline{Q}_\varepsilon) \ge \overline{\mathrm{IBS}}(\tau;P) - \varepsilon/\tau$. Together with part (i), this identifies $\overline{\mathrm{IBS}}(\tau;P)$ as the supremum of the identified set.

As in the proof of Proposition 2.1(ii), any mixture of two distributions in $\mathcal{Q}(P)$ also induces P and belongs to $\mathcal{Q}(P)$. Since the IBS is the expectation of a fixed individual loss, the IBS of the mixture is the same mixture of their IBS values. Thus every value between two attainable IBS values is also attainable, so the identified set is an interval. Its infimum and supremum established above therefore give its closure as $[\underline{\mathrm{IBS}}(\tau;P), \overline{\mathrm{IBS}}(\tau;P)]$.

To establish the claim in part (ii) that the bounds need not be attained, we give two examples. The lower bound is unattained in the first example, and the upper bound is unattained in the second. Take $\tau = 1$ and let $P(Y = 0, \Delta = 0, X = x_0) = 1$ for a fixed covariate value

x_0. Compatibility is then equivalent to $C = 0$, $T > 0$, and $X = x_0$ almost surely. Thus every compatible Q has $m = \mathbb{E}_Q \min(T, 1) \in (0, 1]$: a strictly positive random variable cannot have expectation zero. Conversely, every $m \in (0, 1]$ is realized by setting $C = 0$ and $T = m$ deterministically.

Both constant predictions below are nonincreasing in t. If $\widehat{S} \equiv 0$, definition (2) gives $L(r, x_0; 1) = r$ and $\mathrm{IBS}(1; Q) = m$. If $\widehat{S} \equiv 1$, it gives $L(r, x_0; 1) = 1 - r$ and $\mathrm{IBS}(1; Q) = 1 - m$. Hence the two identified sets are, respectively,

$$\big\{\mathrm{IBS}(1; Q) : Q \in \mathcal{Q}(P)\big\} = \begin{cases} (0, 1], & \widehat{S} \equiv 0, \\ [0, 1), & \widehat{S} \equiv 1. \end{cases}$$

The lower bound 0 in the first example and the upper bound 1 in the second are therefore unattained under every compatible Q. They are approached by $C = 0$, $T = 1/n$ as the positive integer n tends to infinity.

Part (iii). The definitions of $\underline{\mathrm{IBS}}(\tau; P)$ and $\overline{\mathrm{IBS}}(\tau; P)$ differ only in the term for $Y < \tau$ and $\Delta = 0$, and subtracting them gives the stated width.

Part (iv). By the definition in Proposition 2.1 and Tonelli's theorem, $\int_0^\tau \underline{\mathrm{BS}}(t; P)\, dt$ is the expectation under P of

$$\int_0^\tau \Big[\mathbf{1}(Y > t)\big\{1 - \widehat{S}(t \mid X)\big\}^2 + \mathbf{1}(Y \le t, \Delta = 1)\, \widehat{S}(t \mid X)^2 \\ + \mathbf{1}(Y \le t, \Delta = 0) \min\big\{\widehat{S}(t \mid X)^2, \big\{1 - \widehat{S}(t \mid X)\big\}^2\big\}\Big] dt.$$

Tonelli's theorem applies because the integrand is nonnegative and measurable in (t, O). If $Y \ge \tau$, only the first term remains on $[0, \tau)$, so this integral equals $L(\tau, X; \tau)$ by (2). If $Y < \tau$ and $\Delta = 1$, the first term remains on $[0, Y)$ and the second on $[Y, \tau]$, so the integral equals $L(Y, X; \tau)$. If $Y < \tau$ and $\Delta = 0$, the first term remains on $[0, Y)$ and the minimum on $[Y, \tau]$, so the integral is the middle expression in (6) with $y = Y$ and $x = X$, which equals $L(r^*(Y, X), X; \tau)$. These three cases agree with the expression defining $\underline{\mathrm{IBS}}(\tau; P)$, which proves the equality in part (iv).

