Nonparametric Methods for Integration of Survival Analysis and Machine Learning

Charles Wolock

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Marco Carone, Chair
Noah Simon, Chair
Ting Ye

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Nonparametric Methods for Integration
of Survival Analysis and Machine Learning

Charles Wolock

Co-Chairs of the Supervisory Committee:
Marco Carone
Department of Biostatistics

Noah Simon
Department of Biostatistics

This dissertation develops practical methodology incorporating modern machine learning techniques into statistical inference, with a particular focus on the analysis of time-to-event data. Time-to-event data are commonly encountered in biomedical studies, where incomplete follow-up and truncation-induced sampling bias may preclude the use of standard analysis procedures. The primary intended application of this work is variable importance, although the methods developed here are appropriate for a wider range of problems. Chapter 1 serves as an introduction to the dissertation. The three methodological chapters overlap but function as distinct, standalone units. In Chapter 2, we propose an algorithm-agnostic, nonparametric procedure for assessing variable importance for right-censored time-to-event outcomes. In the Chapter 3, we develop a framework in which arbitrary machine learning algorithms can be applied to estimate personalized survival curves from data subject to both censoring and truncation. Chapter 4 addresses the use of sample splitting to provide inference on variable importance when the true importance lies on the boundary of the parameter space.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Marco Carone and Noah Simon, for their endless patience, encouragement, and statistical insight. I cannot imagine having done this work without them. Thanks to Ting Ye, Yates Coley, and Christine Khosropour for serving on my committee. I have been lucky to learn from many dedicated teachers here at UW, and to collaborate with many brilliant colleagues. I have learned so much from all of them.

I could never fully express my gratitude to my cohort: Si Cheng, Avi Kenny, Pearl Liu, Taylor Okonek, and María Valdez Cabrera (and Ian Waudby-Smith, unofficially). The ups and downs of graduate school have been made much more bearable because of them. Many thanks to Minh Vo and Gitana Garofalo for taking care of the details along the way.

I am indebted to Andrew Allen for encouraging me to study biostatistics and helping me to believe that I was capable of doing so. Thanks to my parents and siblings for being by my side throughout the last 22 years, more or less, of academic study. Thanks to my grandmother, Isabel Wolock, for setting an example for me to follow. Thanks to the friends from all stages of my life, who have provided welcome and deeply fulfilling distractions from graduate work.

Finally, thanks to Becca, whose support has meant the most of all.

DEDICATION

To Becca.

Chapter 1
INTRODUCTION

The analysis of time-to-event data presents a unique set of complications to scientific investigators. As a result, survival analysis often involves the use of a specific analytic toolkit, with estimators and inferential procedures tailored to handle data structures commonly encountered in the time-to-event setting (e.g., Kaplan and Meier, 1958; Cox, 1972). Perhaps the quintessential feature of survival data in prospective biomedical studies is right censoring, in which a participant may not experience the event of interest during the follow-up period. There are many potential causes of right censoring, including loss to follow-up and termination of the study. When right censoring is present, great care must be taken to determine the conditions under which a quantity of scientific interest is, in fact, identified by the observed data. Even a simple quantity such as the mean survival time may require a strong set of assumptions in order to be identified, which may be unrealistic outside of specific settings (Ding and Nan, 2015). In some study designs, censoring may be accompanied by truncation, a form of biased sampling in a participant may be ineligible for recruitment into a study if their event time falls above or below certain values. Truncation induces selection bias.

This dissertation is largely focused on the task of prediction within the context of survival data. In recent decades, machine learning methods have been been at the forefront of predictive modeling for many data types, including time-to-event data. The literature on survival-specific machine learning methods is extensive — see Wang et al. (2019) and Sonabend (2021) for comprehensive reviews. For a particular prediction task, given a set of available features, it may be of interest to assess how important a specific feature or group of features is in predicting the outcome of interest. Variable importance may be viewed as

an algorithm-specific quantity, i.e., a measure that is tied to a fitted algorithm, or as an population quantity, i.e., a measure that describes the intrinsic relationship between features and outcome under a particular data-generating mechanism. There exists a diverse array of methods for estimating variable importance for fixed algorithms (Breiman, 2001; Lundberg and Lee, 2017; Fisher et al., 2019; Murdoch et al., 2019). There are also model-agnostic approaches that nevertheless estimate a quantity that conditions on the training data (Gan et al., 2022). Our work concerns model-agnostic, population-level variable importance, which is sometimes referred to simply as intrinsic variable importance (Williamson et al., 2021b). Methods for estimating specific types of intrinsic importance measures date back at least to work by (van der Laan, 2006) and recently have rapidly grown in number (Lei et al., 2018; Williamson et al., 2021a; Zhang and Janson, 2022; Boileau et al., 2023).

In Chapter 2 of this dissertation, we propose a framework for leveraging machine learning tools to estimate and make inference on a broad class of intrinsic variable importance measures in the setting of right-censored time-to-event data. We apply our method to estimate the importance of features used to predict the probability of seroconversion in an HIV vaccine clinical trial. Specific difficulties that arise within the broad problem of variable importance for survival data motivate the other two chapters of the dissertation. In Chapter 3, we present a flexible procedure for estimating the conditional distribution of a time-to-event outcome subject to both censoring and truncation, which is a form of biased sampling. Our proposed procedure allows use of off-the-shelf machine learning tools not specially designed for survival analysis. In Chapter 4, we discuss the use of sample splitting for inference on intrinsic variable importance. Sample splitting entails randomly partitioning the data set into non-overlapping segments, each of which is used to estimate a different component of an overall estimator. For variable importance, sample splitting can be used to overcome the degeneracy that occurs in the debiased machine learning method from Chapter 2 when the feature of interest has zero importance. We investigate the question of whether the drawbacks of sample splitting can be mitigated by aggregating the results of multiple iterations of the procedure. The appendices contain all technical details, as well as simulation and data

analysis results not included in the main text.

Chapter 2

NONPARAMETRIC VARIABLE IMPORTANCE FOR TIME-TO-EVENT OUTCOMES

Charles J. Wolock, Peter B. Gilbert, Noah Simon & Marco Carone

2.1 Introduction

Statistical tasks often involve applying an algorithm to learn the relationship between a set of features and an outcome, and subsequently leveraging that information to make predictions about future observations. As algorithms have become more complex, it has in turn become more difficult to understand how individual features contribute to those predictions. The contribution of a feature or set of features to a prediction algorithm is known as variable importance. The proliferation of machine learning algorithms has given rise to a substantial literature focused on defining, estimating, and making inference on variable importance. In this work, we focus on intrinsic variable importance: the population-level predictiveness potential of a feature (Williamson et al., 2021b). This differs from extrinsic variable importance, which is aimed at understanding how a particular, fixed algorithm uses features to make predictions. While extrinsic approaches can add interpretability to the fitted prediction algorithm, they may not reflect the intrinsic importance of the feature in a general prediction setting.

The question of intrinsic variable importance arises naturally in the context of HIV vaccine trials. To achieve desired statistical power in efficacy trials, trial administrators often aim to recruit individuals who are thought to have a high probability of HIV seroconversion during the intended follow-up period. HIV seroconversion prediction models have been proposed in various populations (see Menza et al., 2009; Smith et al., 2012; Balkus et al., 2016;

Wand et al., 2018). Previous analyses have identified several features associated with the time to seroconversion, ranging from demographic information (e.g., age) to laboratory assays (e.g., prevalent sexually transmitted infection (STI)) to behavioral questionnaire responses (e.g., number of sex partners over a specified time frame). Understanding the importance of these features may inform participant recruitment.

HVTN 702 was a phase 2b-3 trial conducted in South Africa to investigate the safety and efficacy of a vaccine regimen consisting of a recombinant canarypox vector containing HIV-1 subtype C envelope ALVAC-HIV and an MF59-adjuvanted subtype C bivalent glycoprotein 120 vaccine (Gray et al., 2021). The trial began in October 2016 and was terminated in 2020 when prespecified nonefficacy stopping criteria were met. Over this time period, around 5% of subjects experienced seroconversion, while around 7% were lost to follow-up. Both loss to follow-up and study termination may preclude observation of the event time, resulting in right censoring. Right censoring is ubiquitous in prospective biomedical studies of time-to-event outcomes. When right censoring is present, extra care must be taken in evaluating the performance of a prediction algorithm, since outcomes and predictions cannot be directly compared. The literature on model-free, algorithm-agnostic variable importance measures (VIMs) (e.g., Williamson et al., 2021b; Verdinelli and Wasserman, 2021; Zhang and Janson, 2022) has largely bypassed survival data and censoring, although specific measures for variable importance have recently been proposed for evaluating treatment effect modification using right-censored data (Boileau et al., 2023).

Intrinsic variable importance relies on the concept of a predictiveness measure, which quantifies the performance of a prediction algorithm. A natural first step in developing a statistical framework for variable importance in survival analysis is to define a population predictiveness measure in terms of the full or ideal data distribution. The estimation of such ideal data predictiveness measures, and associated VIMs, is complicated by censoring. The available data are not sampled from the ideal data distribution, but rather from an observed data distribution implied by the censoring mechanism. Many widely used methods for evaluating algorithm predictiveness in survival analysis fail to properly account for censoring and

may converge to population parameters that (undesirably) depend on the censoring mechanism. Those that do account for censoring often require strong assumptions on the censoring mechanism or event time distribution. For example, many predictiveness estimation methods are based on the popular semiparametric Cox proportional hazards model (Cox, 1972), which, while convenient, is likely to be misspecified in many applications. The growing use of machine learning prediction models in survival analysis motivates the development of correspondingly flexible methods for evaluating variable importance.

In this chapter, we present a general framework for estimating variable importance using informatively right-censored time-to-event data. Our approach allows us us to construct efficient estimators and perform statistical inference. Our specific contributions include the following.

1. We define a class of predictiveness measures for time-to-event outcomes, encompassing many measures used in practice. These measures are identified under covariate-induced censoring.

2. We derive the nonparametric efficient influence function for measures in this class.

3. We propose a debiased estimation procedure, including an implementation employing cross-fitting, that gives an asymptotically linear estimator of variable importance. We show how to use this estimator to make inference.

4. We analyze variable importance for predicting time to seroconversion in HVTN 702.

The chapter is organized as follows. In Section 2.2, we review the concepts underpinning predictiveness and intrinsic variable importance. In Section 2.3, we describe a class of survival VIMs and provide identification results under right censoring. In Section 2.4, we outline procedures for estimating and performing inference on survival VIMs and provide theoretical results. In Section 2.5, we perform experiments to evaluate the performance of our proposed procedures. In Section 2.6, we analyze variable importance in the HVTN 702 study. In

Section 2.7, we provide concluding remarks. Appendix A contains all technical details, as well as additional simulation results.

2.2 Variable importance

2.2.1 Data structure and notation

We observe a vector of baseline covariates X taking values in $\mathcal{X} \subset \mathbb{R}^p$. Interest lies in studying the time between an initiating event (e.g., randomization into the HVTN 702 study) and a terminating event (e.g., diagnosis of HIV seroconversion). The time between initiating and terminating events, referred to as the event time, is denoted $T \in (0, \infty)$. The ideal data unit is therefore $Z^* := (X, T)$. We use P_0^* to denote the distribution of Z^* and assume P_0^* belongs to a nonparametric model $\mathcal{M}^*$.

We observe a version of Z^* subject to right censoring, by which follow-up on a participant may conclude before the participant has experienced the event, potentially due to loss-to-follow-up or termination of the study. Let $C \in (0, \infty)$ denote the time between the initiating event and censoring. For each participant, we observe $Y := \min\{T, C\}$, the observed follow-up time, and $\Delta := \mathbb{1}(T \leq C)$, the event indicator, resulting in observed data unit $Z := (X, Y, \Delta)$. Our sample consists of n independent and identically distributed observations $Z_1, \ldots, Z_n$ drawn from P_0, the observed data distribution implied by P_0^* and the censoring mechanism. We use $\mathcal{M}$ to denote the observed data model in which P_0 lies.

The subscript 0 denotes a functional of P_0, e.g., $\mathbb{E}_0[f(Z)] := \mathbb{E}_{P_0}[f(Z)]$. The addition of the superscript $*$ denotes a functional of P_0^*, e.g., $\mathbb{E}_0^*[f(Z^*)] := \mathbb{E}_{P_0^*}[f(Z^*)]$. The expectation of a random function of a data unit, e.g., $f_n(Z)$, is taken with respect to the random data unit Z and not f_n. We use $\mathbb{P}_n$ to denote the empirical measure of $Z_1, \ldots, Z_n$. We sometimes use the empirical process notation $Pf := E_P[f(Z)]$ for probability measure P and P-measurable function f.

We use $a \wedge b$ to denote $\min\{a, b\}$. For vectors $u = (u_1, \ldots, u_p)$ and $v = (v_1, \ldots, v_p)$, we take inequalities to be component-wise, i.e., $\{u \leq v\} = \{u_1 \leq v_1, \ldots, u_p \leq v_p\}$. We use

$\|\cdot\|_\infty$, $\|\cdot\|_v$, and $\|\cdot\|_v^*$ to denote the supremum norm, variation norm, and uniform sectional variation norm, respectively. We use $s \subset \{1, \ldots, p\}$ to denote the index set of a covariate subgroup. For any vector v, we use to v_s to denote the elements of v with index in s and v_{-s} the elements of v with index not in s. The sample spaces of X_s and X_{-s} are denoted by $\mathcal{X}_s$ and $\mathcal{X}_{-s}$, respectively.

We use $\mathcal{F}$ to denote the class of potential prediction functions, endowed with norm $\|\cdot\|_{\mathcal{F}}$. The functions in $\mathcal{F}$ map from $\mathcal{X}$ to a context-specific codomain $\mathcal{Y}$, depending on the predictiveness measure being studied. The subset $\mathcal{F}_s := \{f \in \mathcal{F} : f(u) = f(v) \text{ for all } u, v \in \mathcal{X} \text{ satisfying } u_{-s} = v_{-s}\}$ characterizes prediction functions in $\mathcal{F}$ that ignore features with index in s. We allow $\mathcal{F}$ to be largely unrestricted up to regularity conditions.

2.2.2 Predictiveness and variable importance

In this section we give a brief overview of predictiveness and variable importance. We define $\mathbb{V}(f, P^*)$ to be the ideal data predictiveness measure; it quantifies how predictive $f \in \mathcal{F}$ is under P^*, with larger values indicating higher predictiveness. The oracle prediction function f_0 corresponding to the true data-generating mechanism P_0^* is defined by

$$f_0 \in \text{argmax}_{f \in \mathcal{F}} \mathbb{V}(f, P_0^*) \ .$$

The oracle prediction function represents the optimal prediction function possible under P_0^*, as measured by $\mathbb{V}$. If $\mathcal{F}$ is sufficiently rich, the oracle should not depend on the choice of function class. The oracle predictiveness is then defined as $v_0 := \mathbb{V}(f_0, P_0^*)$, which measures the combined predictive potential of X under P_0^*. We analogously define the residual oracle predictiveness of X_s as $v_{0,s} := \mathbb{V}(f_{0,s}, P_0^*)$, where $f_{0,s} \in \text{argmax}_{f \in \mathcal{F}_s} \mathbb{V}(f, P_0^*)$ is the residual oracle prediction function. This quantifies the combined predictive potential of X_{-s}.

For $r \subset s$ a strict subset of s, we define $\psi_{0,s-r}$ as the intrinsic importance of $X_{s \setminus r}$ relative to X_r, i.e., $\psi_{0,s-r} = v_{0,r} - v_{0,s}$. This is the decrease in oracle predictiveness when s is excluded compared to when only r is excluded. Without loss of generality, we focus on $\psi_{0,s} := \psi_{0,s-\emptyset} = v_0 - v_{0,s}$, the intrinsic importance of X_s relative to the full covariate vector

X. Due to covariate-induced censoring, it may be of interest to include features in the full vector X without analyzing their predictive potential. Such variables could be included in the index set r.

2.2.3 Common predictiveness measures in survival analysis

The choice of predictiveness measure should depend on the purpose of the prediction function f. For example, if f is a risk score intended to stratify participants into risk categories, an appropriate measure quantifies how well f discriminates between high-risk and low-risk participants. On the other hand, if f is a predicted survival probability at landmark time τ, an appropriate measure quantifies the accuracy of f as an estimator of the true survival probability at time τ. Our framework is broadly applicable, but the practitioner must choose an appropriate predictiveness measure. We give several example predictiveness measures below, with additional examples in Appendix A.3. For a discussion of predictiveness measures for survival data, see Korn and Simon (1990).

Example 1: AUC. Heagerty and Zheng (2005) describe several variants of time-varying area under the receiver operating characteristic curve (AUC). The most natural analog to binary outcome AUC is the "cumulative/dynamic AUC," defined as

$$\mathbb{V}(f, P_0^*) := P_0^*(f(X_1) > f(X_2) \,|\, T_1 \leq \tau, T_2 > \tau) \ ,$$

where (X_1, T_1) and (X_2, T_2) are independent draws from P_0^*. The cumulative/dynamic AUC at landmark time τ measures the probability that a participant who fails at or before time τ has a higher risk score $f(x)$ compared to a participant who has not failed by time τ. For survival data, AUC is sometimes estimated using inverse-probability-of-censoring weights (IPCW) (Uno et al., 2007; Hung and Chiang, 2010), a Kaplan-Meier approach valid under noninformative censoring (Chambless and Diao, 2006), or the Cox model (Song and Zhou, 2008).

Example 2: C-index. The concordance index (c-index) is a predictiveness measure that does not require the choice of a landmark time and is often considered to be a global measure of

discriminative performance. The population c-index $P_0^*(f(X_1) > f(X_2) \,|\, T_1 < T_2)$ measures the probability that, between a randomly selected pair of participants, the participant with the higher risk score fails earlier. For identifiability under right censoring, we restrict the comparison of times to those falling before some user-specified time τ:

$$\mathbb{V}(f, P_0^*) := P_0^*(f(X_1) > f(X_2) \,|\, T_1 < T_2, T_1 \leq \tau) \ .$$

Besides the c-index statistic proposed by Harrell et al. (1982) — which converges to a population parameter that depends on the censoring distribution — other existing estimation methods are based on the Cox model (Gonen and Heller, 2005), the Pareto distribution (Brentnall and Cuzick, 2018), IPCW (Uno et al., 2011), and the stratified Kaplan-Meier estimator (Efron, 1967).

Example 3: Brier score. It may be of interest to predict a participant's probability of remaining event-free by landmark time τ. In this case, the predictiveness of f can be quantified using any loss function for predicting the binary outcome $\mathbb{1}(T > \tau)$. Let $L : \mathcal{F} \times \{0, 1\} \to [0, \infty)$ denote such a loss function. The predictiveness measure can be defined as the negative expected loss $-\mathbb{E}_0^* [L(f(X), \mathbb{1}(T > \tau))]$. Possible loss functions include log loss, binary classification loss, and squared error loss. The expected squared-error loss for a binary outcome is referred to as the Brier score (Brier, 1950) and is often estimated using IPCW (Gerds and Schumacher, 2006). Here, we focus on the negative Brier score (i.e., negative mean squared error (MSE)):

$$\mathbb{V}(f, P_0^*) := -\mathbb{E}_0^* \left[\{f(X) - \mathbb{1}(T > \tau)\}^2 \right] .$$

The Brier score is related to the proportion of explained variance (PEV) measure given by $1 - \left(\mathbb{E}_0^* [\{f(X) - \mathbb{1}(T > \tau)\}^2]\right) / \text{Var}_0^*[\mathbb{1}(T > \tau)]$ (Schemper and Henderson, 2000). When f equals the conditional mean, this quantifies the proportion of the variance of $\mathbb{1}(T > \tau)$ explained by X.

2.3 Adapting VIMs for survival analysis

2.3.1 Identification and support

Under right censoring, the predictiveness measures and resulting VIMs given in Section 2.2.3 are not functionals of the observed data distribution P_0. Standard plug-in estimation techniques are therefore infeasible, and we must proceed by identifying the predictiveness measures, i.e., writing them in terms of P_0. Our identification approach relies on the hazard function. We define the conditional cumulative hazard of T given $X = x$ at t as $\Lambda_0^*(t \,|\, x) := \int_0^t \frac{F_0^*(du \,|\, x)}{1-F_0^*(u^- \,|\, x)}$, where $F_0^*(t \,|\, x) := P_0^*(T \leq t \,|\, X = x)$. The hazard and survival functions are linked via the product integral mapping $\alpha \mapsto \prod_{(0,t]} \{1 - \alpha(du \,|\, x)\}$, as established by Gill and Johansen (1990):

$$P_0^*(T > t \,|\, X = x) = \prod_{(0,t]} \{1 - \Lambda_0^*(du \,|\, x)\} := S_0^*(t \,|\, x). \tag{2.1}$$

We identify $S_0^*(t \,|\, x)$ under the following assumptions:

(A1) *(conditionally independent censoring)* T is conditionally independent of C given X.

(A2) *(positive event probability)* $P_0^*(C \geq t \,|\, X) > 0$ P_0^*-almost surely.

Under (A1), T and C may depend on each other as long as they are conditionally independent given the set of recorded covariates X. This allows for informative censoring, and the ability to handle informative censoring is a key component of our variable importance framework. In light of condition (A1), it is necessary to include in the full feature vector X any covariate which is thought to inform the censoring mechanism, in addition to features that are of interest for prediction. We note here that conditionally independent censoring is a form of coarsening-at-random, which implies that the model $\mathcal{M}$ is nonparametric (van der Laan and Robins, 2003). This fact will be vital in our efficiency calculations in Section 2.4. Condition (A2) states that there is positive probability of remaining uncensored at time t in nearly all

covariate strata. In essence, at a population level there are no covariate strata for which all individuals are censored at time t.

Next, we define the observed data hazard $\Lambda_0(t \,|\, x) := \int_0^t \frac{H_{0,1}(du \,|\, x)}{1-H_0(u^- \,|\, x)}$, where $H_{0,1}(t \,|\, x) := P_0(Y \leq t, \Delta = 1 \,|\, X = x)$ is the subdistribution function of Y among uncensored individuals and $H_0(t \,|\, x) := P_0(Y \leq t \,|\, X = x)$ is the distribution function of Y. The usual identification of the hazard under dependent censoring (Beran, 1981, see Appendix A.2) yields that $\Lambda_0^*(t \,|\, x) = \Lambda_0(t \,|\, x)$. Defining $S_P(t \,|\, x) := \prod_{(0,t]} \{1 - \Lambda_P(du \,|\, x)\}$, it follows by (2.1) that $S_0^*(t \,|\, x) = S_0(t \,|\, x)$, i.e., S_0 is the observed data equivalent of S_0^* under condition (A1). For convenience, we define $F_P := 1 - S_P$, the identified version of the conditional distribution function of T given X. We emphasize that Λ_P, S_P, and F_P are each in one-to-one correspondence with one another.

Many predictiveness measures can be written as an expectation taken with respect to the joint distribution of X and T. Using the tower property, the joint distribution function evaluated at (x_0, t_0) can be written as $\mathbb{F}_0^*(x_0, t_0) = \int \mathbb{1}(u \leq x_0) F_0^*(t_0 \,|\, u) Q_0^*(du)$, where Q_0^* is the distribution function of X under P_0^*. Letting Q_0 denote the distribution function of X under P_0, we note that $Q_0^* = Q_0$ since X is fully observed. Combining this with the fact that, under (A1) and (A2), $F_0^*(t \,|\, x) = F_0(t \,|\, x)$, we can write the identified version of $\mathbb{F}_0^*$ as

$$\mathbb{F}_0(x_0, t_0) = \int \mathbb{1}(u \leq x_0) F_0(t_0 \,|\, u) Q_0(du) \ .$$

We now define the class of predictiveness measures that is the focus of this chapter.

Definition 1. Let $\mathcal{R} \subseteq \mathcal{X} \times (0, \infty)$ denote a region of integration for the joint distribution function $\mathbb{F}_0^*$. An ideal data predictiveness measure $\mathbb{V}(f, P_0^*)$ is called a standardized survival V-measure if it can be written in the form $\mathbb{V}(f, P_0^*) = \mathbb{V}_1(f, \mathbb{F}_0^*)/\mathbb{V}_2(\mathbb{F}_0^*)$, where

$$\mathbb{V}_1(f, \mathbb{F}_0^*) = \int_{\mathcal{R}} \cdots \int_{\mathcal{R}} \Phi\left((f(x_1), t_1), \ldots, (f(x_m), t_m)\right) \prod_{j=1}^{m} \mathbb{F}_0^*(dx_j, dt_j) \ ,$$

$$\mathbb{V}_2(\mathbb{F}_0^*) = \int_{\mathcal{R}} \cdots \int_{\mathcal{R}} \Theta(t_1, \ldots, t_m) \prod_{j=1}^{m} \mathbb{F}_0^*(dx_j, dt_j) \ ,$$

for symmetric kernel functions $\Phi : \{\mathcal{Y} \times \mathbb{R}\}^m \to \mathbb{R}$ and $\Theta : \mathbb{R}^m \to \mathbb{R}$ and $m \geq 1$ an integer.

A more general definition could allow $\mathbb{V}$ to have an additive dependence on some constant $\eta \in \mathbb{R}$, which would encompass predictiveness measures such as PEV. For variable importance, where interest lies in the difference between full and residual oracle predictiveness, the constant η cancels, and we omit it for simplicity.

Identifiability of $\mathbb{V}$ depends on the region $\mathcal{R}$ over which $\mathbb{F}_0^*(x,t)$ is integrated, due to the fact that $F_0^*(t \mid x)$ is unidentified at times t for which condition (A2) fails to hold. We let $\mathcal{T} \subseteq (0,\infty)$ denote the region of integration for t, such that $\mathcal{R} = \mathcal{X} \times \mathcal{T}$. We place conditions on $\mathcal{T}$ and on the kernel functions Φ and Θ to ensure identifiability.

(A3) *(region of integration)* $\mathcal{T}$, Φ, and Θ satisfy one of the three following conditions:

- (A3a) $\mathcal{T} = [a,b]$ for $a, b \in [0,\infty)$ such that b satisfies condition (A2);
- (A3b) $\mathcal{T} = [0,\infty)$ and there exists a τ satisfying condition (A2) such that Φ and Θ depend on t only via the indicator function $t \mapsto \mathbb{1}(t \leq \tau)$;
- (A3c) $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$, where each of $\mathcal{T}_1$ and $\mathcal{T}_2$ satisfy one of the two previous conditions.

For the remainder of this chapter, the region of integration for all integrals with integrator $\mathbb{F}(x_0,t_0)$ is taken to be $\mathcal{R}$. Table 2.1 gives the form of Φ and Θ for each of the examples introduced in Section 2.2.3. The c-index satisfies (A3a) with $a = 0$ and $b = \tau$ if τ is chosen such that (A2) holds. The Brier score and AUC satisfy condition (A3b) if τ is chosen such that (A2) holds.

The components $\mathbb{V}_1$ and $\mathbb{V}_2$ of a standardized survival V-measure depend on P_0^* only via the joint distribution function $\mathbb{F}_0^*$. Under (A1) - (A3), $\mathbb{F}_0^*(x,t) = \mathbb{F}_0(x,t)$ over $\mathcal{R}$, and so $\mathbb{V}_1(f,\mathbb{F}_0^*)$ and $\mathbb{V}_2(\mathbb{F}_0^*)$ are identified. We use V_1, V_2, and V to denote the observed data counterparts of $\mathbb{V}_1$, $\mathbb{V}_2$, and $\mathbb{V}$, respectively.

2.3.2 Characterizing the oracle prediction function

Because $\mathbb{V}(f,P_0^*) = V(f,P_0)$, we have that $\operatorname{argmax}_{f\in\mathcal{F}}\mathbb{V}(f,P_0^*) = \operatorname{argmax}_{f\in\mathcal{F}}V(f,P_0)$, i.e., the maximizer of the ideal data predictiveness measure is also the maximizer of the ob-

VIM	m	Φ	Θ
AUC	2	$\mathbb{1}(f(x_1) > f(x_2), t_2 \leq \tau, t_1 > \tau)$	$\mathbb{1}(t_1 \leq \tau, t_2 > \tau)$
c-index	2	$\mathbb{1}(f(x_1) > f(x_2), t_1 \leq \tau, t_2 > t_1)$	$\mathbb{1}(t_1 \leq \tau, t_2 > t_1)$
Brier score	1	$2f(x)\mathbb{1}(t > \tau) - f(x)^2 - \mathbb{1}(t > \tau)$	1

Table 2.1: Degree and kernel functions for example standardized survival V-measures. The AUC and c-index kernels can be symmetrized by adding a second evaluation of the kernel with arguments exchanged, and dividing by two.

served data predictiveness measure. In some cases, the ideal data maximizer may already be characterized and must only be written in an appropriately identified form, as seen below.

Example 1: AUC. The ideal data maximizer of $f \mapsto \mathbb{V}(f, P_0^*)$ is given by the population conditional mean of the indicator variable $\mathbb{1}(T > \tau)$ given X, i.e., $f_0 : x \mapsto \mathbb{E}_0^*[\mathbb{1}(T > \tau) \,|\, X = x]$ (Williamson et al., 2021b). We can write this in terms of F_0 as $f_0 : x \mapsto 1 - F_0(\tau \,|\, x)$.

Example 2: C-index. It is not straightforward to characterize the oracle prediction function for the c-index. The oracle is any prediction function f that gives a higher risk to X_1 than to X_2 when $\{T_1 < T_2, T_1 \leq \tau\}$ occurs with P_0^*-probability larger than the P_0^*-probability of $\{T_2 < T_1, T_2 \leq \tau\}$ (see Appendix A.2). For nonparametric $\mathcal{M}$, f_0 does not appear to be available in closed form. This necessitates a direct optimization procedure to estimate f_0, which we outline in Appendix A.4.

Example 3: Brier score. The Brier score at time τ is equivalent to MSE for the binary outcome $\mathbb{1}(T > \tau)$, and so the oracle prediction function is the conditional mean of the indicator variable, i.e., $f_0 : x \mapsto \mathbb{E}_0^*[\mathbb{1}(T > \tau) \,|\, X = x]$. This can be written in terms of F_0 as $f_0 : x \mapsto 1 - F_0(\tau \,|\, x)$.

2.4 Estimation and inference

2.4.1 Overview

The observed data predictiveness measure relies on the unknown nuisance functions f_0 and F_0. Using flexible machine learning methods to estimate these nuisances maximizes the

chances of consistent estimation without making strong distributional assumptions on T given X. Given estimators F_n and f_n, and using the empirical distribution of $(X_1, \ldots, X_n)$ as an estimate Q_n, we might consider the plug-in estimator $\tilde{v}_n := V_1(f_n, \tilde{\mathbb{F}}_n)/V_2(\tilde{\mathbb{F}}_n)$, where

$$\tilde{\mathbb{F}}_n(x_0, t_0) := \int \mathbb{1}(u \leq x_0) F_n(t_0 \,|\, u) Q_n(du) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x_0) F_n(t_0 \,|\, X_i) \ .$$

In general, we cannot expect f_n and F_n to converge at $n^{1/2}$ rate. The fact that f_0 is a maximizer of $f \mapsto V(f, P_0)$ implies that its estimation has no first-order contribution to the asymptotic behavior of the plug-in estimator (Williamson et al., 2021b). Nonetheless, we need to pursue a debiased estimator to account for the behavior of F_n in the integrator of the survival V-measure.

There are several possible approaches to debiasing the plug-in estimator; we focus on one-step debiasing at the level of the joint distribution function of (X, T). We first characterize the asymptotic bias $B_{\mathbb{F},n}(x_0, t_0)$ of the plug-in distribution function estimator $\tilde{\mathbb{F}}_n(x_0, t_0)$. This leads to the debiased estimator $V_1(f_n, \tilde{\mathbb{F}}_n - B_{\mathbb{F},n})/V_2(\tilde{\mathbb{F}}_n - B_{\mathbb{F},n})$, which under regularity conditions is asymptotically linear and nonparametric efficient. In addition, it enjoys a form of double robustness. We outline this procedure in Algorithm 2.1 and provide details in the remainder of this section.

Algorithm 2.1 VIM estimation procedure outline

1: Compute estimator F_n of the conditional distribution function of $T \,|\, X$. Use this to construct the plug-in joint distribution function estimator $\tilde{\mathbb{F}}_n$.
2: Compute estimator $B_{\mathbb{F},n}$ of the asymptotic bias of $\tilde{\mathbb{F}}_n$.
3: Construct one-step debiased estimator $\mathbb{F}_n^{\text{os}} = \tilde{\mathbb{F}}_n - B_{\mathbb{F},n}$.
4: Compute estimators f_n and $f_{n,s}$ of the full and residual oracle prediction functions.
5: Plug in nuisances estimators f_n, $f_{n,s}$, and $\mathbb{F}_n^{\text{os}}$ to yield the overall VIM estimator $V_1(f_n, \mathbb{F}_n^{\text{os}})/V_2(\mathbb{F}_n^{\text{os}}) - V_1(f_{n,s}, \mathbb{F}_n^{\text{os}})/V_2(\mathbb{F}_n^{\text{os}})$.

2.4.2 Efficiency

We first present the EIF of $V(f_0, P_0)$, which plays a role in our proposed inferential procedure. This EIF involves the conditional survival function of C given X, identified under (A1) as

$$P_0^*(C \geq t \,|\, X = x) = G_0(t \,|\, x) := \prod_{[0,u)} \left\{1 - \Lambda_0^C(du \,|\, x)\right\},$$

where $\Lambda_P^C(t \,|\, x) := \int_0^t \left\{\frac{S_P(u^- \,|\, x)}{S_P(u \,|\, x)}\right\} \frac{H_{P,0}(du \,|\, x)}{1 - H_P(u^- \,|\, x)}$ and $H_{P,0}(t \,|\, x) := P(Y \leq t, \Delta = 0 \,|\, X = x)$. This nuisance parameter appears in the key quantity

$$z \mapsto \varphi_{P,z}(t) := -S_P(t \,|\, x) \left\{\frac{\delta \mathbb{1}_{[0,t]}(y)}{S_P(y \,|\, x) G_P(y \,|\, x)} - \int_0^{t \wedge y} \frac{\Lambda_P(du \,|\, x)}{S_P(u \,|\, x) G_P(u \,|\, x)}\right\},$$

which resembles the influence function of the stratified Kaplan-Meier estimator (Reid, 1981). This function plays a prominent role in nonparametric survival analysis. We also define the mappings

$$\begin{aligned}
&\Phi_{P,l}((x_1, t_1), \ldots, (x_l, t_l)) \\
&\quad := \int \cdots \int \Phi\left((f_P(x_1), t_1), \ldots, (f_P(x_l), t_l), (f_P(x_{l+1}), t_{l+1}), \ldots, (f_P(x_m), t_m)\right) \\
&\qquad\qquad \times \prod_{j=l+1}^{m} \mathbb{F}_P(dx_j, dt_j); \\
&\Theta_{P,l}(t_1, \ldots, t_l) := \int \cdots \int \Theta\left(t_1, \ldots, t_l, t_{l+1}, \ldots, t_m\right) \prod_{j=l+1}^{m} \mathbb{F}_P(dx_j, dt_j)\ .
\end{aligned}$$

Before presenting the EIF, we introduce some regularity conditions, which are discussed in greater depth in Williamson et al. (2021b).

(B1) There exists a dense subset $\mathcal{H}$ of $L_2^0(P_0)$ such that, for each $h \in \mathcal{H}$ and regular univariate parametric submodel $\{P_{0,\epsilon}\} \subset \mathcal{M}$ through P_0 at $\epsilon = 0$ and with score for ϵ equal to h at $\epsilon = 0$, the following conditions hold, with $f_{0,\epsilon}$ denoting $f_{P_{0,\epsilon}}$:

(B1a) (*second-order perturbations*) $V(f_{0,\epsilon}, P_\epsilon) - V(f_{0,\epsilon}, P_0) = V(f_0, P_\epsilon) - V(f_0, P_0) + o(\epsilon)$;

(B1b) (*differentiability*) $\epsilon \mapsto V(f_{0,\epsilon}, P_0)$ is differentiable in a neighborhood of $\epsilon = 0$;

(B1c) (*richness of function class*) the optimizer $f_{0,\epsilon}$ is in $\mathcal{F}$ for small enough ϵ.

Theorem 1. *If there exists $\nu \in (0, \infty)$ such that $G_0(\tau \,|\, x) \geq \nu$ for P_0-almost every x and if condition (B1) holds, then $P \mapsto V(f_P, P)$ is pathwise differentiable at P_0 relative to the nonparametric model $\mathcal{M}$, with EIF given by $D(P_0) : z \mapsto \frac{D_\Phi(P_0)(z)}{V_2(\mathbb{F}_0)} - \frac{V_1(f_0,\mathbb{F}_0)D_\Theta(P_0)(z)}{V_2(\mathbb{F}_0)^2}$, where*

$$D_\Phi(P_0) : z \mapsto m\left[\int \Phi_{0,1}(x,t)\left\{F_0(dt \,|\, x) - \varphi_{0,z}(dt)\right\} - V_1(f_0, \mathbb{F}_0)\right];$$
$$D_\Theta(P_0) : z \mapsto m\left[\int \Theta_{0,1}(t)\left\{F_0(dt \,|\, x) - \varphi_{0,z}(dt)\right\} - V_2(\mathbb{F}_0)\right].$$

If F_0 were known, V_1 and V_2 would be standard generalized moment functionals with uncentered EIFs $x \mapsto m\int \Phi_{0,1}(x,t)F_0(dt \,|\, x)$ and $x \mapsto m\int \Theta_{0,1}(t)F_0(dt \,|\, x)$, per the theory of V-statistics. The components $z \mapsto m\int \Phi_{0,1}(x,t)\varphi_{0,z}(dt)$ and $z \mapsto m\int \Theta_{0,1}(t)\varphi_{0,z}(dt)$ represent the contribution from estimating F_0. The fact that f_0 must be estimated has no impact on the EIF.

2.4.3 Estimation of the joint distribution function

En route to building an estimator of $\psi_{0,s}$, we propose a strategy for debiased estimation of $\mathbb{F}_0(x_0, t_0)$. This strategy uses the EIF of $\mathbb{F}_0$ relative to $\mathcal{M}$, which is given by $z \mapsto \bar{\phi}_{0,z}(x_0, t) := \phi_{0,z}(x_0, t_0) - \mathbb{F}_0(x_0, t_0)$, where $\phi_{0,z}(x_0, t_0) := \mathbb{1}(x \leq x_0)\left\{F_0(t_0 \,|\, x) - \varphi_{0,z}(t_0)\right\}$ (Lemma 1 in Appendix A.1). Given $P \in \mathcal{M}$, a first-order expansion of $\mathbb{F}_P(x_0, t_0)$ about P_0 is given by

$$\mathbb{F}_P(x_0, t_0) - \mathbb{F}_0(x_0, t_0) = -P_0\bar{\phi}_0(x_0, t_0) + R_{x_0,t_0}(P, P_0), \tag{2.2}$$

where $R_{x_0,t_0}(P, P_0) := \mathbb{F}_P(x_0, t_0) - \mathbb{F}_0(x_0, t_0) + P_0\bar{\phi}_0(x_0, t_0)$ is the remainder from a linearization of $\mathbb{F}$ about P_0. We expect $R_{x_0,t_0}(P, P_0)$ to have a second-order contribution due to the strong differentiability of $\mathbb{F}$, in the sense of Pfanzagl (1982).

The one-step approach requires an estimator of P_0. Since the representation of ϕ_0 includes the variation-independent nuisances F_0 and G_0, a natural parametrization of P_0 is given by

(F_0, G_0, Q_0). (We recall that estimating F_0 gives estimates of S_0 and Λ_0.) We define P_n as the estimator of P_0 constructed from estimators F_n and G_n, as well as the empirical covariate distribution Q_n. Replacing F_0 and G_0 with F_n and G_n, we obtain the estimated uncentered and centered influence functions ϕ_n and $\bar{\phi}_n$. We suppose that F_n and G_n converge to limits F_∞ and G_∞, in a manner made concrete in Theorem 2, and we define ϕ_∞ and $\bar{\phi}_\infty$ as the uncentered and centered influence functions evaluated at F_∞ and G_∞. Using (2.2), we can write

$$\tilde{\mathbb{F}}_n(x_0, t_0) - \mathbb{F}_0(x_0, t_0) = \mathbb{P}_n \bar{\phi}_\infty(x_0, t_0) + C_{n,x_0,t_0}(P_n, P_\infty) + R_{x_0,t_0}(P_n, P_0) - \mathbb{P}_n \bar{\phi}_n(x_0, t_0), \tag{2.3}$$

where $C_{n,x_0,t_0}(P_n, P_\infty) := (\mathbb{P}_n - P_0)\left\{\bar{\phi}_n(x_0, t_0) - \bar{\phi}_\infty(x_0, t_0)\right\}$ is an empirical process term. The leading term on the right-hand side of (2.3) is the empirical average of mean-zero transformations of $Z_1, \ldots, Z_n$. The second and third terms are expected to be second-order. The final term represents the excess bias due to flexibly estimating F_0 and G_0, and its presence indicates that the plug-in estimator $\tilde{\mathbb{F}}_n(x_0, t_0)$ may fail to achieve $n^{1/2}$-consistency for $\mathbb{F}_0(x_0, t_0)$. A debiased one-step estimator of $\mathbb{F}_0(x_0, t_0)$ is then given by

$$\mathbb{F}_n^*(x_0, t_0) := \tilde{\mathbb{F}}_n(x_0, t_0) + \mathbb{P}_n \bar{\phi}_n(x_0, t_0) = \mathbb{P}_n \phi_n(x_0, t_0) = \frac{1}{n} \sum_{i=1}^{n} \phi_n(x_0, t_0)(Z_i) \ .$$

In light of (2.3), under regularity conditions this is an asymptotically linear estimator of $\mathbb{F}_0(x_0, t_0)$.

One of the conditions often used to control $C_{n,x_0,t_0}(P_n, P_\infty)$ is the Donsker condition, which constrains the complexity of the algorithms used to estimate F_0 and G_0. Flexible estimators are unlikely to satisfy this requirement, but cross-fitting can circumvent the need for Donsker conditions (Zheng and van der Laan, 2011; Chernozhukov et al., 2018). For K-fold cross-fitting, we randomly partition the dataset into K subsets of roughly equal size. For each $k \in \{1, \ldots, K\}$, we set aside the k^{th} subset as test data and construct estimators $F_{n,k}$ and $G_{n,k}$ using the rest of the data, yielding estimated influence function $\phi_{n,k}(x_0, t_0)$. We then construct and store $\mathbb{F}_{n,k}(x_0, t_0) := \mathbb{P}_{n,k} \phi_{n,k}(x_0, t_0)$, where $\mathbb{P}_{n,k}$ is the empirical distribution

of the data unit Z in the test data. We note that $(F_{n,k}, G_{n,k})$ and $\mathbb{P}_{n,k}$ are estimated using non-overlapping subsets of the data. We repeat this procedure for each of the K subsets and construct the overall cross-fitted estimator $\frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_{n,k}\phi_{n,k}(x_0,t_0)$ of $\mathbb{F}_0(x_0,t_0)$. We use a similar procedure to produce a cross-fitted VIM estimator.