The same calculation for $\overline{\mathrm{BS}}$, with the minimum replaced by the maximum, changes only the contribution for $Y < \tau$ and $\Delta = 0$. That contribution becomes the integral of $\{1 - \widehat{S}(t \mid X)\}^2$ over $[0, Y)$ plus the integral of the larger of the two candidate losses over $[Y, \tau]$. The corresponding term of $\overline{\mathrm{IBS}}(\tau; P)$ is $\max\{L(Y, X; \tau), L(\tau, X; \tau)\}$. By (2), $L(Y, X; \tau)$ and $L(\tau, X; \tau)$ share the same integral over $[0, Y)$ and integrate $\widehat{S}(t \mid X)^2$ and $\{1 - \widehat{S}(t \mid X)\}^2$, respectively, over $[Y, \tau]$. Since the two candidate losses differ by $\{1 - \widehat{S}(t \mid X)\}^2 - \widehat{S}(t \mid X)^2 = 1 - 2\widehat{S}(t \mid X)$, the larger one exceeds $\widehat{S}(t \mid X)^2$ by $\{1 - 2\widehat{S}(t \mid X)\}_+$ and exceeds $\{1 - \widehat{S}(t \mid X)\}^2$ by $\{2\widehat{S}(t \mid X) - 1\}_+$. Integrating over $[Y, \tau]$, the integrated pointwise contribution therefore exceeds $L(Y, X; \tau)$ by $\int_Y^\tau \{1 - 2\widehat{S}(t \mid X)\}_+\, dt$ and exceeds $L(\tau, X; \tau)$ by $\int_Y^\tau \{2\widehat{S}(t \mid X) - 1\}_+\, dt$. Since $c - \max\{u, v\} = \min\{c - u, c - v\}$ for real c, u, v, the difference between the two contributions is the smaller of these two integrals, which is $g(Y, X)$.

Multiplying by $\mathbf{1}(Y < \tau, \Delta = 0)$, taking expectations under P, and dividing by τ give the stated difference. The middle inequality in part (iv) is $\underline{\mathrm{IBS}}(\tau; P) \le \overline{\mathrm{IBS}}(\tau; P)$, which follows from (6) with $r = \tau$. Since $g \ge 0$, the stated difference gives the last inequality. The difference is τ^{-1} times the expectation of the nonnegative random variable $\mathbf{1}(Y < \tau, \Delta = 0)\, g(Y, X)$, so it

is positive if and only if $P(Y < \tau, \Delta = 0, g(Y, X) > 0) > 0$. □

Proof of Proposition 2.3. Preliminaries. The function $x \mapsto D_{jk}(r, x)$ is measurable because each model's prediction is measurable. Also, $|D_{jk}(r, x)| \leq \tau$ because $0 \leq L_j(r, x; \tau), L_k(r, x; \tau) \leq \tau$, so the expectations below are finite. By the preliminaries in the proof of Proposition 2.2, each model's loss is Lipschitz in r with constant one, so $D_{jk}(r, x)$ is continuous in r. Unlike L_j, however, D_{jk} need not be convex in r even though both predictions are nonincreasing. We therefore approximate its infimum and supremum through a countable set of candidates.

Part (i). Let $Q \in \mathcal{Q}(P)$. Then O follows P under Q, and by the representation in Section 2.1, $\mathrm{IBS}_{jk}(\tau; Q) = \tau^{-1}\mathbb{E}_Q D_{jk}(\min(T, \tau), X)$. If $\Delta = 1$, or if $\Delta = 0$ and $Y \geq \tau$, then $\min(T, \tau) = \min(Y, \tau)$, so both $\underline{d}_{jk}(O)$ and $\overline{d}_{jk}(O)$ equal $\tau^{-1}D_{jk}(\min(T, \tau), X)$. If $Y < \tau$ and $\Delta = 0$, then $\min(T, \tau) \in (Y, \tau]$, so $D_{jk}(\min(T, \tau), X)$ lies between the infimum and the supremum of $D_{jk}(r, X)$ over this interval. Hence, Q-almost surely,

$$\underline{d}_{jk}(O) \leq \frac{1}{\tau} D_{jk}(\min(T, \tau), X) \leq \overline{d}_{jk}(O).$$

Since $\underline{d}_{jk}$ and $\overline{d}_{jk}$ are functions of O alone, taking expectations under Q and replacing them with expectations under P give part (i).