2.4.4 VIM estimation

To estimate the oracle predictiveness $V(f_0, P_0)$, we consider the cross-fitted estimator

$$v_n := \frac{v_{n,1}}{v_{n,2}} := \frac{\frac{1}{K}\sum_{k=1}^{K} V_1(f_{n,k}, \mathbb{F}_{n,k})}{\frac{1}{K}\sum_{k=1}^{K} V_2(\mathbb{F}_{n,k})} \ ,$$

where $f_{n,k}$ denotes an estimator of the oracle prediction function f_0 constructed using the same data used to construct $F_{n,k}$ and $G_{n,k}$. In some cases, $f_{n,k}$ can be written as a function of $F_{n,k}$, and so requires no additional fitting. When f_0 is not available in closed form, $f_{n,k}$ may be obtained using a direct optimization approach, as outlined in Appendix A.4 for the c-index.

We study the behavior of v_n by separately considering the numerator and denominator. A first-order expansion of $v_{n,1}$ about $V_1(f_0, \mathbb{F}_0)$ yields

$$\begin{aligned} &v_{n,1} - V_1(f_0, \mathbb{F}_0) \\ &\quad = \frac{1}{K}\sum_{k=1}^{K}\{V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\} + \frac{1}{K}\sum_{k=1}^{K}\{V_1(f_{n,k}, \mathbb{F}_0) - V_1(f_0, \mathbb{F}_0)\} + r_n \ , \end{aligned}$$

where $r_n := \frac{1}{K}\sum_{k=1}^{K}[\{V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0)\} - \{V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\}]$. In the first term on the right-hand side above, the prediction function argument is fixed at f_0, and so the behavior of this term is determined by the behavior of $\mathbb{F}_{n,k}$ via the mapping $\mathbb{F} \mapsto V_1(f_0, \mathbb{F})$. The second term is the contribution from the estimation of f_0. The final term r_n is a difference-in-differences term that will be second-order under regularity conditions. A similar expansion of $v_{n,2}$ yields $v_{n,2} - V_2(\mathbb{F}_0) = \frac{1}{K}\sum_{k=1}^{K}\{V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0)\}$. Because V_2 does not involve f_0, we need only consider the behavior of $\mathbb{F}_{n,k}$ via the mapping $\mathbb{F} \mapsto V_2(\mathbb{F})$.

In Theorem 2, we provide conditions under which v_n is consistent for v_0.

(B2) *(bounded away from zero)* There exists $\nu \in (0, \infty)$ such that, with probability tending to 1 and for P_0-almost all x, $G_{n,k}(\tau \,|\, x) \geq 1/\nu$ and $G_0(\tau \,|\, x) \geq 1/\nu$ for each $k \in \{1, 2, \ldots, K\}$.

(B3) *(limits of nuisance estimators)*

(C3a) There exists G_∞ such that $\max_k \mathbb{E}_0 \left[\sup_{u \in \mathcal{T}} \left| \frac{1}{G_{n,k}(u \,|\, X)} - \frac{1}{G_\infty(u \,|\, X)} \right| \right]^2 \xrightarrow{\text{p}} 0$.

(C3b) There exists S_∞ such that $\max_k \mathbb{E}_0 \left[\sup_{u \in \mathcal{T}} \sup_{v \in [0,u]} \left| \frac{S_{n,k}(u \,|\, X)}{S_{n,k}(v \,|\, X)} - \frac{S_\infty(u \,|\, X)}{S_\infty(v \,|\, X)} \right| \right]^2 \xrightarrow{\text{p}} 0$.

(C3c) $\max_k \left\| f_{n,k}(X) - f_0(X) \right\|_{\mathcal{F}} \xrightarrow{\text{p}} 0$.

(B4) (*single correctly specified nuisance*) For P_0-almost all x, there exist sets $\mathcal{S}_x, \mathcal{G}_x \subseteq \mathcal{T}$ such that $\mathcal{S}_x \cup \mathcal{G}_x = \mathcal{T}$, $\Lambda_\infty(u \,|\, x) = \Lambda_0(u \,|\, x)$ for all $u \in \mathcal{S}_x$, and $G_\infty(u \,|\, x) = G_0(u \,|\, x)$ for all $u \in \mathcal{G}_x$.

(B5) (*continuity at f_0*) There exists some constant J_1 such that, for each sequence $f_1, f_2, \cdots \in \mathcal{F}$ such that $\|f_j - f_0\|_{\mathcal{F}} \to 0$, $|V(f_j, \mathbb{F}_0) - V(f_0, \mathbb{F}_0)| \leq J_1 \|f_j - f_0\|_{\mathcal{F}}$ for each j large enough.

(B6) (*variation norm*) The functions Φ and Θ are bounded on $\mathcal{R}$, and additionally, for all $l \leq m$,

$$\sup_{(x_1,t_1),\ldots,(x_{l-1},t_{l-1})} \left\| (x,t) \mapsto \Phi_{0,l}\left((f_0(x_1),t_1),\ldots,(f_0(x_{l-1}),t_{l-1}),(f_0(x),t)\right) \right\|_v^* < \infty \,;$$

$$\sup_{t_1,\ldots,t_{l-1}} \left\| t \mapsto \Theta_{0,l}\left(t_1,\ldots,t_{l-1},t\right) \right\|_v < \infty \,.$$

Theorem 2. *If conditions (B2)–(B5) hold, then v_n converges in probability to v_0.*

Condition (B2) requires that G_n and G_0 are uniformly bounded away from zero. Condition (B3) requires that F_n and G_n converge to fixed limits, and that f_n converges to f_0. Conditions (B3) and (B4) together imply that, for almost all $t \in \mathcal{T}$ and $x \in \mathcal{X}$, either S_n or G_n is consistent. Because of this, v_n enjoys a form of double robustness; unsurprisingly, though,

f_n must converge to f_0. Condition (B5) essentially requires that the map $f \mapsto V(f, \mathbb{F}_0)$ is continuous about f_0 with respect to the $\|\cdot\|_{\mathcal{F}}$ norm. Condition (B6) requires that the kernel functions Φ and Θ are bounded and also places restrictions on the uniform sectional variation norm of $\Phi_{0,l}$ and $\Theta_{0,l}$ for all $l \leq m$, with respect to a single data unit. This is weaker than requiring finite uniform sectional variation norm of the kernel with respect to all arguments. Because Φ and Θ depend on x only via $f_0(x)$, it is relatively straightforward to impose restrictions on f_0 such that (B6) holds (see Appendix A.2).

To establish asymptotic linearity of v_n, we require additional conditions. Conditions (B9) and (B10) involve random functions $h_{n,k,1}, h_{n,k,2}, h_{n,k,3}$, and $h_{n,k,t}$, defined in Appendix A.1. We let $\bar{h}_{n,k,1} := P_0^{m-1} h_{n,k,1}$ and $\bar{h}_{n,k,2} := P_0^{m-1} h_{n,k,2}$.

(B7) *(optimality)* There exists some constant J_2 such that, for each sequence $f_1, f_2, \cdots \in \mathcal{F}$ such that $\|f_j - f_0\|_{\mathcal{F}} \to 0$, $|V(f_j, P_0) - V(f_0, P_0)| \leq J_2 \|f_j - f_0\|_{\mathcal{F}}^2$ for each j large enough.

(B8) *(minimum rate of convergence)* $\max_k \|f_n - f_0\|_{\mathcal{F}} = o_P\left(n^{-1/4}\right)$.

(B9) *(weak consistency)* $\max_k \mathbb{E}_0\left[\bar{h}_{n,k,1}^2(Z)\right] = o_P(1)$ and $\max_k \mathbb{E}_0\left[\bar{h}_{n,k,2}^2(Z)\right] = o_P(1)$.

(B10) *(negligibility of higher-order remainder terms)* $\max_k \sup_{t \in \mathcal{T}} \mathbb{E}_0 |h_{n,k,t}(X)| = o_P\left(n^{-1/2}\right)$ and $\max_k \mathbb{E}_0\left[h_{n,k,3}(Z_1, \ldots, Z_m)\right] = o_P\left(n^{-1/2}\right)$.

Theorem 3. *If conditions (B2), (B3), and (B7)–(B10) hold, and in addition $F_\infty = F_0$ and $G_\infty = G_0$, then $v_n - v_0 = \frac{1}{n}\sum_{i=1}^n D(P_0)(Z_i) + o_P\left(n^{-1/2}\right)$. If condition (B1) holds, then $D(P_0)$ is the EIF of $P \mapsto V(f_P, P)$ at P_0 relative to $\mathcal{M}$, and v_n is nonparametric efficient.*

Condition (B7) formalizes the requirement that estimation of f_0 has no first-order contribution. Whether this condition is satisfied in a particular setting depends on the regularity conditions placed on P_0 and $\mathcal{F}$. We provide details for the examples in Appendix A.2. Condition (B8) requires that f_0 be estimated at a sufficiently fast rate. Condition (B9) ensures

asymptotic negligibility of a set of empirical process terms. Condition (B10) requires that two remainder terms, depending on $(G_{n,k} - G_0)(\Lambda_{n,k} - \Lambda_0)$ and $(f_{n,k} - f_0)(\bar{\phi}_{n,k} - \bar{\phi}_0)$, tend to zero at rate faster than $n^{-1/2}$.

By substituting $f_{n,k}$, f_0, and $\mathcal{F}$ for $f_{n,k,s}$, $f_{0,s}$, and $\mathcal{F}_s$, respectively, a similar result to Theorem 3 holds for the cross-fitted one-step estimator $v_{n,s}$ of $v_{0,s}$, the residual oracle predictiveness. We use $D_s(P_0)$ to denote the resulting influence function. To estimate the variable importance $\psi_{0,s}$, we use the estimator $\psi_{n,s} := v_n - v_{n,s}$, which will itself be asymptotically linear and nonparametric efficient with influence function $D(P_0) - D_s(P_0)$. Under non-null importance ($\psi_{0,s} \neq 0$) and $P_0 \left\{D(P_0) - D_s(P_0)\right\}^2 < \infty$, we have that $n^{1/2}(\psi_{n,s} - \psi_{0,s})$ converges weakly to a mean-zero normal random variable with variance $\sigma_{0,s}^2 = P_0 \left\{D(P_0) - D_s(P_0)\right\}^2$. We estimate $\sigma_{0,s}^2$ by $\sigma_{n,s}^2 := \frac{1}{K} \sum_{k=1}^{K} \sigma_{n,k,s}^2$, whose form is given in Algorithm 2.2 along with the full procedure for inference under non-null importance. Letting z_q denote the qth quantile of the standard normal distribution, a confidence interval with asymptotic coverage $1-\alpha$ is given by

$$\left(\psi_{n,s} - z_{1-\alpha/2}\sigma_{n,s}n^{-1/2}, \psi_{n,s} + z_{1-\alpha/2}\sigma_{n,s}n^{-1/2}\right).$$

When $\psi_{0,s} = 0$, the influence function of $\psi_{n,s}$ is also zero. In this scenario, $\psi_{n,s}$ converges at a rate faster than $n^{1/2}$, and our proposed inferential procedure will be invalid. This parameter space boundary problem has been a topic of recent study (Dai et al., 2022; Lundborg et al., 2022). The solution identified by Williamson et al. (2021b) and Dai et al. (2022) to construct confidence intervals that remain valid under the null hypothesis that $\psi_{0,s} = 0$, and to perform hypothesis tests, involves sample splitting. The sample splitting procedure (detailed in Algorithm 2.3) entails estimating v_0 and $v_{0,s}$ using non-overlapping portions of the data. By constructing separate asymptotically linear estimators of the full and residual oracle predictiveness and taking the difference between them, we ensure that the resulting estimator has non-degenerate behavior under the null. However, sample splitting results in decreased in power for testing the null hypothesis that $\psi_{0,s} = 0$. Dai et al. (2022) propose to perform the sample splitting procedure multiple times and aggregate the resulting p-values

Algorithm 2.2 Cross-fitted inference on VIM value $\psi_{0,s}$ (non-zero importance)

1: Select approximation time grid $\mathcal{B} := \{t_1, \ldots, t_J\}$, where $t_K \geq \tau$.
2: Generate $M_n \in \{1, \ldots, K\}^n$ by sampling uniformly from $\{1, \ldots, K\}$ with replacement. For $k = 1, \ldots, K$, denote by $\mathcal{D}_k$ the subset of observations with index in $\{i : M_{n,i} = k\}$.
3: **for** $k = 1, \ldots, K$ **do**
4: Using only data in $\cup_{j \neq k} \mathcal{D}_j$, construct estimators $F_{n,k}$ and $G_{n,k}$ of F_0 and G_0, respectively, on $\mathcal{B}$. In addition, construct estimators $f_{n,k}$ and $f_{n,k,s}$ of f_0 and $f_{0,s}$, respectively.
5: Plug in $F_{n,k}$ and $G_{n,k}$ to obtain $\phi_{n,k}(x_0, t_0)$, using $\mathcal{B}$ to approximate integrals as Riemann sums.
6: Using only data in $\mathcal{D}_k$, construct empirical distribution estimator $\mathbb{P}_{n,k}$ of P_0. Denote by $\mathbb{F}_{n,k}$ the estimator of $\mathbb{F}_0$ constructed using $F_{n,k}$, $G_{n,k}$, and $\mathbb{P}_{n,k}$.
7: Compute $v_{n,1,k} := V_1(f_{n,k}, \mathbb{F}_{n,k})$, $v_{n,2,k} := V_2(\mathbb{F}_{n,k})$, $v_{n,1,k,s} := V_1(f_{n,k,s}, \mathbb{F}_{n,k})$. Compute $\sigma^2_{n,k,s} := \mathbb{P}_{n,k} \{D(P_{n,k}) - D(P_{n,k,s})\}^2$.
8: **end for**
9: Compute estimator $\psi_{n,s} := \left\{ \frac{1}{K} \sum_{k=1}^K (v_{n,1,k} - v_{n,1,k,s}) \right\} / \left\{ \frac{1}{K} \sum_{k=1}^K v_{n,2,k} \right\}$ of $\psi_{0,s}$.
10: Compute estimator $\sigma^2_{n,s} := \frac{1}{K} \sum_{k=1}^K \sigma^2_{n,k,s}$ of $\sigma^2_{0,s}$.

in order to recoup power. We adapt this to our setting by performing Algorithm 2.3 U times, producing p-values $\{p_1, \ldots, p_U\}$. These p-values can be aggregated via any multiple testing procedure that controls of the family-wise error rate, e.g., the Bonferroni method. Confidence intervals may be obtained by inverting the resulting hypothesis test.

2.4.5 Estimation of nuisance parameters

Our proposed procedure requires estimation of F_0 and G_0, the conditional distribution/survival functions of the event and censoring times, respectively. For the examples we consider, F_0 and G_0 must be estimated over an interval $(0, \tau]$. Because of this, we recommend using an approach that targets the entire distribution function, rather than the distribution function as a single time point. Furthermore, to reduce the risk of biased estimation due to misspecified models we focus on data-adaptive, flexible methods, examples of which are included in the simulations in Section 2.5.

Estimation of f_0 must be handled on a case-by-case basis. For VIMs such as AUC and

Algorithm 2.3 Sample-split, cross-fitted inference on VIM value $\psi_{0,s}$

Select approximation time grid $\mathcal{B} := \{t_1, \ldots, t_J\}$, where $t_K \geq \tau$.
Generate $M_n \in \{1, \ldots, 2K\}^n$ by sampling uniformly from $\{1, \ldots, 2K\}$ with replacement. For $k = 1, \ldots, 2K$, denote by $\mathcal{D}_k$ the subset of observations with index in $\{i : M_{n,i} = k\}$. Let n_s denote the number of observations in $\cup_{k \text{ even}} \mathcal{D}_k$.
for $k = 1, \ldots, 2K$ **do**
    Using only data in $\cup_{j \neq k} \mathcal{D}_j$, construct estimators $F_{n,k}$ and $G_{n,k}$ of F_0 and G_0, respectively, on $\mathcal{B}$. In addition, construct estimators $f_{n,k}$ and $f_{n,k,s}$ of f_0 and $f_{0,s}$, respectively.
    Plug in $F_{n,k}$ and $G_{n,k}$ to obtain $\phi_{n,k}(x_0, t_0)$, using $\mathcal{B}$ to approximate integrals as Riemann sums.
    Using only data in $\mathcal{D}_k$, construct empirical distribution estimator $\mathbb{P}_{n,k}$ of P_0. Denote by $\mathbb{F}_{n,k}$ the estimator of $\mathbb{F}_0$ constructed using $F_{n,k}$, $G_{n,k}$, and $\mathbb{P}_{n,k}$.
    if k is odd **then**
        Compute $v_{n,1,k} := V_1(f_{n,k}, \mathbb{F}_{n,k})$ and $v_{n,2,k} := V_2(\mathbb{F}_{n,k})$. Compute $\tau^2_{n,k} = \mathbb{P}_{n,k} D(P_{n,k})^2$.
    else
        Compute $v_{n,1,k,s} := V_1(f_{n,k,s}, \mathbb{F}_{n,k})$ and $v_{n,2,k} := V_2(\mathbb{F}_{n,k})$. Compute $\tau^2_{n,k,s} = \mathbb{P}_{n,k} D(P_{n,k,s})^2$.
    end if
end for
Compute estimator $\psi^*_{n,s} := \frac{\frac{1}{K}\sum_{k=1}^K v_{n,1,2k-1}}{\frac{1}{K}\sum_{k=1}^K v_{n,2,2k-1}} - \frac{\frac{1}{K}\sum_{k=1}^K v_{n,1,2k,s}}{\frac{1}{K}\sum_{k=1}^K v_{n,2,2k,s}}$ of $\psi_{0,s}$.
Compute estimator $\tau^2_{n,s,*} := \frac{1}{K(n-n_s)} \sum_{k=1}^K \tau^2_{n,2k-1} + \frac{1}{Kn_s} \sum_{k=1}^K \tau^2_{n,2k,s}$ of the variance of $\psi^*_{n,s}$.

Brier score, estimating F_0 directly yields an estimate of f_0. (It is important to note that, in the case of these landmark time VIMs, f_0 depends on F_0 only at time τ, and not over the entire interval $\mathcal{T}$.) When the oracle is not available in closed form, as is the case for the c-index, an estimate of F_0 can be leveraged in an additional optimization scheme to produce an estimate of f_0. Using F_n to directly or indirectly construct f_n, as we describe here, is only one of several possible strategies for estimating f_0. We discuss other approaches in Section 2.7.

Due to the possibility of informative censoring, we must take care in estimating the residual oracle prediction function $f_{0,s}$. In many cases, $f_{0,s}$ is a conditional mean of a function of T given the reduced feature vector, e.g., $\mathbb{E}_0\left[g(T) \,|\, X_{-s}\right]$ with $g : t \mapsto \mathbb{1}(t > \tau)$ in Examples 1 and 3. This quantity can be written as a function of $F_0(\cdot \,|\, X_{-s})$, and it may be tempting to directly estimate $F_0(\cdot \,|\, X_{-s})$ with the same procedure used for $F_0(\cdot \,|\, X)$. However, this could lead to bias due to covariate-induced censoring if the set s includes variables that inform the censoring mechanism. We leverage the fact that $\mathbb{E}_0\left[g(T) \,|\, X_{-s}\right] = \mathbb{E}_0\left[\mathbb{E}_0\left[g(T) \,|\, X\right] \,|\, X_{-s}\right]$, which suggests that we construct full oracle predictions and then regress those predictions on the reduced feature vector. For example, $F_0(\tau \,|\, X_{-s})$ can be estimated by first estimating $F_0(\tau \,|\, X)$ and regressing the resulting estimates on X_{-s}. For VIMs requiring direct optimization of V, the class of potential optimizers can be restricted to those depending only on X_{-s}, with $F_n(\cdot \,|\, X)$ used in the optimization procedure.

2.5 Numerical experiments

2.5.1 Simulation setup

We conducted numerical studies to evaluate the performance of our proposed estimation procedure for survival VIMs. In all experiments, we began by generating independent replicates of (X, T, C), where X is a p-dimensional covariate vector generated from a multivariate normal distribution with mean vector $(0, \ldots, 0)$ and covariance matrix Σ. Given covariate vector $X = x$, the event time T and censoring time C were simulated from

the log-normal accelerated failure time models $\log(T) = \beta_{1,T}x_1 + \cdots + \beta_{p,T}x_p + \varepsilon_T$ and $\log(C) = \beta_{0,C} + \beta_{1,C}x_1 + \cdots + \beta_{p,C}x_p + \varepsilon_C$, where ε_T and ε_C were independent standard normal random variables, and where $\beta_{0,C}$ was chosen to achieve the desired censoring rate in each simulation setting. For each observation, the observed follow-up time Y was set to $\min\{T, C\}$ and the event indicator Δ was set to $\mathbb{1}(T \leq C)$. The simulation scenarios are summarized in Table 2.2, and the true VIM values are given in Table A.1.

We evaluated our estimation procedure using three different estimators for the nuisance functions F_0 and G_0: random survival forests (RSF) (Ishwaran et al., 2008), global survival stacking (Wolock et al., 2022), and survival Super Learner (Westling et al., 2023). Global survival stacking and survival Super Learner are different forms of stacked regression, while RSF is a tree-based ensemble learner. Details on nuisance estimation, including algorithm libraries and selection of tuning parameters, are given in Appendix A.5. To approximate the integrals appearing in v_n and $v_{n,s}$, each nuisance function was estimated on the grid of observed event times.

For landmark time VIMs, oracle prediction functions were estimated using $F_n(\tau \,|\, x)$. Residual oracle prediction functions were estimated by regressing $\{F_n(\tau \,|\, X_i)\}_{i=1}^n$ on the reduced covariate vectors $\{X_{i,-s}\}_{i=1}^n$ using Super Learner with the same library included in global survival stacking. For the c-index, the $F_n(\cdot \,|\, x)$ was plugged into the gradient boosting method described in Appendix A.4 to estimate the oracle and residual oracle prediction functions. Five-fold cross-validation was used to select tuning parameters for the boosting procedure. In all scenarios, we implemented our estimation procedure with five-fold cross-fitting and also included non-cross-fitted comparators (using the procedures outlined in Algorithms 2.2 and 2.3 with $K = 1$).

2.5.2 Simulation results

We first investigated the overall performance of our procedure under two representative scenarios. In Scenario 1, we set $p = 2$, $\beta_T = (0.5, -0.3)$, $\beta_C = (-0.2, 0.2)$, and $\Sigma = \boldsymbol{I}_2$, the identity matrix. In Scenario 2, we set $p = 25$, $\beta_T = (0.5, -0.3, 0, \ldots, 0)$, $\beta_C =$

Scen.	p	Corr. features	Censoring rate	Samp. split	Results
1	2	No	50%	No	Main text & Appendix
2	25	No	50%	Yes	Main text & Appendix
3	2	No	$\{30\%, 40\%, \ldots, 70\%\}$	No	Appendix
4	5	Yes	50%	Yes	Appendix

Table 2.2: Summary of simulation scenarios.

$(-0.2, 0.2, 0, \ldots, 0)$, and $\Sigma = \boldsymbol{I}_{25}$. We set $\beta_{0,C} = 0$ to achieve a 50% censoring rate. In both scenarios, X_1 and X_2 had non-zero importance, while other covariates (where applicable) had zero importance. Both X_1 and X_2 inform the censoring mechanism. For each scenario, we generated 500 random datasets of size $n \in \{500, 750, \ldots, 1500\}$. In Scenario 1, we considered the importance of X_1 and X_2. In Scenario 2, we considered the importance of X_1 and X_4. We considered VIMs based on AUC, Brier score, and c-index. Brier score results, as well as results for X_2 in Scenario 1 and X_1 in Scenario 2, are shown in Appendix A.5. Landmark times were set to $\tau \in \{0.5, 0.9\}$, corresponding to the 50^{th} and 75^{th} quantiles of observed event times. The restriction time for the c-index was set to $\tau = 0.9$.

In Scenario 1, we used Algorithm 2.2 to compute point and standard error estimates, from which we computed nominal 95% Wald-type confidence intervals. We evaluated performance using the empirical bias scaled by $n^{1/2}$, the empirical variance scaled by n and divided by the theoretical asymptotic variance, the empirical confidence interval coverage, and the average confidence interval width. In Scenario 2, where X_4 has null importance, we used the sample-splitting procedure described in Algorithm 2.3. As in Scenario 1, we computed point and standard error estimates, and used these to construct nominal 95% Wald-type confidence intervals. In addition, we computed p-values corresponding to the null hypothesis of zero importance versus the one-sided alternative. We evaluated performance using scaled empirical bias, scaled empirical variance, and empirical confidence interval coverage. For X_4, we computed the empirical rejection probability of the test of the null importance hypothesis, while for X_1 we computed average confidence interval width.

In Figure 2.1, we show the results for estimating the importance of X_1 in Scenario 1, where both features have non-zero importance. From column A, we observe that the scaled bias for the cross-fitted estimators using global stacking and survival Super Learner is near 0. The cross-fitted RSF estimator has larger bias. For RSF and global stacking, the non-cross-fitted estimators for AUC have substantially inflated bias compared to their cross-fitted counterparts, and the bias generally does not tend to 0 at a rate faster than $n^{1/2}$. Interestingly, the non-cross-fitted c-index estimator performs on par with the cross-fitted version. In column B, we see that the scaled variances of the cross-fitted estimators are stable with increasing sample size, and the estimators using global stacking and survival Super Learner have variance near the expected asymptotic variance (represented by the black horizontal line) for all three VIMs. For AUC VIM, the variances of the other estimators tend to be larger than the expected asymptotic variance. Column C shows that the confidence intervals constructed using cross-fitting achieve near-nominal coverage, although the RSF procedure is mildly anticonservative. The non-cross-fitted procedure using RSF is substantially anti-conservative for AUC. The c-index estimator shows good coverage with and without cross-fitting. Column D shows that the width of all confidence intervals decreases with increasing sample size, as expected.

In Figure 2.2, we show the results for estimating the null importance of X_4 in Scenario 2 using sample splitting. The results are similar as in the previous experiment. The cross-fitted estimators have reduced bias compared to their non-cross-fitted analogs, except in the case of the c-index. All estimators have variance roughly proportional to sample size, with all but the non-cross-fitted RSF estimator having variance near the expected asymptotic variance. The confidence interval coverage approaches the nominal level with increasing sample size for all but the non-cross-fitted intervals using RSF, which are substantially anti-conservative for AUC. Similarly, the type I error is generally controlled near the 0.05 level when using cross-fitting.

These empirical results show that our proposed procedure has strong performance in samples of realistic size. The simulations underscore the importance of cross-fitting when

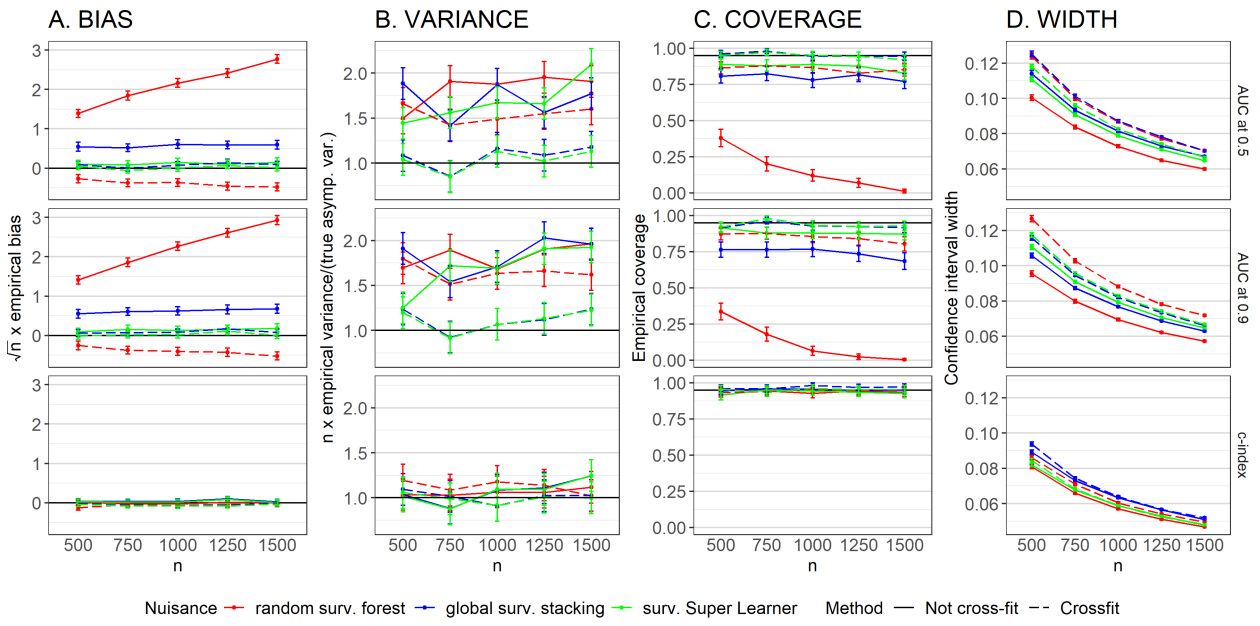

Figure 2.1: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 1 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

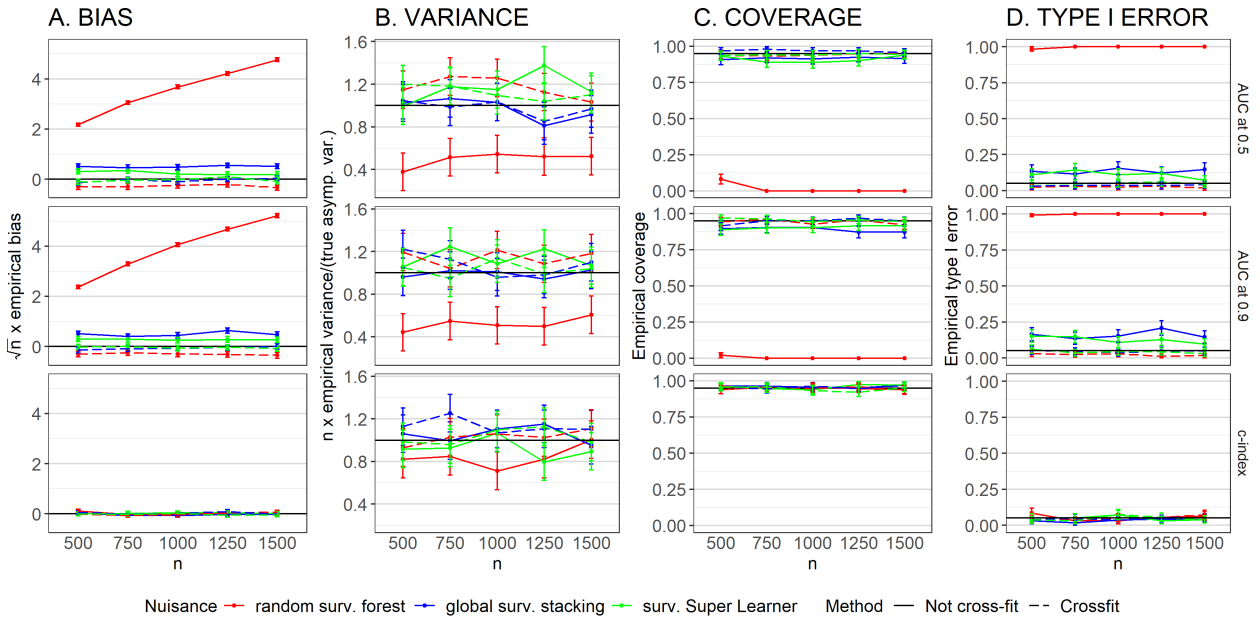

Figure 2.2: Performance of the one-step VIM estimator for the (zero) importance of X_4 in Scenario 2 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

using flexible machine learning nuisance estimators. Under the null hypothesis, the sample splitting procedure is well calibrated. The results also show that, even when the oracle prediction function is not available in closed form, a direct optimization approach can yield good results. We note also that, excluding the non-cross-fitted RSF estimator, all other estimators performed better in estimating the null importance of X_4 in Scenario 2 than in estimating the non-null importance of X_1 in Scenario 1.

In Appendix A.5, we present additional numerical results. These include results in Scenarios 1 and 2 for the features not included in the main text and for the Brier score VIM, as well as all results in Scenario 3, in which we vary the censoring rate, and Scenario 4, which

includes correlated features. Overall, we see that the operating characteristics of the procedure are largely consistent across censoring levels. Unsurprisingly, the impact of censoring on estimator variance and confidence interval width is larger at the later landmark time, when the censoring rate is higher. The inclusion of correlated features has little discernible impact on the performance of our proposed procedure, although it does change the population-level VIM value. This could render interpretation more difficult but does not appear detrimental to statistical performance.

2.6 Variable importance in HVTN 702

The HVTN 702 trial included 5404 individuals, with 2704 assigned to the vaccine group and 2700 to the placebo group. Participants were healthy adults ranging from 18 to 35 years of age. One component of the secondary analyses of the trial was the development of a "baseline risk score" for the 24-month probability of HIV seroconversion using covariates measured at the time of enrollment, and a subsequent analysis of vaccine efficacy within strata defined by that risk score. The baseline risk score was constructed using a regularized Cox model. A formal assessment of the importance of the baseline covariates in predicting the probability of seroconversion was not conducted.

In order to investigate variable importance in predicting the probability of HIV seroconversion in HVTN 702, we chose to measure predictiveness in terms of landmark time AUC. This aligns with the goal of baseline risk score development in HIV trials: viable time horizons in most trials will be short enough that only a small percentage of participants will undergo seroconversion, and interest lies in discriminating between those who are at higher versus lower risk of seroconversion over those time horizons. The AUC can be evaluated at several landmark times to assess if and how variable importance varies over different time horizons.

We analyzed variable importance in the modified intention-to-treat cohort, consisting of the 5384 participants who underwent randomization and were HIV-1 negative at baseline. The median follow-up time in this cohort was 623 days. Because the vaccine efficacy was

estimated to be near 0 (estimated incidence rates per 100 person-years were 3.4 and 3.3 in the vaccine and placebo arms, respectively), we conducted a pooled analysis of both treatment arms. We chose landmark times of 18 months, 24 months, and 30 months post-randomization, corresponding to approximately 60%, 36%, and 13% of participants still at-risk, respectively. (The original baseline risk score analysis in Gray et al., 2021 used a landmark time of 24 months.) The percentage of participants diagnosed with HIV at each of these landmark times were 4.1%, 5.1%, and 5.4%, respectively. The majority of censored participants were subject to administrative censoring due to the termination of the trial. The estimated rates of loss to follow-up (early study termination for any reason, including death) at 18, 24, and 30 months were roughly 6.4%, 7.3%, and 7.5%, respectively.

The baseline risk scores in Gray et al. (2021) were constructed separately for male and female sex assigned at birth. We investigated variable importance in the pooled cohort consisting of both males and females, and among females only, due to the fact that only 37 seroconversions were observed among males. The baseline variables of interest are summarized in Table 2.3 and detailed fully in Gray et al. (2021). We consider the individuals features sex assigned at birth, body mass index (BMI), and age, as well as feature groups consisting of geographic confounders, variables related to sexual health, variables related to behavior, and variables related to housing. Missing covariate values were imputed using `missforest` software package, following the procedure detailed in Gray et al. (2021).

The baseline risk scores in Gray et al. (2021) were constructed separately for male and female sex assigned at birth. In addition to performing an analysis stratified by sex assigned at birth, we also investigated variable importance in the pooled cohort consisting of both males and females. The baseline variables of interest are summarized in Table 2.3 and detailed fully in Gray et al. (2021). We consider the individuals features sex assigned at birth (in the combined cohort only), body mass index (BMI), and age, as well as feature groups consisting of geographic confounders, variables related to sexual health, variables related to behavior, and variables related to housing. Missing covariate values were imputed using `missforest` software package, following the procedure detailed in Gray et al. (2021).

Feature group	Feature(s)
1	Sex assigned at birth
2	Age
3	Body mass index
4	Prevalent STI, genital sores, genital discharge
5	Sexual orientation, married or have main sex partner, live with partner, partner has other partners, anal sex, condom use, unprotected sex with alcohol use, sex with HIV+ partner, unprotected sex with HIV+ partner, exchange services for sex
6	Urban/rural, formal dwelling, home has 3+ services
7	Geographic region

Table 2.3: Features included in the HVTN 702 VIM analysis.

We implemented our procedure using five fold cross-fitting, with global survival stacking for nuisance estimation. The algorithm library was as described for the numerical simulations in Section 2.5. To make inference, we used sample splitting with 20 random splits and the Bonferroni method for combining p-values, as described in Section 2.4.4. The point estimate presented is the average of the point estimates over random splits.

We present results for two types of VIM. The first is conditional VIM, where the full feature vector includes all available covariates and the reduced feature vector excludes the feature or feature group of interest. The second is marginal VIM relative to geographic confounders, where the reduced feature vector includes only geographic variables, and the full feature vector includes geographic variables plus the feature or feature group of interest. The former considers the loss in predictiveness from exclusion of the feature(s) of interest; the latter considers the gain in predictiveness from inclusion of the feature(s) of interest relative to a simple base model.

In Figure 2.3, we display the results of the conditional VIM analysis. For the combined cohort, across all time horizons, sex assigned at birth is estimated to be the most important feature for predicting the probability of seroconversion. This is unsurprising, as the HIV incidence estimates in Gray et al. (2021) were 4.3 infections per 100 person-years in females

and 1.3 infections per 100 person-years in males. In this cohort, the ranking of feature groups across the three VIMs is similar, with behavioral features ranking behind sex assigned at birth, followed by the remaining five feature groups. The magnitude of the estimated VIMs is similar or slightly larger at the later landmark times, which suggests that the importance of sex assigned at birth and sexual behavior features, while not particularly large, are stable over the time horizons of interest in this context. Among participants assigned female sex at birth, the feature ranking is different than that observed in the combined cohort. In particular, sexual behavior features appear to have decreased importance among females compared to the combined cohort. As in the combined cohort, the estimated VIMs and associated confidence intervals are similar among the three time horizons. Among participants assigned male sex at birth, sexual behavior features have the highest estimated importance, with a point estimate of around 0.1 at the 18-month time horizon. In this cohort, the magnitude of estimated VIMs is generally lower at later time horizons. Notably, as there were only 37 seroconversions observed among males, the confidence intervals in the male cohort are substantially wider than those in the combined and female cohorts. We performed tests of the one-sided null hypothesis of zero importance for each feature group, none of which achieved statistical significance at a 0.05 level after adjusting for multiplicity using a Bonferroni correction.

In Figure 2.4, we present the results of the marginal VIM analysis. For this analysis, we considered the importance of each feature group relative to geographic confounders; the reduced prediction model included geographic features only, while the full prediction model included geographic features plus the feature group of interest. In the combined cohort, sex assigned at birth is again the most important feature and is deemed statistically significant at all three time horizons after adjusting for multiplicity ($p = 1.5 \times 10^{-5}$, $p = 0.00015$, and $p = 0.00022$ at 18, 24, and 30 months, respectively). Sexual health features, sexual behavior features, and age are of secondary importance after sex assigned at birth, with all three achieving significance at 18 months of follow-up ($p = 0.0061$, $p = 0.0026$, and $p = 0.00047$, respectively). Housing features and BMI rank last at all time horizons. As in

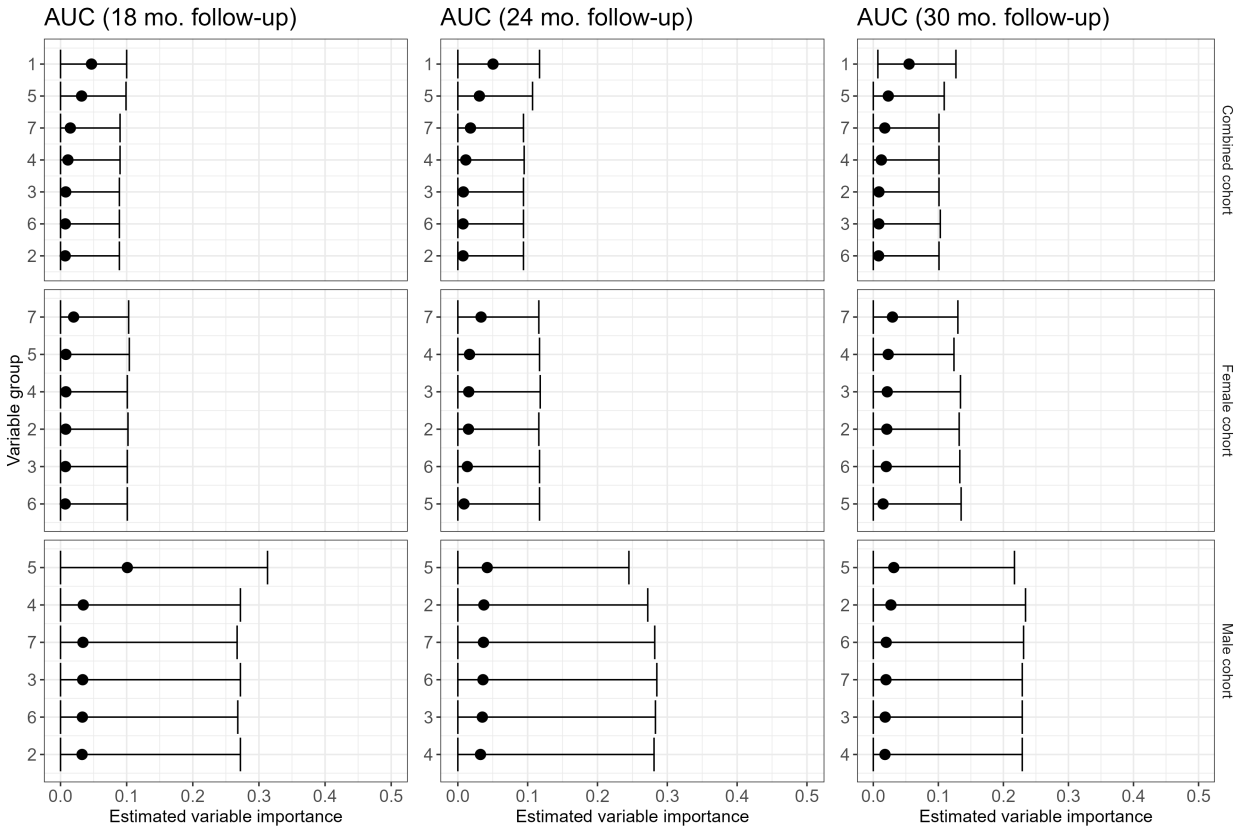

Figure 2.3: Conditional VIM analysis. Rows correspond, from top to bottom, to the combined male and female cohort, female cohort, and male cohort. Columns correspond, from left to right, to AUC VIM evaluated at 18, 24, and 30 months of follow-up. Feature groups are given by (1) sex assigned at birth; (2) age; (3) BMI; (4) sexual health features; (5) sexual behavior features; (6) housing features; (7) geographic confounders.

the conditional VIM analysis, the feature group rankings and estimated VIM magnitudes are relatively stable over the three horizons. The relatively larger importance of sexual health features in the marginal analysis versus the conditional analysis could be due in part to an observed association of sex assigned at birth with having a prevalent STI in this dataset. (Roughly 17% of males and 29% of females had a prevalent STI.) This emphasizes that considering correlations between features should inform the interpretation of any VIM analysis.