Part (ii). Fix $\varepsilon > 0$. By continuity of D_{jk} in r, the infimum over $(Y, \tau]$ can be computed using a countable set of candidates. Enumerate $(\mathbb{Q} \cap [0, \tau]) \cup \{\tau\}$ as $q_1, q_2, \ldots$, and call a candidate q_h admissible if $q_h > Y$. On $Y < \tau$, the admissible candidates are dense in $(Y, \tau]$ and include τ, so every $r \in (Y, \tau]$ is approached by admissible candidates whose values of D_{jk} converge to $D_{jk}(r, X)$. For each h, let $V_h = D_{jk}(q_h, X)$ if $q_h > Y$ and $V_h = +\infty$ otherwise. Each V_h is measurable, and on $Y < \tau$,

$$\inf_{r \in (Y, \tau]} D_{jk}(r, X) = \inf_{h \geq 1} V_h,$$

which is measurable as the infimum of countably many measurable functions and finite because $|D_{jk}| \leq \tau$. By this equality, some V_h is smaller than $\inf_{r \in (Y, \tau]} D_{jk}(r, X) + \varepsilon$; such a V_h is finite, so its candidate q_h is admissible.

On $Y < \tau$, let $r_\varepsilon(O)$ be the first admissible q_h such that $D_{jk}(q_h, X) < \inf_{r \in (Y, \tau]} D_{jk}(r, X) + \varepsilon$, and set $r_\varepsilon(O) = \tau$ on $Y \geq \tau$. The event that index h is selected is the measurable event that q_h satisfies this inequality and every earlier candidate fails it or is inadmissible, so r_ε is measurable. Given O, define $C = Y$ and

$$T = \begin{cases} Y, & \Delta = 1, \\ Y + 1, & \Delta = 0,\ Y \geq \tau, \\ r_\varepsilon(O), & \Delta = 0,\ Y < \tau, \end{cases}$$

and let Q_ε be the distribution of (T, C, X) when O follows P. In every case, $T \leq C$ if and only if $\Delta = 1$, and $\min(T, C) = Y$; hence $Q_\varepsilon \in \mathcal{Q}(P)$. This single value of T is shared by both models, and taking expectations under P gives $\mathrm{IBS}_{jk}(\tau; Q_\varepsilon) \leq \underline{\mathrm{IBS}}_{jk}(\tau; P) + \varepsilon/\tau$. Together with part (i), the arbitrariness of ε identifies $\underline{\mathrm{IBS}}_{jk}(\tau; P)$ as the infimum of the identified set.

For the upper bound, use the same candidates, assign $-\infty$ to inadmissible candidates, and choose the first admissible q_h with $D_{jk}(q_h, X) > \sup_{r \in (Y, \tau]} D_{jk}(r, X) - \varepsilon$ on $Y < \tau$. The same measurability and compatibility arguments give $\overline{Q}_\varepsilon \in \mathcal{Q}(P)$ with $\mathrm{IBS}_{jk}(\tau; \overline{Q}_\varepsilon) \geq \overline{\mathrm{IBS}}_{jk}(\tau; P) - \varepsilon/\tau$, which identifies $\overline{\mathrm{IBS}}_{jk}(\tau; P)$ as the supremum. Mixtures of distributions compatible with P still induce P, and the IBS difference of a mixture is the same mixture of their IBS differences.

Hence the identified set is an interval, and its closure is the stated closed interval.