In the female cohort, sexual health features have the largest estimated marginal importance across all three time horizons, although the magnitude is small. Among males, sexual behavior features are estimated to be the most important feature group, with substantially larger point estimates than other feature groups at all three time horizons. Hypothesis tests of non-zero importance in the female and male cohorts do not reach statistical significance.

2.7 Discussion

In this chapter, we propose a framework for nonparametric estimation and inference on variable importance in survival analysis. We define a broad class of predictiveness measures tailored to time-to-event outcomes, which includes many existing predictiveness measures. As long as a rich enough set of covariates is observed, this class is identified by the observed data, even under informative censoring. We provide a closed-form EIF for all measures within this class and leverage this EIF in a one-step estimation procedure using flexible estimation of nuisance parameters. The estimator depends on consistent estimation of the oracle prediction function f_0 but enjoys doubly robust consistency with regard to the nuisance functions F_0 and G_0. It is asymptotically linear and nonparametric efficient when both F_0 and G_0 are consistently estimated. Numerical experiments show that our method possesses good operating characteristics in finite samples.

Analyses of the HVTN 702 trial suggests that both the rank and magnitude of VIMs for predicting HIV seroconversion are relatively stable over time horizons between 18 and 30 months. However, there is a substantial amount of uncertainty in the VIM estimates,

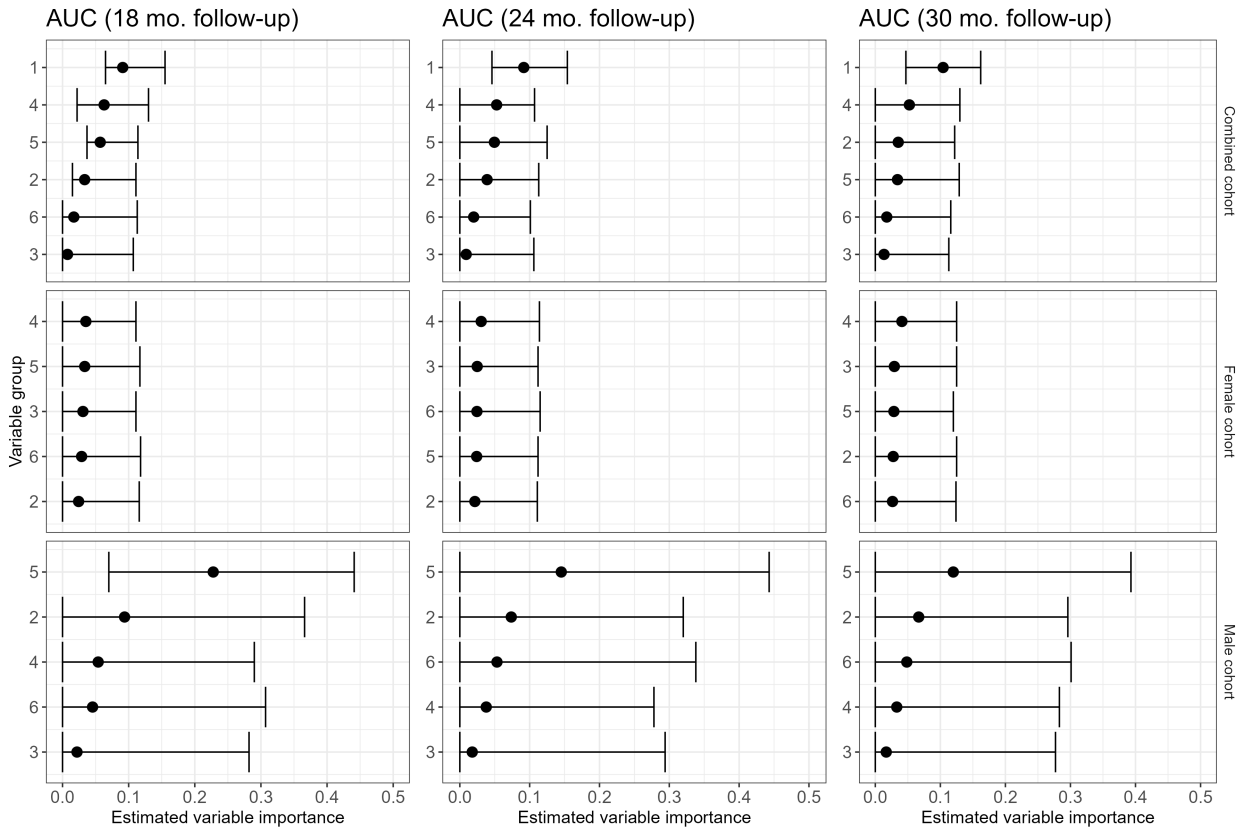

Figure 2.4: Marginal VIM analysis. Rows correspond, from top to bottom, to the combined male and female cohort, female cohort, and male cohort. Columns correspond, from left to right, to AUC VIM evaluated at 18, 24, and 30 months of follow-up. Feature groups are given by (1) sex assigned at birth; (2) age; (3) BMI; (4) sexual health features; (5) sexual behavior features; (6) housing features. Predictiveness is evaluated relative to a base model that uses only geographic confounders.

and similar analyses of future trials are warranted. It may also be of interest to investigate VIMs in other populations in which HIV vaccine trials are planned. For populations in which baseline features are more strongly predictive of seroconversion risk, we may expect to find more signal of variable importance.

As mentioned in Section 2.4.5, the task of estimating f_0 must be handled on a case-by-case basis. For many examples, an estimate of F_0 can be used to produce an estimate of f_0 without fitting any additional algorithms. Another viable approach involves IPCW. When f_0 can be written as a conditional mean of a transformation of T given covariates X, the IPCW pseudo-outcome mapping proposed by Koul et al. (1981) provides one means of estimation. This method relies on consistent estimation of G_0 and is a two-step procedure: first, the IPCW weights are estimated, and then an additional regression is fit to estimate the desired conditional mean. Rubin and van der Laan (2007) propose a doubly robust pseudo-outcome regression procedure to estimate a conditional mean of a transformation of T, which in some contexts could produce a doubly robust estimate of f_0.

2.8 Acknowledgements

The authors thank the study participants and investigators of the HVTN 702 trial conducted by the HIV Vaccine Trials Network. The authors also thank Dr. Michele Andrasik for offering valuable scientific guidance. This work was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2140004. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Chapter 3

A FRAMEWORK FOR LEVERAGING MACHINE LEARNING TOOLS TO ESTIMATE PERSONALIZED SURVIVAL CURVES

Charles J. Wolock, Peter B. Gilbert, Noah Simon & Marco Carone

3.1 Introduction

In the analysis of time-to-event data, the conditional survival function is a key quantity of interest. Within the biomedical field, the conditional survival function, which describes the distribution of an outcome variable conditional on a set of covariates, is especially relevant for prediction. For example, the survival function of a clinical outcome, such as death or disease recurrence, conditional on baseline characteristics may allow a clinician to better understand a patient's medical prognosis. The conditional survival function also appears as a function-valued nuisance parameter in nonparametric and semiparametric survival analysis problems (see, for example, Díaz, 2019 and Westling et al., 2023).

Typically, the analysis of survival data is complicated by the fact that the data are subject to censoring, truncation, or both, depending on the study design. In prospective studies, participants are sampled from the population of interest and followed over time, ideally until experiencing the event of interest. However, loss to follow-up or study termination may preclude observation of the event time. Participants who do not experience the event of interest during follow-up are considered right-censored. Additionally, individuals who have already experienced the event at study initiation are not eligible for recruitment. This sampling constraint is referred to as left truncation. Conversely, in retrospective studies, individuals must have already experienced the event in order to be recruited into the study, leading to right truncation. In this design, censoring is generally not a concern. All forms

of truncation lead to systematic selection bias.

There is a substantial literature focused on estimating the conditional survival function, which we briefly review in Section 3.2. Many existing methods directly or indirectly aim to minimize an empirical risk based on one of two loss functions: the inverse probability of censoring weighted (IPCW) loss for survival function estimation at a single time-point, or the hazard loss for estimation of a discrete-time hazard function (Polley and van der Laan, 2011). The IPCW loss requires estimation of the conditional survival function of the censoring variable and does not correct for truncation-induced sampling bias, while the discrete-time hazard loss does not apply to events occurring in continuous time. Methods employing objective functions for risk stratification, such as the Cox partial likelihood (Cox, 1972), do not explicitly target survival function estimation but may produce estimates as a byproduct.

In this chapter, we consider decompositions of the conditional survival function that allow use of standard loss-based estimation of functionals of the observed data distribution. These decompositions underlie our proposed method, called *global survival stacking*, which involves estimating a small number of binary regression functions using tools neither specially designed to handle censoring nor truncation. The strengths of this approach include:

1. it is a general framework in which practitioners can employ any off-the-shelf learner designed for binary regression or classification;

2. it can be applied in prospective (left truncation and right censoring) and retrospective (right truncation) settings without assuming a discrete-time process;

3. it simultaneously yields estimates of both the event time and censoring time conditional survival functions using the same fitted regressions.

The chapter is organized as follows: In the remainder of this section, we review existing methods for conditional survival function estimation. In Section 3.3, we describe the data

structures emerging from survival studies, provide identification results that form the basis for our estimation framework, and propose the global survival stacking procedure. In Section 3.4, we evaluate the performance of our class of estimators and demonstrate their use on data from the STEP HIV vaccine trial. In Section 3.5, we provide concluding remarks. Technical details and additional results can be found in Appendix B. We have implemented global survival stacking in the `R` package `survML` (`https://github.com/cwolock/survML`). Code to reproduce all results is available online at `https://github.com/cwolock/stack_supplementary`.

3.2 Review of related work

3.2.1 Classical methods

Even under right censoring alone, standard regression techniques cannot be directly applied to estimation of the conditional survival function. However, a number of survival-specific approaches have been proposed. Parametric methods, such as exponential or Weibull regression, are straightforward to use and automatically yield inference. Since their validity relies on strong distributional assumptions, they are less widely used than semiparametric methods. The most common regression model used to study survival outcomes is the Cox proportional hazards model (Cox, 1972). Hazard ratio estimates from the estimated Cox model can be combined with an estimate of the baseline cumulative hazard function (e.g., the Breslow (1972) estimator) to yield a conditional survival function estimate (Lin et al., 1994). A common alternative to the Cox model is the accelerated failure time model (Wei, 1992), which is usually implemented in a fully parametric manner. Semiparametric implementations (Buckley and James, 1979; Lin and Chen, 2013) exist but are seldom used because they are complicated and can be unstable (Zeng and Lin, 2007).

Under independent censoring (i.e., independence of the event and censoring times), the Kaplan-Meier estimator (Kaplan and Meier, 1958) is the nonparametric maximum likelihood estimator of the marginal survival function. If the covariates of interest are low-dimensional

and discrete-valued, a stratified Kaplan-Meier approach may be reasonable. This method breaks down in moderate dimensions, or when the covariates include continuous variables. As such, it is of limited use in most applications. Beran (1981) introduced a conditional Kaplan-Meier estimator using kernel smoothing. However, kernel-based methods tend to perform poorly as the number of covariates grows. Furthermore, the smoothing bandwidth can, in general, be allowed to vary for each covariate; selection of an optimal set of bandwidths can be computationally expensive.

3.2.2 Risk stratification methods

Fortunately, machine learning methods offer many strategies for estimating complex functions of moderate or large numbers of covariates in a flexible manner. For this reason, there has been a recent proliferation of machine learning methods in survival analysis — see Sonabend (2021) for a comprehensive review. This motivates a discussion of the precise objectives of these methods. Estimation of the conditional survival function and risk stratification are distinct tasks that are often conflated. The Cox proportional hazards model (and related machine learning methods) are based on the partial likelihood. Maximization of the partial likelihood is equivalent to maximization of the expected concordance between estimated risk scores and survival times (Tarkhan and Simon, 2022). Indeed, the partial likelihood has no dependence on actual event times and relies only on the relative ordering of events. Due to this fact, methods based on the Cox proportional hazards model might be best understood as risk stratification techniques. It is often the case that conditional survival function estimates can be derived from the resulting risk stratification algorithm (e.g., combining the Breslow baseline hazard estimate with a fitted Cox model), but ultimately these are a byproduct rather than the core goal of stratification approaches.

Along these lines, flexible and high-dimensional implementations of the Cox model have become increasingly common. Regularized regression methods such as LASSO, ridge, and elastic net have been implemented in a proportional hazards framework (Tibshirani, 1997; Simon et al., 2011). Flexible modeling of the proportional hazards risk score has been imple-

mented within the framework of generalized additive models (Hastie and Tibshirani, 1986). In addition, many deep learning survival analysis methods are built on the proportional hazards model and have largely been used to estimate complex interactions between covariates. Examples of non-linear proportional hazards neural networks include DeepSurv (Katzman et al., 2018) and Cox-nnet (Ching et al., 2018).

Tree-based methods are another popular machine learning strategy. Random survival forests (Ishwaran et al., 2008) encompass several previously proposed approaches; they are built similarly as standard random forests, but tree splits are determined based on stratification objective functions that can be evaluated with censored data (e.g., the log-rank test statistic). One difficulty with tree methods is evaluating predictions, which is important for choosing tuning parameters (or more generally, discriminating between candidate learners). Ishwaran et al. (2008) suggest using Harrell's concordance index (Harrell et al., 1982) to evaluate out-of-bag prediction error; this statistic measures discrimination rather than predictive accuracy per se. Once the random survival forest is constructed, cumulative hazard estimates for an individual with covariate vector x are obtained by dropping x down each tree, and using the Nelson-Aalen cumulative hazard estimator (Nelson, 1969; Aalen, 1978) in the resulting leaf node. Averaging the Nelson-Aalen estimates over all trees yields an overall cumulative hazard estimate, from which a survival function estimate can be obtained. In conditional inference forests (Hothorn et al., 2006b), predictors on which to split are chosen using hypothesis testing rather than exhaustive search, but otherwise a similar splitting principle as in random survival forests is used. Relative risk forests (Ishwaran et al., 2004) generalize the random forest algorithm to survival outcomes using Poisson regression under a proportional hazards model.

3.2.3 Methods based on censoring-weighted loss functions

Stratification objective functions circumvent the hurdle of comparing observed and predicted outcomes when some observations are censored. An alternative strategy is to use IPCW to connect a full-data loss function, such as squared-error loss, to a loss that can be evaluated

using the observed data (van der Laan and Robins, 2003). Defining an observed-data loss function allows the use of many established learning methods. Molinaro et al. (2004) use IPCW in the context of regression trees, where inverse-weighted loss functions can be used to both grow trees and evaluate performance. Hothorn et al. (2006a) use IPCW to extend ensemble learning techniques such as random forests and boosting to survival data. These methods rely on estimation of the conditional survival function of the censoring variable in order to construct the inverse probability weights. When the censoring variable can be assumed to be independent of covariates (for example, in studies where censoring times are controlled by the investigator), a simple marginal survival function estimator, such as Kaplan-Meier, can be used. But in general, estimation of the conditional survival function of the censoring variable is no easier than estimation of the conditional survival function of the outcome. Furthermore, in the presence of truncation, a candidate loss function must be adapted to account for the induced sampling bias.

Recently, Westling et al. (2023) framed the conditional event time and censoring survival functions as minimizers of oracle risks, which allows for iterative empirical risk minimization in order to combine multiple candidate estimators in a Super Learner approach (van der Laan et al., 2007). This method, termed survival Super Learner, is appealing because unlike usual IPCW loss functions, which are evaluated at a single time-point, it targets the entire survival function and simultaneously provides estimates of the outcome and censoring distributions. However, the candidate survival function estimators comprising the Super Learner are limited to existing survival-specific methods. In addition, in their current implementations, the oracle risk functions do not account for truncation.

3.2.4 Discrete-time methods

Many methods aimed at estimation of the conditional survival function, as opposed to risk stratification, rely on the assumption that events occur in discrete time. For discrete time-to-event variables, the hazard function at a single time is a conditional probability whose estimation can be framed as a binary regression problem in terms of the observed data

distribution: among those who have not experienced the event by time t, what proportion experience the outcome at that time? Reframing survival function estimation as a binary regression or classification problem allows use of a wider array of machine learning algorithms. Estimation of the survival function at time t involves computing the product of one minus the hazard at each time-point up to and including t.

For some approaches, time is discretized and the conditional hazard is estimated at each time based on a separate binary regression. Methods in this category include multitask logistic regression (MTLR), in which the hazards in discrete-time bins are modeled as separate logistic regressions (Yu et al., 2011). Penalized regression enforces smoothness over time, since event status in one time bin depends on what occurred in previous time bins. Censored observations are handled by marginalizing over possible sequences of survival statuses for all remaining times after censoring. In order to better learn nonlinear hazard-predictor relationships, Fotso (2018) proposed a deep neural network adaptation of the MTLR framework. Other discretization-based neural network implementations include RNN-SURV (Giunchiglia et al., 2018) and Nnet-survival (Gensheimer and Narasimhan, 2019), which estimate the hazard in user-specified time bins. In these implementations, guidelines for choosing time bins, as well as handling individuals that are censored within a particular time bin, are provided. Friedman (1982) proposed a piecewise constant exponential hazards model, where time is treated as continuous, but with a constant hazard in user-specified time bins.

Perhaps the most flexible discretization method is what has recently been referred to as "survival stacking" (Craig et al., 2021), and which we hereafter call *local survival stacking*. The approach dates back at least to work by Polley and van der Laan (2011). Craig et al. (2021) propose to implement the approach by discretizing at each observed event time. With time discretized, a survival dataset can be transformed into a longitudinal data set, where each individual appears in the data set at each observed event time until exiting the risk set. Time itself is included as a covariate, often as a dummy variable indicating whether an individual is in the risk set at a given time, or as a continuous predictor. Fitting a logistic regression without interactions on this "stacked" data set (with time treated as a

dummy variable) approximates fitting the Cox proportional hazards model using the partial likelihood as the grid of time-points becomes increasingly fine (D'Agostino et al., 1990). In contrast to methods discussed above, local survival stacking does not involve estimating separate regressions at each discrete time, but rather estimates a single regression including time as a covariate. The tradeoff is that the size of the stacked data grows at rate $O(n^2)$ for a sample of size n, and so computational issues may be a concern. It is unclear how performance depends on the choice of time discretization. Beyond right censoring, left truncation is dealt with naturally in discrete-time models, since each individual appears in the longitudinal data for as many time-points as they remain in the risk set. Left truncation is handled by only including the individual at those time-points after which they have entered the study.

Outside of the discrete-time framework, few proposed methods have explicitly viewed conditional survival function estimation in terms of the observed data distribution. One such method uses generative adversarial networks to learn the joint distribution of the observed data (Zhou et al., 2022); however, it is tied to a specific machine learning architecture and does not handle truncation.

3.3 Materials and methods

3.3.1 Ideal data and parameter of interest

Suppose that X is a vector of baseline covariates taking values in $\mathcal{X} \subset \mathbb{R}^p$, and $T \in (0, \infty)$ is the event time of interest. The ideal data unit is $O^* := (X, T)$. We use P^* to denote the distribution of O^*. In reality, O^* is observed subject to both censoring and truncation, which are determined by the study design. The observed data consist of n independent and identically distributed observations $O_1, O_2, \ldots, O_n$ drawn from P, the observed data distribution implied by P^*. The relationship between P^* and P is determined by the censoring and sampling mechanisms. Our goal is estimation of the conditional survival function of T given X, defined as $S(t \,|\, x) := P^*(T > t \,|\, X = x)$. Because T is not directly observed, this parameter is not a functional of the observed data distribution. However, with an addi-

tional assumption, the conditional hazard function (and through it, the conditional survival function) can be identified. Our method relies on a reformulation of standard identification results in order to write the hazard function in terms of observable regression functions.

Let $\Lambda(t \,|\, x) := \int_0^t \frac{F(du \,|\, x)}{1-F(u^- \,|\, x)}$ denote the conditional cumulative hazard of T given $X = x$ at time t, where $F := 1 - S$ is the conditional distribution function of T. Identification of S in full generality requires the use of product integrals via the mapping $S(t \,|\, x) = \prod_{u \in (0,t]} \{1 - \Lambda(du \,|\, x)\}$. When the mapping $t \mapsto F(t \,|\, x)$ is differentiable everywhere, the product integral simplifies to the exponential form $S(t \,|\, x) = \exp\{-\Lambda(t \,|\, x)\}$.

Epidemiological studies are often conducted to learn characteristics of the distribution of time from an initiating event (e.g., disease onset) until a terminating event (e.g., death). Here, we treat the time of the initiating event as $t = 0$, and use the event time T to refer to the time between initiating and terminating events.

3.3.2 Identification

To start, we consider prospective studies in which individuals who have not yet experienced the event of interest are sampled and followed over time. Ideally, every participant is followed until the event has occurred, but right censoring is essentially inevitable in prospective biomedical studies. Participants who do not experience the event during follow-up are considered right-censored. This may be due to loss to follow-up or to termination of the study. Let $C \in (0, \infty)$ denote the right censoring time. For each participant in the study, we observe $Y := \min\{T, C\}$, the observed follow-up time, and $\Delta := \mathbb{1}(T \leq C)$, the event indicator.

Common prospective observational study designs include: (a) the incident cohort — people who have not experienced the initiating event upon entering the study, and can be followed from the initiating event onward; and (b) the prevalent cohort — people who experienced the initiating event prior to entering the study. A study sample may also contain both prevalent and incident cases.

Because prevalent cases have already experienced the initiating event upon study entry, observation of these participants does not begin at $t = 0$. This phenomenon is commonly

referred to as delayed entry, and it implies that the event times are observed subject to left truncation. Left truncation induces sampling bias, since individuals with larger event times are more likely to enter the sample. Let $W \in (0, \infty)$ denote the time from the initiating event until entry into the study. Under left truncation, an individual can only enter the study (i.e., be observed) if $W \leq Y$. The observed data for participants in the sample are $O := (X, Y, \Delta, W)$, and the sampling criterion is $W \leq Y$. If a prospective study consists only of incident cases, there is no left truncation. In that special case, $W = 0$ for all participants.

To identify $\Lambda(\cdot \,|\, x)$ in the prospective setting, we rely on the following assumption:

Assumption C: T and (C, W) are conditionally independent given X.

Let $F_\delta(y \,|\, x) := P(Y \leq y \,|\, \Delta = \delta, X = x, W \leq Y)$ denote the conditional distribution function of Y among observed participants with $\Delta = \delta$. Let $\pi(x) := P(\Delta = 1 \,|\, X = x, W \leq Y)$ denote the probability of a random observed individual being uncensored. In addition, define $G_\delta(y \,|\, x) := P(W \leq y \,|\, \Delta = \delta, X = x, Y \geq y, W \leq Y)$. We note that these regressions are all functionals of the observed data distribution P. As detailed in Section B.2, we then have that $\Lambda(\cdot \,|\, x)$ can be identified at generic time t by

$$\Lambda^{\text{obs}}(t \,|\, x) := \int_0^t \frac{\pi(x) F_1(du \,|\, x)}{G_1(u \,|\, x)\pi(x)\left\{1 - F_1(u^- \,|\, x)\right\} + G_0(u \,|\, x)\left\{1 - \pi(x)\right\}\left\{1 - F_0(u^- \,|\, x)\right\}} \,. \quad (3.1)$$

When there is no left truncation, the distribution of W is degenerate at 0, so that $G_1(u \,|\, x) = G_0(u \,|\, x) = 1$ for all u. In that case, Λ^{obs} is a function only of the conditional distributions of Y given (Δ, X) and Δ given X.

Assumption C can be considered when C is defined for all individuals in the target population. However, in some settings, censoring may only act on enrolled participants. It may then be more appropriate to consider an alternative assumption expressed in terms of residual censoring (Qian and Betensky, 2014). In Section B.2, we show that (3.1) still holds, and so our proposed estimation strategy is still valid, under one such alternative assumption.

In the retrospective setting, we consider studies in which investigators only sample individuals who have experienced the terminating event prior to the end of the sampling period. For example, in autopsy studies where death is the terminating event, an individual who did not die prior to the end of the study could not enter the sample. This sampling scheme results in right truncation, which, similarly as left truncation, induces sampling bias. In Section B.1, we show that the above identification results can be directly applied to the retrospective setting by simply considering the time scale to be reversed.

3.3.3 Estimation procedure

The results of Section 3.3.2 suggest that we can construct an estimator of $\Lambda(\cdot \,|\, x)$ by estimating a small number of regression functions based on the observed data. This hazard estimator can then be mapped to a survival function estimator via either the product integral or exponential mappings. The regression functions that appear in the identification results constitute either: (i) a conditional probability (specifically, $\pi(x)$), or (ii) a conditional cumulative probability function ($F_1, F_0, G_1,\ G_0$). These regression functions can be estimated using standard machine learning techniques, without requiring any adaptation for the censoring or sampling mechanisms.

Let $t_{\max}$ denote the maximum time at which the survival function is to be estimated, and let $t \in (0, t_{\max}]$ be a generic time-point of interest. Below we outline the steps to estimate $S(t \,|\, x)$. Our proposed procedure is as follows:

1. Select an approximation grid: Choose a partition $\mathcal{B} := \{t_0, t_1 \ldots, t_{\max}\}$ of the interval $[0, t_{\max}]$ (t_0 will often be 0).

2. Estimate the cumulative hazard: For each $t_j \in \mathcal{B}$, obtain estimators $F_{1,n}(t_j \,|\, x)$, $F_{0,n}(t_j \,|\, x)$, $\pi_n(x)$, $G_{1,n}(t_j \,|\, x)$, and $G_{0,n}(t_j \,|\, x)$ of $F_1(t_j \,|\, x)$, $F_0(t_j \,|\, x)$, $\pi(x)$, $G_1(t_j \,|\, x)$, and $G_0(t_j \,|\, x)$ respectively.

3. Approximate a mapping from the hazard to the survival function: Let $t_k := \max\{t' \in$

$\mathcal{B} : t' \leq t\}$. Define the estimated differential of the cumulative hazard at t_i as $M_n(t_i, x)$, given by

$$\frac{\pi_n(x)\left\{F_{1,n}(t_i \,|\, x) - F_{1,n}(t_{i-1} \,|\, x)\right\}}{G_{1,n}(t_i \,|\, x)\pi_n(x)\left\{1 - F_{1,n}(t_{i-1} \,|\, x)\right\} + G_{0,n}(t_i \,|\, x)\left\{1 - \pi_n(x)\right\}\left\{1 - F_{0,n}(t_{i-1} \,|\, x)\right\}},$$

where $M_n(t_0, x) := 0$. For the product integral form, approximate the product integral using the product $S_{n,p}(t \,|\, x) := \prod_{i=1}^{k}\left\{1 - M_n(t_i, x)\right\}$. For the exponential form, approximate the exponentiated negative cumulative hazard using the Riemann sum approximation $S_{n,e}(t \,|\, x) := \exp\left\{-\sum_{i=1}^{k} M_n(t_i, x)\right\}$.

The product integral form of the estimator is the more natural option, since the product integral mapping holds whether T has a discrete, continuous, or mixed distribution. However, in practice, $S_{n,p}$ can have numerical issues, particularly in the right tail of the distribution of Y. We discuss this in depth in Section B.5.

Approximating the product integral and the cumulative hazard requires choosing an approximation partition $\mathcal{B}$ of the interval $[0, t_{\max}]$. A simple option for this is the set of observed follow-up times $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$, where $Y_{(j)}$ denotes the jth order statistic. Alternatively, $\mathcal{B}$ could be set to an evenly spaced grid of times between 0 and $t_{\max}$. In large samples, it may be more computationally practical to use a grid of fixed size rather than including every observed follow-up time.

3.3.4 Constructing the constituent regressions

Our procedure requires estimation of π, F_1, F_0, and, when truncation is present, G_1 and G_0. Estimating the conditional event probability function $\pi(x)$ is a simple binary regression problem, for which there are numerous flexible methods. In practice, we recommend using a boosted classifier, such as boosted trees (Friedman, 2001), or an ensemble regression method such as Super Learner (van der Laan et al., 2007). The observed follow-up time distribution functions $F_1(\cdot \,|\, x)$ and $F_0(\cdot \,|\, x)$ are slightly more complicated to estimate. At any fixed time t, these can be viewed as a binary regression on the indicator variable $\mathbb{1}(Y \leq t)$. However,

we must estimate these distribution functions on the grid $\mathcal{B}$ of times in order to approximate the product- or sum-integral of the hazard function. A simple and flexible approach is to perform pooled binary regression on a user-specified time grid $\mathcal{C}$. This grid could be the same as the approximation grid $\mathcal{B}$, but in general it need not be. A natural choice for $\mathcal{C}$ would be the observed follow-up times. A coarser grid speeds computation at the cost of increased bias. At each time-point t in the grid, the available data are baseline covariates, an indicator outcome variable $\mathbb{1}(Y \leq t)$, and time t. These data are pooled across time into a single dataset, which serves as training data for binary regression. This approach differs from local survival stacking in that the risk set at each time-point consists of all participants, and the outcome is cumulative across times. To ensure monotonicity in time, we recommend isotonizing the distribution function estimates using isotonic regression (see for example Westling et al., 2020).

The conditional entry time regression functions $G_1(\cdot \,|\, x)$ and $G_0(\cdot \,|\, x)$ are similar to conditional distribution functions, although they are each conditioned on being at-risk for an event at time t. Similarly as the conditional distribution functions, these functions can be easily estimated using pooled binary regression. Given a time grid, the data at each time-point t consist of all individuals who remain under follow-up at time t, along with covariates, time t, and the outcome $\mathbb{1}(W \leq t)$.

Because pooled binary regression involves "stacking" datasets across time-points, we refer to the resulting estimation procedure as global survival stacking, which we differentiate from the discrete-time hazard approach of local survival stacking. The global stacking procedure is detailed in Algorithm 3.1.

3.3.5 Comparison to local survival stacking

Local survival stacking (Polley and van der Laan, 2011; Craig et al., 2021) is a natural alternative to the proposed framework since it allows practitioners to draw upon a wide array of general machine learning techniques. When using local survival stacking, the user must choose how to discretize time. Local survival stacking assumes a discrete survival

Algorithm 3.1 Global survival stacking

Choose grid $\mathcal{B} := \{t_0, t_1, \ldots, t_{\max}\}$ for approximation of product- or sum-integral.
Construct estimator $\pi_n(x)$ of $\pi(x)$ using binary regression.

Estimate F_1 and F_0

for $\delta \in \{0, 1\}$ **do**
    Choose grid of time-points $\mathcal{C} := \{t_1^*, t_2^*, \ldots, t_k^*\}$ on which to discretize F_δ.
    Choose how to include time in model (continuous, dummy variable, etc.).
    for $t_j^* \in \mathcal{C}$ **do**
        Including only participants with $\Delta = \delta$, construct dataset $D_{t_j^*}$ consisting of participant baseline covariates, outcomes $\mathbb{1}(Y \leq t_j^*)$, and time using chosen basis.
    end for
    Construct full stacked dataset by combining $\{D_{t_1^*}, D_{t_2^*}, \ldots, D_{t_k^*}\}$.
    Fit binary regression or classification algorithm of choice.
    Generate predictions $\{F_{\delta,n}(t_0 \,|\, x), F_{\delta,n}(t_1 \,|\, x), \ldots, F_{\delta,n}(t_{\max} \,|\, x)\}$.
end for

Estimate G_1 and G_0 (if truncation is present)

for $\delta \in \{0, 1\}$ **do**
    Choose grid of time-points $\mathcal{C} := \{t_1^*, t_2^*, \ldots, t_k^*\}$ on which to discretize G_δ.
    Choose how to include time in model (continuous, dummy variable, etc.).
    for $t_j^* \in \mathcal{C}$ **do**
        Including only participants with $\Delta = \delta$ and $Y \geq t_j^*$, construct dataset $D_{t_j^*}$ consisting of participant baseline covariates, outcomes $\mathbb{1}(Y \leq t_j^*)$, and time using chosen basis.
    end for
    Construct full stacked dataset by combining $\{D_{t_1^*}, D_{t_2^*}, \ldots, D_{t_k^*}\}$.
    Fit binary regression or classification algorithm of choice.
    Generate predictions $\{G_{\delta,n}(t_0 \,|\, x), G_{\delta,n}(t_1 \,|\, x), \ldots, G_{\delta,n}(t_{\max} \,|\, x)\}$.
end for

Combine constituent estimators

Compute $\{M_n(t_0, x), M_n(t_1, x), \ldots, M_n(t_{\max}, x)\}$, as detailed in Section 3.3.3.
Compute $S_{n,p}(t \,|\, x)$ or $S_{n,e}(t \,|\, x)$ as detailed in Section 3.3.3.

process, so that the conditional hazard takes the form of a conditional probability that can be estimated for each time-point in the grid. The discretization is usually chosen on the basis of the observed event times $\mathcal{R}_n := \{Y_i : \Delta_i = 1, i = 1, 2, \ldots, n\}$. In an illustrative data analysis, Polley and van der Laan (2011) choose 30 time-points based on quantiles of $\mathcal{R}_n$, while Craig et al. (2021) define local survival stacking based on discretizing at each time in $\mathcal{R}_n$. The fineness of the time grid determines the number of events used to estimate the conditional probability of an observed event at each time-point, and we would expect the grid choice to affect performance. The fineness of the time grid may also be relevant for global survival stacking, although we emphasize that the outcome is cumulative over time, meaning that the probability of an outcome at any given time does not shrink as the grid becomes finer. The experiments in Section 3.4 explore the performance of these methods under various grid sizes. In Section B.4, we provide an operational description of local survival stacking.

3.4 Results

3.4.1 Primary simulation studies

We conducted several simulation studies to evaluate the performance of our proposed method. In addition to overall estimation performance, i.e., mean squared error, we aimed to assess the sensitivity of global survival stacking to the choice of time grid used for estimating F and G via pooled binary regression. As discussed in Section 3.3.5, we expected global survival stacking to be less sensitive to the grid choice compared to local survival stacking due to the fact that the regression outcome is cumulative over time.

The methods compared in our simulations are described in Table 3.1, with full details given in Section B.4. Briefly, we included global survival stacking, local survival stacking, survival Super Learner, random forests (specifically, LTRC conditional inference forests, Fu and Simonoff, 2017), a linear Cox model, and a generalized additive Cox model (Hastie and Tibshirani, 1986). For both global and local survival stacking, binary regressions were estimated using a Super Learner consisting of the marginal mean, logistic regression with all

Method	Package	Truncation?	Description
Global surv. stacking	`survML`	Left, right	Proposed method
Local surv. stacking	`survML`	Left, right	Discrete hazard approach
surv. Super Learner	`survSuperLearner`	No	Ensemble survival regression
LTRC forests	`LTRCforests`	Left	Conditional inference forest
Linear Cox	`survival`	Left	Linear PH model
Gen. additive Cox	`mgcv`	No	Gen. additive PH model

Table 3.1: Estimators included in simulation studies. PH indicates proportional hazards.

pairwise interactions, generalized additive models, multivariate adaptive regression splines, random forests, and gradient-boosted trees. For global stacking, we considered three time grids $\mathcal{C}$ for the pooled regression: a grid made up of every observed follow-up time and grids of 10 or 40 cutpoints evenly spaced on the quantile scale of observed follow-up times. For local stacking, the same three time grids were included, based on observed event times $\mathcal{R}_n$.

The simulation scenarios are summarized in Table 3.2, with Scenarios 1 and 2 described here. We simulated a covariate vector $X := (X_1, X_2, \ldots, X_{10})$ of 10 independent components. These components included continuous covariates $X_1, X_2 \sim \text{Uniform}(-1, 1)$, discrete covariates $X_3, X_4 \sim \text{Uniform}(\{-1, 1\})$, and continuous covariate $X_5 \sim N(0, 1)$. The five additional covariates were independent standard normal noise, i.e., $(X_6, X_7, \ldots, X_{10}) \sim \text{MVN}(0, \mathbf{I}_5)$. Given covariate vector $X = x$, we simulated the censoring time C from a Weibull distribution with shape 1.5 and scale $\lambda_C = \exp\left\{\beta_{0C} + \frac{1}{2}(x_1 + x_2) + \frac{1}{5}(x_3 + x_4 + x_5)\right\}$, where in each simulation setting β_{0C} was chosen to give a censoring rate of 25%. Given covariate vector $X = x$, we independently simulated the event time T to be distributed as $100Z_1$. In the left-skewed scenario, Z_1 was a $\text{Beta}(a(x) + 2, 2)$ random variable with $\log a(x) = x_1 + x_2 + x_3 + x_4 + x_5 + x_1x_2 + x_3x_4 + x_1x_5$. In the right-skewed scenario, Z_1 was a $\text{Beta}(2, a(x) + 2)$ random variable. Density plots for T given X for 10 random draws from the covariate distribution are given in Section B.4. These distributions do not meet the proportional hazards assumption. Given covariate vector $X = x$, the study entry time variable W was distributed as $100Z_2$, where Z_2 was a $\text{Beta}\big(1 + \frac{1}{2}\mathbb{1}(x_1 > 0), 1 + \frac{1}{2}\mathbb{1}(x_1 < 0)\big)$

Scenario	Study design	Description
1	Prospective	right-censored, non-proportional hazards
2	Prospective	left-truncated, right-censored, non-proportional hazards
3	Retrospective	right-truncated, no censoring, non-proportional hazards
4	Prospective	left-truncated, right-censored, proportional hazards
5	Prospective	left-truncated, right-censored, time observed on discrete grid

Table 3.2: Simulation scenarios. Appendix B contains results for Scenarios 3 – 5.

random variable. In Scenario 2, in which left truncation was present, only observations with $Y \geq W$ were sampled. In Scenario 1, in which there was no truncation, all observations were sampled. The average truncation rates for all simulation settings are given in Section B.4.

We evaluated performance using Monte Carlo approximations of mean squared error (MSE) at three landmark times and mean integrated squared error (MISE) over the interval $[0, 100]$. We computed MSE at landmark times corresponding to the 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times. To calculate the MISE, we computed the MSE at each time on an evenly spaced grid of 1000 points from $t = 0.1$ to $t = 100$, and took a simple average over times. We estimated the performance metrics using a test set of size 1000. The test data were generated without truncation in order to evaluate performance across the marginal distribution of covariates in the target population.

Figures 3.1 and 3.2 display a subset of results for Scenarios 1 and 2, with global and local stacking implemented using a 40 cutpoint grid. The full results, including comparisons of global and local stacking with various grid sizes, are given in Figures B.2 and B.3. From Figures 3.1 and 3.2, we observe that global survival stacking performs well both with and without truncation. The performance of local survival stacking is more variable, but it performs particularly well in the right-skewed setting without truncation. Without truncation, survival Super Learner performs reasonably well, although it is outperformed by global survival stacking in the left-skewed setting and by both global and local survival stacking in the right-skewed setting. The LTRC forests method performs similarly with or without

truncation, with performance on par with local stacking in Scenario 2. The Cox model is misspecified, and the performances of both linear and generalized additive Cox models do not improve substantially with sample size.

From Figures B.2 and B.3, we see that for global survival stacking, finer grids generally yield slight to moderate decreases in MISE and MSE compared to the 10 cutpoint grids. As expected, the performance of local survival stacking appears to be more sensitive to grid size choice. Among local stacking implementations, the grid of 40 cutpoints performs the best in general. The 10 cutpoint grid appears too coarse for optimal performance, while the finest grid performs well in the right-skewed settings but poorly in the left-skewed settings.

3.4.2 Additional simulation studies

In Section B.5, we present additional results. In Scenario 3, we evaluated our procedure under a retrospective study design with right truncation. As in the prospective study design, global stacking demonstrates strong performance, with the finer grids generally outperforming the coarsest grid. In Scenario 4, in which the data were generated from a distribution satisfying the proportional hazards assumption, we found that the correctly specified Cox model yields moderately better performance than the machine learning comparators. Global stacking shows generally good performance, with relatively small differences between different choices of grid size compared with local stacking. While we expect the proportional hazards assumption to rarely hold in practice, predictably, the Cox model would be the preferred method in this situation but with only modest performance loss from more flexible methods. Finally, in Scenario 5, Y and W were observed on a discrete grid of times, rather than in continuous time. When Y and W are observed on a discrete grid of 10 or 20 times, global and local stacking demonstrate similar performance, while global survival stacking performs the best overall when the data are observed on a grid of 50 times. These experiments show that there is relatively little difference between global and local survival stacking when times are observed on a coarse grid, and the advantages of global survival stacking become more pronounced on a finer grid.

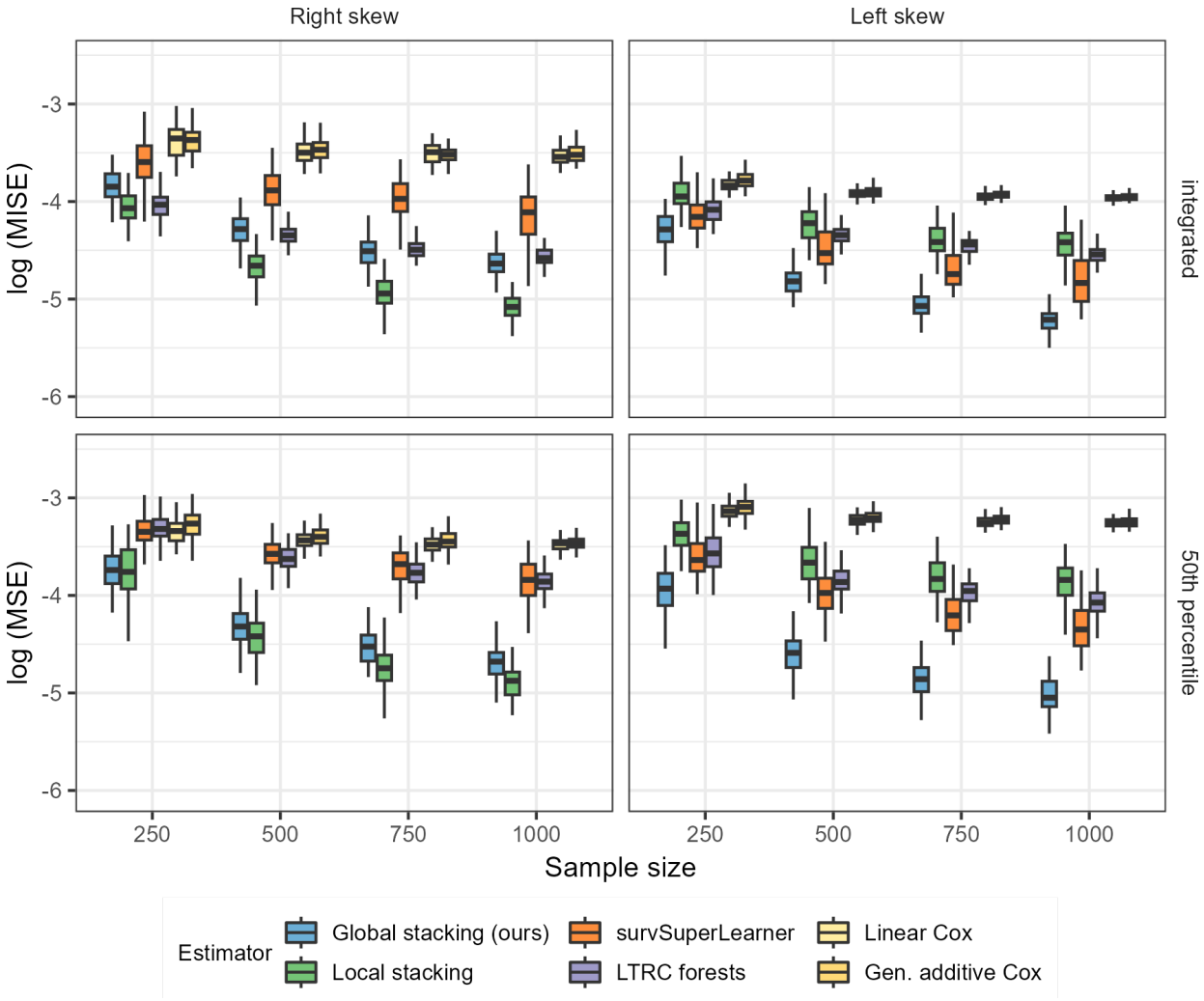

Figure 3.1: Performance of conditional survival estimators with right-censored data (Scenario 1). The methods compared were global survival stacking, local survival stacking, survival Super Learner, LTRC forests, a main-terms linear Cox proportional hazards model with Breslow baseline hazard estimator, and a main-terms generalized additive Cox proportional hazards model with Breslow baseline hazard estimator. Rows correspond to MISE (top) and MSE at the 50^{th} percentile of observed event times (bottom).

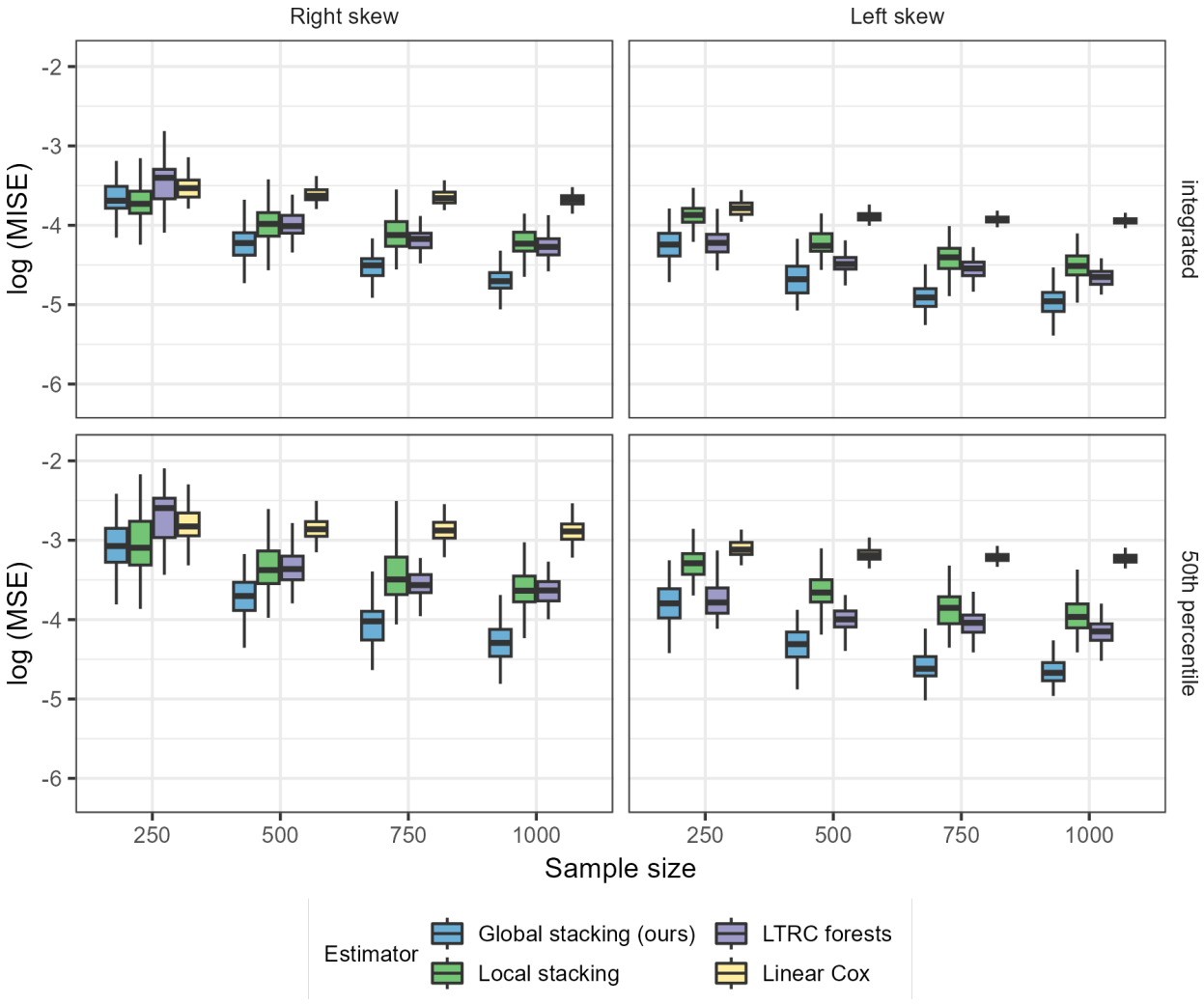

Figure 3.2: Performance of conditional survival estimators with left-truncated, right-censored data (Scenario 2). The methods compared were global survival stacking, local survival stacking, LTRC forests, and a main-terms linear Cox proportional hazards model with Breslow baseline hazard estimator. Rows correspond to MISE (top) and MSE at the 50^{th} percentile of observed event times (bottom).

We also conducted a computational benchmarking experiment, finding that local survival stacking, survival Super Learner, and LTRC forests were faster than global survival stacking, and that, unsurprisingly, the computation time required for both global and local survival stacking increases as the grid becomes finer (i.e., as the number of times in $\mathcal{C}$ increases).

3.4.3 Predictive performance on time-to-event datasets

We also evaluated our proposed method on several publicly available datasets with right-censored time-to-event outcomes, which are described in Section B.6. We predicted the survival probability at three landmark times, corresponding to the 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times in each dataset, with performance evaluated using the Brier score. To account for censoring, we used the IPCW Brier score given by Gerds and Schumacher (2006), with Kaplan-Meier censoring weights. For each of the five datasets, we compared global survival stacking, local survival stacking, and LTRC forests. Both global and local stacking were implemented with 40 cutpoints using the same algorithm library as in the simulations. We also included a naïve approach in which, for predicting the survival probability at time t, the outcome $\mathbb{1}(Y > t)$ was regressed on X. We used the Super Learner, with the same algorithm library as in global and local stacking, to fit this binary regression. We used five-fold cross-validation to estimate the Brier score of each of the methods under consideration. The performance of each method was evaluated relative to the performance of the marginal model constructed without covariates using the Kaplan-Meier estimator (i.e., using the same prediction for every observation in the test set).

Global survival stacking performs well in all five datasets (Table 3.3). The naïve model typically has relatively poor performance but does slightly outperform the other methods at two landmark times in the SUPPORT dataset, where the censoring rate is zero. This is unsurprising: for prediction at landmark time t, observations censored after t provide the same information as uncensored observations. When the censoring rate is higher — for example, in the METABRIC dataset — the naïve approach is outperformed by the methods that account for censoring.

Dataset	Quantile	Censoring	Performance relative to KM			
			Global stacking	Local stacking	LTRC forests	Naïve
FLCHAIN	50^{th}	0.07	**0.748**	0.752	0.773	0.758
	75^{th}	0.19	**0.685**	0.690	0.707	0.694
	90^{th}	0.31	**0.646**	0.658	0.672	0.663
GBSG	50^{th}	0.03	**0.858**	0.883	0.876	0.892
	75^{th}	0.07	**0.828**	0.842	0.860	0.967
	90^{th}	0.17	**0.841**	0.853	0.864	1.120
METABRIC	50^{th}	0.07	**0.892**	0.910	0.912	0.914
	75^{th}	0.19	**0.886**	0.888	0.899	0.980
	90^{th}	0.30	0.877	**0.869**	0.876	1.049
NWTCO	50^{th}	0.02	**0.859**	0.862	0.918	0.860
	75^{th}	0.04	**0.866**	0.875	0.906	0.896
	90^{th}	0.11	**0.865**	0.871	0.900	0.995
SUPPORT	50^{th}	0.00	0.931	0.952	0.948	**0.927**
	75^{th}	0.00	0.909	0.923	0.919	**0.908**
	90^{th}	0.08	**0.879**	0.892	0.885	0.904

Table 3.3: Predictive performance of candidate methods on publicly available survival datasets. The performance metric is the Brier score standardized by the Brier score of the Kaplan-Meier (KM) estimator (i.e., predicting survival probability without using covariate information). The Brier score was evaluated at three landmark times corresponding to the 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times. Lower values are preferred. Boldface font indicates the best performance for each dataset and landmark time. The methods compared were global survival stacking (our proposed method), local survival stacking, LTRC forests, and a naïve binary regression approach ignoring censoring.

3.4.4 Assessing risk of HIV infection in the STEP trial

Between December 2004 and March 2007, 3,000 HIV-negative individuals were enrolled in the STEP study (HVTN 502/Merck 023), a randomized, placebo-controlled phase 2b trial that tested the efficacy of a candidate HIV vaccine to prevent acquisition of HIV-1 infection. The vaccine contains an adenovirus serotype 5 (Ad5) vector that expresses subtype B HIV-1 *gag/pol/nef* proteins. Participants were at high risk of HIV-1 acquisition. Participants were unblinded in October 2007 after the prespecified monitoring boundary for efficacy futility was crossed at the first interim analysis (Buchbinder et al., 2008). Data analyses suggested an increased risk of HIV-1 infection among vaccine recipients versus placebo recipients, particularly among participants who were uncircumcised or had neutralizing antibodies against the Ad5 vector at enrollment ("baseline Ad5 titer").

In order to assess the risk of HIV-1 infection conditional on circumcision status and baseline Ad5 titer, we estimated the conditional survival function of the time-to-infection-diagnosis variable within randomized treatment arms at landmark times of one year and two years of follow-up, corresponding to approximately 60% and 10% of participants still at-risk in each treatment arm. We limited our analyses to the 1,836 participants with male sex assigned at birth in the modified intention-to-treat cohort, which included all vaccinated participants except those diagnosed as HIV-1 positive on or before the day 1 visit. At one year of follow-up, 41 participants in the vaccine arm (4.6%) and 27 participants in the placebo arm (3.0%) had been diagnosed with HIV-1; at two years, 51 participants in the vaccine arm (5.7%) and 35 participants in the placebo arm (3.9%) had been diagnosed. We implemented global survival stacking using Super Learner with the same algorithm library as in the simulations, using a grid of 40 cutpoints based on quantiles of observed follow-up times (i.e., the number of days from randomization until the end of follow-up, for each participant) with five-fold cross-validation for tuning. For comparison, we also fit a Cox model including the two-way circumcision/baseline Ad5 titer interaction and estimated the baseline cumulative hazard function using the Breslow estimator. Both models were fit separately in the two treatment

arms. Baseline Ad5 titer was log-transformed (using the natural logarithm), and titers under the assay detection limit of 18 were treated as equal to 18 for analysis (Duerr et al., 2012). We calculated the risk difference conditional on circumcision status and baseline Ad5 titer by taking the difference of the estimated conditional survival functions in the two treatment arms at each landmark time. Using global survival stacking, we also computed representative survival curves for individuals in each treatment arm, circumcised and uncircumcised, at log baseline Ad5 titer values of 3, 5, and 7.

The estimated survival curves (Figure 3.3) show that, in the vaccine group, the probability of HIV-1 diagnosis through day 730 tends to be higher for individuals with higher baseline Ad5 titers. The probability of HIV-1 diagnosis was higher in the vaccine arm than the placebo arm, as estimated by both global stacking and the Cox model, except at low baseline Ad5 titers among circumcised participants (Figure 3.4). The estimated excess risk in the vaccine arm tends to increase with baseline Ad5 titer and is generally higher among uncircumcised participants, although the Cox model fit suggests that circumcised participants may have slightly larger excess risk at high baseline Ad5 titers. Overall, these results agree with the original analysis in Duerr et al. (2012), which did not explicitly account for right censoring.

3.5 Discussion

In this chapter, we proposed a framework for estimating a conditional survival function in both prospective and retrospective settings using flexible machine learning tools. This framework, which we call global survival stacking, relies on an identification of the hazard function in terms of observable regressions that can be estimated using standard methods for binary outcomes, without the need to explicitly account for censoring or truncation. Similarly as with local survival stacking, our approach recasts conditional survival function estimation as a statistical learning task that does not require specially tailored survival analysis tools. These methods not only enable practitioners to take advantage of the myriad machine learning methods currently available, but also to harness the improved performance of new methods as they are developed. Numerical experiments show that global survival

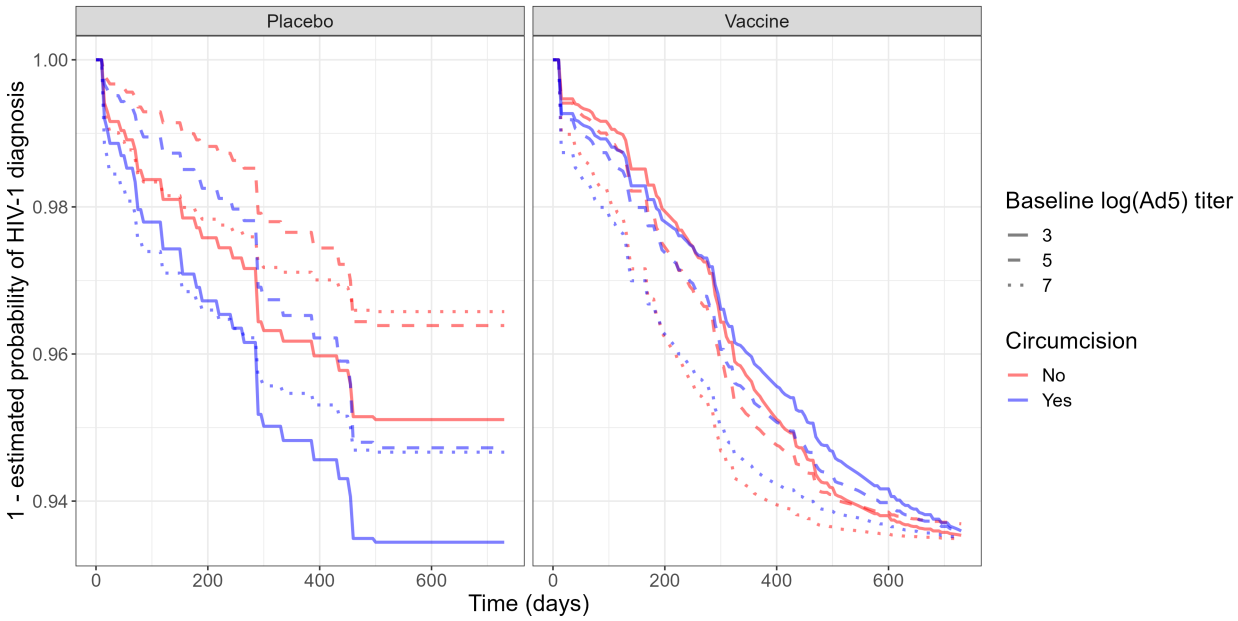

Figure 3.3: Estimated survival curves for time to HIV-1 diagnosis in the STEP study. The curves were estimated separately in each treatment arm, conditional on baseline Ad5 titer and circumcision status.

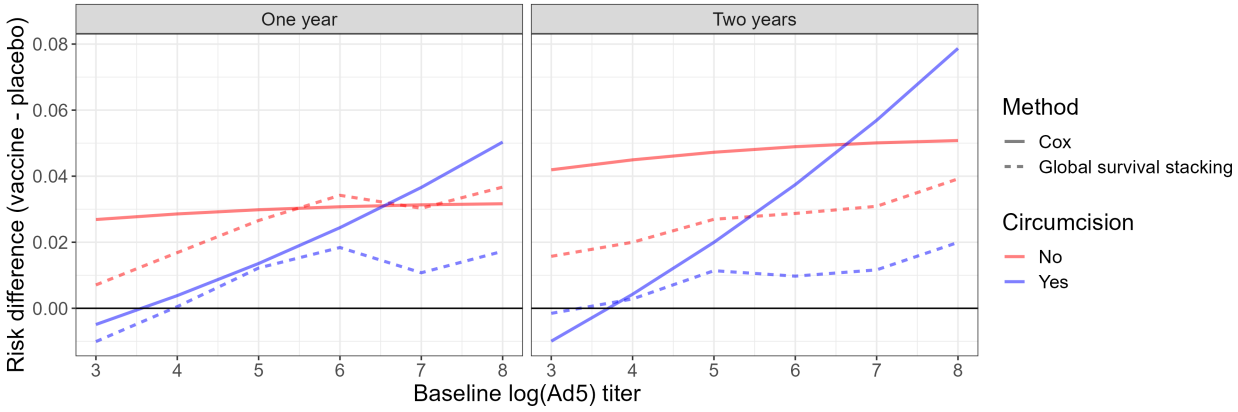

Figure 3.4: Estimated risk difference (vaccine - placebo) of HIV-1 infection diagnosis in the STEP study conditional on baseline Ad5 titer and circumcision status at one year and two years of follow-up. The estimators compared were the Cox model with first-order interaction, and global survival stacking.

stacking works well across a variety of settings, performing on par with or better than competing methods when the proportional hazards assumption fails to hold. Global survival stacking is relatively insensitive to the choice of grid size in the pooled binary regression.

Both global and local survival stacking can be computationally expensive, particularly when the number of cutpoints in the time grid is allowed to grow with sample size. Global survival stacking requires fitting multiple regressions on data sets that are generally larger than those that arise in local survival stacking. Based on the performance of global survival stacking, there appears to be no harm in using as fine a time grid as computational resources and time allow. If an analysis is only performed once on a dataset of modest size, using a grid of every observed follow-up time may be reasonable. However, in our experiments there was little gain, if any, for the computational cost, and global stacking suffered virtually no decrease in performance using a relatively coarse grid of fixed size. In practice, of course, the computational resources required for using any ensemble regression method will depend on which algorithms are included in the library.

Because our method involves estimating regression functions within strata defined by the event indicator, we can use the same procedure to obtain an estimate of the conditional censoring distribution, simply replacing $\pi_n(x)$ with $1-\pi_n(x)$ and $F_{1,n}(t \mid x)$ with $F_{0,n}(t \mid x)$ in the numerator of $M_n(t, x)$ in Section 3.3.3. While any conditional survival function estimation algorithm can be repurposed by reversing the roles of T and C, our approach requires no refitting, resulting in greater computational efficiency when both distributions are desired.

Unlike right censoring, interval censoring is a common type of data coarsening that remains unaddressed by many survival function estimators. Interval-censored event times are known only to lie in a particular interval, rather than being observed exactly. Data that are truly subject to interval censoring (e.g., data from biomedical studies with periodic follow-up) are often treated as subject only to right censoring. Whether or not the flexible machine learning methods presented here can be adapted to handle interval censored data remains an open question.

3.6 Acknowledgments

The authors thank the study participants and investigators of the STEP HVTN 502/Merck 023 trial conducted by the HIV Vaccine Trials Network. The authors also thank Alex Luedtke for his insightful comments. Research reported in this publication was supported by National Institute Of Allergy And Infectious Diseases grants UM1-AI068635 and R37-AI029168, and by National Heart, Lung, and Blood Institute grant R01-HL137808. This work was also supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2140004. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Chapter 4

MULTIPLE SAMPLE SPLITTING FOR ALGORITHM-AGNOSTIC VARIABLE IMPORTANCE

Charles J. Wolock, Marco Carone & Noah Simon

4.1 Introduction

Model-agnostic variable importance methods provide a framework for assessing the value of a feature or group of features in predicting an outcome. The strong predictive performance of black box algorithms such as random forests (Breiman, 2001) and gradient-boosted trees (Friedman, 2001) across diverse settings While black box prediction algorithms such as random forests (Breiman, 2001) and gradient-boosted trees (Friedman, 2001) demonstrate strong predictive performance in diverse settings, unlike classical regression methods, they explicitly decouple the tasks of prediction and estimation of the association between features and outcome. The desire to understand the role of a feature in a prediction task has led to use of variable importance measures (VIMs) based on nonparametric predictiveness or goodness-of-fit (Williamson et al., 2021a,b; Zhang and Janson, 2022; Verdinelli and Wasserman, 2021; Hudson, 2023). Within this general framework, the importance of a feature is defined as the decrease in predictiveness when that feature is excluded from the prediction model (alternatively, the increase in predictiveness when the feature is added to a prespecified base model). The literature on model-agnostic variable importance is closely connected to the topic of conditional independence testing, which can also be framed as a comparison of the goodness of fit of nested prediction models (see, e.g., Dai et al., 2022; Lundborg et al., 2022).

To assess the importance of a feature, we define a pair of nested function classes, which differ only by whether they contain functions that are permitted to depend on the feature of

interest. Without loss of generality, we define the full class as containing prediction functions which may depend on the feature of interest along with all other available features, and the reduced class as containing prediction functions that may not depend on the feature of interest. The maximum predictiveness achievable by a function in the full class is called the full oracle predictiveness, while the residual oracle predictiveness is the maximum predictiveness achievable without the feature of interest.

Several previous works on variable importance have framed the difference between full and residual predictiveness as a pathwise differentiable parameter in a nonparametric model, thereby allowing the use of tools from semiparametric efficiency theory (e.g., one-step estimation procedures, Pfanzagl, 1982) in order to achieve $n^{1/2}$-rate inference (Williamson et al., 2021a,b; Verdinelli and Wasserman, 2021). Difficulty arises when the feature of interest has importance exactly equal to zero, as this implies that the pathwise derivative of the importance parameter is equal to zero as well. This occurs because the full and residual oracle prediction functions are identical, and the resulting degeneracy renders invalid the theoretical guarantees of inferential techniques based on the first-order pathwise derivative. Williamson et al. (2021b) and Dai et al. (2022) independently proposed a sample splitting procedure to circumvent this problem, in which the full and residual predictiveness are estimated on separate segments of the data set. These segments are often halves, i.e., the data set is divided evenly in two. However, other splitting proportions still provide asymptotically valid inference and may in fact be preferable (Lundborg et al., 2022), although we do not attempt do address that question here.

There are two major drawbacks to the existing sample splitting approach.

1. Sub-optimal efficiency: There are two senses in which sample splitting is not optimal. First, sample splitting reduces the effective sample size used to estimate each of the full and residual oracle predictiveness. When the true importance in non-zero, then, we can expect the sample split estimator to have a larger asymptotic variance compared to the estimator constructed without sample splitting, leading to wider confidence

intervals. Moreover, when the true importance is zero, the sample split estimator converges at rate $n^{-1/2}$, while faster rates are often achievable. Hudson (2023) proposes an estimator of variable importance that achieves n-consistency estimators under the null, and Lundborg et al. (2022) detail a hypothesis testing procedure that, in certain models, has power against local alternatives converging to zero at a rate faster than $n^{-1/2}$. While sample splitting is appealing both for its simplicity and for the fact that it yields inference (both confidence intervals and hypothesis tests) valid under null and alternative hypotheses, it does not necessarily yield optimal procedures.

2. Introduction of additional randomness: Sample splitting requires randomly partitioning the data into two segments and is therefore a randomized statistical procedure. To be precise, conditional on the data, the output of the sample splitting procedure is random rather than fixed. This additional randomness is generally viewed as undesirable, as it may cause the substantive result of an analysis to depend on the seed selected by the investigator. In the context of variable importance, there are other potential sources of extra randomness besides the sample splitting procedure. For example, the algorithms used to estimate the oracle prediction functions may themselves be randomized, whether due to selection of tuning parameters via cross-validation or due to the use of subsampling (as in random forests). Furthermore, debiased machine learning estimators, like those use to estimate variable importance, generally demonstrate improved performance when cross-fitting is employed (Zheng and van der Laan, 2011; Chernozhukov et al., 2018). Cross-fitting entails further subdivision of the data set in a random manner.

Naturally, for any randomized statistical procedure, it is possible to perform the analysis multiple times, ensuring that the randomness of each iteration — conditional on the data — is independent of the others. This can be achieved in practice by using a different seed for each iteration. The collected results can then be aggregated in some manner. For example, if the output of a procedure is a test statistic, then one may take, for example, the arithmetic

mean or the median of the collection of the test statistics. Using the mean as an example, the variance of the aggregated test statistic conditional on the data scales as the reciprocal of the number of iterations. For a large number of iterations, then, the excess randomness can be made as small as desired. Figure 4.1 demonstrates this phenomenon when using the cross-fit, sample-split survival VIM estimator detailed in Chapter 2 in order to estimate the importance of BMI in predicting the two-year probability of HIV seroconversion in the combined male and female cohort of HVTN 702 (Gray et al., 2021). The histograms show the distribution of the test statistic (the point estimate scaled by $n^{1/2}$ and divided by its estimated asymptotic variance) across 10000 iterations of the procedure, each using a different seed. Without aggregation, there is substantial variation across iterations, which, because the data are fixed, is entirely due to the randomized procedure. But by averaging the test statistics across iterations, the excess variation in the point estimate can be greatly reduced.

There is a substantial literature on methods for aggregating exchangeable test statistics or p-values in order to perform inference. Guo and Shah (2023) provides a thorough review of such methods. Some approaches are meant to handle statistics with arbitrary joint distributions; for example, applying a Bonferroni correction to the minimum of a collection of p-values guarantees control of the family-wise error rate and hence the probability of a type I error. Alternatively, one may attempt to learn the distribution of the aggregated statistic under the null hypothesis (Guo and Shah, 2023). Construction of confidence intervals additionally requires knowledge of this distribution across non-null parameter values.

In this chapter, we discuss approaches for aggregating multiple iterations of sample splitting in the context of asymptotically linear variable importance estimators. Of particular interest is the behavior of multiple sampling splitting when the number of iterations is permitted to grow with the sample size. We provide a rate condition on the number of splitting iterations, finding that the maximum allowable rate depends on the convergence rate of a second-order remainder term arising in an analysis of the single-split estimator. We assess the performance of multiple sample splitting in numerical experiments and find that in many scenarios, it can drastically increase power over a single split while maintaining control of

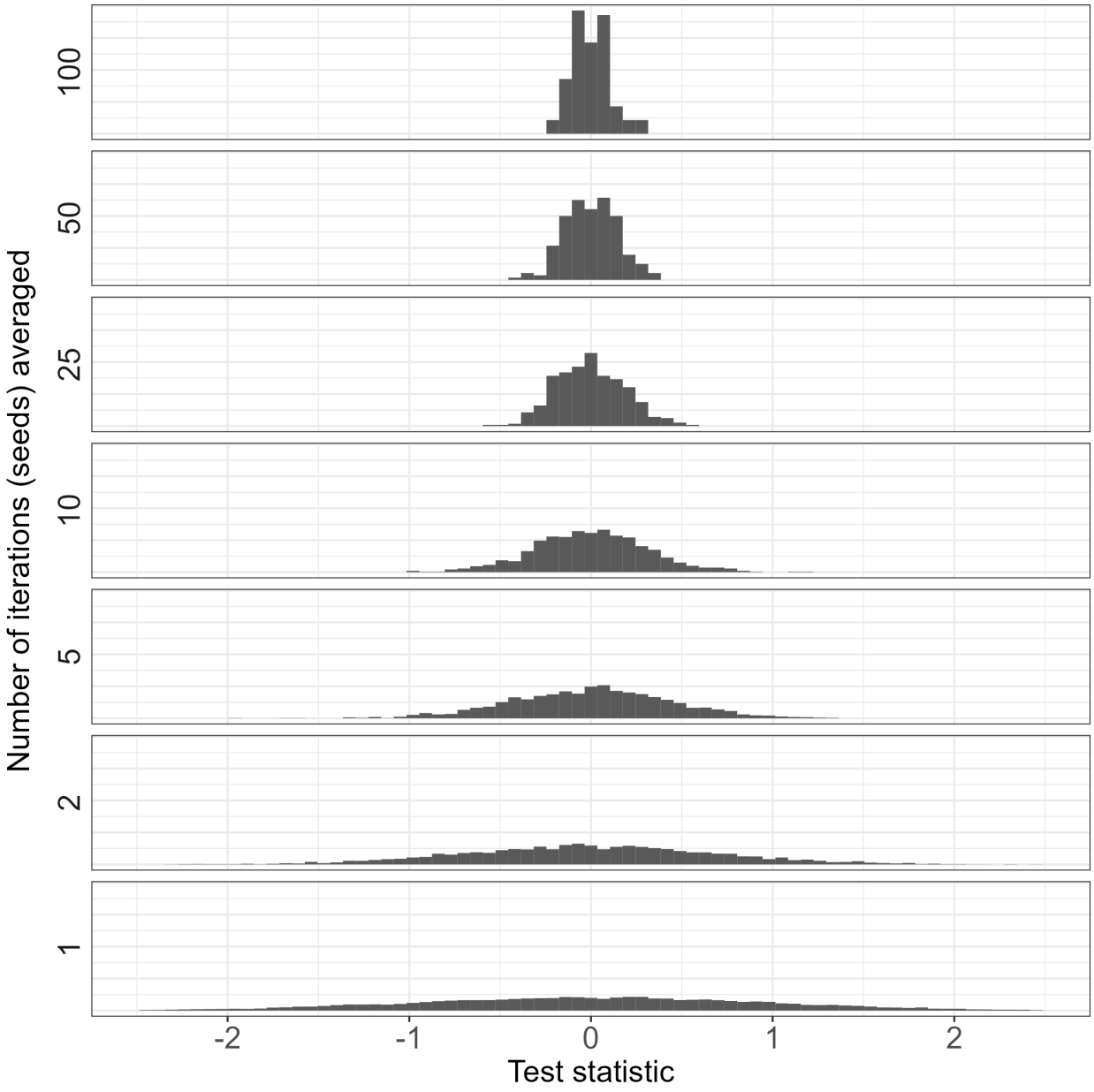

Figure 4.1: Histograms showing the distribution of test statistics across 10000 iterations of the cross-fit, sample-split procedure used to estimate the importance of BMI in predicting the risk of HIV seroconversion in HVTN 702.

type I error. This result demonstrates the promise of this approach in real-world variable importance analyses.

4.2 Variable importance

4.2.1 Data structure and notation

We observe a vector of baseline covariates X taking values in $\mathcal{X} \subseteq \mathbb{R}^p$. Interest lies in predicting an outcome Y taking values in $\mathcal{Y} \subseteq \mathbb{R}$. We use P_0 to denote the distribution of the data unit $Z := (X, Y)$ and assume that P_0 belongs to a nonparametric model $\mathcal{M}$. Our sample consists of n independent and identically distributed observations $Z_1, Z_2, \ldots, Z_n$ draws from P_0. For the sake of exposition, we assume that the data are fully observed, i.e., there is no missingness or censoring. However, we note that the general problem and methods we discuss in this chapter apply to more complex settings than that which we describe here.

We use the subscript 0 to denote a functional of P_0, e.g. $\mathbb{E}_0[f(Z)] := \mathbb{E}_{P_0}[f(Z)]$. For a random function of a data unit, e.g. $f_n(Z)$, the expectation is taken with respect to the random data unit Z and not f_n. We use $\mathbb{P}_n$ to denote the empirical measure of $Z_1, Z_2, \ldots, Z_n$. We use the empirical process notation $Pf := \mathbb{E}_P[f(Z)]$ for probability measure P and P-measurable function f.

In discussing variable importance, we often refer to subgroups of covariates. We use $s \subset \{1, \ldots, p\}$ to denote the index set of a covariate subgroup. Then, for any vector v, v_s denotes the elements of v with index in s and v_{-s} denotes the elements of v with index not in s. The sample spaces of X_s and X_{-s} are denoted by $\mathcal{X}_s$ and $\mathcal{X}_{-s}$, respectively.

We use $\mathcal{F}$ to denote the class of potential prediction functions. The subset $\mathcal{F}_s := \{f \in \mathcal{F} : f(u) = f(v) \text{ for all } u, v \in \mathcal{X} \text{ satisfying } u_{-s} = v_{-s}\}$ characterizes prediction functions in $\mathcal{F}$ that ignore features with index in s. We allow $\mathcal{F}$ to be largely unrestricted up to regularity conditions.

4.2.2 Predictiveness and variable importance

For a prediction function $f \in \mathcal{F}$ and distribution $P \in \mathcal{M}$, we define $V(f, P)$ to be the predictiveness of f under P, with larger values indicating higher predictiveness. Throughout this section, we use as an example the binary classification accuracy predictiveness measure, defined as $V(f, P) := P(Y = f(X))$ for Y binary and $f : \mathcal{X} \to \{0, 1\}$.

In studying variable importance, the key prediction function is the oracle prediction function f_0 corresponding to the data-generating mechanism P_0, defined as

$$f_0 \in \text{argmax}_{f \in \mathcal{F}} V(f, P_0) \ .$$

If $\mathcal{F}$ is a rich function class, we can expect f_0 to depend only on the choice of predictiveness measure and on the distribution P_0. The oracle predictiveness is then defined as $v_0 := V(f_0, P_0)$. This quantifies the total predictive potential of the covariate vector X under P_0. For index set s, we define $f_{0,s} \in \text{argmax}_{f \in \mathcal{F}_s} V(f, P_0)$ as the residual oracle prediction function, with corresponding residual oracle predictiveness $v_{0,s} := V(f_{0,s}, P_0)$. This measures the predictive potential of X_{-s}, i.e., the maximum predictiveness achievable when features with index in s are not used in the prediction task. Our target of inference in the importance of feature(s) X_s, which we define as $\psi_{0,s} := v_0 - v_{0,s}$. There is particular interest in testing the null hypothesis $H_0 : \psi_{0,s} = 0$, i.e., that $v_0 = v_{0,s}$, versus the alternative $H_1 : \psi_{0,s} > 0$.

For classification accuracy predictiveness, Williamson et al. (2021b) showed that the oracle and residual oracle prediction functions are given by

$$f_0(x) = \mathbb{1}(\mu_0(x) > 0.5) \qquad \text{and} \qquad f_{0,s}(x) = \mathbb{1}(\mu_{0,s}(x) > 0.5) \ ,$$

where $\mu_0(x) := \mathbb{E}_0 [Y \,|\, X = x]$ and $\mu_{0,s}(x) := \mathbb{E}_0 [Y \,|\, X_{-s} = x_{-s}]$. These conditional mean functions play a key role in many example predictiveness functions, including R^2, deviance, and area under the receiver operating characteristic curve (AUC).

4.2.3 Plug-in estimation

Using the definition of $\psi_{0,s}$, the natural plug-in VIM estimator is given by $\psi^*_{n,s} := V(f_n, \mathbb{P}_n) - V(f_{n,s}, \mathbb{P}_n)$, where f_n and $f_{n,s}$ are estimators of the full and residual oracle prediction functions, respectively. Under regularity conditions, the contribution of the estimation of f_0 and $f_{0,s}$ to the asymptotic bias of $\psi^*_{n,s}$ is second-order. Because of this, the plug-in estimator, even if it involves flexible estimation of nuisance parameters, achieves asymptotic linearity. In particular, the representation

$$\psi^*_{n,s} - \psi_{0,s} = \frac{1}{n}\sum_{i=1}^{n}\{\phi_0(Z_i) - \phi_{0,s}(Z_i)\} + o_P\left(n^{-1/2}\right) \tag{4.1}$$

holds, where ϕ_0 is the efficient influence function (EIF) of $P \mapsto V(f_P, P)$ at P_0 relative to $\mathcal{M}$, and $\phi_{0,s}$ is the EIF of $P \mapsto V(f_{P,s}, P)$ at P_0 relative to $\mathcal{M}$. In other words, $\psi^*_{n,s}$ is $n^{1/2}$-consistent and nonparametric efficient, without requiring debiasing. The plug-in estimator for binary classification accuracy VIM is given by

$$\psi^*_{n,s} := \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{1}\left(Y_i = f_n(X_i)\right) - \mathbb{1}\left(Y_i = f_{n,s}(X_i)\right)\right\}.$$

In order to reduce the chance of systematic bias due to model misspecification in estimating f_0 and $f_{0,s}$, it is advisable to use flexible, data-adaptive algorithms, many of which are randomized. For example, some algorithms require the selection of tuning parameters, such as the regularization parameter in the lasso (Friedman et al., 2010). Cross-validation, which involves partitioning the data into folds in order to train and then evaluate the algorithm over a grid of possible tuning parameter choices, is a widely used procedure for tuning parameter selection. Bootstrap aggregation, or bagging, entails training an ensemble of learners on random subsamples of a data set, and is a key feature of algorithms such as random forests (Breiman, 1996, 2001). When the nuisance estimators f_n and $f_{n,s}$ involve randomized procedures, such as cross-validation or bagging, the overall estimation procedure becomes randomized as well. This is one stage at which additional randomness is injected.

Randomness may also be introduced as a byproduct of cross-fitting. A key regularity condition ensuring asymptotic negligibility of the remainder term in (4.1) involves constraints

on the complexity of the algorithms used to construct f_n and $f_{n,s}$. This is often referred to as a Donsker condition, and it may fail for flexible machine learning algorithms. In many settings, the use of cross-fitting obviates the need for Donsker conditions, thereby weakening the requirements for asymptotic linearity of $\psi^*_{n,s}$. As described in Chapter 2, cross-fitting entails splitting the data into K folds. For each $k \in 1, 2, \ldots, K$, the k^{th} fold is held out, and the remaining $k-1$ folds are used to construct f_n and $f_{n,s}$. Using these fitted algorithms, the plug-in estimator is constructed from the k fold. Repeating this procedure K times, once for each fold, yields K estimates $\psi_{n,s,1}, \psi_{n,s,2}, \ldots, \psi_{n,s,K}$. These K estimates are averaged to yield an overall estimator which shares the same first-order asymptotic behavior of $\psi^*_{n,s}$.

Like cross-validation, cross-fitting involves randomly partitioning the data into folds, which adds additional randomness. However, neither the use of randomized algorithms in constructing f_n and $f_{n,s}$ nor the use of cross-fitting changes the first-order asymptotic behavior of the VIM estimator. Because of this fact, and because the primary focus of this chapter is sample splitting, in the following sections we do not explicitly consider the effects of randomized prediction algorithms and cross-fitting.

4.3 Sample splitting

4.3.1 Existing sample splitting approach

As described in Section 4.2, previous work on nonparametric variable importance has provided procedures for constructing asymptotically linear plug-in estimators $V(f_n, \mathbb{P}_n)$ and $V(f_{n,s}, \mathbb{P}_n)$ of v_0 and $v_{0,s}$, respectively, with influence functions ϕ_0 and $\phi_{0,s}$. When H_0 holds, however, $\phi_0 = \phi_{0,s}$, implying that $\psi^*_{n,s} = V(f_n, \mathbb{P}_n) - V(f_{n,s}, \mathbb{P}_n)$ may fail to be asymptotically linear and that inferential procedures based on the influence function $\phi_0(z) - \phi_{0,s}(z)$ may be invalid.

One approach to achieve $n^{1/2}$-rate inference for $\psi_{0,s}$ under both H_0 and H_1 involves sample splitting (Williamson et al., 2021b; Dai et al., 2022). To perform sample splitting, we generate n independent Bernoulli(1/2) random variables $S_1, S_2, \ldots, S_n$. We use the data

$\mathcal{D}_n := \{Z_i : S_i = 1\}$ to construct an asymptotically linear plug-in estimator v_n of v_0 using the procedure described in Section 4.2, and the data $\bar{\mathcal{D}}_n := \{Z_i : S_i = 0\}$ to construct an asymptotically linear plug-in estimator $v_{n,s}$ of $v_{0,s}$. The overall sample-split estimator is then given by $\psi_{n,s} := v_n - v_{n,s}$, which, due to the independence of S and Z, admits the representation

$$\psi_{n,s} - \psi_{0,s} = \frac{1}{n}\sum_{i=1}^{n} 2\left\{S_i\phi_0(Z_i) - (1-S_i)\phi_{0,s}(Z_i)\right\} + r_n \ , \tag{4.2}$$

where $r_n = o_P\left(n^{-1/2}\right)$. This sample-split estimator converges weakly to a mean-zero Gaussian random variable with variance

$$\sigma^2_{0,s} := 4\mathbb{E}_0\left[\left\{S\phi_0(Z) - (1-S)\phi_{0,s}(Z)\right\}^2\right] = 2P_0\left(\phi_0^2 + \phi_{0,s}^2\right) \ . \tag{4.3}$$

4.3.2 Multiple sample splitting: fixed number of splits

A natural extension of the sample-splitting approach involves performing the procedure multiple times and aggregating the results. The sample splitting indicators $S_1, S_2, \ldots, S_n$ are randomly generated under control of the investigator, and so the sample-splitting procedure could be repeated by, for example, using different seeds in the generation of the splitting indicators. We let $1, 2, \ldots, B$ denote the indices of the seeds used to generate the splitting indicators. For observation i, the B splitting indicators are denoted $S_{i1}, S_{i2}, \ldots, S_{iB}$. We let $\psi_{n1,s}, \psi_{n2,s}, \ldots, \psi_{nB,s}$ denote the estimators constructed from the B iterations. In light of (4.2), the vector of B estimators constructed using the B iterations of the sample splitting procedure satisfy

$$\begin{pmatrix} \psi_{n1,s} - \psi_{0,s} \\ \vdots \\ \psi_{nB,s} - \psi_{0,s} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} 2\left\{S_{i1}\phi_0(Z_i) - (1-S_{i1})\phi_{0,s}(Z_i)\right\} \\ \vdots \\ 2\left\{S_{iB}\phi_0(Z_i) - (1-S_{iB})\phi_{0,s}(Z_i)\right\} \end{pmatrix} + \begin{pmatrix} r_{n1} \\ \vdots \\ r_{nB} \end{pmatrix}, \tag{4.4}$$

where $r_{n1}, r_{n2}, \ldots, r_{nB}$ denote remainder terms from each iteration.