Non-attainment examples. To establish the claim in part (ii) that the bounds need not be attained, we give two examples. The lower bound is unattained in the first example, and the upper bound is unattained in the second. Take $\tau = 1$, $P(Y = 0, \Delta = 0, X = x_0) = 1$, and $\widehat{S}_k \equiv 1/2$. Every compatible Q has $C = 0$, $T > 0$, and $m = \mathbb{E}_Q \min(T, 1) \in (0, 1]$, and every such m is realized by $T = m$, $C = 0$. Definition (2) gives $L_k(r, x_0; 1) = 1/4$. Taking $\widehat{S}_j \equiv 0$ gives $D_{jk}(r, x_0) = r - 1/4$ and $\mathrm{IBS}_{jk}(1; Q) = m - 1/4$. Taking $\widehat{S}_j \equiv 1$ instead gives $D_{jk}(r, x_0) = 3/4 - r$ and $\mathrm{IBS}_{jk}(1; Q) = 3/4 - m$. Thus

$$\big\{\mathrm{IBS}_{jk}(1; Q) : Q \in \mathcal{Q}(P)\big\} = \begin{cases} (-1/4, 3/4], & \widehat{S}_j \equiv 0, \\ [-1/4, 3/4), & \widehat{S}_j \equiv 1. \end{cases}$$

The lower bound $-1/4$ in the first example and the upper bound $3/4$ in the second are not attained by any compatible Q, but are approached by $C = 0$, $T = 1/n$ as $n \to \infty$.

Part (iii). The definitions of $\underline{d}_{jk}(O)$ and $\overline{d}_{jk}(O)$ differ only in the term for $Y < \tau$ and $\Delta = 0$, and subtracting their expectations gives the stated width.

Part (iv). By part (ii) and Proposition 2.2(ii) applied to models j and k, and since $\inf(f - g) \geq \inf f - \sup g$ on a common domain,

$$\begin{aligned} \underline{\mathrm{IBS}}_{jk}(\tau; P) &= \inf_{Q \in \mathcal{Q}(P)} \big\{\mathrm{IBS}_j(\tau; Q) - \mathrm{IBS}_k(\tau; Q)\big\} \\ &\geq \inf_{Q \in \mathcal{Q}(P)} \mathrm{IBS}_j(\tau; Q) - \sup_{Q \in \mathcal{Q}(P)} \mathrm{IBS}_k(\tau; Q) = \underline{\mathrm{IBS}}_j(\tau; P) - \overline{\mathrm{IBS}}_k(\tau; P). \end{aligned}$$

The inequality for the upper bound follows in the same way from $\sup(f - g) \leq \sup f - \inf g$. The middle inequality follows directly from $\inf_{r \in (Y, \tau]} D_{jk}(r, X) \leq \sup_{r \in (Y, \tau]} D_{jk}(r, X)$ in the individual contributions. □

Proof of Proposition 2.5. By Proposition 2.3(i), every $Q \in \mathcal{Q}(P)$ satisfies

$$M_j(\tau; Q) = \max_{k \neq j} \mathrm{IBS}_{jk}(\tau; Q) \leq \max_{k \neq j} \overline{\mathrm{IBS}}_{jk}(\tau; P) = \overline{M}_j(\tau; P).$$

For each $k \neq j$, the inequality $M_j(\tau; Q) \geq \mathrm{IBS}_{jk}(\tau; Q)$ and Proposition 2.3(ii) give

$$\sup_{Q \in \mathcal{Q}(P)} M_j(\tau; Q) \geq \sup_{Q \in \mathcal{Q}(P)} \mathrm{IBS}_{jk}(\tau; Q) = \overline{\mathrm{IBS}}_{jk}(\tau; P).$$

Maximizing the right-hand side over $k \neq j$ gives the reverse inequality and proves sharpness. □

B Proofs for Section 3

Proof of Proposition 3.1. Parts (i) and (ii). The individual contributions are measurable and bounded, as established in Propositions 2.2 and 2.3. Each estimator is the sample average of independent and identically distributed copies of the corresponding contribution. Its expectation is the bound being estimated, so the strong law of large numbers proves parts (i) and (ii).