Using this representation, we see that the vector of B sample-split estimators, centered and scaled by $n^{1/2}$, converges weakly to a B-dimensional Gaussian random vector with

covariance matrix denoted $\Sigma_s^{(B)}$. The diagonal elements of this covariance matrix are given by $\Sigma_{s,jk}^{(B)} = 2P_0(\phi_0^2 + \phi_{0,s}^2)$ for $j = k$, and the off-diagonal elements are given by

$$\Sigma_{s,jk}^{(B)} = P_0\phi_0^2 - 2P_0\left(\phi_0\phi_{0,s}\right) + P_0\phi_{0,s}^2 \ ,$$

for $j \neq k$. Thus, given a consistent estimator $\Sigma_{n,s}^{(B)}$ of the covariance matrix, we can use the asymptotic joint distribution of $\{\psi_{n1,s}, \psi_{n2,s}, \ldots, \psi_{nB,s}\}$, combined with a continuous transformation $G : \mathbb{R}^B \to \mathbb{R}$, to perform tests and construct confidence intervals simply by simulating draws from a multivariate Gaussian random vector with covariance matrix $\Sigma_n^{(B)}$ and applying the transformation G. Possible choices for G include the arithmetic mean and the maximum.

When G is the arithmetic mean, i.e., $G(u_1, u_2, \ldots, u_B) = \frac{1}{B}\sum_{b=1}^{B} u_b$, we can easily derive the limiting distribution of the aggregated statistic. Letting $\psi_{n,s}^{(B)} := G(\psi_{n1,s}, \psi_{n2,s}, \ldots, \psi_{nB,s})$, $S_i^{(B)} = G(S_{i1}, S_{i2}, \ldots, S_{iB})$, and $r_n^{(B)} := G(r_{n1}, r_{n2}, \ldots, r_{nB})$, we have that

$$\psi_{n,s}^{(B)} - \psi_{0,s} = \frac{1}{n}\sum_{i=1}^{n} 2\left\{S_i^{(B)}\phi_0(Z_i) - (1 - S_i^{(B)})\phi_{0,s}(Z_i)\right\} + r_n^{(B)}. \tag{4.5}$$

Using this representation, we can conclude that the aggregated estimator $\psi_{n,s}^{(B)}$, centered and scaled by $n^{1/2}$, converges weakly to a Gaussian random variable with variance $\sigma_{0,s,B}^2 := B^{-1}\tau_{0,s}^2 + \nu_{0,s}^2$, where $\tau_{0,s}^2 := P_0\left(\phi_0^2 + 2\phi_0\phi_{0,s} + \phi_{0,s}^2\right)$ and $\nu_{0,s}^2 := P_0\left(\phi_0^2 - 2\phi_0\phi_{0,s} + \phi_{0,s}^2\right)$. We note that $\sigma_{0,s,1}^2 = \sigma_{0,s}^2$, so this representation agrees with (4.3) when only a single split is performed. Under H_0, $\nu_{0,s}^2$ vanishes, and so the asymptotic variance of the estimator is equal to $B^{-1}\tau_{0,s}^2$.

4.3.3 Multiple sample splitting: growing number of splits

The results of Section 4.3.2 hold when B is fixed and n tends to infinity. In this section, we consider the case where B is allowed to grow with n. To make this dependence explicit, we now use B_n to denote the number of sample splitting iterations. For practical purposes, we may wonder how large B_n can be chosen without sacrificing performance, as under H_0 the distribution of $n^{1/2}\psi_{n,s}^{(B_n)}$ approaches degeneracy as B_n increases. (Heuristically, for fixed

B, when H_0 holds, $\sigma^2_{0,s,B} = B^{-1}\tau^2_{0,s}$, which will be near zero for B large.) Furthermore, it is of interest to explicitly characterize how much of the efficiency lost in sample splitting can be recovered by the multiple splitting approach, which necessitates consideration of the behavior of $\psi^{(B_n)}_{n,s}$ as both n and B_n grow. To this end, we begin by deriving the asymptotic distribution of $\psi^{(B_n)}_{n,s}$.

Theorem 4. *Suppose that* $\sup_{b\in\{1,2,\ldots,B_n\}} r_{nb} = O_P\left(n^{-\alpha}\right)$. *If* $B_n \asymp n^{\delta}$ *with* $1/2 < \alpha < (1+\delta)/2$, *then the random variable*

$$\left(\frac{n}{\sigma^2_{0,s,B_n}}\right)^{1/2}\left(\psi^{(B_n)}_{n,s} - \psi_{0,s}\right)$$

converges weakly to a standard Gaussian random variable.

The key implication of Theorem 4 is that the rate at which B_n may grow while preserving asymptotic normality is determined by the convergence rate of the remainder term $r^{(B_n)}_n$. Specifically, the more quickly $r^{(B_n)}_n$ tends to zero, the more quickly B_n may be allowed to grow while still allowing for valid inference based on a Gaussian limiting distribution.

Theorem 4 provides a guideline for constructing asymptotically valid hypothesis tests of H_0. Doing so requires a consistent estimator of $\tau^2_{0,s}$, which we denote $\tau^2_{n,s}$. For example, this estimator could be constructed as follows: (1) use the full sample $(Z_1, Z_2, \ldots, Z_n)$ to construct estimators $\mathbb{P}_n$, f_n, and $f_{n,s}$; (2) plug in $\mathbb{P}_n$ for P_0, f_n for f_0, and $f_{n,s}$ for $f_{0,s}$ in the definitions of ϕ_0 and $\phi_{0,s}$; (3) set $\tau^2_{n,s} := \mathbb{P}_n\left(\phi_n^2 + 2\phi_n\phi_{n,s} + \phi^2_{n,s}\right)$. We note that sample splitting is not required for variance estimation, since all we require is consistency. We then define

$$\sigma^2_{n,s,*} := B_n^{-1}\tau^2_{n,s}\ .$$

For testing purposes, there is no need to estimate $\nu^2_{0,s}$, since under H_0, $\nu^2_{0,s}$ is known to be identically zero. Under H_0 the test statistic $T_n := \left(\frac{n}{\sigma^2_{n,s,B_n,*}}\right)^{1/2}\psi^{(B_n)}_{n,s}$ converges weakly to a standard normal random variable. Comparing this test statistic to the α quantile of a standard normal random variable will yield an asymptotically valid one-sided level α test

of H_0. Algorithm 4.1 details the procedure for performing a hypothesis test using multiple sample splitting.

Algorithm 4.1 Multiple sample splitting with mean aggregation

1: Using all data, construct estimators f_n and $f_{n,s}$.
2: Using all data, construct estimators $\tau^2_{n,s} := \mathbb{P}_n\left(\phi_n^2 + 2\phi_n\phi_{n,s} + \phi_{n,s}^2\right)$.
3: Select number of splitting iterations B.
4: **for** $b = 1, \ldots, B$ **do**
5: Generate Bernoulli sample splitting indicators $S_1, \ldots, S_n$. Let n_s denote the size of $\mathcal{D}_0 := \{(X_i, Y_i) : S_i = 0\}$.
6: Using only data in $\mathcal{D}_0$, construct empirical distribution estimator $\mathbb{P}_{n,0}$ of P_0. Construct full predictiveness plug-in estimator $v_{nb} := V(f_n, \mathbb{P}_{n,0})$.
7: Using only data not in $\mathcal{D}_0$, construct empirical distribution estimator $\mathbb{P}_{n,1}$ of P_0. Construct residual predictiveness plug-in estimator $v_{nb,s} := V(f_{n,s}, \mathbb{P}_{n,1})$.
8: Compute $\psi_{nb,s} := v_{nb} - v_{nb,s}$.
9: **end for**
10: Compute overall VIM estimator $\psi_{n,s}^{(B)} := \frac{1}{B}\sum_{b=1}^{B} \psi_{nb,s}$.
11: To test $H_0 : \psi_{0,s} = 0$ vs. $H_1 : \psi_{0,s} > 0$ at level $1 - \alpha$, reject H_0 iff $p_n := 1 - \Phi(T_n) < \alpha$ with $T_n := \left(\frac{n}{\sigma^2_{n,s,B,*}}\right)^{1/2} \psi_{n,s}^{(B)}$.

4.4 Simulation studies

We conducted a numerical experiment to investigate the properties of the multi-split hypothesis testing procedure. In particular, we aimed to assess how the performance of multi-splitting depends on the number of splitting iterations and how the performance compares to that of a simple Bonferroni correction, which does not leverage the joint asymptotic normality of the sample split estimators.

In all experiments, we generated independent replicates of (X, Y), where X is a p-dimensional feature vector generated from a multivariate normal distribution with mean vector $(0, 0, \ldots, 0)$ and identity covariance matrix. Given covariate vector $X = x$, the outcome Y was generated from a Bernoulli distribution with $P(Y = 1 \,|\, X = x) = \Phi(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$, where Φ is the standard normal distribution function. (In other words, Y satisfies a probit regression model.) In Scenario 1, we set $p = 5$ and $\beta = (\beta_1, 3.5, 0, 0, 0)$;

in Scenario 2, we set $p = 100$ and $\beta = (\beta_1, 3.5, 0, \ldots, 0)$. We varied $\beta_1 \in \{0, 0.2, 0.4, \ldots, 1\}$, where $\beta_1 = 0$ corresponds to the null hypothesis H_0. We considered VIMs based on R^2, AUC, and binary classification accuracy.

We used two different estimators for the nuisance functions f_0 and $f_{0,s}$: a correctly specified probit regression regression model and gradient boosted trees (Friedman, 2001, implemented in the `xgboost` software package). In Scenario 2, we used a lasso probit regression model with five-fold cross-validation using logistic loss in order to select the regularization parameter (Tibshirani, 1996). For gradient boosting, we used 500 boosting iterations and a learning rate of 0.01. Five-fold cross-validation with logistic loss was used to select the maximum tree depth among possible values $\{1, 2, 3\}$. We varied the number of sample splitting iterations over $B \in \{1, 10, 50, 100, 200\}$. For $B = 1$, no aggregation was necessary. For all other values of B, hypothesis tests were performed using either a Bonferroni correction applied to the minimum p-value across iterations (i.e., $p = B \min(p_1, p_2, \ldots, p_B)$) or using Algorithm 4.1. For each scenario considered, we generated 500 random datasets of size $n = 1000$. In all experiments, we used five-fold cross-fitting within data splits. We evaluated performance in terms of empirical rejection probability.

We display results for AUC variable importance in the main text; results for R^2 and accuracy are similar and appear in Appendix C.2. From Figures 4.2 and 4.3, we see that in both scenarios, all procedures control type I error below the nominal level of the test, although the Bonferroni correction is slightly conservative compared to a single-split test or multi-split aggregation using Algorithm 4.1. When the effect size is non-zero, we observe that our method results in higher power than the single-split test, with substantial increases even at relatively small effect sizes. At least up to 200 splitting iterations, power increases monotonically, while type I error remains near 0.05. On the other hand, with small effect sizes, the Bonferroni correction tends to be conservative when the number of splitting iterations is large; indeed, a strategy using 100 or 200 splitting iterations is sometimes outperformed by a single split. These patterns are the same for the two nuisance estimators.

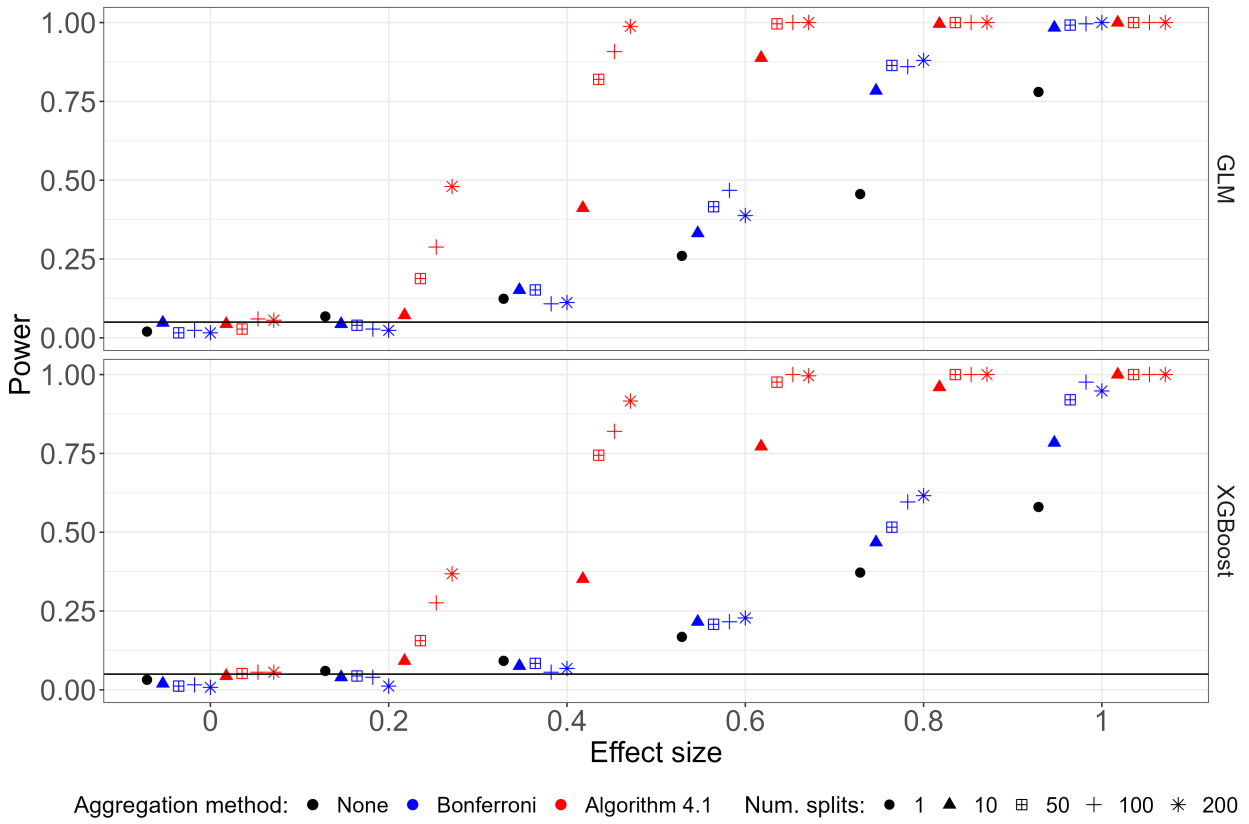

Figure 4.2: Performance of sample splitting approaches for testing the hypothesis of zero importance using AUC predictiveness in Scenario 1 ($p = 5$). When the effect size is zero, the null hypothesis holds. Oracle prediction functions were estimated using either probit regression (top row) or gradient-boosted trees (bottom row). The number of splitting iterations ranged from 1 to 200. Where more than one splitting iteration was used, results from multiple iterations were combined using either a Bonferroni correction (blue) or our proposed method, Algorithm 4.1 (red).

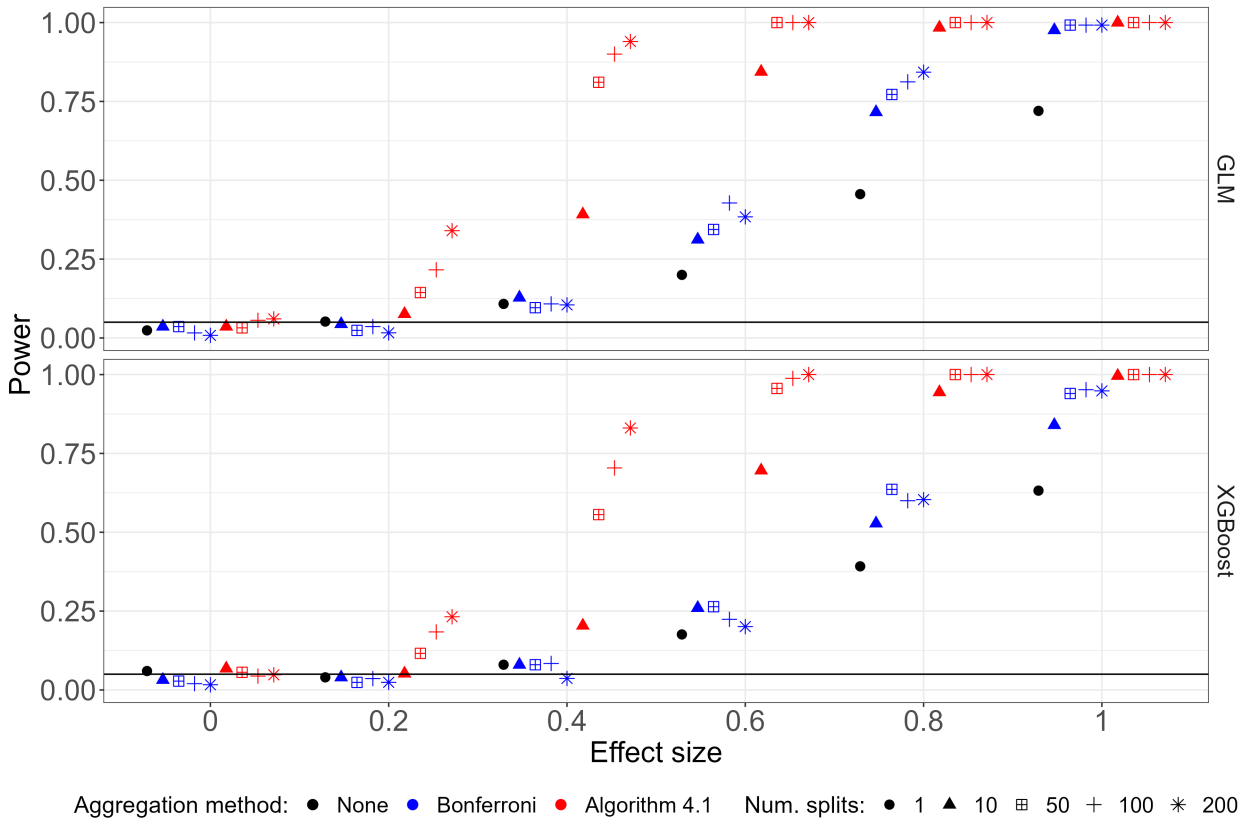

Figure 4.3: Performance of sample splitting approaches for testing the hypothesis of zero importance using AUC predictiveness in Scenario 2 ($p = 100$). When the effect size is zero, the null hypothesis holds. Oracle prediction functions were estimated using either probit regression (top row) or gradient-boosted trees (bottom row). The number of splitting iterations ranged from 1 to 200. Where more than one splitting iteration was used, results from multiple iterations were combined using either a Bonferroni correction (blue) or our proposed method, Algorithm 4.1 (red).

4.5 Discussion

We have outlined a pragmatic approach to both increase efficiency and decrease undesirable randomness due to sample splitting in nonparametric estimation of variable importance. Our procedure is simple to implement and embarrassingly parallel, which means that its computational footprint need not be substantially larger than standard sample splitting approaches. Multiple sample splitting performs well in simulated datasets, demonstrating substantial increases in power over single splitting while maintaining type I error control.

Further work is required to devise a procedure for selecting the number of sample splitting iterations. The theoretical analysis suggests that the remainder term in (4.5) which may dominate the first-order behavior of $\psi_{n,s}^{(B_n)}$ under H_0 if the number of splitting iterations increases too quickly. A conservative approach using a relatively small number of iterations can have a major impact on power, but a method for selecting B_n, ideally in a data-adaptive manner, would greatly increase the practical appeal of our proposal.

Another avenue of future research concerns the power of our proposed hypothesis test against local alternatives. Lundborg et al. (2022) note that the single sample splitting approach of Williamson et al. (2021b) is powerless against local alternatives $\psi_{0,n}$ such that $n^{1/2}\psi_{0,n} \to 0$, i.e., against local alternatives converging to zero a rate faster than $n^{-1/2}$. This is unsurprising, as the sample splitting procedure is intended to yield $n^{1/2}$-rate inference in light of the degeneracy of the non-split estimator. One recently proposed estimator (Hudson, 2023) achieves n-rate consistency for $\psi_{0,s}$ when H_0 holds, although the power of the resulting test remains an open question.

Construction of confidence intervals is another area of possible future research. While the distributional result of Theorem 4 holds under both H_0 and H_1, constructing a symmetric Wald interval of the form

$$\left(\psi_{n,s}^{(B_n)} - z_{1-\alpha/2}\sigma_{n,s,B_n}n^{-1/2}, \psi_{n,s}^{(B_n)} - z_{1-\alpha/2}\sigma_{n,s,B_n}n^{-1/2}\right)$$

is likely to be conservative, as the component $\nu_{0,s}^2$ of σ_{0,s,B_n}^2 is known to be zero under H_0. Leveraging this relationship between the VIM parameter $\psi_{0,s}$ and the variance component

$\nu_{0,s}^2$ could yield improved confidence intervals.

4.6 Acknowledgments

This work was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2140004. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

BIBLIOGRAPHY

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726.

Balkus, J. E., Brown, E., Palanee, T., Nair, G., Gafoor, Z., Zhang, J., Richardson, B. A., Chirenje, Z. M., Marrazzo, J. M., and Baeten, J. M. (2016). An empiric HIV risk scoring tool to predict HIV-1 acquisition in African women. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 72(3):333.

Beran, R. (1981). Non-parametric regression with censored survival time data. Technical report, University of California, Berkeley.

Boileau, P., Leng, N., Hejazi, N. S., van der Laan, M., and Dudoit, S. (2023). A nonparametric framework for treatment effect modifier discovery in high dimensions. *arXiv preprint arXiv:2304.05323*.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brentnall, A. R. and Cuzick, J. (2018). Use of the concordance index for predictors of censored survival data. *Statistical Methods in Medical Research*, 27:2359–2373.

Breslow, N. E. (1972). Discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 34:216–217.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Buchbinder, S. P., Mehrotra, D. V., Duerr, A., Fitzgerald, D. W., Mogg, R., Li, D., Gilbert, P. B., Lama, J. R., Marmor, M., del Rio, C., McElrath, M. J., Casimiro, D. R., Gottesdiener, K. M., Chodakewitz, J. A., Corey, L., and Robertson, M. N. (2008). Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *The Lancet*, 372:1881–1893.

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

Chambless, L. E. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25:3474–3486.

Chen, L., Lin, D. Y., and Zeng, D. (2012). Predictive accuracy of covariates for event times. *Biometrika*, 99:615–630.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.

Ching, T., Zhu, X., and Garmire, L. (2018). Cox-nnet: an artificial neural network method for prognosis prediction on high-throughput omics data. *PLoS Computational Biology*, 14:e1006076.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220.

Craig, E., Zhong, C., and Tibshirani, R. (2021). Survival stacking: casting survival analysis as a classification problem. *arXiv:2107.13480*.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

D'Agostino, R. B., Lee, M.-L., Belanger, A. J., Cupples, L. A., Anderson, K., and Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The framingham heart study. *Statistics in Medicine*, 9:1501–1515.

Dai, B., Shen, X., and Pan, W. (2022). Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*.

D'Angio, G. J., Evans, A. E., Breslow, N., Beckwith, B., Bishop, H., Feigl, P., Goodwin, W., Leape, L. L., Sinks, L. F., Sutow, W., Tefft, M., and Wolff, J. (1976). The treatment of Wilms' tumor. results of the national Wilms' tumor study. *Cancer*, 38:633–646.

Ding, Y. and Nan, B. (2015). Estimating mean survival time: when is it possible? *Scandinavian Journal of Statistics*, 42(2):397–413.

Duerr, A., Huang, Y., Buchbinder, S., Coombs, R. W., Sanchez, J., Rio, C. D., Casapia, M., Santiago, S., Gilbert, P., Corey, L., and Robertson, M. N. (2012). Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus hiv vaccine (Step study). *Journal of Infectious Diseases*, 206:258–266.

Díaz, I. (2019). Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Statistics in Medicine*, 38:2735–2748.

Efron, B. (1967). The two sample problem with censored data. In Cam, L. L. and Neyman, J., editors, *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 831–853. University of California Press.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Fleming, T. R. and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics: Theory and Methods*, 13:2469–2486.

Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv:1801.05512*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

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

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10:101–113.

Fu, W. and Simonoff, J. S. (2017). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 18(2):352–369.

Gan, L., Zheng, L., and Allen, G. I. (2022). Inference for interpretable machine learning: Fast, model-agnostic confidence intervals for feature importance. *arXiv preprint arXiv:2206.02088*.

Gensheimer, M. F. and Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, pages 1–19.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48:1029–1040.

Gill, R. D. (1993). Multivariate survival analysis. *Theory of Probability & Its Applications*, 37(2):284–301.

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18:1501–1555.

Giunchiglia, E., Nemchenko, A., and van der Schaar, M. (2018). Rnn-surv: A deep recurrent model for survival analysis. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 23–32. Springer.

Gonen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92:965–970.

Gray, G. E., Bekker, L.-G., Laher, F., Malahleha, M., Allen, M., Moodie, Z., Grunenberg, N., Huang, Y., Grove, D., Prigmore, B., et al. (2021). Vaccine efficacy of ALVAC-HIV and bivalent subtype C gp120–MF59 in adults. *New England Journal of Medicine*, 384(12):1089–1100.

Guo, F. R. and Shah, R. D. (2023). Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values. *arXiv preprint arXiv:2301.02739*.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247:2543–6.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1:297–318.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105.

Hothorn, T., Bühlmann, P., Dudoit, S., and van der Laan, M. J. (2006a). Survival ensembles. *Biostatistics*, 7:355–373.

Hothorn, T., Hornik, K., and Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Hudson, A. (2023). Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space. *arXiv preprint arXiv:2306.07492*.

Hung, H. and Chiang, C.-T. (2010). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38:8–26.

Ishwaran, H., Blackstone, E. H., Pothier, C. E., and Lauer, M. S. (2004). Relative risk forests for exercise heart rate recovery as a predictor of mortality. *Journal of the American Statistical Association*, 99:591–600.

Ishwaran, H. and Kogalur, U. (2022). *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 3.1.0.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860.

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

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:1–12.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N., et al. (1995). The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203.

Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9:487–503.

Koul, H., Susarla, V., and Ryzin, J. V. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9.

Kyle, R. A., Therneau, T. M., Rajkumar, S. V., Larson, D. R., Plevak, M. F., Offord, J. R., Dispenzieri, A., Katzmann, J. A., and Melton III, L. J. (2006). Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 354(13):1362–1369.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lin, D., Fleming, T., and Wei, L. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika*, 81:73–81.

Lin, Y. and Chen, K. (2013). Efficient estimation of the censored linear regression model. *Biometrika*, 100:525–530.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lundborg, A. R., Kim, I., Shah, R. D., and Samworth, R. J. (2022). The projected covariance measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039*.

Mayr, A. and Schmid, M. (2014). Boosting the concordance index for survival data - a unified framework to derive and evaluate biomarker combinations. *PLoS ONE*, 9.

Menza, T. W., Hughes, J. P., Celum, C. L., and Golden, M. R. (2009). Prediction of HIV acquisition among men who have sex with men. *Sexually transmitted diseases*, pages 547–555.

Molinaro, A. M., Dudoit, S., and van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90:154–177.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1:27–52.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*. Springer.

Polley, E. C. and van der Laan, M. J. (2011). *Super Learning for Right-Censored Data*, pages 249–258. Springer.

Qian, J. and Betensky, R. A. (2014). Assumptions regarding right censoring in the presence of left truncation. *Statistics and Probability Letters*, 87:12–17.

Raykar, V. C., Steck, H., Krishnapuram, B., Dehing-Oberije, C., and Lambin, P. (2008). On ranking in survival analysis: Bounds on the concordance index. *Advances in Neural Information Processing Systems 20*, pages 1209–1216.

Reid, N. (1981). Influence functions for censored data. *The Annals of Statistics*, 9:78–92.

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

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics*, 56:249–255.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., and Rauschecker, H. (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39:1–13.

Smith, D. K., Pals, S. L., Herbst, J. H., Shinde, S., and Carey, J. W. (2012). Development of a clinical screening index predictive of incident HIV infection among men who have sex with men in the United States. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 60(4):421–427.

Sonabend, R. E. B. (2021). *A Theoretical and Methodological Framework for Machine Learning in Survival Analysis*. PhD thesis, University College London.

Song, X. and Zhou, X.-H. (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica*, 18:947–965.

Tarkhan, A. and Simon, N. (2022). An online framework for survival analysis: reframing Cox proportional hazards model for large data sets and neural networks. *Biostatistics*.

Therneau, T. M. (2022). *A Package for Survival Analysis in R*. R package version 3.3-1.

Tian, L., Zhao, L., and Wei, L. J. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15:222–233.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30:1105–1117.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102:527–537.

van der Laan, M., Polley, E., and Hubbard, A. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Online Article 25.

van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1).

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer New York, 1st edition.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag.

Verdinelli, I. and Wasserman, L. (2021). Decorrelated variable importance. *arXiv:2111.10853*.

Wand, H., Reddy, T., Naidoo, S., Moonsamy, S., Siva, S., Morar, N. S., and Ramjee, G. (2018). A simple risk prediction algorithm for HIV transmission: results from HIV prevention trials in KwaZulu Natal, South Africa (2002–2012). *AIDS and Behavior*, 22:325–336.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

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

Westling, T., Luedtke, A., Gilbert, P. B., and Carone, M. (2023). Inference for treatment-specific survival curves using machine learning. *Journal of the American Statistical Association*, 0(0):1–26.

Westling, T., van der Laan, M. J., and Carone, M. (2020). Correcting an estimator of a multivariate monotone function with isotonic regression. *Electronic Journal of Statistics*, 14:3032–3069.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021a). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77:9–22.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021b). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*.

Wolock, C. J., Gilbert, P. B., Simon, N., and Carone, M. (2022). A framework for leveraging machine learning tools to estimate personalized survival curves. *arXiv:2211.03031*.

Yu, C. N., Greiner, R., Lin, H. C., and Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, 24:1845–1853.

Zeng, D. and Lin, D. Y. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102:1387–1396.

Zhang, L. and Janson, L. (2022). Floodgate: inference for model-free variable importance. *arXiv:2007.01283*.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In van der Laan, M. J. and Rose, S., editors, *Targeted Learning: Causal Inference for Observational Data*, pages 459–474. Springer.

Zhou, X., Su, W., Liu, C., Jiao, Y., Zhao, X., and Huang, J. (2022). Deep generative survival analysis: Nonparametric estimation of conditional survival function. *arXiv:2205.09633*.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Proofs of theorems

A.1.1 Efficiency calculations

Proof of Theorem 1. In the following, we let $P_0(y, \delta \,|\, x)$ denote the joint distribution function of (Y, Δ) given $X = x$.

Let $\{P_\epsilon\}$ be a suitably smooth and bounded Hellinger differentiable path with $P_{\epsilon=0} = P_0$ and score function $\dot{\ell}_0$ at $\epsilon = 0$. We will repeatedly make use of several properties of score functions; namely, that $\dot{\ell}_0(x, y, \delta) = \dot{\ell}_0(x) + \dot{\ell}_0(y, \delta \,|\, x)$, and that scores are mean-zero:

$$\iiint b(x)\dot{\ell}_0(y, \delta \,|\, x)P_0(dy, d\delta \,|\, x)Q_0(dx) = 0$$

for any function b.

We begin by using the quotient rule to write

$$\frac{d}{d\epsilon}\left\{V(f_0, P_\epsilon)\right\}\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\left\{\frac{V_1(f_0, \mathbb{F}_\epsilon)}{V_2(\mathbb{F}_\epsilon)}\right\}\Big|_{\epsilon=0} = \frac{\frac{d}{d\epsilon}V_1(f_0, \mathbb{F}_\epsilon)\Big|_{\epsilon=0}}{V_2(\mathbb{F}_0)} - \frac{V_1(f_0, \mathbb{F}_0)\frac{d}{d\epsilon}V_2(\mathbb{F}_\epsilon)\Big|_{\epsilon=0}}{V_2(\mathbb{F}_0)^2}. \tag{A.1}$$

We study the two derivatives above separately.

Under appropriate boundedness conditions, we have that

$$\begin{aligned}
\frac{d}{d\epsilon}V_1(f_0,\mathbb{F}_\epsilon)\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}\int \Phi\left(((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)\prod_{j=1}^m \mathbb{F}_\epsilon(dx_j,dt_j)\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\int \Phi\left(((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)\prod_{j=1}^m F_\epsilon(dt_j\,|\,x_j)Q_\epsilon(dx_j)\Big|_{\epsilon=0} \\
&= \int \Phi\left((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)\frac{d}{d\epsilon}\prod_{j=1}^m F_\epsilon(dt_j\,|\,x_j)\Big|_{\epsilon=0}\prod_{j=1}^m Q_0(dx_j) \\
&+ \int \Phi\left((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)\prod_{j=1}^m F_0(dt_j\,|\,x_j)\frac{d}{d\epsilon}\prod_{j=1}^m Q_\epsilon(dx_j)\Big|_{\epsilon=0}.
\end{aligned} \tag{A.2}$$

Using the product rule and symmetry of Φ, the second term in (A.2) is equal to

$$\begin{aligned}
&m\int \Phi\left((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)\prod_{j=1}^m F_0(dt_j\,|\,x_j)\prod_{j=2}^m Q_0(dx_j)\frac{d}{d\epsilon}Q_\epsilon(dx_1)\Big|_{\epsilon=0} \\
&= m\int \Phi\left((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)\prod_{j=1}^m F_0(dt_j\,|\,x_j)\prod_{j=2}^m Q_0(dx_j)\dot{\ell}_0(x)Q_0(dx_1) \\
&= m\int\Bigg\{\Phi\left((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right) \\
&\qquad\times\prod_{j=1}^m F_0(dt_j\,|\,x_j)\prod_{j=2}^m Q_0(dx_j)\dot{\ell}_0(x_1,y,\delta)P_0(dy,d\delta\,|\,x_1)Q_0(dx_1)\Bigg\},
\end{aligned}$$

where the final equality follows from properties of score functions. This expression can be rewritten as

$$\begin{aligned}
&m\int \Phi\left((f_0(x_1),t_1),\ldots,(f_0(x_m),t_m)\right)F_0(dt_1\,|\,x_1)\prod_{j=2}^m \mathbb{F}_0(dx_j,dt_j)\dot{\ell}_0(x,y,\delta)P_0(dy,d\delta,dx_1) \\
&= m\int \Phi_{0,1}\left(x_1,t_1\right)F_0(dt_1\,|\,x_1)\dot{\ell}_0(x,y,\delta)P_0(dy,d\delta,dx_1).
\end{aligned}$$

Therefore, this term contributes

$$x \mapsto m\int \Phi_{0,1}\left(x,t\right)F_0(dt\,|\,x) \tag{A.3}$$

to the EIF.

Now, by definition we have that

$$\frac{d}{d\epsilon}F_\epsilon(t\,|\,x)\bigg|_{\epsilon=0} = -\frac{d}{d\epsilon}\prod_{(0,t]}\{1-\Lambda_\epsilon(du\,|\,x)\}\bigg|_{\epsilon=0}.$$

By Theorem 8 of Gill and Johansen (1990), the product integral map $H \mapsto S_H(t)$ is Hadamard differentiable with respect to the supremum norm with derivative $\alpha \mapsto S_H(t)\int_0^t \frac{S_H(u^-)}{S_H(u)}\alpha(du)$. By the chain rule for functional derivatives, we have that

$$\frac{d}{d\epsilon}F_\epsilon(t\,|\,x)\bigg|_{\epsilon=0} = -\frac{d}{d\epsilon}\prod_{(0,t]}\{1-\Lambda_\epsilon(du\,|\,x)\}\bigg|_{\epsilon=0} = S_0(t\,|\,x)\int_0^t \frac{S_0(u^-\,|\,x)}{S_0(u\,|\,x)}\frac{d}{d\epsilon}\Lambda_\epsilon(du\,|\,x)\bigg|_{\epsilon=0}.$$

Now, using the quotient rule, we can write

$$\begin{aligned}\frac{d}{d\epsilon}\Lambda_\epsilon(t\,|\,x)\bigg|_{\epsilon=0} &= \int_0^t \frac{d}{d\epsilon}\frac{F_{\epsilon,1}(du\,|\,x)}{1-F_\epsilon(u^-\,|\,x)}\bigg|_{\epsilon=0} \\ &= \int \frac{\mathbb{1}_{[0,t]}(u)\frac{d}{d\epsilon}F_{\epsilon,1}(du\,|\,x)\Big|_{\epsilon=0}}{1-F_0(u^-\,|\,x)} + \int \frac{\mathbb{1}_{[0,t]}(u)\frac{d}{d\epsilon}F_\epsilon(u^-\,|\,x)\Big|_{\epsilon=0}F_{0,1}(du\,|\,x)}{\{1-F_0(u^-\,|\,x)\}^2}.\end{aligned}$$

Computing each of the derivatives above, we have that

$$\begin{aligned}\frac{d}{d\epsilon}F_{\epsilon,1}(u\,|\,x)\bigg|_{\epsilon=0} = \frac{d}{d\epsilon}P_\epsilon(Y\le u,\Delta=1\,|\,X=x)\bigg|_{\epsilon=0} &= \frac{d}{d\epsilon}\iint \mathbb{1}_{[0,u]}(y)\delta P_\epsilon(dy,d\delta\,|\,x)\bigg|_{\epsilon=0} \\ &= \iint \mathbb{1}_{[0,u]}(y)\delta\dot{\ell}_0(y,\delta\,|\,x)P_0(dy,d\delta\,|\,x),\end{aligned}$$

and

$$\begin{aligned}\frac{d}{d\epsilon}F_\epsilon(u^-\,|\,x)\bigg|_{\epsilon=0} &= \frac{d}{d\epsilon}P_\epsilon(Y<u\,|\,x)\bigg|_{\epsilon=0} = \frac{d}{d\epsilon}\int \mathbb{1}_{[0,u)}(y)P_\epsilon(dy\,|\,x)\bigg|_{\epsilon=0} \\ &= \int \mathbb{1}_{[0,u)}(y)\dot{\ell}_0(y\,|\,x)P_0(dy\,|\,x) \\ &= \iint \mathbb{1}_{[0,u)}(y)\dot{\ell}_0(y\,|\,x)P_0(d\delta\,|\,y,x)P_0(dy\,|\,x) + \iint \mathbb{1}_{[0,u)}(y)\dot{\ell}_0(\delta\,|\,y,x)P_0(d\delta\,|\,y,x)P_0(dy\,|\,x) \\ &= \iint \mathbb{1}_{[0,u)}(y)\dot{\ell}_0(y,\delta\,|\,x)P_0(dy,d\delta\,|\,x),\end{aligned}$$

where the last two equalities follow from properties of score functions. Therefore,

$$\begin{aligned}
\frac{d}{d\epsilon}\Lambda_\epsilon(t \,|\, x)\bigg|_{\epsilon=0} &= \int \frac{\mathbb{1}_{[0,t]}(u) \int \delta \dot{\ell}_0(u, \delta \,|\, x) P_0(du, d\delta \,|\, x)}{1 - F_0(u^- \,|\, x)} \\
&\quad + \int \frac{\mathbb{1}_{[0,t]}(u) \mathbb{1}_{[0,u)}(y) \iint \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) F_{0,1}(du \,|\, x)}{\{1 - F_0(u^- \,|\, x)\}^2} \\
&= \iint \frac{\mathbb{1}_{[0,t]}(y) \delta \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x)}{1 - F_0(y^- \,|\, x)} \\
&\quad + \iiint \frac{\mathbb{1}_{[0,t]}(u) \mathbb{1}_{[0,u)}(y) \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) F_{0,1}(du \,|\, x)}{\{1 - F_0(u^- \,|\, x)\}^2} \\
&= \iint \frac{\mathbb{1}_{[0,t]}(y) \delta \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x)}{1 - F_0(y^- \,|\, x)} \\
&\quad + \iiint \frac{\mathbb{1}_{[0,t]}(u) \mathbb{1}_{[0,u)}(y) \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) \Lambda_0(du \,|\, x)}{1 - F_0(u^- \,|\, x)}.
\end{aligned}$$

Again using the fact that scores are mean-zero and observing that $\mathbb{1}(y > u) = 1 - \mathbb{1}(u \leq y)$, we plug this expression into the derivative of the product integral to obtain

$$\begin{aligned}
\frac{d}{d\epsilon}F_\epsilon(t \,|\, x) &= S_0(t \,|\, x) \iint \frac{\mathbb{1}_{[0,t]}(u) S_0(u^- \,|\, x)}{S_0(u \,|\, x) \{1 - F_0(u^- \,|\, x)\}} \delta \dot{\ell}_0(u, \delta \,|\, x) P_0(du, d\delta \,|\, x) \\
&\quad S_0(t \,|\, x) \iiint \frac{\mathbb{1}_{[0,t]}(u) \mathbb{1}_{[0,u)}(y) S_0(u^- \,|\, x)}{S_0(u \,|\, x) \{1 - F_0(u^- \,|\, x)\}} \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) \Lambda_0(du \,|\, x) \\
&= S_0(t \,|\, x) \iint \frac{\delta \mathbb{1}_{[0,t]}(y) S_0(y- \,|\, x)}{S_0(y \,|\, x) \{1 - F_0(y- \,|\, x)\}} \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) \\
&\quad - S_0(t \,|\, x) \iint \int_0^{t \wedge y} \frac{S_0(u^- \,|\, x) \Lambda_0(du \,|\, x)}{S_0(u \,|\, x) \{1 - F_0(u^- \,|\, x)\}} \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) \\
&= S_0(t \,|\, x) \iint \frac{\delta \mathbb{1}_{[0,t]}(y)}{S_0(y \,|\, x) G_0(y \,|\, x)} \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) \\
&\quad - S_0(t \,|\, x) \iint \int_0^{t \wedge y} \frac{\Lambda_0(du \,|\, x)}{S_0(u \,|\, x) G_0(u \,|\, x)} \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) \\
&= - \iint \varphi_{0,z}(t) \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x).
\end{aligned}$$