Part (iii). By the definitions of the upper bound and its estimator,

$$\left|\widehat{\overline{M}}_j(\tau) - \overline{M}_j(\tau; P)\right| \leq \max_{k \neq j} \left|\widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) - \overline{\mathrm{IBS}}_{jk}(\tau; P)\right|.$$

Because the number of competitors is finite, part (ii) makes the right-hand side converge to zero almost surely. □

Proof of Theorem 3.2. The contributions are bounded, as noted at the beginning of Section 3.2, so they have finite second moments. Since the estimators are sample averages of independent and identically distributed random vectors, parts (i) and (ii) follow from the multivariate central limit theorem. Indeed, the scalar central limit theorem applies to every fixed linear combination of the contributions; if that combination has variance zero, its centered value is zero almost surely. The Cramér–Wold argument therefore also covers singular covariance matrices. Part (iii) follows from the strong law of large numbers applied to the contributions and to their products. □

Proof of Theorem 3.3. The vector of pairwise upper-bound estimators is a sample average of independent and identically distributed bounded vectors. The multivariate central limit theorem therefore gives the stated normal limit and covariance matrix, including when that matrix is singular.

Any competitor outside $\mathcal{K}_j^*$ has a population upper bound strictly smaller than $\overline{M}_j(\tau; P)$. There are finitely many such competitors, and Proposition 3.1(ii) gives simultaneous almost sure convergence of all pairwise upper-bound estimators. Consequently, with probability one, no competitor outside $\mathcal{K}_j^*$ maximizes the estimated upper bounds once n is sufficiently large. It follows that, eventually almost surely,

$$\sqrt{n}\left\{\widehat{\overline{M}}_j(\tau) - \overline{M}_j(\tau; P)\right\} = \max_{k \in \mathcal{K}_j^*} \sqrt{n}\left\{\widehat{\overline{\mathrm{IBS}}}_{jk}(\tau) - \overline{\mathrm{IBS}}_{jk}(\tau; P)\right\}.$$

The maximum over this fixed finite set is a continuous function, so the continuous mapping theorem proves the result. □

Proof of Proposition 3.4. We first show that the estimated standard deviations and correlation converge to their population counterparts. By Theorem 3.2(iii) and continuity of the square root, $\widehat{\underline{\sigma}} \to \underline{\sigma} > 0$ and $\widehat{\overline{\sigma}} \to \overline{\sigma} > 0$ almost surely. Since almost sure convergence implies convergence in probability, the probability that $\widehat{\underline{\sigma}} > 0$ and $\widehat{\overline{\sigma}} > 0$ tends to one, and the convention $\mathcal{C}_{1-\alpha} = \mathbb{R}$ does not affect the limit. Correlation is a continuous function of the covariance matrix when both marginal variances are positive, so $\widehat{\rho} \to \rho$ almost surely as well.

Next, we reduce coverage to an event about the larger of the two standardized errors. On the event that $\widehat{\underline{\sigma}} > 0$ and $\widehat{\overline{\sigma}} > 0$, $[\underline{\theta}(P), \overline{\theta}(P)] \subseteq \mathcal{C}_{1-\alpha}$ holds if and only if

$$\max\left\{\frac{\sqrt{n}(\widehat{\underline{\theta}} - \underline{\theta}(P))}{\widehat{\underline{\sigma}}}, -\frac{\sqrt{n}(\widehat{\overline{\theta}} - \overline{\theta}(P))}{\widehat{\overline{\sigma}}}\right\} \leq c_{1-\alpha}(\widehat{\rho}).$$

By Theorem 3.2 and Slutsky's theorem, the vector inside the maximum converges in distribution to $(Z_1, -Z_2)$, where (Z_1, Z_2) has standard normal margins and correlation ρ. For each $\rho \in [-1, 1]$, let F_ρ be the distribution function of $\max\{Z_1, -Z_2\}$. It therefore remains to show that F_ρ is continuous and that $c_{1-\alpha}(\widehat{\rho}) \to c_{1-\alpha}(\rho)$ almost surely.