Next, we note that

$$\begin{aligned}
&\mathbb{E}_0[\varphi_{0,Z}(t) \,|\, X = x] \\
&= -S_0(t \,|\, x) \int \left[\frac{\delta \mathbb{1}_{[0,t]}(y)}{S_0(y \,|\, x) G_0(y \,|\, x)} + \int_0^{t \wedge y} \frac{\Lambda_0(du \,|\, x)}{S_0(u \,|\, x) G_0(u \,|\, x)} \right] P_0(dy, d\delta \,|\, x) \\
&= -S_0(t \,|\, x) \left[\int \frac{F_{0,1}(dy \,|\, x)}{S_0(y \,|\, x) G_0(y \,|\, x)} + \int_0^t \frac{\{1 - F_0(u^- \,|\, x)\} \Lambda_0(du \,|\, x)}{S_0(u \,|\, x) G_0(u \,|\, x)} \right] \\
&= -S_0(t \,|\, x) \left[\int \frac{F_{0,1}(dy \,|\, x)}{S_0(y \,|\, x) G_0(y \,|\, x)} + \int_0^t \frac{F_{0,1}(du \,|\, x)}{S_0(u \,|\, x) G_0(u \,|\, x)} \right] \\
&= 0.
\end{aligned}$$

By properties of score functions, this implies that

$$\begin{aligned}
\frac{d}{d\epsilon} F_\epsilon(t \,|\, x) &= -\iint \varphi_{0,z}(t) \dot{\ell}_0(y, \delta \,|\, x) P_0(dy, d\delta \,|\, x) - \underbrace{\iint \varphi_{0,z}(t) \dot{\ell}_0(x) P_0(dy, d\delta \,|\, x)}_{=0} \\
&= -\iint \varphi_{0,z}(t) \dot{\ell}_0(y, \delta, x) P_0(dy, d\delta \,|\, x).
\end{aligned}$$

Hence, using the product rule and symmetry of Φ for the leading term in (A.2), we have

$$\begin{aligned}
&- m \int \Bigg\{ \Phi\left((f_0(x_1), t_1), \ldots, (f_0(x_m), t_m)\right) \varphi_{0,z_1}(dt_1) \\
&\qquad\qquad \times \prod_{j=2}^{m} \mathbb{F}_0(dx_j, dt_j) \dot{\ell}_0(y_1, \delta_1, x_1) P_0(dy_1, d\delta_1 \,|\, x_1) Q_0(dx_1) \Bigg\} \\
&= -m \int \Phi_{0,1}\left(x_1, t_1\right) \varphi_{0,z_1}(dt_1) \dot{\ell}_0(y_1, \delta_1, x_1) P_0(dy_1, d\delta_1 \,|\, x_1) Q_0(dx_1),
\end{aligned}$$

yielding that the EIF contribution from this term is

$$z \mapsto -m \int \Phi_{0,1}\left(x, t\right) \varphi_{0,z}(dt). \tag{A.4}$$

Next, we note that

$$\begin{aligned}
\varphi_{0,z}(dt) = &- \left\{ \frac{\delta \mathbb{1}_{[0,t]}(y)}{S_0(y \,|\, x) G_0(y \,|\, x)} - \int_0^{t \wedge y} \frac{\Lambda_0(du \,|\, x)}{S_0(u \,|\, x) G_0(u \,|\, x))} \right\} S_0(dt \,|\, x) \\
&- S_0(t \,|\, x) \frac{\delta \gamma(t - y)}{S_0(y \,|\, x) G_0(y \,|\, x)} + S_0(t \,|\, x) \frac{\mathbb{1}_{[0,y]}(t) \Lambda_0(dt \,|\, x)}{S_0(t \,|\, x) G_0(t \,|\, x)},
\end{aligned}$$

where $\gamma(\cdot)$ is the Dirac delta function. Therefore,

$$\begin{aligned}
&\int \Phi_{0,1}(x,t)\,\varphi_{0,z}(dt) \\
&= -\int \Phi_{0,1}(x,t)\left\{\frac{\delta}{S_0(y\,|\,x)G_0(y\,|\,x)} - \int_0^{t\wedge y}\frac{\Lambda_0(du\,|\,x)}{S_0(u\,|\,x)G_0(u\,|\,x))}\right\}S_0(dt\,|\,x) \\
&\quad - \Phi_{0,1}(x,t)\,\frac{S_0(y\,|\,x)\delta}{S_0(y\,|\,x)G_0(y\,|\,x)} + \int \Phi_{0,1}(x,t)\,S_0(t\,|\,x)\frac{\mathbb{1}_{[0,y]}(t)\Lambda_0(dt\,|\,x)}{S_0(t\,|\,x)G_0(t\,|\,x)} \\
&= -\int \Phi_{0,1}(x,t)\left\{\frac{\delta\mathbb{1}_{[0,t]}(y)}{S_0(y\,|\,x)G_0(y\,|\,x)} - \int_0^{t\wedge y}\frac{\Lambda_0(du\,|\,x)}{S_0(u\,|\,x)G_0(u\,|\,x))}\right\}S_0(dt\,|\,x) \\
&\quad - \Phi_{0,1}(x,t)\,\frac{\delta}{G_0(y\,|\,x)} + \int \Phi_{0,1}(x,t)\,\frac{\mathbb{1}_{[0,y]}(t)\Lambda_0(dt\,|\,x)}{G_0(t\,|\,x)}.
\end{aligned}$$

Taking an expectation over (Y,Δ) given $X=x$, we have

$$\begin{aligned}
&\mathbb{E}_0\left[\int \Phi_{0,1}(x,t)\,\varphi_{0,Z}(dt)\,|\,X=x\right] \\
&= -\int \Phi_{0,1}(x,t)\left\{\int_0^t \frac{F_{0,1}(dy\,|\,x)}{S_0(y\,|\,x)G_0(y\,|\,x)} - \int_0^t \frac{\{1-F_0(u^-\,|\,x)\}\,\Lambda_0(du\,|\,x)}{S_0(u\,|\,x)G_0(u\,|\,x))}\right\}S_0(dt\,|\,x) \\
&\quad - \int \Phi_{0,1}(x,t)\,\frac{F_{0,1}(dy\,|\,x)}{G_0(y\,|\,x)} + \int \Phi_{0,1}(x,t)\,\frac{\{1-F_0(t^-\,|\,x)\}\,\Lambda_0(dt\,|\,x)}{G_0(t\,|\,x)} \\
&= -\int \Phi_{0,1}(x,t)\left\{\int_0^t \frac{F_{0,1}(dy\,|\,x)}{S_0(y\,|\,x)G_0(y\,|\,x)} - \int_0^t \frac{F_{0,1}(du\,|\,x)}{S_0(u\,|\,x)G_0(u\,|\,x))}\right\}S_0(dt\,|\,x) \\
&\quad - \int \Phi_{0,1}(x,t)\,\frac{F_{0,1}(dy\,|\,x)}{G_0(y\,|\,x)} + \int \Phi_{0,1}(x,t)\,\frac{F_{0,1}(dt\,|\,x)}{G_0(t\,|\,x)} = 0.
\end{aligned}$$

The tower rule then implies that (A.4) is mean-zero. Thus, this portion of the EIF is already centered. To center the EIF overall, we subtract $mV_1(f_0,\mathbb{F}_0)$ from (A.3), yielding the centered gradient $D_\Phi(P_0)$. The derivation of the gradient of V_2 is identical, so combining these results with (A.1) yields the overall gradient

$$D(P_0): z \mapsto \frac{D_\Phi(P_0)(z)}{V_2(\mathbb{F}_0)} - \frac{V_1(f_0,\mathbb{F}_0)D_\Theta(P_0)(z)}{V_2(\mathbb{F}_0)^2}.$$

It remains to show that $P \mapsto V(f_P,P)$ and $P \mapsto V(f_0,P)$ have the same EIF at P_0. We note that

$$V(f_\epsilon,P_\epsilon) - V(f_0,P_0) = V(f_\epsilon,P_\epsilon) - V(f_0,P_\epsilon) + V(f_0,P_\epsilon) - V(f_0,P_0).$$

Under condition (B1a), we have that $V(f_\epsilon, P_\epsilon) - V(f_0, P_\epsilon) = V(f_\epsilon, P_0) - V(f_0, P_0) + o(\epsilon)$. Furthermore, under conditions (B1b) and (B1c), we have that

$$\frac{d}{d\epsilon} V(f_\epsilon, P_0)\bigg|_{\epsilon=0} = 0,$$

and so Taylor's theorem yields that $V(f_\epsilon, P_0) - V(f_0, P_0) = o(\epsilon)$. Since the parameter $P \mapsto V(f_0, P)$ is pathwise differentiable at P_0 with canonical gradient $D_1(P_0)$, we can write that $V(f_0, P_\epsilon) - V(f_0, P_0) = \epsilon \int D(P_0)(z)\dot{\ell}_0(z)P_0(dz) + O(\epsilon^2)$. Combining all these observations, we have that $V_1(f_\epsilon, P_\epsilon) - V(f_0, P_0) = \epsilon \int D(P_0)(z)\dot{\ell}_0(z)P_0(dz) + o(\epsilon)$. Hence, $P \mapsto V(f_P, P)$ is pathwise differentiable at P_0 with EIF equal to $D(P_0)$. □

Lemma 1. *If there exists $\nu \in (0, \infty)$ such that $G_0(t_0 \,|\, x) \geq 1/\nu$ for P_0-almost every x such that $x \leq x_0$, then $P \mapsto \mathbb{F}_P(x_0, t_0)$ is pathwise differentiable at P_0 relative to the nonparametric model $\mathcal{M}$, with EIF given by $z \mapsto \bar{\phi}_{0,z}(x_0, t) := \phi_{0,z}(x_0, t_0) - \mathbb{F}_0(x_0, t_0)$, where*

$$\phi_{0,z}(x_0, t_0) := \mathbb{1}(x \leq x_0) \left\{ F_0(t_0 \,|\, x) - \varphi_{0,z}(t_0) \right\}.$$

Proof of Lemma 1. As in the proof of Theorem 1, we proceed by direct calculation of the pathwise derivative of the parameter. Again we let $\{P_\epsilon\}$ be a suitably smooth and bounded Hellinger differentiable path with $P_{\epsilon=0} = P_0$ and score function $\dot{\ell}_0$ at $\epsilon = 0$. The parameter evaluated at P_ϵ is given by

$$\mathbb{F}_\epsilon(x_0, t_0) = \int \mathbb{1}(x \leq x_0) F_\epsilon(t_0 \,|\, x) Q_\epsilon(dx).$$

Under appropriate boundedness conditions, we have that

$$\begin{aligned} \frac{d}{d\epsilon} \mathbb{F}_\epsilon(x_0, t_0)\bigg|_{\epsilon=0} &= \frac{d}{d\epsilon} \int \mathbb{1}(x \leq x_0) F_\epsilon(t_0 \,|\, x) Q_\epsilon(dx)\bigg|_{\epsilon=0} \\ &= \int \mathbb{1}(x \leq x_0) \frac{d}{d\epsilon} F_\epsilon(t_0 \,|\, x)\bigg|_{\epsilon=0} Q_0(dx) + \int \mathbb{1}(x \leq x_0) F_0(t_0 \,|\, x) \frac{d}{d\epsilon} Q_\epsilon(dx)\bigg|_{\epsilon=0}. \end{aligned} \tag{A.5}$$

Using the same argument as in the proof of Theorem 1, the second term in (A.5) contributes $x \mapsto \mathbb{1}(x \leq x_0) F_0(t_0 \,|\, x)$ to the the EIF.

For the first term, we recall that $\frac{d}{d\epsilon} F_\epsilon(t_0 \,|\, x)\Big|_{\epsilon=0} = -\iint \varphi_{0,z}(t_0)\dot{\ell}_0(y, \delta, x) P_0(dy, d\delta, dx)$. Therefore the first term in (A.5) contributes $z \mapsto -\mathbb{1}(x \leq x_0)\varphi_{0,z}(t_0)$. This portion of the

EIF is already centered. To center the EIF overall, we subtract $\mathbb{F}_0(x_0, t_0)$, yielding the overall EIF

$$\bar{\phi}_0(x_0, t_0) : z \mapsto \mathbb{1}(x \leq x_0) \{F_0(t_0 \,|\, x) - \varphi_{0,z}(t_0)\} - \mathbb{F}_0(x_0, t_0).$$

□

A.1.2 Additional theoretical results

The functions ϕ and φ are both indexed by the nuisance functions $S(\cdot \,|\, x)$ and $G(\cdot \,|\, x)$, which represent generic conditional event time and censoring survival functions, respectively. We let

$$\varphi_{S,G,z}(t) := -S(t \,|\, x) \left[\frac{\delta \mathbb{1}_{[0,t]}(y)}{S(y \,|\, x) G(y \,|\, x)} - \int_0^{t \wedge y} \frac{\Lambda(du \,|\, x)}{S(u \,|\, x) G(u \,|\, x)} \right],$$

where Λ is the conditional cumulative hazard function corresponding to S. We then let $\phi_{S,G,z}(t) := 1 - S(t \,|\, x) - \varphi_{S,G,z}(t)$. As shorthand, we let $\varphi_{n,k,z} := \varphi_{S_{n,k},G_{n,k},z}$, $\phi_{n,k,z} := \phi_{S_{n,k},G_{n,k},z}$, $\varphi_{\infty,z} := \varphi_{S_\infty,G_\infty,z}$, and $\phi_{\infty,z} := \phi_{S_\infty,G_\infty,z}$.

Lemma 2. *For any conditional event distribution S and corresponding cumulative hazard Λ, and any conditional censoring distribution G,*

$$\begin{aligned} &P_0 \phi_{S,G}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) \\ &\quad = \mathbb{E}_0 \left[\mathbb{1}(X \leq x_0) S(t_0 \,|\, X) \int_0^{t_0} \frac{S_0(u^- \,|\, X)}{S(u \,|\, X)} \left\{ \frac{G_0(u \,|\, X)}{G(u \,|\, X)} - 1 \right\} (\Lambda - \Lambda_0)(du \,|\, X) \right]. \end{aligned}$$

Proof of Lemma 2. First, we note that

$$\begin{aligned} &\mathbb{E}_0 \left[\frac{\delta \mathbb{1}_{[0,t_0]}(y)}{S(y \,|\, x) G(y \,|\, x)} - \int_0^{t_0 \wedge y} \frac{\Lambda(du \,|\, x)}{S(u \,|\, x) G(u \,|\, x)} \,|\, X = x \right] \\ &= - \int_0^{t_0} \frac{S_0(y^- \,|\, x) G_0(y \,|\, x)}{S(y \,|\, x) G(y \,|\, x)} (\Lambda - \Lambda_0)(du \,|\, x). \end{aligned}$$

. Thus,

$$\mathbb{E}_0 \left[\varphi_{S,G,Z}(t) \,|\, X = x \right] = S(t \,|\, x) \int_0^{t_0} \frac{S_0(y^- \,|\, x) G_0(y \,|\, x)}{S(y \,|\, x) G(y \,|\, x)} (\Lambda - \Lambda_0)(du \,|\, x).$$

On the other hand, $F(t_0 \,|\, x) - F_0(t_0 \,|\, x) = S_0(t_0 \,|\, x) - S(t_0 \,|\, x)$, and, applying the Duhamel equation (Theorem 6 of Gill and Johansen, 1990), we have

$$S_0(t_0 \,|\, x) - S(t_0 \,|\, x) = S(t_0 \,|\, x) \int_0^{t_0} \frac{S_0(u^- \,|\, x)}{S(u \,|\, x)} (\Lambda - \Lambda_0)(du \,|\, x).$$

Combining these results, we have that

$$\begin{aligned}
&F(t_0 \,|\, x) - F_0(t_0 \,|\, x) - \mathbb{E}_0 \left[\varphi_{S,G,Z}(t_0) \,|\, X = x \right] \\
&= S(t_0 \,|\, x) \int_0^{t_0} \frac{S_0(u^- \,|\, x)}{S(u \,|\, x)} (\Lambda - \Lambda_0)(du \,|\, x) - S(t_0 \,|\, x) \int_0^{t_0} \frac{S_0(u^- \,|\, x) G_0(u \,|\, x)}{S(u \,|\, x) G(u \,|\, x)} (\Lambda - \Lambda_0)(du \,|\, x) \\
&= -S(t_0 \,|\, x) \int_0^{t_0} \frac{S_0(u^- \,|\, x)}{S(u \,|\, x)} \left\{ \frac{G_0(u \,|\, x)}{G(u \,|\, x)} - 1 \right\} (\Lambda - \Lambda_0)(du \,|\, x).
\end{aligned}$$

Finally, applying the tower property

$$\begin{aligned}
&P_0 \phi_{S,G}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) \\
&\qquad = \mathbb{E}_0 \left[\mathbb{1}(X \le x_0) \left\{ F(t_0 \,|\, X) - \varphi_{S,G,Z}(t_0) \right\} \right] - \mathbb{E}_0 \left[\mathbb{1}(X \le x_0) F_0(t_0 \,|\, X) \right] \\
&\qquad = \mathbb{E}_0 \left[\mathbb{1}(X \le x_0) \left\{ F(t_0 \,|\, X) - \varphi_{S,G,Z}(t_0) - F_0(t_0 \,|\, X) \right\} \right] \\
&\qquad = \mathbb{E}_0 \left[\mathbb{1}(X \le x_0) \mathbb{E}_0 \left[F(t_0 \,|\, X) - \varphi_{S,G,Z}(t_0) - F_0(t_0 \,|\, X) \,|\, X = x \right] \right] \\
&\qquad = -\mathbb{E}_0 \left[\mathbb{1}(X \le x_0) S(t_0 \,|\, X) \int_0^{t_0} \frac{S_0(u^- \,|\, X)}{S(u \,|\, X)} \left\{ \frac{G_0(u \,|\, X)}{G(u \,|\, X)} - 1 \right\} (\Lambda - \Lambda_0)(du \,|\, X) \right].
\end{aligned}$$

□

Lemma 3. *If condition (B4) holds, then* $P_0 \phi_\infty(x_0, t_0) = \mathbb{F}_0(x_0, t_0)$.

Proof of Lemma 3. By Lemma 2, we have that

$$\begin{aligned}
&P_0 \phi_\infty(x_0, t_0) - \mathbb{F}_0(x_0, t_0) \\
&\qquad = \mathbb{E}_0 \left[\mathbb{1}(X \le x_0) S(t_0 \,|\, X) \int_0^{t_0} \frac{S_0(u^- \,|\, X)}{S_\infty(u \,|\, X)} \left\{ \frac{G_0(u \,|\, X)}{G_\infty(u \,|\, X)} - 1 \right\} (\Lambda_\infty - \Lambda_0)(du \,|\, X) \right].
\end{aligned}$$

As long as $t_0 \in \mathcal{T}$, for each value $X = x$, we can use condition (B4) to decompose the interval $[0, t_0]$ into $\mathcal{S}_x \cup \mathcal{G}_w$. For any $u \in \mathcal{S}_x$, we have that $\Lambda_\infty = \Lambda_0$, so that $(\Lambda_\infty - \Lambda_0)(du \,|\, x) = 0$. For any $u \notin \mathcal{S}_x$, by assumption $u \in \mathcal{G}_x$, and so we have that $G_\infty = G_0$, so that $\frac{G_0(u \,|\, x)}{G_\infty(u \,|\, x)} - 1 = 0$. Therefore, the integral over $\mathcal{T} = \mathcal{S}_x \cup \mathcal{S}_x^C$ is equal to 0. □

We now give some results regarding the large-sample behavior of the cross-fitted estimator. We denote by $M_n \in \{1, \ldots, K\}^n$ a random vector generated by sampling uniformly from $\{1, \ldots, K\}$ with replacement. We let $\mathcal{D}_k$ denote the subset of observations with index in $\{i : M_{n,i} = k\}$ for $k = 1, \ldots, K$. We let $S_{n,k}$ denote an estimator of S_0 constructed using the data $\cup_{j \neq k} \mathcal{D}_j$, and likewise for $G_{n,k}$ and $f_{n,k}$. We then use $\mathbb{P}_{n,k}$ to denote the empirical distribution of P_0 based on data in $\mathcal{D}_k$.

Lemma 4. *Under condition (B2), there exists a universal constant M_ν, depending only on ν, such that for all n and k,*

$$\begin{aligned}
&\left[P_0 \left\{ \sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} |\phi_{n,k}(x_0, t_0) - \phi_\infty(x_0, t_0)| \right\}^2 \right]^{1/2} \\
&\leq M_\nu \Bigg\{ \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \sup_{u \in [0, t_0]} \left| \frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(u \,|\, X)} - \frac{S_\infty(t_0 \,|\, X)}{S_\infty(u \,|\, X)} \right| \right]^2 \\
&\qquad + \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \left| \frac{1}{G_{n,k}(t_0 \,|\, X)} - \frac{1}{G_\infty(t_0 \,|\, X)} \right| \right]^2 \Bigg\}.
\end{aligned}$$

Proof of Lemma 4. We start with the decomposition $\phi_{n,k}(x_0, t_0)(z) - \phi_\infty(x_0, t_0)(z) = \mathbb{1}(x \leq x_0) \sum_{j=1}^5 A_{n,k,j}(t_0)$, where

$$\begin{aligned}
A_{n,k,1}(t_0)(z) &:= \{F_{n,k}(t_0 \,|\, x) - F_\infty(t_0 \,|\, x)\} \\
A_{n,k,2}(t_0)(z) &:= \frac{\delta \mathbb{1}_{[0,t_0]}(y)}{G_\infty(y \,|\, x)} \left\{ \frac{S_{n,k}(t_0 \,|\, x)}{S_{n,k}(y \,|\, x)} - \frac{S_\infty(t_0 \,|\, x)}{S_\infty(y \,|\, x)} \right\} \\
A_{n,k,3}(t_0)(z) &:= \frac{\delta \mathbb{1}_{[0,t_0]}(y) S_{n,k}(t_0 \,|\, x)}{S_{n,k}(y \,|\, x)} \left\{ \frac{1}{G_{n,k}(y \,|\, x)} - \frac{1}{G_\infty(y \,|\, x)} \right\} \\
A_{n,k,4}(t_0)(z) &:= -\int_0^{t_0 \wedge y} \left\{ \frac{1}{G_{n,k}(u \,|\, x)} - \frac{1}{G_\infty(u \,|\, x)} \right\} \frac{S_\infty(t_0 \,|\, x) \Lambda_\infty(du \,|\, x)}{S_\infty(u \,|\, x)} \\
A_{n,k,5}(t_0)(z) &:= -\int_0^{t_0 \wedge y} \frac{1}{G_{n,k}(u \,|\, x)} \left\{ \frac{S_{n,k}(t_0 \,|\, x) \Lambda_{n,k}(du \,|\, x)}{S_{n,k}(u \,|\, x)} - \frac{S_\infty(t_0 \,|\, x) \Lambda_\infty(du \,|\, x)}{S_\infty(u \,|\, x)} \right\}.
\end{aligned}$$

Using the triangle inequality, we have that

$$
\begin{aligned}
&P_0 \left\{ \sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} |\phi_{n,k}(x_0, t_0) - \phi_\infty(x_0, t_0)| \right\}^2 \\
&\leq \left\{ \sum_{j=1}^{5} \left(\mathbb{E}_0 \left[\sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} \left| \mathbb{1}(X \leq x_0) A_{n,k,j}^2(t_0)(Z) \right| \right]^2 \right)^{1/2} \right\}^2 \\
&\leq \left\{ \sum_{j=1}^{5} \left(\mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} A_{n,k,j}^2(t_0)(Z) \right] \right)^{1/2} \right\}^2 = \left\{ \sum_{j=1}^{5} \left\{ P_0 \left(\sup_{t_0 \in \mathcal{T}} A_{n,k,j}^2(t_0) \right) \right\}^{1/2} \right\}^2 .
\end{aligned}
$$

We proceed by bounding each term individually. First, since $S_{n,k}(0 \,|\, x) = S_\infty(0 \,|\, x) = 1$, we have that

$$
\begin{aligned}
P_0 \left(\sup_{t_0 \in \mathcal{T}} A_{n,k,1}^2(t_0) \right) &= \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \{ S_\infty(t_0 \,|\, X) - S_{n,k}(t_0 \,|\, X) \}^2 \right] \\
&= \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \left\{ \frac{S_\infty(t_0 \,|\, X)}{S_\infty(0 \,|\, X)} - \frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(0 \,|\, X)} \right\}^2 \right] \\
&\leq \sup_{u \in [0, t_0]} \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \left\{ \frac{S_\infty(t_0 \,|\, X)}{S_\infty(u \,|\, X)} - \frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(u \,|\, X)} \right\}^2 \right] \\
&\leq \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}, u \in [0, t_0]} \left\{ \frac{S_\infty(t_0 \,|\, X)}{S_\infty(u \,|\, X)} - \frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(u \,|\, X)} \right\}^2 \right] \\
&\leq \mathbb{E}_0 \left[\sup_{u \in \mathcal{T}, v \in [0, u]} \left| \frac{S_\infty(u \,|\, X)}{S_\infty(v \,|\, X)} - \frac{S_{n,k}(u \,|\, X)}{S_\infty(v \,|\, X)} \right| \right]^2 .
\end{aligned}
$$

For $A_{n,k,2}$, we use the fact that $1/G_\infty \leq \nu$, so that

$$
\begin{aligned}
P_0 \left(\sup_{t_0 \in \mathcal{T}} A_{n,k,2}^2(t_0) \right) &= \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \left\{ \frac{\delta \mathbb{1}_{[0, t_0]}(Y)}{G_\infty(Y \,|\, X)} \left(\frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(Y \,|\, X)} - \frac{S_\infty(t_0 \,|\, X)}{S_\infty(Y \,|\, X)} \right) \right\}^2 \right] \\
&\leq \nu^2 \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}} \left\{ \frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(Y \,|\, X)} - \frac{S_\infty(t_0 \,|\, X)}{S_\infty(Y \,|\, X)} \right\}^2 \right] \\
&\leq \nu^2 \mathbb{E}_0 \left[\sup_{t_0 \in \mathcal{T}, u \in [0, t_0]} \left\{ \frac{S_\infty(t_0 \,|\, X)}{S_\infty(u \,|\, X)} - \frac{S_{n,k}(t_0 \,|\, X)}{S_{n,k}(u \,|\, X)} \right\}^2 \right] \\
&\leq \nu^2 \mathbb{E}_0 \left[\sup_{u \in \mathcal{T}, v \in [0, u]} \left| \frac{S_\infty(u \,|\, X)}{S_\infty(v \,|\, X)} - \frac{S_{n,k}(u \,|\, X)}{S_\infty(v \,|\, X)} \right| \right]^2 .
\end{aligned}
$$

Next, for $A_{n,k,3}$, we note that $y \leq t_0$ implies $S_{n,k}(y \mid x) \geq S_{n,k}(t_0 \mid x)$, and so $\frac{\mathbb{1}_{[0,t_0]}(y) S_{n,k}(t_0 \mid x)}{S_{n,k}(y)} \leq 1$. Hence,

$$\begin{aligned} P_0\left(\sup_{t_0 \in \mathcal{T}} A_{n,k,3}^2(t_0)\right) &= \mathbb{E}_0\left[\sup_{t_0 \in \mathcal{T}}\left\{\frac{\delta \mathbb{1}_{[0,t_0]}(Y) S_{n,k}(t_0 \mid X)}{S_{n,k}(Y \mid X)}\left(\frac{1}{G_{n,k}(Y \mid X)} - \frac{1}{G_\infty(Y \mid X)}\right)\right\}^2\right] \\ &= \mathbb{E}_0\left[\left\{\frac{1}{G_{n,k}(Y \mid X)} - \frac{1}{G_\infty(Y \mid X)}\right\}^2\right] \\ &\leq \mathbb{E}_0\left[\sup_{u \in \mathcal{T}}\left|\frac{1}{G_{n,k}(u \mid X)} - \frac{1}{G_\infty(u \mid X)}\right|\right]^2. \end{aligned}$$

For $A_{n,k,4}$, we have

$$\begin{aligned} &P_0\left(\sup_{t_0 \in \mathcal{T}} A_{n,k,4}^2(t_0)\right) \\ &= \mathbb{E}_0\left[\sup_{t_0 \in [0,\tau]}\left\{\int_0^{t_0 \wedge Y}\left(\frac{1}{G_{n,k}(u \mid X)} - \frac{1}{G_\infty(u \mid X)}\right)\frac{S_\infty(t_0 \mid X)\Lambda_\infty(du \mid X)}{S_\infty(u \mid X)}\right\}^2\right] \\ &\leq \mathbb{E}_0\left[\sup_{u \in [0,\tau]}\left(\frac{1}{G_{n,k}(u \mid X)} - \frac{1}{G_\infty(u \mid X)}\right)^2 \sup_{t_0 \in \mathcal{T}}\left\{\int_0^{t_0 \wedge Y}\frac{S_\infty(t_0 \mid X)\Lambda_\infty(du \mid X)}{S_\infty(u \mid X)}\right\}^2\right]. \end{aligned}$$

Using the backwards equation (Theorem 5 of Gill and Johansen, 1990), we have that $\int_0^t \frac{S(t)\Lambda(du)}{S(u)} = 1 - S(t)$ for any survival function S and corresponding cumulative hazard Λ. Continuing from above, we can then write

$$\begin{aligned} P_0\left(\sup_{t_0 \in \mathcal{T}} A_{n,k,4}^2(t_0)\right) &\leq \mathbb{E}_0\left[\sup_{u \in [0,\tau]}\left(\frac{1}{G_{n,k}(u \mid X)} - \frac{1}{G_\infty(u \mid X)}\right)^2 \sup_{t_0 \in \mathcal{T}}\left\{1 - S_\infty(t_0 \wedge T \mid X)\right\}^2\right] \\ &\leq \mathbb{E}_0\left[\sup_{u \in [0,\tau]}\left(\frac{1}{G_{n,k}(u \mid X)} - \frac{1}{G_\infty(u \mid X)}\right)^2\right] \\ &\leq \mathbb{E}_0\left[\sup_{u \in \mathcal{T}}\left|\frac{1}{G_{n,k}(u \mid X)} - \frac{1}{G_\infty(u \mid X)}\right|\right]^2. \end{aligned}$$

For $A_{n,k,5}$, we define $\alpha_{n,k,t}(u \mid x) = \frac{S_{n,k}(t \mid x)}{S_{n,k}(u \mid x)}$, and likewise $\alpha_{\infty,t}(u \mid x) = \frac{S_\infty(t \mid x)}{S_\infty(u \mid x)}$. Again using the backwards equation, we have that $\alpha_{n,k,t}(du \mid x) = \frac{S_{n,k}(t \mid x)\Lambda_{n,k}(du \mid x)}{S_{n,k}(u \mid x)}$ and $\alpha_{\infty,t}(du \mid x) =$

$\frac{S_\infty(t\,|\,x)\Lambda_\infty(du\,|\,x)}{S_\infty(u\,|\,x)}$. Using integration by parts, we can then write

$$
\begin{aligned}
P_0\left(\sup_{t_0\in\mathcal{T}} A^2_{n,k,5}(t_0)\right) &= \mathbb{E}_0\left[\sup_{t_0\in[0,\tau]}\left\{\int_0^{t_0\wedge Y}\frac{1}{G_{n,k}(u\,|\,X)}\left(\alpha_{n,k,t_0}(du\,|\,X)-\alpha_{\infty,t_0}(du\,|\,X)\right)\right\}^2\right] \\
&= \mathbb{E}_0\Bigg[\sup_{t_0\in[0,\tau]}\Bigg\{\frac{\alpha_{n,k,t_0}(t_0\wedge Y\,|\,X)-\alpha_{\infty,t_0}(t_0\wedge Y\,|\,X)}{G_{n,k}(t_0\wedge Y\,|\,X)} - \frac{\alpha_{n,k,t_0}(0\,|\,X)-\alpha_{\infty,t_0}(0\,|\,X)}{G_{n,k}(0\,|\,X)} \\
&\qquad - \int_0^{t\wedge Y}\left(\alpha_{n,k,t_0}(u\,|\,X)-\alpha_{\infty,t_0}(u\,|\,X)\right)\frac{G_{n,k}(du\,|\,X)}{G_{n,k}(u\,|\,X)^2}\Bigg\}^2\Bigg] \\
&= \mathbb{E}_0\Bigg[\sup_{t_0\in[0,\tau]}\Bigg\{\frac{\alpha_{n,k,t_0}(t_0\wedge Y\,|\,X)-\alpha_{\infty,t_0}(t_0\wedge Y\,|\,X)}{G_{n,k}(t_0\wedge Y\,|\,X)} - S_{n,k}(t_0\,|\,X)+S_\infty(t_0\,|\,X) \\
&\qquad - \int_0^{t\wedge Y}\left(\alpha_{n,k,t_0}(u\,|\,X)-\alpha_{\infty,t_0}(u\,|\,X)\right)\frac{G_{n,k}(du\,|\,X)}{G_{n,k}(u\,|\,X)^2}\Bigg\}^2\Bigg] \\
&\le \mathbb{E}_0\Bigg[\sup_{t_0\in[0,\tau]}\Bigg\{\nu^2\left|\frac{S_{n,k}(t_0\,|\,X)}{S_{n,k}(t_0\wedge Y\,|\,X)}-\frac{S_\infty(t_0\,|\,X)}{S_\infty(t_0\wedge Y\,|\,X)}\right| + \left|\frac{S_{n,k}(t_0\,|\,X)}{S_{n,k}(0\,|\,X)}-\frac{S_\infty(t_0\,|\,X)}{S_\infty(0\,|\,X)}\right| \\
&\qquad + \nu^2\left|\int_0^{t\wedge Y}\left(\alpha_{n,k,t_0}(u\,|\,X)-\alpha_{\infty,t_0}(u\,|\,X)\right)G_{n,k}(du\,|\,X)\right|\Bigg\}^2\Bigg] \\
&\le \mathbb{E}_0\Bigg[\sup_{t_0\in[0,\tau]}\Bigg\{\nu^2\sup_{u\in[0,t_0]}\left|\frac{S_{n,k}(t_0\,|\,X)}{S_{n,k}(u\,|\,X)}-\frac{S_\infty(t_0\,|\,X)}{S_\infty(u\,|\,X)}\right| + \sup_{u\in[0,t_0]}\left|\frac{S_{n,k}(t_0\,|\,X)}{S_{n,k}(u\,|\,X)}-\frac{S_\infty(t_0\,|\,X)}{S_\infty(u\,|\,X)}\right| \\
&\qquad + \nu^2\sup_{u\in[0,t_0]}\left|\frac{S_{n,k}(t_0\,|\,X)}{S_{n,k}(u\,|\,X)}-\frac{S_\infty(t_0\,|\,X)}{S_\infty(u\,|\,X)}\right|\Bigg\}^2\Bigg] \\
&\le (1+\nu^2)\mathbb{E}_0\left[\sup_{u\in\mathcal{T},v\in[0,u]}\left|\frac{S_\infty(u\,|\,X)}{S_\infty(v\,|\,X)}-\frac{S_{n,k}(u\,|\,X)}{S_\infty(v\,|\,X)}\right|\right]^2.
\end{aligned}
$$

The claim therefore holds with $M_\nu = \max\left\{(2+2\nu^2)^{1/2}, 2^{1/2}\right\}$. □

Lemma 5. *If conditions (B2) and (B3) hold, then for each k*

$$\sup_{x_0\in\mathcal{X},t\in\mathcal{T}}\left|n_k^{1/2}(\mathbb{P}_{n,k}-P_0)\left\{\phi_{n,k}(x_0,t_0)-\phi_\infty(x_0,t_0)\right\}\right| = o_P(1).$$

Proof of Lemma 5. We begin by defining $\pi_{n,k,z}(t_0) := F_{n,k}(t_0\,|\,x) - \varphi_{n,k,z}(t_0)$ and $\pi_{\infty,z}(t_0) := F_\infty(t_0\,|\,x) - \varphi_{\infty,z}(t_0)$. Using this notation, we then have that $\phi_{n,k,z}(x_0,t_0) = \mathbb{1}(x\le x_0)\pi_{n,k,z}(t_0)$ and $\phi_{\infty,z}(x_0,t_0) = \mathbb{1}(x\le x_0)\pi_{\infty,z}(t_0)$. We use $\mathbb{G}_{n,k}$ to denote the empirical process $n_k^{1/2}(\mathbb{P}_{n,k}-P_0)$.

We begin by using the tower property to write

$$\begin{aligned}
&\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}}\left|\mathbb{G}_{n,k}\left(\phi_{n,k}(x_0,t_0)-\phi_\infty(x_0,t_0)\right)\right|\right]\\
&=\mathbb{E}_0\left[\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}}\left|\mathbb{G}_{n,k}\left(\phi_{n,k}(x_0,t_0)-\phi_\infty(x_0,t_0)\right)\right|\mid\cup_{j\neq k}\mathcal{D}_j\right]\right]\\
&=\mathbb{E}_0\left[\mathbb{E}_0\left[\sup_{h\in\mathcal{H}_{n,k,x_0,t_0}}\left|\mathbb{G}_{n,k}|h\mid\cup_{j\neq k}\mathcal{D}_j\right]\right],
\end{aligned}$$

where $\mathcal{H}_{n,k}:=\{z\mapsto\phi_{n,k,z}(x_0,t_0)-\phi_{\infty,z}(x_0,t_0):x_0\in\mathcal{X},t_0\in\mathcal{T}\}$. Conditioning on the training data $\cup_{j\neq k}\mathcal{D}_j$, the nuisance estimators $S_{n,k}$ and $G_{n,k}$ are fixed functions. Furthermore, for any $h\in\mathcal{H}_{n,k}$, we can write $h_z(x_0,t_0)=\mathbb{1}(x\leq x_0)\{\pi_{n,k,z}(t_0)-\pi_{\infty,z}(t_0)\}$.

Next, we define $p+1$ function classes $\mathcal{H}_{n,k,\tau}:=\{z\mapsto\pi_{n,k,z}(t_0)-\pi_{\infty,z}(t_0):t_0\in\mathcal{T}\}$ and, for $j=1,\ldots,p$, $\mathcal{H}_j:=\{x_j\mapsto\mathbb{1}(x_j\leq x_{0,j}):x_{0,j}\in\mathcal{X}_j\}$. For a generic function class $\mathcal{H}$, positive real number ϵ, and norm $\|\cdot\|$ we let the covering number $N(\epsilon,\mathcal{H},\|\cdot\|)$ denote the number of $\|\cdot\|$ balls of radius no larger than ϵ needed to cover $\mathcal{F}$. Let Π denote a generic distribution on the sample space of the observed data.

For fixed S_n and G_n, Lemma 5 of Westling et al. (2023) implies that for any $\epsilon\in(0,1]$,

$$\sup_\Pi N(\epsilon\|H_\tau\|_{\Pi,2},\mathcal{H}_{n,k,t_0},L^2(\Pi))<32/\epsilon^{10},$$

where $H_\tau:=2(1+\nu)$ is a natural envelope for $\mathcal{H}_{n,k,\tau}$. Furthermore, for each $j\in1,\ldots,p$, we have that $\mathcal{H}_j$ is a VC-subgraph class, and so for ϵ small, there exists a constant C_j such that $\sup_\Pi N(\epsilon,\mathcal{H}_j,L^2(\Pi))<C_j/\epsilon$. The natural envelope for each $\mathcal{H}_j$ is $H_j:=1$.