For each $\rho \in [-1,1]$, F_ρ is continuous on $\mathbb{R}$ and satisfies $F_\rho(c_{1-\alpha}(\rho)-\varepsilon) < 1-\alpha < F_\rho(c_{1-\alpha}(\rho)+\varepsilon)$ for every $\varepsilon > 0$. For $|\rho| < 1$, both properties hold because (Z_1, Z_2) has a positive density on $\mathbb{R}^2$. At $\rho = 1$, we have $Z_2 = Z_1$, so the maximum equals $|Z_1|$, whose distribution function is continuous on $\mathbb{R}$ and strictly increasing on $(0,\infty)$, which contains $c_{1-\alpha}(1) > 0$. At $\rho = -1$, we have $Z_2 = -Z_1$, so the maximum equals Z_1, which is standard normal.

These two properties make $c_{1-\alpha}(\cdot)$ continuous on $[-1,1]$. To compare different values of ρ, including $\rho = \pm 1$, write $(Z_1, Z_2) = (X, \rho X + \sqrt{1-\rho^2}Y)$ with independent standard normal X and Y. For fixed X and Y, $\max\{X, -\rho X - \sqrt{1-\rho^2}Y\}$ is continuous in ρ, so $\rho' \to \rho$ implies $F_{\rho'} \to F_\rho$ pointwise, by continuity of F_ρ. Evaluating at $c_{1-\alpha}(\rho) \pm \varepsilon$ shows that $F_{\rho'}(c_{1-\alpha}(\rho)-\varepsilon) < 1-\alpha < F_{\rho'}(c_{1-\alpha}(\rho)+\varepsilon)$ for ρ' close to ρ, and hence $|c_{1-\alpha}(\rho') - c_{1-\alpha}(\rho)| \le \varepsilon$. Since $\widehat{\rho} \to \rho$ almost surely, $c_{1-\alpha}(\widehat{\rho}) \to c_{1-\alpha}(\rho)$ almost surely, and the continuous mapping theorem and Slutsky's theorem give the limit $\Pr(\max\{Z_1, -Z_2\} \le c_{1-\alpha}(\rho)) = 1-\alpha$. □

Proof of Proposition 3.5. All three parts follow from the asymptotic normality of the estimated upper bound on the IBS difference. Assumption 4 fixes j, k, and τ before the evaluation sample is examined, so the following limits apply to a fixed pair and horizon. Write $\theta = \overline{\mathrm{IBS}}_{jk}(\tau; P)$, $\widehat{\theta} = \widehat{\overline{\mathrm{IBS}}}_{jk}(\tau)$, and $\sigma = \overline{\sigma}_{jk}$. By Theorem 3.2, $\sqrt{n}(\widehat{\theta} - \theta)$ converges in distribution to σZ with Z standard normal, and $\widehat{\overline{\sigma}}_{jk} \to \sigma$ almost surely.

Part (i). We bound the rejection probability under H_0, treating the cases $\sigma > 0$ and $\sigma = 0$ separately. Under $\theta \ge 0$, the event $U_{jk} < 0$ implies $\sqrt{n}(\widehat{\theta} - \theta) + z_{1-\alpha}\widehat{\overline{\sigma}}_{jk} < -\sqrt{n}\,\theta \le 0$. Suppose first that $\sigma > 0$. The left-hand side converges in distribution to $\sigma(Z + z_{1-\alpha})$, whose distribution is continuous at zero, so $\limsup_n \Pr\{U_{jk} < 0\} \le \Pr\{\sigma(Z + z_{1-\alpha}) < 0\} = \alpha$. Suppose next that $\sigma = 0$. Since $\theta = \mathbb{E}_P\, \overline{d}_{jk}(O)$, this means $\mathbb{E}_P[\{\overline{d}_{jk}(O) - \theta\}^2] = \sigma^2 = 0$. A nonnegative random variable with mean zero equals zero almost surely, so $\overline{d}_{jk}(O) = \theta$ almost surely. Hence $\widehat{\theta} = \theta$ and $\widehat{\overline{\sigma}}_{jk} = 0$ almost surely, and $U_{jk} = \theta \ge 0$, so H_0 is never rejected.