Next, we note that

$$\mathcal{H}_{n,k}\subseteq\left\{h^*h_1h_2\cdots h_p:h^*\in\mathcal{H}_{n,k,\tau},h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2,\ldots,h_p\in\mathcal{H}_p\right\}.$$

Given that $P_0\left(H_\tau^2H_1^2\ldots H_p^2\right)=4(1+\nu)^2<\infty$ and all $p+1$ function classes are uniformly bounded, Theorem 2.10.20 of van der Vaart and Wellner (1996) implies that

$$\int_0^1\sup_\Pi\left\{N\left(4\epsilon\left(1+\nu\right)^2,\mathcal{H}_{n,k},L_2(\Pi)\right\}^{1/2}d\epsilon\right.$$

is bounded above by a constant not depending on n or k. This in turn implies that

$$\sup_{\Pi} \int_0^1 \left[1 + N(4\epsilon(1+\nu)^2, \mathcal{H}_{n,k}, L_2(\Pi))\right]^{1/2} d\epsilon$$

is bounded above by a constant not depending on n or k. By Theorem 2.14.2 of van der Vaart and Wellner (1996), there exists a constant C^* not depending on n or k such that

$$\begin{aligned}
&\mathbb{E}_0\left[\mathbb{E}_0\left[\sup_{h\in\mathcal{H}_{n,k,}} |\mathbb{G}_{n,k}|h \mid \cup_{j\neq k}\mathcal{D}_j\right]\right] \\
&\leq C^*\mathbb{E}_0\left[\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \{\phi_{n,k,Z}(x_0,t_0) - \phi_{\infty,Z}(x_0,t_0)\}^2 \mid \cup_{j\neq k}\mathcal{D}_j\right]^{1/2}\right] \\
&\leq C^*\left\{\mathbb{E}_0\left[\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \{\phi_{n,k,Z}(x_0,t_0) - \phi_{\infty,Z}(x_0,t_0)\}^2 \mid \cup_{j\neq k}\mathcal{D}_j\right]\right]\right\}^{1/2}. \qquad \text{(A.6)}
\end{aligned}$$

Next, let $U_{n,k}$ denote

$$\begin{aligned}
M_\nu\Bigg\{&\mathbb{E}_0\left[\sup_{t_0\in\mathcal{T}}\sup_{u\in[0,t_0]}\left|\frac{S_{n,k}(t_0\,|\,X)}{S_{n,k}(u\,|\,X)} - \frac{S_{\infty}(t_0\,|\,X)}{S_{\infty}(u\,|\,X)}\right|\right]^2 \\
&+ \mathbb{E}_0\left[\sup_{t_0\in\mathcal{T}}\left|\frac{1}{G_{n,k}(t_0\,|\,X)} - \frac{1}{G_{\infty}(t_0\,|\,X)}\right|\right]^2\Bigg\}.
\end{aligned}$$

By Lemma 4, we have that

$$\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \{\phi_{n,k,Z}(x_0,t_0) - \phi_{\infty,Z}(x_0,t_0)\}^2 \mid \cup_{j\neq k}\mathcal{D}_j\right] \leq U_{n,k}.$$

Combining this with (A.6)

$$\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |\mathbb{G}_{n,k}\left(\phi_{n,k}(x_0,t_0) - \phi_{\infty}(x_0,t_0)\right)|\right] \leq C^*\left\{\mathbb{E}_0\left[U_{n,k}\right]\right\}^{1/2}.$$

We note that $U_{n,k}$ is a uniformly bounded sequence of random variables converging in probability to 0 under condition (B3). This implies that $C^*\left\{\mathbb{E}_0\left[U_{n,k}\right]\right\}^{1/2} \to 0$. Finally, applying Markov's inequality, for any $\epsilon > 0$ we have that

$$\begin{aligned}
&P_0\left(\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |\mathbb{G}_{n,k}\left(\phi_{n,k}(x_0,t_0) - \phi_{\infty}(x_0,t_0)\right)| > \epsilon\right) \\
&\qquad \leq \frac{\mathbb{E}_0\left[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |\mathbb{G}_{n,k}\left(\phi_{n,k}(x_0,t_0) - \phi_{\infty}(x_0,t_0)\right)|\right]}{\epsilon} \to 0.
\end{aligned}$$

□

Lemma 6. *If conditions (B2)–(B4) hold, then for each k*

$$\sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} |\mathbb{F}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0)| \xrightarrow{p} 0.$$

Proof of Lemma 6. For each k, we define the plug-in estimator of $\mathbb{F}_0(x_0, t_0)$ as $\tilde{\mathbb{F}}_{n,k}(x_0, t_0) := \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \mathbb{1}(X_i \leq x_0) F_n(t_0 \,|\, X_i)$. We then have the following expansion of $\tilde{\mathbb{F}}_{n,k}$ about $\mathbb{F}_0$:

$$\begin{aligned}
&\tilde{\mathbb{F}}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) \\
&\qquad = \mathbb{P}_{n,k} \bar{\phi}_\infty(x_0, t_0) + C_{n,k,x_0,t_0}(P_{n,k}, P_\infty) + R_{x_0,t_0}(P_{n,k}, P_0) - \mathbb{P}_{n,k} \bar{\phi}_{n,k}(x_0, t_0),
\end{aligned}$$

where

$$\begin{aligned}
C_{n,k,x_0,t_0}(P_{n,k}, P_\infty) &= n_k^{-1/2} \mathbb{G}_{n,k}(\bar{\phi}_{n,k}(x_0, t_0) - \bar{\phi}_\infty(x_0, t_0)); \\
R_{x_0,t_0}(P_{n,k}, P_0) &= \tilde{\mathbb{F}}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) + P_0 \bar{\phi}_{n,k}(x_0, t_0).
\end{aligned}$$

We can therefore write

$$\mathbb{F}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) = \mathbb{P}_{n,k} \bar{\phi}_\infty(x_0, t_0) + C_{n,k,x_0,t_0}(P_{n,k}, P_\infty) + R_{x_0,t_0}(P_{n,k}, P_0). \tag{A.7}$$

Applying the triangle inequality yields

$$\begin{aligned}
&\sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} |\mathbb{F}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0)| \\
&\leq \sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} \left| \mathbb{P}_{n,k} \bar{\phi}_\infty(x_0, t_0) \right| + \sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} |C_{n,k,x_0,t_0}(P_{n,k}, P_\infty)| + \sup_{x_0 \in \mathcal{X}, t_0 \in \mathcal{T}} |R_{x_0,t_0}(P_{n,k}, P_0)|.
\end{aligned}$$

Lemma 5 implies that $\left\{ \bar{\phi}_\infty(x_0, t_0) : x_0 \in \mathcal{X}, t_0 \in \mathcal{T} \right\}$ is a P_0-Donsker class, and so the leading term on the right-hand side above is $O_P(n_k^{-1/2})$. We note also that $\mathbb{G}_{n,k}(\bar{\phi}_{n,k}(x_0, t_0) - \bar{\phi}_\infty(x_0, t_0)) = \mathbb{G}_{n,k}(\phi_{n,k}(x_0, t_0) - \phi_\infty(x_0, t_0))$, and so Lemma 5 implies that under conditions (B2) and (B3), the second term on the right-hand side above is $o_P(1)$. Finally, because $P_0 \bar{\phi}_{n,k}(x_0, t_0) = P_0 \phi_{n,k}(x_0, t_0) - \tilde{\mathbb{F}}_{n,k}(x_0, t_0)$, we note that

$$\begin{aligned}
R_{x_0,t_0}(P_{n,k}, P_0) &= \tilde{\mathbb{F}}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) + P_0 \phi_{n,k}(x_0, t_0) - \tilde{\mathbb{F}}_{n,k}(x_0, t_0) \\
&= P_0 \phi_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0).
\end{aligned} \tag{A.8}$$

Under condition (B4), we have that $\mathbb{F}_0(x_0, t_0) = P_0\phi_\infty(x_0, t_0)$, and so Lemma 3 implies that $R_{x_0,t_0}(P_{n,k}, P_0) = P_0\left\{\phi_{n,k}(x_0, t_0) - \phi_\infty(x_0, t_0)\right\}$. Under conditions (B2) and (B3), Lemma 4 implies that $\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |R_{x_0,t_0}(P_{n,k}, P_0)| = o_P(1)$. □

Lemma 7. *If conditions (B2) and (B3) hold with $S_\infty = S_0$ and $G_\infty = G_0$, and condition (B10) holds, then for each k*

$$\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \left|\mathbb{F}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) - \mathbb{P}_{n,k}\bar{\phi}_0(x_0, t_0)\right| = o_P(n_k^{-1/2}).$$

In particular, $\left\{n_k^{1/2}\left(\mathbb{F}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0)\right) : x_0 \in \mathcal{X}, t_0 \in \mathcal{T}\right\}$ converges weakly to a tight mean-zero Gaussian process with covariance $(x_1, t_1), (x_2, t_2) \mapsto P_0(\bar{\phi}_0(x_1, t_1)\bar{\phi}_0(x_2, t_2))$.

Proof of Lemma 7. We again use decomposition (A.7). Because $F_\infty = F_0$, $S_\infty = S_0$, $\Lambda_\infty = \Lambda_0$, and $G_\infty = G_0$, we have that $\bar{\phi}_\infty = \bar{\phi}_0$, and so we can use the triangle inequality to write

$$\begin{aligned}
&\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \left|\mathbb{F}_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) - \mathbb{P}_{n,k}\bar{\phi}_0(x_0, t_0)\right| \\
&\leq \sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |C_{n,k,x_0,t_0}(P_{n,k}, P_0)| + \sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |R_{x_0,t_0}(P_{n,k}, P_0)|.
\end{aligned}$$

Under conditions (B2) and (B3), Lemma 5 implies that the leading term on the right hand side above is $o_P(n_k^{-1/2})$. For the second term, we can use (A.8) and Lemma 2 to write

$$\begin{aligned}
R_{x_0,t_0}(P_{n,k}, P_0) &= P_0\phi_{n,k}(x_0, t_0) - \mathbb{F}_0(x_0, t_0) \\
&= \mathbb{E}_0\left[\mathbb{1}(X \leq x_0)S_{n,k}(t_0 \,|\, X)\int_0^{t_0} \frac{S_0(u^- \,|\, X)}{S_{n,k}(u \,|\, X)}\left\{\frac{G_0(u \,|\, X)}{G_{n,k}(u \,|\, X)} - 1\right\}(\Lambda_{n,k} - \Lambda_0)(du \,|\, X)\right].
\end{aligned}$$

Therefore

$$
\begin{aligned}
&\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} |R_{x_0,t_0}(P_{n,k},P_0)| \\
&\leq \sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \Bigg| \mathbb{E}_0\Big[\mathbb{1}(X\leq x_0) S_{n,k}(t_0\,|\,X) \\
&\qquad\qquad \times \int_0^{t_0} \frac{S_0(u^-\,|\,X)}{S_{n,k}(u\,|\,X)} \left\{ \frac{G_0(u\,|\,X)}{G_{n,k}(u\,|\,X)} - 1 \right\} (\Lambda_{n,k}-\Lambda_0)(du\,|\,X) \Big] \Bigg| \\
&\leq \mathbb{E}_0 \Bigg[\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \Bigg| \mathbb{1}(X\leq x_0) S_{n,k}(t_0\,|\,X) \\
&\qquad\qquad \int_0^{t_0} \frac{S_0(u^-\,|\,X)}{S_{n,k}(u\,|\,X)} \left\{ \frac{G_0(u\,|\,X)}{G_{n,k}(u\,|\,X)} - 1 \right\} (\Lambda_{n,k}-\Lambda_0)(du\,|\,X) \Bigg| \Bigg] \\
&\leq \mathbb{E}_0 \left[\sup_{t_0\in\mathcal{T}} \left| S_{n,k}(t_0\,|\,X) \int_0^{t_0} \frac{S_0(u^-\,|\,X)}{S_{n,k}(u\,|\,X)} \left\{ \frac{G_0(u\,|\,X)}{G_{n,k}(u\,|\,X)} - 1 \right\} (\Lambda_{n,k}-\Lambda_0)(du\,|\,X) \right| \right].
\end{aligned}
$$

This upper bound is $o_P(n_k^{-1/2})$ under condition (B10).

We conclude that $\sup_{x_0\in\mathcal{X},t_0\in\mathcal{T}} \left|\mathbb{F}_{n,k}(x_0,t_0) - \mathbb{F}_0(x_0,t_0) - \mathbb{P}_{n,k}\bar{\phi}_0(x_0,t_0)\right| = o_P(n_k^{-1/2})$. Lemma 5 implies that the class of influence functions $\left\{\bar{\phi}_0(x_0,t_0) : x_0\in\mathcal{X}, t_0\in\mathcal{T}\right\}$ is a uniformly bounded P_0-Donsker class, and so $\left\{ n_k^{1/2} \left(\mathbb{F}_{n,k}(x_0,t_0) - \mathbb{F}_0(x_0,t_0)\right) : x_0\in\mathcal{X}, t_0\in\mathcal{T} \right\}$ converges weakly to a tight mean-zero Gaussian process with covariance $(x_1,t_1),(x_2,t_2)\mapsto P_0(\bar{\phi}_0(x_1,t_1)\bar{\phi}_0(x_2,t_2))$, as claimed. □

A.1.3 Asymptotic analysis of VIM estimator

In the following analysis, we use $\pi_{n,k,z}(t_0)$ to denote $F_{n,k}(t_0\,|\,x) - \varphi_{n,k,z}(t_0)$ and $\pi_{0,z}(t_0)$ to denote $F_0(t_0\,|\,x) - \varphi_{0,z}(t_0)$. With this notation, we have $\phi_{n,k,z}(x_0,t_0) = \mathbb{1}(x\leq x_0)\pi_{n,k,z}(t_0)$ and $\phi_{0,z}(x_0,t_0) = \mathbb{1}(x\leq x_0)\pi_{0,z}(t_0)$.

Lemma 8. *For each k, define the remainder terms*

$$r_{n,\Phi,k} := V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0) - m \iint \Phi_{0,1}\left(f_0(x_1), t_1\right)\left\{\mathbb{F}_{n,k} - \mathbb{F}_0\right\}(dx_1, dt_1);$$

$$r_{n,\Theta,k} := V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0) - m \int \Theta_{0,1}\left(t_1\right)\left\{\mathbb{F}_{n,k} - \mathbb{F}_0\right\}(dx_1, dt_1).$$

If conditions (B2) and (B3) hold with $S_\infty = S_0$ and $G_\infty = G_0$, and in addition (B6) and (B10) hold, then $r_{n,\Phi,k}$ and $r_{n,\Theta,k}$ are $o_P(n_k^{-1/2})$.

Proof of Lemma 8. Throughout this proof, we repeatedly make use of several facts. First, for any càdlàg functions β and α, there exists a constant C such that

$$\left|\int \beta(u)\alpha(du)\right| \leq C\|\alpha\|_\infty \|\beta\|_v^*.$$

See Gill (1993) for details.

Next, for a function $\beta : \mathbb{R}^{d+p} \to \mathbb{R}$ and measure α, we define

$$\beta_{-p}(u_1, \ldots, u_d) := \int \beta(u_1, \ldots, u_d, u_{d+1}, \ldots, u_{d+p})\alpha(du_{d+1}, \ldots, du_{d+p}).$$

We have that $\|\beta_{-p}\|_v^* = \sup_s \sup_{r_s} \|\beta_{-1}\|_{v,r_s}$, where $\|\beta_{-p}\|_{v,r_s} := \sup_{u_{-r_s}} \int \left|\beta_{-p}(du_{r_s}, u_{r_{-s}})\right|$. Hence

$$\begin{aligned}
\|\beta_{-p}\|_v^* &= \sup_s \sup_{r_s} \sup_{u_{-r_s}} \int \left|\int \beta(du_{r_s}, u_{r_{-s}}, u_{d+1}, \ldots, u_{d+p})\alpha(du_{d+1}, \ldots, du_{d+p})\right| \\
&\leq \sup_s \sup_{r_s} \sup_{u_{-r_s}} \sup_{u_{d+1}, \ldots, u_{d+p}} \left\{\int \left|\beta(du_{r_s}, u_{r_{-s}}, u_{d+1}, \ldots, u_{d+p})\right|\right\} \int \left|\alpha(du_{d+1}, \ldots, du_{d+p})\right| \\
&= \sup_{u_{d+1}, \ldots, u_{d+p}} \|\beta(\cdot, u_{d+1}, \ldots, u_{d+p})\|_v^* \|\alpha\|_v.
\end{aligned}$$

Next, let $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^p$ be vectors, and suppose $g : \mathbb{R}^{d+p} \to \mathbb{R}$ can be written as the product $g(u, v) = g_1(u)g_2(v)$. We claim that $\|g\|_v^* \leq \|g_1\|_v^* \|g_2\|_v^*$. We have that

$$\begin{aligned}
\|g\|_v^* &= \sup_s \sup_{r_s} \sup_{u_{-r_s}, v_{-r_s}} \int \left|g_1(du_{r_s}, du_{-r_s})g_2(dv_{r_s}, dv_{-r_s})\right| \\
&\leq \sup_s \sup_{r_s} \sup_{u_{-r_s}, v_{-r_s}} \int \left|g_1(du_{r_s}, du_{-r_s})\right| \left|g_2(dv_{r_s}, dv_{-r_s})\right| \\
&\leq \sup_s \sup_{r_s} \sup_{u_{-r_s}} \int \left|g_1(du_{r_s}, du_{-r_s})\right| \sup_s \sup_{r_s} \sup_{v_{-r_s}} \int \left|g_2(dv_{r_s}, dv_{-r_s})\right| = \|g_1\|_v^* \|g_2\|_v^*.
\end{aligned}$$

Next, we claim that $n_k^{-1/2}\left\|\mathbb{G}_{n,k}\bar{\phi}_0\right\|_v^* = O_P(1)$. Because we can write $(\mathbb{P}_{n,k} - P_0)\bar{\phi}_0(x_0,t_0) = \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\bar{\phi}_{0,Z_i}(x_0,t_0)$, the triangle inequality yields that $n_k^{-1/2}\left\|\mathbb{G}_{n,k}\bar{\phi}_0\right\|_v^* \leq \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\left\|\bar{\phi}_{0,Z_i}\right\|_v^*$. For fixed x, the uniform sectional variation norm of $x_0 \mapsto \mathbb{1}(x \leq x_0)$ is 1, so $\left\|\bar{\phi}_{0,z}\right\|_v^* \leq \|\pi_{0,z}\|_v^* = \|\pi_{0,z}\|_v$, where we have replaced the uniform sectional variation norm with the variation norm since π_0 is univariate. Now, for any fixed $Z = z$, we have

$$
\begin{aligned}
&\|\pi_{0,z}\|_v = \int |\{F_0(dt\,|\,x) - \varphi_{0,z}(dt)\}| \\
&\leq \int |F_0(dt\,|\,x)| + \int \left|\left\{\frac{\delta}{S_0(y\,|\,x)G_0(y\,|\,x)} - \int_0^{t\wedge y}\frac{\Lambda_0(du\,|\,x)}{S_0(u\,|\,x)G_0(u\,|\,x)}\right\}S_0(dt\,|\,x)\right| \\
&+ \frac{\delta}{G_0(y\,|\,x)} + \int\left|\frac{\mathbb{1}_{[0,y]}(t)\Lambda_0(dt\,|\,x)}{G_0(t\,|\,x)}\right| \\
&\leq 1 + \frac{\delta}{S_0(y\,|\,x)G_0(y\,|\,x)} + \sup_t\left|\int_0^{t\wedge y}\frac{\Lambda_0(du\,|\,x)}{G_0(u\,|\,x)}\right|\int|S_0(dt\,|\,x)| + \frac{\delta}{G_0(y\,|\,x)} + \int_0^y\left|\frac{\Lambda_0(dt\,|\,x)}{G_0(t\,|\,x)}\right| \\
&\leq 1 + \frac{\delta}{S_0(y\,|\,x)G_0(y\,|\,x)} + \left|\int_0^y\frac{\Lambda_0(du\,|\,x)}{G_0(u\,|\,x)}\right| + \frac{\delta}{G_0(y\,|\,x)} + \int_0^y\left|\frac{\Lambda_0(dt\,|\,x)}{G_0(t\,|\,x)}\right|.
\end{aligned}
$$

In light of condition (B2), the above function of Z has finite mean and variance, and so

$$\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\|\pi_{0,Z_i}\|_v = O_P(1).$$

Next, we note that since $\mathbb{F}_{n,k}(x_0,t_0) - \mathbb{F}_0(x_0,t_0) = n_k^{-1/2}\mathbb{G}_{n,k}\bar{\phi}_0(x_0,t_0) + r_{n,\mathbb{F}_0}(x_0,t_0)$, then by the triangle inequality

$$\|r_{n,\mathbb{F}_0}\|_v^* \leq \left\|n_k^{-1/2}\mathbb{G}_{n,k}\bar{\phi}_0\right\|_v^* + \|\mathbb{F}_{n,k} - \mathbb{F}_0\|_v^* \leq n_k^{-1/2}\left\|\mathbb{G}_{n,k}\bar{\phi}_0\right\|_v^* + 2 = O_P(1).$$

We now analyze the remainder terms $r_{n,\Phi,k}$ and $r_{n,\Theta,k}$. We begin by analyzing $r_{n,\Theta,k}$. For $m = 1$, $r_{n,\Theta,k} = 0$, and for $m \geq 2$, we have that

$$r_{n,\Theta,k} = \sum_{l=2}^m A_{l,m}\int\cdots\int\Theta_{0,l}(t_1,\ldots,t_l)\left\{\prod_{j=1}^l(\mathbb{F}_{n,k} - \mathbb{F}_0)(dx_j,dt_j)\right\},$$

where the coefficients are defined recursively via the relationships

$$
\begin{aligned}
&A_{1,m} = m; \\
&A_{l+1,m} = \sum_{i=l}^{m-1}A_{l,i} \text{ for } j = 1,2,\ldots,m-1.
\end{aligned}
$$

For any $l \geq 2$, we can write the corresponding term in the above sum as

$$
\begin{aligned}
&A_{l,m} n_k^{-l/2} \int \Theta_{0,l}(t_1, \ldots, t_l) \prod_{j=1}^{l} \left\{ \mathbb{G}_{n,k} \bar{\phi}_0(dx_j, dt_j) \right\} \\
&+ A_{l,m} \sum_{j=1}^{l-1} \binom{l}{j} \int \Theta_{0,l}(t_1, \ldots, t_l) \prod_{s=1}^{l-j} \left\{ n_k^{-1/2} \mathbb{G}_{n,k} \bar{\phi}_0(dx_j, dt_s) \right\} \prod_{s=l-j+1}^{l} r_{n,\mathbb{F}_0}(dx_s, dt_s) \\
&+ A_{l,m} \int \Theta_{0,l}(t_1, \ldots, t_l) \prod_{j=1}^{l} r_{n,\mathbb{F}_0}(dx_j, dt_j) \\
&:= r_{n,\Theta,k,1} + r_{n,\Theta,k,2} + r_{n,\Theta,k,3}.
\end{aligned}
$$

We bound each of these terms separately. The leading term $r_{n,\Theta,k,1}$ is a degenerate V-statistic and hence is $o_P\left(n^{-1/2}\right)$ (Serfling, 1980). For $r_{n,\Theta,k,2}$, there exists a constant C_1 such that for all $1 \leq j \leq l-1$,

$$
\begin{aligned}
A_{l,m} \binom{l}{j} \int \Theta_{0,l}(t_1, \ldots, t_l) \prod_{s=1}^{l-j} \left\{ n_k^{-1/2} \mathbb{G}_{n,k} \bar{\phi}_0(dx_s, dt_s) \right\} \prod_{s=l-j+1}^{l} r_{n,\mathbb{F}_0}(dx_s, dt_s) \\
\leq C_1 \|r_{n,\mathbb{F}_0}\|_\infty \|q_{n,j}\|_v^*, \qquad \text{(A.9)}
\end{aligned}
$$

where $q_{n,j}(t_l) := \int \Theta_{0,l}(t_1, \ldots, t_l) \prod_{s=1}^{l-j} \left\{ n_k^{-1/2} \mathbb{G}_{n,k} \bar{\phi}_0(dt_s) \right\} \prod_{s=l-j+1}^{l-1} r_{n,\mathbb{F}_0}(dx_s, dt_s)$. Now, we note that

$$
\|q_{n,j}\|_v^* \leq n_k^{-(l-j)/2} \sup_{t_1, \ldots, t_{l-1}} \|\Theta_{0,l}(t_1, t_2, \ldots, t_{l-1}, \cdot)\|_v \left\| \mathbb{G}_{n,k} \bar{\phi}_0 \right\|_v^{l-j} \|r_{n,\mathbb{F}_0}\|_v^{j-1}.
$$

Under condition (B6), this upper bound is $O_P(1)$, and so the upper bound in (A.9) is $\|r_{n,\mathbb{F}_0}\|_\infty O_P(1)$.

For $r_{n,\Theta,k,3}$, we apply an analogous argument, replacing $n_k^{-1/2} \mathbb{G}_{n,k} \pi_0$ with $r_{n,\mathbb{F}_0}$, to conclude that $r_{n,\Theta,k,3} = \|r_{n,\mathbb{F}_0}\|_\infty O_P(1)$. Under conditions (B2), (B3) (with $S_\infty = S_0$ and $G_\infty = G_0$), and (B10), Lemma 7 yields that $\|r_{n,\mathbb{F}_0}\|_\infty = o_P\left(n^{-1/2}\right)$, and so $r_{n,\Theta,k,2} = o_P\left(n^{-1/2}\right) O_P(1) = o_P\left(n^{-1/2}\right)$ and $r_{n,\Theta,k,3} = o_P\left(n^{-1/2}\right) O_P(1) = o_P\left(n^{-1/2}\right)$.

Next, for $r_{n,\Phi,k}$, we again have that $r_{n,\Phi,k} = 0$ for $m = 1$, and for $m \geq 2$, we have

$$
r_{n,\Phi,k} = \sum_{l=2}^{m} A_{l,m} \int \cdots \int \Phi_{0,l}((f_0(x_1), t_1), \ldots, (f_0(x_l), t_l)) \prod_{j=2}^{l} (\mathbb{F}_{n,k} - \mathbb{F}_0)(dx_j, dt_j),
$$

For any $l \geq 2$, we can write the corresponding term in the above sum as

$$
\begin{aligned}
& A_{l,m} n_k^{-l/2} \int \Phi_{0,l}\left((f_0(x_1), t_1), \ldots, (f_0(x_l), t_l)\right) \prod_{j=1}^{l} \left\{\mathbb{G}_{n,k} \bar{\phi}_0(dx_j, dt_j)\right\} \\
& + A_{l,m} \sum_{j=1}^{l-1} \binom{l}{j} \int \left\{ \Phi_{0,l}\left((f_0(x_1), t_1), \ldots, (f_0(x_l), t_l)\right) \prod_{s=1}^{l-j} \left\{ n_k^{-1/2} \mathbb{G}_{n,k} \bar{\phi}_0(dx_s, dt_s) \right\} \right. \\
& \left. \times \prod_{s=l-j+1}^{l} r_{n,\mathbb{F}_0}(dx_s, dt_s) \right\} \\
& + A_{l,m} \int \Phi_{0,l}\left((f_0(x_1), t_1), \ldots, (f_0(x_l), t_l)\right) \prod_{j=1}^{l} r_{n,\mathbb{F}_0}(dx_j, dt_j) \\
& := r_{n,\Phi,k,1} + r_{n,\Phi,k,2} + r_{n,\Phi,k,3}.
\end{aligned}
$$

Each of these three terms is analyzed identically as their analogs above. We have that $r_{n,\Phi,k,1} = o_P\left(n^{-1/2}\right)$ because it is a degenerate V-statistic. Under conditions (B2), (B3) (with $S_\infty = S_0$ and $G_\infty = G_0$), and (B10), Lemma 7 implies that $r_{n,\Phi,k,2} = o_P\left(n^{-1/2}\right) O_P(1) = o_P\left(n^{-1/2}\right)$ and $r_{n,\Phi,k,3} = o_P\left(n^{-1/2}\right) O_P(1) = o_P\left(n^{-1/2}\right)$.

□

Proof of Theorem 2. We begin with an expansion of $v_{n,1}$ about $v_{0,1}$:

$$
v_{n,1} - V_1(f_0, \mathbb{F}_0) = \frac{1}{K} \sum_{k=1}^{K} \left\{V_1(f_{n,k}, \mathbb{F}_0) - V_1(f_0, \mathbb{F}_0)\right\} + \frac{1}{K} \sum_{k=1}^{K} \left\{V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0)\right\}.
$$

Under condition (B5), for each k we have that $|V_1(f_{n,k}, \mathbb{F}_0) - V_1(f_0, \mathbb{F}_0)| \leq J_1 \|f_{n,k} - f_0\|_{\mathcal{F}} = o_P(1)$, and hence $\frac{1}{K} \sum_{k=1}^{K} \left\{V_1(f_{n,k}, \mathbb{F}_0) - V_1(f_0, \mathbb{F}_0)\right\} = o_P(1)$. Next, we note that

$$
\begin{aligned}
& |V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0)| \\
& = \left| \int \cdots \int \Phi\left((f_{n,k}(x_1), t_1), \ldots, (f_{n,k}(x_m), t_m)\right) \left\{ \prod_{j=1}^{m} \mathbb{F}_{n,k}(dx_j, dt_j) - \prod_{j=1}^{m} \mathbb{F}_0(dx_j, dt_j) \right\} \right| \\
& \leq \int \cdots \int \left|\Phi\left((f_{n,k}(x_1), t_1), \ldots, (f_{n,k}(x_m), t_m)\right)\right| \left| \left\{ \prod_{j=1}^{m} \mathbb{F}_{n,k}(dx_j, dt_j) - \prod_{j=1}^{m} \mathbb{F}_0(dx_j, dt_j) \right\} \right| \\
& \leq \|\Phi\|_\infty \int \cdots \int \left| \left\{ \prod_{j=1}^{m} \mathbb{F}_{n,k}(dx_j, dt_j) - \prod_{j=1}^{m} \mathbb{F}_0(dx_j, dt_j) \right\} \right|.
\end{aligned}
$$

We use a telescoping sum to write

$$\prod_{j=1}^{m} \mathbb{F}_{n,k}(dx_j, dt_j) - \prod_{j=1}^{m} \mathbb{F}_0(dx_j, dt_j)$$
$$= \sum_{j=1}^{K+1} \left\{ \prod_{i=1}^{j} \mathbb{F}_{n,k}(dx_i, dt_i) \prod_{i=j+1}^{K} \mathbb{F}_0(dx_i, dt_i) - \prod_{i=1}^{j-1} \mathbb{F}_{n,k}(dx_i, dt_i) \prod_{i=j}^{K} \mathbb{F}_0(dx_i, dt_i) \right\}.$$

For any j, we have

$$\int \cdots \int \left| \prod_{i=1}^{j-1} \mathbb{F}_{n,k}(dx_i, dt_i) \left\{ \mathbb{F}_{n,k}(dx_j, dt_j) - \mathbb{F}_0(dx_j, dt_j) \right\} \prod_{i=j+1}^{K} \mathbb{F}_0(dx_i, dt_i) \right|$$
$$\leq \left\| \prod_{i=1}^{j-1} \mathbb{F}_{n,k} \left\{ \mathbb{F}_{n,k} - \mathbb{F}_0 \right\} \prod_{i=j+1}^{K} \mathbb{F}_0 \right\|_{\infty} \leq \left\| \mathbb{F}_{n,k} - \mathbb{F}_0 \right\|_{\infty} = o_P(1).$$

Therefore, $V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0) = o_P(1)$ for each k, and so we conclude that

$$\frac{1}{K} \sum_{k=1}^{k} \left\{ V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0) \right\} = o_P(1).$$

For $v_{n,2}$, we have that $v_{n,2} - V_2(\mathbb{F}_0) = \frac{1}{K} \sum_{k=1}^{K} \left\{ V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0) \right\}$. We apply Lemma 8 to yield that $V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0) = \mathbb{P}_{n,k} D_\Theta + o_P(1) = o_P(1) + o_P(1) = o_P(1)$. This holds for all k, and so $\frac{1}{K} \sum_{k=1}^{K} \left\{ V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0) \right\} = o_P(1)$. Therefore $v_{n,2} \xrightarrow{\text{p}} V_2(\mathbb{F}_0)$.

Slutsky's lemma yields that $v_{n,1}/v_{n,2} \xrightarrow{\text{p}} V(f_0, \mathbb{F}_0)$. □

Before proving Theorem 3, we define for any f and π

$$\Gamma(f, \pi)(z_1, \ldots, z_j) = \int \cdots \int \Phi\left((f(x_1), t_1), \ldots, (f(x_m), t_m) \right) \prod_{j=1}^{m} \pi_{z_j}(dt_j).$$

Using this notation, we define the following random functions for each $k \in \{1, \ldots, K\}$:

$$h_{n,k,1}(z_1, \ldots, z_m) := \Gamma(f_{n,k}, \pi_{n,k})(z_1, \ldots, z_m) - \Gamma(f_0, \pi_0)(z_1, \ldots, z_m)$$
$$h_{n,k,2}(z_1, \ldots, z_m) := \Gamma(f_0, \pi_{n,k})(z_1, \ldots, z_m) - \Gamma(f_0, \pi_0)(z_1, \ldots, z_m);$$
$$h_{n,k,3}(z_1, \ldots, z_m) := \int \cdots \int \{ \Phi\left((f_{n,k}(x_1), t_1), \ldots, (f_{n,k}(x_m), t_m) \right);$$
$$- \Phi\left((f_0(x_1), t_1), \ldots, (f_0(x_m), t_m) \right) \} \left\{ \pi^m_{n,k,z_j} - \pi^m_{0,z_j} \right\} (dt_1, \ldots, dt_m);$$
$$h_{n,k,t}(x) := S_{n,k}(t \,|\, x) \int_0^t \left\{ \frac{G_0(u \,|\, x)}{G_{n,k}(u \,|\, x)} - 1 \right\} \left\{ \Lambda_{n,k} - \Lambda_0 \right\} (du \,|\, x).$$

Proof of Theorem 3. We begin with the following expansion of $v_{n,1}$ about $v_{0,1}$:

$$\begin{aligned} & v_{n,1} - V_1(f_0, \mathbb{F}_0) \\ & \quad = \frac{1}{K}\sum_{k=1}^{K}\{V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\} + \frac{1}{K}\sum_{k=1}^{K}\{V_1(f_{n,k}, \mathbb{F}_0) - V_1(f_0, \mathbb{F}_0)\} + r_n, \end{aligned}$$

where $r_n := \frac{1}{K}\sum_{k=1}^{K}\left[\{V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0)\} - \{V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\}\right]$. Under conditions (B7) and (B8), for each k we have that $|V_1(f_{n,k}, \mathbb{F}_0) - V_1(f_0, \mathbb{F}_0)| \leq J_3\|f_{n,k} - f_0\|_{\mathcal{F}}^2 = o_P\left(n^{-1/2}\right)$. Next, applying Lemma 8, we have that

$$\begin{aligned} & V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0) = m\iint \Phi_{0,1}\left(f_0(x_1), t_1\right)\{\mathbb{F}_{n,k} - \mathbb{F}_0\}\left(dx_1, dt_1\right) + o_P(n_k^{-1/2}) \\ & = m\Big\{\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\int \Phi_{0,1}\left(f_0(X_i), t_1\right)\pi_{0,Z_i}(dt_1) \\ & \qquad + \iint \Phi_{0,1}\left(f_0(x_1), t_1\right) r_{n,\mathbb{F}_0}(dx_1, dt_1) - V_1(f_0, \mathbb{F}_0)\Big\} + o_P(n_k^{-1/2}) \\ & \overset{\text{(b)}}{=} \frac{m}{n_k}\sum_{i\in\mathcal{D}_k}\left\{\int \Phi_{0,1}\left(f_0(X_i), t_1\right)\left(F_0(dt_1 \mid X_i) - \varphi_{0,Z_i}(dt_1)\right) - V_1(f_0, \mathbb{F}_0)\right\} + o_P(n_k^{-1/2}) \\ & = \mathbb{P}_{n,k}D_\Phi + o_P(n_k^{-1/2}), \end{aligned}$$

where (b) follows by noting that, under condition (B6), there exists a constant C such that

$$\left|m\iint \Phi_{0,1}\left(f_0(x_1), t_1\right) r_{n,\mathbb{F}_0}(dx_1, dt_1)\right| \leq C\|\Phi_{0,1}\|_v^*\|r_{n,\mathbb{F}_0}\|_\infty = o_P(n_k^{-1/2}).$$

Then, using the triangle inequality,

$$\begin{aligned} \left|\left\{\frac{1}{K}\sum_{k=1}^{K} V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\right\} - \mathbb{P}_n D_\Phi\right| & \leq \max_k\left|\frac{n}{Kn_k} - 1\right| \cdot \mathbb{P}_n D_\Phi + \frac{1}{K}\sum_{k=1}^{K} o_P(n_k^{-1/2}) \\ & = O_P\left(n^{-1}\right) + o_P\left(n^{-1/2}\right) = o_P\left(n^{-1/2}\right). \end{aligned}$$

To analyze r_n, we note that $P_0^m\Gamma(f_0, \pi_0) = V_1(f_0, \mathbb{F}_0)$. We also use the fact that $\int b(x,t)\mathbb{F}_{n,k}(dx, dt) = \frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\int b(X_i, t)\pi_{n,k}(dt)$ for any function b. Then, for any k we

have

$$
\begin{aligned}
&\{V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0)\} - \{V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\} \\
&= \mathbb{P}_{n,k}^m \Gamma(f_{n,k}, \pi_{n,k}) - P_0^m \Gamma(f_{n,k}, \pi_0) - \mathbb{P}_{n,k}^m \Gamma(f_0, \pi_{n,k}) + P_0^m \Gamma(f_0, \pi_0) \\
&= (\mathbb{P}_{n,k}^m - P_0^m) \{\Gamma(f_{n,k}, \pi_{n,k}) - \Gamma(f_0, \pi_0)\} - (\mathbb{P}_{n,k}^m - P_0^m) \{\Gamma(f_0, \pi_{n,k}) - \Gamma(f_0, \pi_0)\} \\
&\qquad + P_0^m \{\Gamma(f_{n,k}, \pi_{n,k}) - \Gamma(f_{n,k}, \pi_0) - \Gamma(f_0, \pi_{n,k}) + \Gamma(f_0, \pi_0)\} \\
&= (\mathbb{P}_{n,k}^m - P_0^m) \{\Gamma(f_{n,k}, \pi_{n,k}) - \Gamma(f_0, \pi_0)\} - (\mathbb{P}_{n,k}^m - P_0^m) \{\Gamma(f_0, \pi_{n,k}) - \Gamma(f_0, \pi_0)\} \\
&\qquad + \mathbb{E}_0 \Bigg[\int \cdots \int \Big\{ \Phi\left((f_{n,k}(X_1), t_1), \ldots, (f_{n,k}(X_m), t_m)\right) \\
&\qquad\qquad - \Phi\left((f_0(X_1), t_1), \ldots, (f_0(X_m), t_m)\right) \Big\} \left\{ \pi_{n,k,Z}^m - \pi_{0,Z}^m \right\} (dt_1, \ldots, dt_m) \Bigg] \\
&= (\mathbb{P}_{n,k}^m - P_0^m) h_{n,1,k} - (\mathbb{P}_{n,k}^m - P_0^m) h_{n,2,k} \\
&\qquad + \mathbb{E}_0 \Bigg[\int \cdots \int \Big\{ \Phi\left((f_{n,k}(X_1), t_1), \ldots, (f_{n,k}(X_m), t_m)\right) \\
&\qquad\qquad - \Phi\left((f_0(X_1), t_1), \ldots, (f_0(X_m), t_m)\right) \Big\} \left\{ \pi_{n,k,Z}^m - \pi_{0,Z}^m \right\} (dt_1, \ldots, dt_m) \Bigg].
\end{aligned}
$$

The trailing term is $o_P\left(n^{-1/2}\right)$ under condition (B10).

Next, for any function $h : \mathcal{X}^m \to \mathbb{R}$, we let $\mathbb{P}_{n,k,*}^m h$ denote the U-statistic analog of $\mathbb{P}_{n,k}^m h$, i.e.

$$
\mathbb{P}_{n,k,*} h = \binom{n}{m}^{-1} \sum_{\underline{i}_m \in \mathcal{D}_{m,n_k}} h(X_{i_1}, \ldots X_{i_m}),
$$

where $\mathcal{D}_{m,n_k} := \{\underline{i}_m \subseteq \{1, \ldots, n_k\} : i_1 < i_2 < \cdots < i_m\}$. We also let $\bar{h}(x) := P_0^{m-1} h$. Using results from Serfling (1980), if $\mathrm{Var}_0[\bar{h}] > 0$, we have that

$$
\mathrm{Var}_0[n_k^{1/2} \mathbb{P}_{n,k,*}^m h] = m^2 \mathrm{Var}_0[\bar{h}] + O(n_k^{-1}).
$$

For any $\epsilon > 0$, we can therefore apply Chebyshev's inequality to yield

$$\begin{aligned} 0 \leq P_0\left(\left|n_k^{1/2}(\mathbb{P}_{n,k,*}^m - P_0^m)h_{n,1,k}\right| > \epsilon \mid \cup_{j\neq k}\mathcal{D}_j\right) &\leq \frac{\mathrm{Var}_0\left[n_k^{1/2}\mathbb{P}_{n,k,*}^m h_{n,1,k} \mid \cup_{j\neq k}\mathcal{D}_j\right]}{\epsilon^2} \\ &= \frac{m^2\mathrm{Var}_0[\bar{h}_{n,1,k}] + O(n^{-1})}{\epsilon^2} \\ &\leq \frac{m^2 P_0\bar{h}_{n,1,k}^2 + O(n^{-1})}{\epsilon^2}. \end{aligned}$$

In light of condition (B9), then, $P_0\left(\left|n_k^{1/2}(\mathbb{P}_{n,k,*}^m - P_0^m)h_{n,1,k}\right| > \epsilon \mid \cup_{j\neq k}\mathcal{D}_j\right) = o_P(1)$. This holds for any realization of $\cup_{j\neq k}\mathcal{D}_j$, and since probabilities are uniformly bounded

$$\begin{aligned} &\mathbb{E}_0\left[P_0\left(\left|n_k^{1/2}(\mathbb{P}_{n,k,*}^m - P_0^m)h_{n,1,k}\right| > \epsilon \mid \cup_{j\neq k}\mathcal{D}_j\right)\right] \\ &\quad = P_0\left(\left|n_k^{1/2}(\mathbb{P}_{n,k,*}^m - P_0^m)h_{n,1,k}\right| > \epsilon\right) = o(1). \end{aligned}$$

On the other hand, for fixed h, we have that $(\mathbb{P}_{n,k}^m - \mathbb{P}_{n,k,*}^m)h = o_P(n_k^{-1/2})$ (Serfling, 1980). Using the same argument as above, we have

$$\begin{aligned} &\mathbb{E}_0\left[P_0\left(\left|n_k^{1/2}(\mathbb{P}_{n,k}^m - \mathbb{P}_{n,k,*}^m)h_{n,1,k}\right| > \epsilon \mid \cup_{j\neq k}\mathcal{D}_j\right)\right] \\ &\quad = P_0\left(\left|n_k^{1/2}(\mathbb{P}_{n,k}^m - \mathbb{P}_{n,k,*}^m)h_{n,1,k}\right| > \epsilon\right) = o(1). \end{aligned}$$

Finally, we can write that

$$\begin{aligned} (\mathbb{P}_{n,k}^m - P_0^m)h_{n,1,k} &= (\mathbb{P}_{n,k,*}^m - P_0^m)h_{n,1,k} + (\mathbb{P}_{n,k}^m - \mathbb{P}_{n,k,*}^m)h_{n,1,k} \\ &= o_P(n_k^{-1/2}) + o_P(n_k^{-1/2}) = o_P(n_k^{-1/2}). \end{aligned}$$

Since $n/n_k \xrightarrow{\mathrm{p}} K$, we have that $(\mathbb{P}_{n,k}^m - P_0^m)h_{n,1,k} = o_P\left(n^{-1/2}\right)$. An identical argument holds for $(\mathbb{P}_{n,k}^m - P_0^m)h_{n,2,k}$. Thus, for all k,

$$\{V_1(f_{n,k}, \mathbb{F}_{n,k}) - V_1(f_{n,k}, \mathbb{F}_0)\} - \{V_1(f_0, \mathbb{F}_{n,k}) - V_1(f_0, \mathbb{F}_0)\} = o_P\left(n^{-1/2}\right).$$

We conclude that $r_n = o_P\left(n^{-1/2}\right)$. Finally, we have that $v_{n,1} - V_1(f_0, \mathbb{F}_0) = \mathbb{P}_n D_\Phi + o_P\left(n^{-1/2}\right)$.

For $v_{n,2}$, we have that $v_{n,2} - V_2(\mathbb{F}_0) = \frac{1}{K}\sum_{k=1}^{K} V_2(\mathbb{F}_{n,k})$. Applying Lemma 8, we have that

$$\begin{aligned} V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0) &= m \int \cdots \int \Theta_{0,1}(t_1) \{\mathbb{F}_{n,k} - \mathbb{F}_0\}(dx_1, dt_1) + o_P(n_k^{-1/2}) \\ &= m \left\{ \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \int \Theta_{0,1}(t_1) (F_0(dt_1 \mid X_i) - \varphi_{0,Z_i}(dt_1)) - V_2(\mathbb{F}_0) \right\} + o_P(n_k^{-1/2}) \\ &= \mathbb{P}_{n,k} D_\Theta + o_P(n_k^{-1/2}). \end{aligned}$$

Using the same argument as above, we have that $\left| \frac{1}{K}\sum_{k=1}^{K} V_2(\mathbb{F}_{n,k}) - V_2(\mathbb{F}_0) - \mathbb{P}_n D_\Theta \right| = o_P\left(n^{-1/2}\right)$, and hence $v_{n,2} - V_2(\mathbb{F}_0) = \mathbb{P}_n D_\Theta + o_P\left(n^{-1/2}\right)$.