Part (ii). Under H_1, the upper confidence bound converges to the negative value θ, so the test rejects with probability tending to one. Proposition 3.1(ii) gives $\widehat{\theta} \to \theta$ almost surely, and Theorem 3.2(iii) gives $\widehat{\overline{\sigma}}_{jk}/\sqrt{n} \to 0$ almost surely. Hence $U_{jk} \to \theta < 0$ almost surely, so $\Pr\{U_{jk} < 0\} \to 1$.

Part (iii). The coverage of U_{jk} follows from the same limit as in Part (i). The event $\theta \le U_{jk}$ is equivalent to $\sqrt{n}(\widehat{\theta} - \theta) + z_{1-\alpha}\widehat{\overline{\sigma}}_{jk} \ge 0$. If $\sigma > 0$, its probability converges to $\Pr\{\sigma(Z + z_{1-\alpha}) \ge 0\} = 1 - \alpha$. If $\sigma = 0$, then $U_{jk} = \theta$ almost surely, and the probability is one. □

Proof of Proposition 3.6. Each part follows from Proposition 3.5 applied to the pairwise statistics U_{jk}. By Assumption 4, j, the set of compared models, and τ do not depend on the evaluation sample, so each U_{jk} is the statistic in Proposition 3.5 for a fixed pair and horizon.

Part (i). Under H_0, the IUT can reject only if the pairwise test against a competitor that satisfies the null hypothesis rejects. By Proposition 2.5, the condition $\overline{M}_j(\tau; P) \ge 0$ is equivalent to H_0, so there is $k_0 \ne j$ with $\overline{\mathrm{IBS}}_{jk_0}(\tau; P) \ge 0$. The event $U_j^{\mathrm{IUT}} < 0$ implies $U_{jk_0} < 0$, so Proposition 3.5(i) gives $\limsup_n \Pr\{U_j^{\mathrm{IUT}} < 0\} \le \limsup_n \Pr\{U_{jk_0} < 0\} \le \alpha$. This argument does not depend on how many competitors satisfy the null hypothesis or on the correlation among the estimators.

Part (ii). Under H_1, every pairwise test rejects with probability tending to one, and so does the IUT. If $\overline{M}_j(\tau; P) < 0$, then $\overline{\mathrm{IBS}}_{jk}(\tau; P) < 0$ for every $k \ne j$. Proposition 3.5(ii) applied to the finitely many competitors therefore gives $\Pr\{U_{jk} < 0 \text{ for all } k \ne j\} \to 1$, which is $\Pr\{U_j^{\mathrm{IUT}} < 0\} \to 1$.

Part (iii). The bound U_j^{IUT} covers $\overline{M}_j(\tau; P)$ whenever the pairwise bound for a maximizing competitor covers its own target. Fix $k^* \in \mathcal{K}_j^*$, the set of maximizing competitors defined before Theorem 3.3. Since $U_j^{\mathrm{IUT}} \geq U_{jk^*}$ and $\overline{M}_j(\tau; P) = \overline{\mathrm{IBS}}_{jk^*}(\tau; P)$, the event $\overline{\mathrm{IBS}}_{jk^*}(\tau; P) \leq U_{jk^*}$ implies $\overline{M}_j(\tau; P) \leq U_j^{\mathrm{IUT}}$, and Proposition 3.5(iii) gives the result. □