An application of the delta method yields that $v_n - v_0 = (\mathbb{P}_n - P_0)D(P_0) + o_P\left(n^{-1/2}\right)$. Therefore, v_n is an asymptotically linear estimator of v_0 with influence function equal to $D(P_0)$. Under condition (B1), Theorem 1 holds and $D(P_0)$ is the efficient influence function of $P \mapsto V(f_P, P)$ at P_0 relative to $\mathcal{M}$, and so v_n is nonparametric efficient. □

A.2 Additional technical details

A.2.1 Identification of the conditional cumulative hazard function

Under conditions (A1) and (A2), we have that $\Lambda_0^*(t \mid x) = \Lambda_0(t \mid x)$; in other words, we can write the conditional cumulative hazard function of T given X under P_0^* as a functional of the observed data distribution P_0. To see this, we begin by using standard probability rules to write

$$\begin{aligned} H_{0,1}(u \mid x) = P_0(Y \leq u, \Delta = 1 \mid X = x) &= P_0^*(T \leq u, T \leq C \mid X = x) \\ &= \int_0^u P_0^*(C \geq t \mid X = x) F_0^*(dt \mid x). \end{aligned}$$

Here, we have used the conditional independence of T and C given X. On the other hand, using the same conditional independence, we have that $1 - H_0(u^- \mid x) = P_0^*(C \geq u \mid X = x)\{1 - F_0^*(u^- \mid x)\}$. Condition (A2) implies that $P_0^*(C \geq u \mid X = x) > 0$ for all $u \in (0, t]$,

and so we can write

$$\Lambda_0(t \,|\, x) = \int_0^t \frac{H_{0,1}(du \,|\, x)}{1 - H_0(u^- \,|\, x)} = \int_0^t \frac{P_0^*(C \geq u \,|\, X = x) F_0^*(du \,|\, x)}{P_0^*(C \geq u \,|\, X = x) \{1 - F_0^*(u^- \,|\, x)\}}$$
$$= \int_0^t \frac{F_0^*(du \,|\, x)}{1 - F_0^*(u^- \,|\, x)} = \Lambda_0^*(t \,|\, x).$$

A.2.2 Derivation of oracle prediction functions for examples

Example 1: AUC. For a binary outcome D, Williamson et al. (2021b) considered the AUC $V(f, P_0) := P_0(f(X_1) > f(X_2) \,|\, D_1 = 1, D_2 = 0)$ and showed that the oracle prediction function was $f_0 : x \mapsto \mathbb{E}_0[D \,|\, X = x]$. Adapting this to the binary outcome $\mathbb{1}(T > \tau)$, we have the oracle $f_0 : x \mapsto \mathbb{E}_0^*[\mathbb{1}(T > \tau) \,|\, X = x] = S_0(\tau \,|\, x)$.

Example 2: C-index. The class of functions that maximize the c-index can be written as

$$\mathcal{F}_0 := \big\{ f : f(x_1) > f(x_2) \text{ when}$$
$$P_0^*(T_1 < T_2, T_1 \leq \tau \,|\, X_1 = x_1, X_2 = x_2) \geq P_0^*(T_2 < T_1, T_2 \leq \tau \,|\, X_1 = x_1, X_2 = x_2) \big\}.$$

We can write the symmetrized c-index as

$$\mathbb{E}_0^* [\mathbb{1}(f(X_1) > f(X_2)) \mathbb{1}(T_1 < T_2, T_1 \leq \tau) + \mathbb{1}(f(X_2) > f(X_1)) \mathbb{1}(T_2 < T_1, T_2 \leq \tau) \,|\, X_1, X_2]$$
$$= \mathbb{E}_0^* \big[\mathbb{1}(f(X_1) > f(X_2)) \mathbb{1}(T_1 < T_2, T_1 \leq \tau)$$
$$+ (1 - \mathbb{1}(f(X_1) > f(X_2))) \mathbb{1}(T_2 < T_1, T_2 \leq \tau) \,|\, X_1, X_2 \big]$$
$$= \mathbb{E}_0^* \big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau)$$
$$+ \mathbb{1}(f(X_1) > f(X_2)) \{ \mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau) \} \,|\, X_1, X_2 \big].$$

Let f_0 be a function such that $\mathbb{1}(f_0(x_1) > f_0(x_2)) = \mathbb{1}(P_0(T_1 < T_2, T_1 \leq \tau \,|\, X_1 = x_1, X_2 =$

$x_1) > P_0(T_2 < T_1, T_2 \leq \tau \,|\, X_1 = x_1, X_2 = x_1))$. For any other f, we note that

$$
\begin{aligned}
&\mathbb{E}_0^*\big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau) \\
&\qquad + \mathbb{1}(f_0(X_1) > f_0(X_2)) \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\big] \\
&- \mathbb{E}_0^*\big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau) \\
&\qquad + \mathbb{1}(f(X_1) > f(X_2)) \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\big] \\
&= \mathbb{E}_0^*\Big[\big\{\mathbb{1}(f_0(X_1) > f_0(X_2)) \\
&\qquad - \mathbb{1}(f(X_1) > f(X_2))\big\} \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\Big] \\
&\geq 0,
\end{aligned}
$$

by the definition of f_0. Then, by the tower property

$$
\begin{aligned}
&\mathbb{V}(f_0, P_0^*) - \mathbb{V}(f, P_0^*) \\
&= \mathbb{E}_0^*\Big[\mathbb{E}_0^*\big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau) \\
&\qquad + \mathbb{1}(f_0(X_1) > f_0(X_2)) \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\big]\Big] \\
&- \mathbb{E}_0^*\Big[\mathbb{E}_0^*\big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau) \\
&\qquad + \mathbb{1}(f(X_1) > f(X_2)) \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\big]\Big] \\
&= \mathbb{E}_0^*[\mathbb{E}_0^*\big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau) \\
&\qquad + \mathbb{1}(f_0(X_1) > f_0(X_2)) \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\big] \\
&- \mathbb{E}_0^*\big[\mathbb{1}(T_2 < T_1, T_2 \leq \tau) \\
&\qquad + \mathbb{1}(f(X_1) > f(X_2)) \left\{\mathbb{1}(T_1 < T_2, T_1 \leq \tau) - \mathbb{1}(T_2 < T_1, T_2 \leq \tau)\right\} \,|\, X_1, X_2\big]] \\
&\geq 0.
\end{aligned}
$$

Example 3: Brier score. The Brier score at time τ is simply the negative MSE for the binary outcome $\mathbb{1}(T > \tau)$, which is maximized by the conditional mean $x \mapsto \mathbb{E}_0[\mathbb{1}(T > \tau) \,|\, X = x] = S_0(\tau \,|\, x)$.

A.2.3 Explicit form of efficient influence function for examples

Example 1: AUC. We have

$$
\begin{aligned}
&\Phi_{0,1}(x,t) \\
&= \frac{1}{2}\int \left\{\mathbb{1}(f_0(x) > f_0(x_2), t \le \tau, t_2 > \tau) + \mathbb{1}(f_0(x_2) > f_0(x), t_2 \le \tau, t > \tau)\right\} \mathbb{F}_0(dx_2, dt_2) \\
&= \frac{1}{2}\int \Big\{\mathbb{1}(f_0(x) > f_0(x_2), t \le \tau)\left\{1 - F_0(\tau \,|\, x_2)\right\} \\
&\qquad + \mathbb{1}(f_0(x_2) > f_0(x), t > \tau) F_0(\tau \,|\, x_2)\Big\} Q_0(dx_2) \\
&= \frac{1}{2}\mathbb{E}_0\left[\mathbb{1}(f_0(x) > f_0(X), t \le \tau)\left\{1 - F_0(\tau \,|\, X)\right\} + \mathbb{1}(f_0(X) > f_0(x), t > \tau) F_0(\tau \,|\, X)\right]; \\
&\Theta_{0,1}(t) \\
&= \frac{1}{2}\int \left[\mathbb{1}(t \le \tau, t_2 > \tau) + \mathbb{1}(t_2 \le \tau, t > \tau)\right] \mathbb{F}_0(dx_2, dt_2) \\
&= \frac{1}{2}\int \left\{\mathbb{1}(t \le \tau)\left\{1 - F_0(\tau \,|\, x_2)\right\} + \mathbb{1}(t > \tau) F_0(\tau \,|\, x_2)\right\} Q_0(dx_2) \\
&= \frac{1}{2}\mathbb{E}_0\left[\mathbb{1}(t \le \tau)\left\{1 - F_0(\tau \,|\, X)\right\} + F_0(\tau \,|\, X)\mathbb{1}(t > \tau)\right].
\end{aligned}
$$

Noting that $\int_0^\tau F_0(dt \,|\, x) = F_0(\tau \,|\, x)$, $\int_\tau^\infty F_0(dt \,|\, x) = 1 - F_0(\tau \,|\, x)$, $\int_0^\tau \varphi_{0,z}(dt) = \varphi_{0,z}(\tau)$, and $\int_\tau^\infty \varphi_{0,z}(dt) = -\varphi_{0,z}(\tau)$, we have

$$
\begin{aligned}
D_\Phi(z) = \mathbb{E}_0\Big[&\mathbb{1}(f_0(x) > f_0(X))\left\{F_0(\tau \,|\, x) - \varphi_{0,z}(\tau)\right\}\left\{1 - F_0(\tau \,|\, X)\right\} \\
&+ \mathbb{1}(f_0(X) > f_0(x)) F_0(\tau \,|\, X)\left\{1 - F_0(\tau \,|\, x) + \varphi_{0,z}(\tau)\right\}\Big] - 2V_1(f_0, \mathbb{F}_0); \\
D_\Theta(x) = \mathbb{E}_0\Big[&\left\{F_0(\tau \,|\, x) - \varphi_{0,z}(\tau)\right\}\left\{1 - F_0(\tau \,|\, X)\right\} \\
&+ F_0(\tau \,|\, X)\left\{1 - F_0(\tau \,|\, x) + \varphi_{0,z}(\tau)\right\}\Big] - 2V_2(\mathbb{F}_0).
\end{aligned}
$$

Example 2: C-index. We have

$$\Phi_{0,1}(x,t) = \frac{1}{2}\int \Big\{ \mathbb{1}(f_0(x) > f_0(x_2), t \le t_2, t \le \tau) + \mathbb{1}(f_0(x_2) > f_0(x), t_2 \le t, t_2 \le \tau) \Big\} \mathbb{F}_0(dx_2, dt_2)$$
$$= \frac{1}{2}\mathbb{E}_0\left[\int \left[\mathbb{1}(f_0(x) > f_0(X), t \le t_2, t \le \tau) + \mathbb{1}(f_0(X) > f_0(x), t_2 \le t, t_2 \le \tau)\right] F_0(dt_2 \,|\, X)\right];$$
$$\Theta_{0,1}(t) = \frac{1}{2}\int \left[\mathbb{1}(t \le t_2, t \le \tau) + \mathbb{1}(t_2 \le t, t_2 \le \tau)\right] \mathbb{F}_0(dx_2, dt_2)$$
$$= \frac{1}{2}\mathbb{E}_0\left[\int \left[\mathbb{1}(t \le t_2, t \le \tau) + \mathbb{1}(t_2 \le t, t_2 \le \tau)\right] F_0(dt_2 \,|\, X)\right].$$

Therefore

$$D_\Phi(z) = \mathbb{E}_0\Bigg[\int \Big\{\mathbb{1}(f_0(x) > f_0(X), t \le t_2, t \le \tau) + \mathbb{1}(f_0(X) > f_0(x), t_2 \le t, t_2 \le \tau)\Big\} \left\{F_0(dt \,|\, x) - \varphi_{0,z}(dt)\right\} F_0(dt_2 \,|\, x_2)\Bigg] - 2V_1(f_0, \mathbb{F}_0);$$
$$D_\Theta(z) = \mathbb{E}_0\Bigg[\int \Big\{\mathbb{1}(t \le t_2, t \le \tau) + \mathbb{1}(t_2 \le t_1, t_2 \le \tau)\Big\} \left\{F_0(dt \,|\, x) - \varphi_{0,z}(dt)\right\} F_0(dt_2 \,|\, X)\Bigg] - 2V_2(\mathbb{F}_0).$$

Example 3: Brier score. Because $m = 1$ and $\Theta(t) = 1$, we have that $\Phi_{0,1}(x,t) = \Phi(f_0(x), t)$ and $\Theta_{0,1}(t) = 1$. Therefore

$$D_\Phi(z) = \int \left\{2f_0(x)\mathbb{1}(t > \tau) - f_0(x)^2 - \mathbb{1}(t > \tau)\right\} \left\{F_0(dt \,|\, x) - \varphi_{0,z}(dt)\right\}$$
$$= 2f_0(x)\left\{1 - F_0(\tau \,|\, x) + \varphi_{0,z}(\tau)\right\} - f_0(x)^2 - 1 + F_0(\tau \,|\, x) - \varphi_{0,z}(\tau) - V_1(f_0, \mathbb{F}_0).$$

A.2.4 Verification of conditions (B6) and (B7) for examples

Example 1: AUC. Williamson et al. (2021b) considered AUC for binary outcomes, and showed that condition (B7) holds for $C = \frac{2\kappa}{\pi_0(1-\pi_0)}$, where $\pi_0 = P_0^*(T \le \tau)$, and $\|\cdot\|_{\mathcal{F}}$ the supremum norm, under the margin condition

$$P_0\left(|F_0(\tau \,|\, X_1) - F_0(\tau \,|\, X_2)| < s\right) \le \kappa s,$$

for some $0 < \kappa < \infty$ and all s small.

For any fixed t_1, we can write

$$\|\Theta(t_1, \cdot)\|_v^* = \|\Theta(t_1, \cdot)\|_v = \int |\Theta(t_1, dt_2)| = \mathbb{1}(t_1 > \tau),$$

and so $\sup_{t_1} \|\Theta(t_1, \cdot)\|_v^* = 1$.

Next, for any fixed (x_1, t_1), we can write Φ as the product of functions $\mathbb{1}(f_0(x_1) > f_0(x_2))$ and $\mathbb{1}(t_2 \leq \tau)\mathbb{1}(t_1 > \tau)$. The variation norm of each of these functions is 1 for fixed (x_1, t_1), and so $\|\Phi((f_0(x_1), t_1), \cdot)\|_v^* < \infty$.

We note that for any bivariate function $g(u, v)$ that can be written in the form $g_1(u)g_2(v)$, we have that

$$\|g\|_v^* = \max\left\{\sup_{u,v} |g(u, v)|, \sup_u \int |g(u, dv)|, \sup_v \int |g(du, v)|, \int |g(du, dv)|\right\}$$

The first term is bounded above by $\sup_u |g_1(u)| \sup_v |g_2(v)|$. The second is bounded above by $\sup_u |g_1(u)| \int |g_2(dv)|$. The third is bounded above by $\sup_v |g_2(v)| \int |g_1(du)|$. The fourth is bounded above by $\int |g_1(du)| \int |g_2(du)|$. Therefore, if the variation norms of each of g_1 and g_2 are bounded, the uniform sectional variation norm of g is bounded.

Example 2: C-index. For any fixed t_1, we can write

$$\|\Theta(t_1, \cdot)\|_v^* = \|\Theta(t_1, \cdot)\|_v = \int |\Theta(t_1, dt_2)| = \int d\mathbb{1}(t_2 \leq (t_1 \wedge \tau)) = 1,$$

and so $\sup_{t_1} \|\Theta(t_1, \cdot)\|_v^* = 1$.

Next, for any fixed (x_1, t_1), we can write Φ as the product of functions $\mathbb{1}(f_0(x_1) > f_0(x_2))$ and $\mathbb{1}(t_2 \leq t_1, t_2 \leq \tau)$. The variation norm of each of these functions is 1 for fixed (x_1, t_1), and so $\|\Phi((f_0(x_1), t_1), \cdot)\|_v^* < \infty$.

Example 3: Brier score. We have that $|V(f, P_0) - V(f_0, P_0)| = \{f(x) - f_0(x)\}^2$ as long as $F_0(\tau \,|\, x)$ falls in $\mathcal{F}$. Therefore, condition (B7) holds with $C = 1$ and $\|\cdot\|_{\mathcal{F}}$ taken to be the $L_2(P_0)$ or supremum norm.

We have that $\Phi(f_0(x), t) = f_0(x)^2 - 2f_0(x)\mathbb{1}(t > \tau) + \mathbb{1}(t > \tau)$. By the triangle inequality,

$$\|\Phi\|_v^* \leq \|f_0\|_v^* - 2\|f_0\|_v^* + 1.$$

This quantity is finite if $\|f\|_v^*$ is finite, which in this example requires that $\|F_0(\tau \,|\, \cdot)\|_v^*$ be finite. This would be satisfied, for example, if $F_0(\tau \,|\, \cdot)$ were differentiable with uniformly bounded derivatives.

A.3 Additional predictiveness measures

Example 4: MSE for predicted survival time. In some situations, there is interest in using features to predict a participant's actual event time, or a transformation thereof. Let $g : (0, \infty) \to \mathbb{R}$ denote a transformation of T. As in Example 3, predictiveness can be measured as the expectation of a loss function. Letting $L : \mathcal{F} \times \mathbb{R} \to [0, \infty)$ denote the loss function, we have the predictiveness measure $-\mathbb{E}_0^* \left[L(f(X), g(T)) \right]$.

A popular choice for g is the map $g : t \mapsto t \wedge \tau$ for a user-specified constant τ. Using this g amounts to quantifying the loss for prediction of the restricted mean survival time (RMST) given X. RMST has become a popular summary measure for the distribution of a time-to-event variable due to the fact that the mean itself is generally not identifiable (Tian et al., 2014). Using squared-error loss, the predictiveness measure for predicting the restricted survival time is given by

$$\mathbb{V}(f, P_0^*) := -\mathbb{E}_0^* \left[\{ f(X) - (T \wedge \tau) \}^2 \right].$$

Like the c-index, the MSE for predicting the restricted survival time may be considered a global performance metric.

For this example, $m = 1$, $\Theta = 1$, and $\Phi(x, t) = -\{ f(x) - t \wedge \tau \}^2$.

Oracle prediction function: As in Example 3, the negative MSE is maximized by the conditional mean, which is this case is $x \mapsto \mathbb{E}_0[T \wedge \tau \,|\, X = x]$. This can be written in terms of F_0 as $f_0 : x \mapsto \int_0^\tau \{ 1 - F_0(t \,|\, x) \} \, dt$.

Efficient influence function: As in Example 3, we have that $m = 1$ and $\Theta(t) = 1$, and so

$$D_\Phi(z) = -\int \{ f_0(x) - (t \wedge \tau) \}^2 \{ F_0(dt \,|\, x) - \varphi_{0,z}(dt) \} - V_1(f_0, \mathbb{F}_0).$$

Verification of conditions: The MSE for predicting restricted survival time satisfies (A3c) with $\mathcal{T}_1 = [0, \tau)$ and $\mathcal{T}_2 = [\tau, \infty)$.

We have that $|V(f, P_0) - V(f_0, P_0)| = \{f(x) - f_0(x)\}^2$ as long as $E_0[T \wedge \tau \,|\, X = x]$ falls in $\mathcal{F}$. Therefore, condition (B7) holds with $C = 1$ and $\|\cdot\|_{\mathcal{F}}$ taken to be the $L_2(P_0)$ or supremum norm.

We have that $\Phi(f_0(x), t) = f_0(x)^2 - 2f_0(x)(t \wedge \tau) + (t \wedge \tau)$. Note that $\int_0^\tau |dt| = \tau$, and so by the triangle inequality,

$$\|\Phi\|_v^* \leq \|f_0\|_v^* - 2\tau\|f_0\|_v^* + \tau.$$

This quantity is finite if $\|f_0\|_v^*$ is finite. We recall that $f_0(x) = \int_0^\tau S_0(t \,|\, x)dt = E_0[T \wedge \tau \,|\, X = x]$. This regression function will have finite uniform sectional variation norm if, for example, it is differentiable with respect to x with uniformly bounded derivatives.

A.4 Details on gradient-boosted c-index

Existing approaches to c-index maximization have used either the original Harrell concordance estimator (Harrell et al., 1982) or an IPCW version that assumes marginal independence between C and X (Mayr and Schmid, 2014).

Due to the indicator function $\mathbb{1}(f(X_1) > f(X_2))$, the c-index is not amenable to gradient-based optimization techniques, and so optimization-based methods typically use a smoothed, differentiable approximation or lower bound. As suggested by the form of $\mathcal{F}_0$, Raykar et al. (2008) cast maximization of the c-index as a ranking problem and propose a differentiable lower bound on the c-index, which is then optimized using a conjugate gradients algorithm. Chen et al. (2012) propose a smoothed c-index approximation based on the sigmoid function and optimize it via gradient boosted machines. Mayr and Schmid (2014) use a similar smoothed objective function and likewise use gradient boosting with linear models as base learners. Specifically, Mayr and Schmid (2014) consider the smoothed c-index

$$\mathbb{E}_0^* \left[\mathbb{1}(T_1 < T_2) h_\omega(f(X_2) - f(X_1))\right],$$

where $h_\omega : s \mapsto \{1 + \exp(z/\omega)\}$ is the sigmoid function and ω is a tuning parameter determining the smoothness of the approximation. Smaller values of ω lead to a closer approximation of the indicator function, but less stable optimization due to potentially large values

of the gradient. For direct optimization, we consider the identified version of the smoothed restricted c-index given by

$$\frac{\mathbb{E}_0\left[h_\omega(f(X_2)-f(X_1))\iint \mathbb{1}(t_1<t_2,t_1\leq\tau)F_0(dt_2\,|\,X_2)F_0(dt_1\,|\,X_1)\right]}{\mathbb{E}_0\left[\iint \mathbb{1}(t_1<t_2,t_1\leq\tau)F_0(dt_2\,|\,X_2)F_0(dt_1\,|\,X_1)\right]}.$$

Details on the proposed optimization procedure, including selection of the smoothing parameter ω, are given in Appendix A.4. Different versions of the population risk: Our aim is to maximize the smoothed population objective function

$$\mathbb{E}_0\left[h_\omega(f(X_2)-f(X_1))\mathbb{1}(T_1<T_2,T_1\leq\tau)\right].$$

Mayr and Schmid (2014) identify this population objective function using IPCW, under the assumption that C and X are independent, using

$$\frac{\mathbb{E}_0\left[h_\omega(f(X_2)-f(X_1))\frac{\Delta_1\mathbb{1}(Y_1<Y_2,Y_1\leq\tau)}{\tilde{G}_0(Y_1)^2}\right]}{\mathbb{E}_0\left[\frac{\Delta_1\mathbb{1}(Y_1<Y_2,Y_1\leq\tau)}{\tilde{G}_0(Y_1)^2}\right]},$$

where $\tilde{G}_0(\cdot)$ is the marginal survival function of C. Rather than using this version, we leverage our identification of the c-index using the conditional distribution function of T given X, which yields the smoothed objective function

$$\frac{\mathbb{E}_0\left[h_\omega(f(X_2)-f(X_1))\iint \mathbb{1}(t_1<t_2,t_1\leq\tau)F_0(dt_1\,|\,X_1)F_0(dt_2\,|\,X_2)\right]}{\mathbb{E}_0\left[\iint \mathbb{1}(t_1<t_2,t_1\leq\tau)F_0(dt_1\,|\,X_1)F_0(dt_2\,|\,X_2)\right]}.$$

The empirical objective function is then given by

$$\sum_{i=1}^{n}\sum_{j=1}^{n} h_\omega(f(X_j)-f(X_i))w_{i,j}, \tag{A.10}$$

where $w_{i,j} := \frac{\iint \mathbb{1}(t_i<t_j,t_i\leq\tau)F_n(dt_i\,|\,X_i)F_n(dt_j\,|\,X_j)}{\sum_{k=1}^{n}\sum_{l=1}^{n}\iint \mathbb{1}(t_l<t_m,t_l\leq\tau)F_n(dt_l\,|\,X_l)F_n(dt_m\,|\,X_m)}$ is a weight. The gradient of this objective function with respect to $f(X_j)$ is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\frac{\exp\left(\frac{f(X_j)-f(X_i)}{\omega}\right)}{\omega\left\{1+\exp\left(\frac{f(X_j)-f(X_i)}{\omega}\right)\right\}^2}. \tag{A.11}$$

In the numerical experiments in Section 2.5 and Appendix A.5, we use the empirical objective (A.10) and gradient (A.11) to define a custom family in the `mboost` gradient boosting `R` package. The boosting tuning parameters `mtry` (number of boosting iterations) and `nu` (learning rate), as well as the smoothing parameter ω, were selected by five-fold cross-validation. The objective function used for cross-validation was the unsmoothed plug-in estimate of the c-index $\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}(f(X_j)-f(X_i))w_{i,j}$.

A.5 Simulation details and additional results

A.5.1 Details on data-generating mechanism and true VIMs

In all experiments we generate the data as follows:

1. Draw $X \sim MVN(\mathbf{0}, \Sigma)$.
2. Draw $\varepsilon_T \sim N(0,1)$ and $\varepsilon_C \sim N(0,1)$, independent of each other and independent of X.
3. Set $\log(T) = X\beta_T + \varepsilon_T$.
4. Set $\log(C) = \beta_{0,C} + X\beta_C + \varepsilon_C$.
5. Set $Y := \min\{T, C\}$ and $\Delta = \mathbb{1}(T \leq C)$.

The values of p, Σ, β_T, and β_C depend on the scenario. Under this data-generating mechanism, the conditionally independent censoring assumption (A1) holds. The distributions of both T and C fall in the class of accelerated failure time models.

In Scenario 1, we set $p = 2$, $\beta_T = (0.5, -0.3)$, $\beta_C = (-0.2, 0.2)$, $\beta_{0,C} = 0$, and $\Sigma = \boldsymbol{I}_2$. In Scenario 2, we set $p = 25$, $\beta_T = (0.5, -0.3, \mathbf{0}_{23})$, $\beta_C = (-0.2, 0.2, \mathbf{0}_{23})$, $\beta_{0,C} = 0$ and $\Sigma = \boldsymbol{I}_{25}$, where $\mathbf{0}_m$ denotes a 0-vector of length m. In Scenario 3, we set $p = 2$, $\beta_T = (0.5, -0.3)$, $\beta_C = (-0.2, 0.2)$, and $\Sigma = \boldsymbol{I}_2$. To achieve censoring rates of $\{30\%, 40\%, 50\%, 60\%, 70\%\}$,

we set $\beta_{0,C}$ equal to $\{0.75, 0.5, 0, -0.5, -0.75\}$, respectively. In Scenario 4, we set $p = 5$, $\beta_T = (0.5, -0.3, 0, 0, 0)$, $\beta_C = (-0.2, 0.2, 0, 0, 0)$, and

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & \rho_{14} & 0 \\ 0 & 1 & \rho_{23} & 0 & 0 \\ 0 & \rho_{23} & 1 & 0 & 0 \\ \rho_{14} & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where $\rho_{14} = 0.7$ and $\rho_{23} = -0.3$.

The true survival function $S(t \,|\, x)$ is given by

$$\begin{aligned} P(T > t \,|\, X = x) &= P(\log(T) > \log(t) \,|\, X = x) = P(X\beta_T + \varepsilon_T > \log(t) \,|\, X = x) \\ &= P(\varepsilon_T > \log(t) - x\beta_T) = 1 - \Omega(\log(t) - x\beta_T). \end{aligned}$$

where Ω is the standard normal distribution function. The reduced-dimension conditional survival function omitting X_s is given by

$$\begin{aligned} P(T > t \,|\, X_{-s} = x_{-s}) &= P(\log(T) > \log(t) \,|\, X_{-s} = x_{-s}) \\ &= P(X\beta_T + \varepsilon_T > \log(t) \,|\, X_{-s} = x_{-s}) \\ &= P(X_s\beta_{s,T} + \varepsilon_T > \log(t) - x_{-s}\beta_{-s,T} \,|\, X_{-s} = x_{-s}). \end{aligned}$$

We note that $X_s\beta_{T,s} + \varepsilon_T \,|\, X_{-s}$ is a Normal random variable with mean 0 and variance $\sum_{j \in s} \beta_{j,T}^2 + 1$. These calculations hold for Scenarios 1 – 3, where the features are uncorrelated.

When the features are correlated, we can carry out similar calculations using the conditional distributions derived from a multivariate normal distribution. We note that $\varepsilon_T \,|\, X_{-s} \sim N(0, 1)$, while

$$\begin{aligned} X_1 \,|\, X_2, \ldots, X_5 &\sim N(\rho_{14}X_4, (1 - \rho_{14}^2)), \\ X_4 \,|\, X_1, X_2, X_3, X_5 &\sim N(\rho_{14}X_1, (1 - \rho_{14}^2)), \\ aX_1 + bX_4 \,|\, X_2, X_3, X_5 &\sim N(0, a^2 + b^2 + 2ab\rho_{14}), \end{aligned}$$

and therefore

$$X_1\beta_{1,T} \,|\, X_2, \ldots, X_5 \sim N(\beta_{1,T}\rho_{14}X_4, \beta_{1,T}^2(1-\rho_{14}^2),$$
$$X_4\beta_{4,T} \,|\, X_1, X_2, X_3, X_5 \sim N(\beta_{4,T}\rho_{14}X_4, \beta_{T,T}^2(1-\rho_{14}^2)$$
$$X_1\beta 1, T + X_4\beta_{4,T} \,|\, X_2, X_3, X_5 \sim N(0, \beta_{1,T}^2 + \beta_{4,T}^2 + 2\beta_{1,T}\beta_{4,T}\rho_{14}).$$

Hence, the desired conditional distributions are given by

$$X_1\beta_{1,T} + \varepsilon_T \,|\, X_2, \ldots, X_5 \sim N(\beta_{1,T}\rho_{14}X_4 + \beta_{2,T}X_2, \beta_{1,T}^2(1-\rho_{14}^2) + 1),$$
$$X_4\beta_{4,T} + \varepsilon_T \,|\, X_1, X_2, X_3, X_5 \sim N(\beta_{4,T}\rho_{14}X_1 + \beta_{2,T}X_2, \beta_{4,T}^2(1-\rho_{14}^2) + 1),$$
$$X_1\beta_{1,T} + X_4\beta_{4,T} + \varepsilon_T \,|\, X_2, X_3, X_5 \sim N(0, \beta_{1,T}^2 + \beta_{4,T}^2 + 2\beta_{1,T}\beta_{4,T}\rho_{14} + 1).$$

For landmark time VIMs, the true oracle and residual oracle prediction functions are characterized by the conditional survival function $S(t \,|\, x)$. For the c-index, because the distribution of $T \,|\, X$ is normal, a valid oracle prediction function is given by the conditional mean $f_0(x) = \mathbb{E}_0[T \,|\, X = x]$. This is due to the fact that, for two independent normally distributed random variables T_1 and T_2 with respective means μ_1 and μ_2 and equal variances, T_1 stochastically dominates T_2 if $\mu_1 > \mu_2$. This implies that $P(T_1 > T_2) > 1/2$, and hence the mean falls in $\mathcal{F}_0$, the class of oracle prediction functions.

The true VIM values for all simulations are given in Table A.1.

A.5.2 Details on nuisance parameter and oracle prediction function estimation

Table A.2 describes the algorithms used to estimate S_0 and G_0. Tuning parameters for the random survival forest (RSF) were selected to minimize out-of-bag error rate, as measured by one minus Harrell's c-index (the default evaluation metric in the `rfsrc` software package).Table A.3 gives the algorithms included in the Super Learner library for global survival stacking and estimation of the residual oracle prediction function $f_{0,s}$ for landmark VIMs. Five-fold cross-validation was used to determine the optimal convex combination of these learners that minimized cross-validated squared-error loss, as described in (Wolock et al.,

Scenario	Feature	Importance measure				
		AUC at τ		Brier score at τ		c-index
‘		$\boldsymbol{\tau = 0.5}$	**0.9**	**0.5**	**0.9**	
1	X_1	0.118	0.115	0.022	0.030	0.096
	X_2	0.035	0.034	0.008	0.011	0.029
2	X_1	0.118	0.115	0.022	0.030	0.096
	X_4	0	0	0	0	0
3	X_1	0.118	0.115	0.022	0.030	0.096
4	X_1	0.052	0.050	0.012	0.015	0.043
	X_4	0	0	0	0	0
	(X_1, X_4)	0.118	0.115	0.022	0.030	0.096

Table A.1: Approximate values of $\psi_{0,s}$ for numerical experiments. These parameter values were approximated using a Monte Carlo approach with sample size 10^7.

2022). Table A.4 gives the algorithms included in the survival Super Learner library. Five-fold cross-validation was used to determine the optimal convex combination of these learners that minimized cross-validated oracle risk functions detailed in (Westling et al., 2023). The RSF algorithm was fit twice, once to estimate F_0 and once to estimate G_0. For global survival stacking and survival Super Learner, estimates for both distributions are produced simultaneously.

For landmark VIMs, the full oracle prediction function f_0 is a simple transformation of S_0 and was not estimated separately in our experiments. For the c-index, we implemented the boosting procedure details in Appendix A.4 with five-fold cross-validation for tuning parameter selection. The unsmoothed c-index was used as the evaluation metric for cross-validation. The tuning parameters are detailed in Table A.5.

A.5.3 Additional simulation results in Scenarios 1 and 2

In this section, we provide additional results for Scenarios 1 and 2.

In Scenario 1, we set $p = 2$, and both X_1 and X_2 have non-zero importance. We generated 500 random datasets of size $n \in \{500, 750, \ldots, 1500\}$ and used Algorithm 2.2, which is valid

Algorithm	`R` implementation	Tuning parameters
Random survival forest	`rfsrc` (Ishwaran and Kogalur, 2022)	`mtry` $\in \{1, \ldots, \sqrt{p}^{\dagger}\}$ `nodesize` $\in \{5, 15, 25\}$ `ntree` $\in \{500, 1000\}$
Global survival stacking	`stackG` (Wolock et al., 2022)	`SL.library` (see Table A.3) `bin_size = 0.04`
Survival Super Learner	`survSuperLearner` (Westling et al., 2023)	`SL.library` (see Table A.4)

Table A.2: Algorithms used for estimation of nuisance parameters. All options besides those listed here were set to default values. In particular, the random survival forests were grown using sampling without replacement and the log-rank splitting rule. The combination of `mtry`, `nodesize`, and `ntree` minimizing out-of-bag error rate, as measured by one minus Harrell's c-index, was selected. For global survival stacking, `time_basis` was set to `"continuous"` (time included as continuous predictor in the pooled binary regression), and `surv_form` was set to `"PI"` (product-integral mapping from hazard to survival function). For both global stacking and survival Super Learner, five-fold cross-validation was used to determine the optimal convex combination of algorithms in `SL.library`.
†: p denotes the number of predictors.

Algorithm name	Algorithm description	Tuning parameters
`SL.mean`	Marginal mean	NA
`SL.glm`	Logistic regression with all pairwise interactions	NA
`SL.gam`	Generalized additive model	default
`SL.earth`	Multivariate adaptive regression splines	default
`SL.ranger`	Random forest	default
`SL.xgboost`	Gradient-boosted trees	`ntrees` $\in \{250, 500, 1000\}$ `max_depth` $\in \{1, 2\}$

Table A.3: Algorithms included in the Super Learner for global survival stacking and for estimation of the residual oracle prediction function for landmark VIMs. All tuning parameters besides those for `SL.xgboost` were set to default values. In particular, `gam` was implemented with `degree = 2`; `earth` with `degree = 2, penalty = 3, nk` = number of predictors plus 1, `endspan = 0`, `minspan = 0`; and `ranger` with `num.trees = 500, mtry` = the square root of the number of predictors, `min.node.size = 1, sample.fraction = 1` with replacement. For `SL.xgboost`, `shrinkage` was set to 0.01, `minobspernode` was set to 10, and each combination of `ntrees` and `max_depth` was included in the Super Learner library.

when the importance is *a priori* known to be non-zero. Here, we show the results for both features using the Brier score predictiveness measure, as well as for X_2 using AUC and c-index. We assess performance in the same manner as described in the main text.

Figure A.1 displays the AUC and c-index results for X_2. Figures A.2 and A.3 display the results for Brier score VIM for X_1 and X_2, respectively. The results closely match those observed for AUC and c-index for X_1 in Section 2.5 of the main text. In particular, we observe that for AUC and Brier score, cross-fitting is necessary for good performance. The cross-fitted global stacking and survival Super Learner estimators achieves low bias and coverage within Monte Carlo error of the nominal level, while the RSF implementation has larger bias and is somewhat anti-conservative for AUC and Brier score VIMs.

We also present additional results for Scenario 2, in which $p = 25$ and all features besides X_1 and X_2 have zero importance. We generated 500 random datasets of size $n \in \{500, 750, \ldots, 1500\}$ and used Algorithm 2.3, which involves sample splitting and pro-

Algorithm name	Algorithm description
`survSL.km`	Kaplan-Meier estimator
`survSL.expreg`	Survival regression assuming event and censoring times follow an exponential distribution conditional on covariates
`survSL.weibreg`	Survival regression assuming event and censoring times follow a Weibull distribution conditional on covariates
`survSL.loglogreg`	Survival regression assuming event and censoring times follow a log-logistic distribution conditional on covariates
`survSL.AFTreg`	Survival regression assuming event and censoring times follow a log-normal distribution conditional on covariates
`survSL.coxph`	Main-terms Cox proportional hazards estimator with Breslow baseline cumulative hazard
`survSL.rfsrc`	Random survival forest as implemented in the `randomForestSRC` package

Table A.4: Algorithms included in the survival Super Learner. All tuning parameters were set to default values. In particular, `gam` was implemented with `degree = 1`; and `rfsrc` with `ntree = 500`, `mtry =` the square root of the number of predictors, `nodesize = 15`, `splitrule = "logrank"`, `sampsize = 1` with replacement.

Parameter	Description	Possible values
`mstop`	Number of boosting iterations	$\{100, 250, 500, 1000\}$
`nu`	Learning rate	0.01
`sigma`	Smoothing parameter for sigmoid function	$\{0.01, 0.05\}$

Table A.5: Tuning parameters for the c-index boosting procedure.

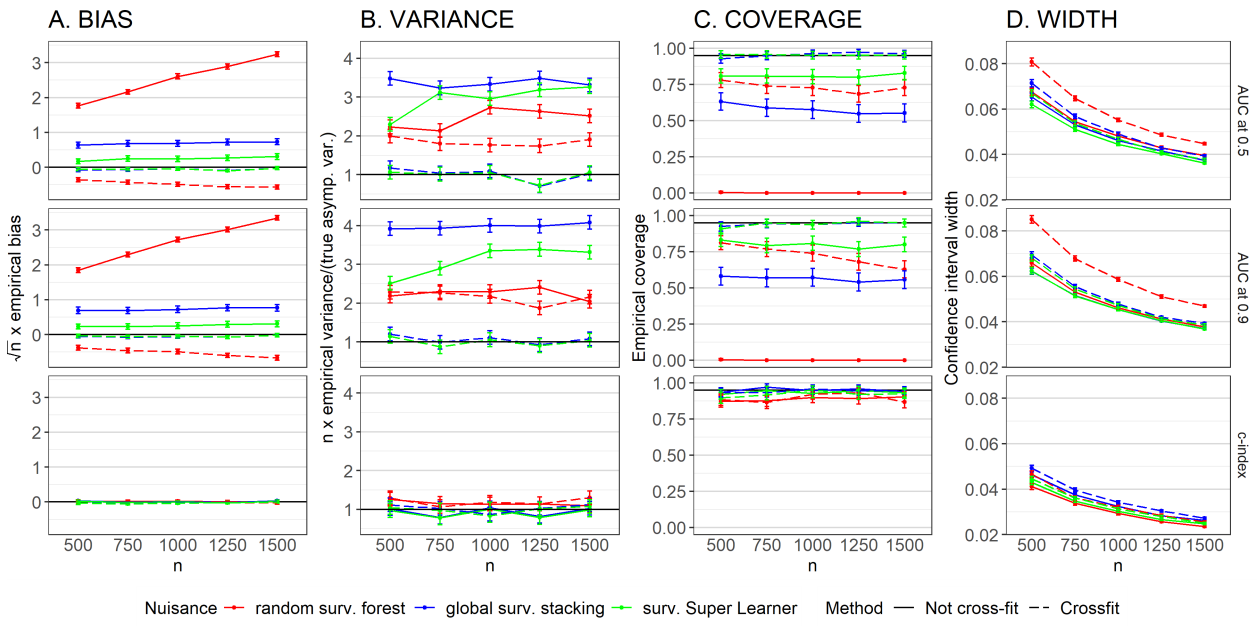

Figure A.1: Performance of the one-step VIM estimator for the importance of X_2 in Scenario 1 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

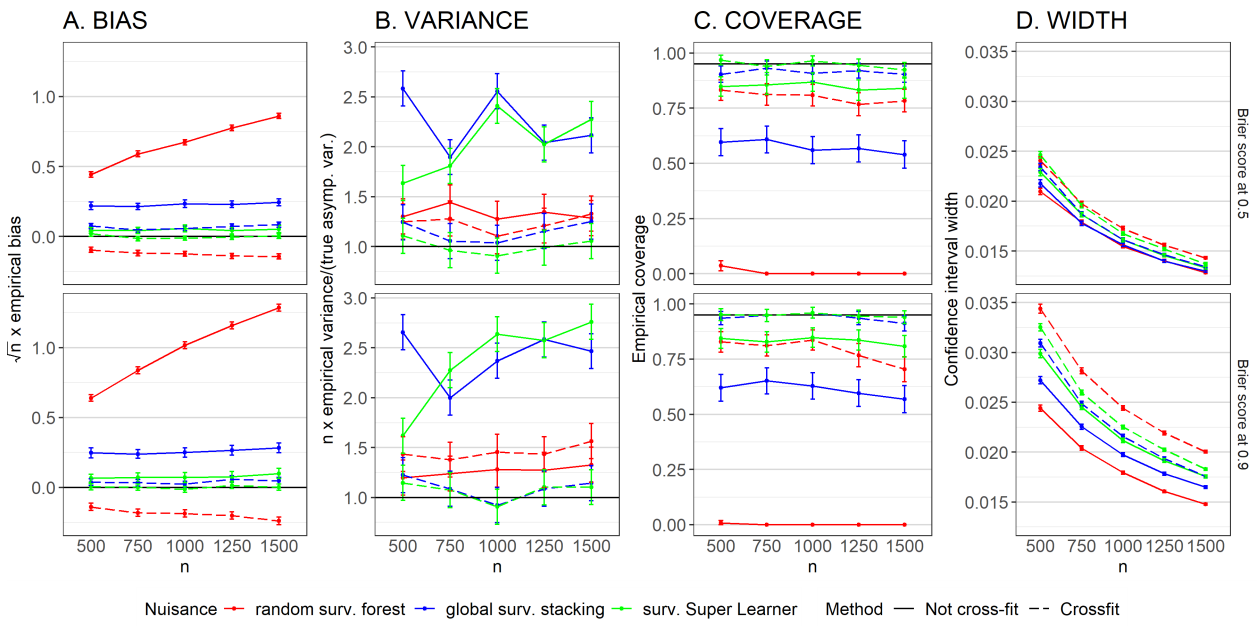

Figure A.2: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 1 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

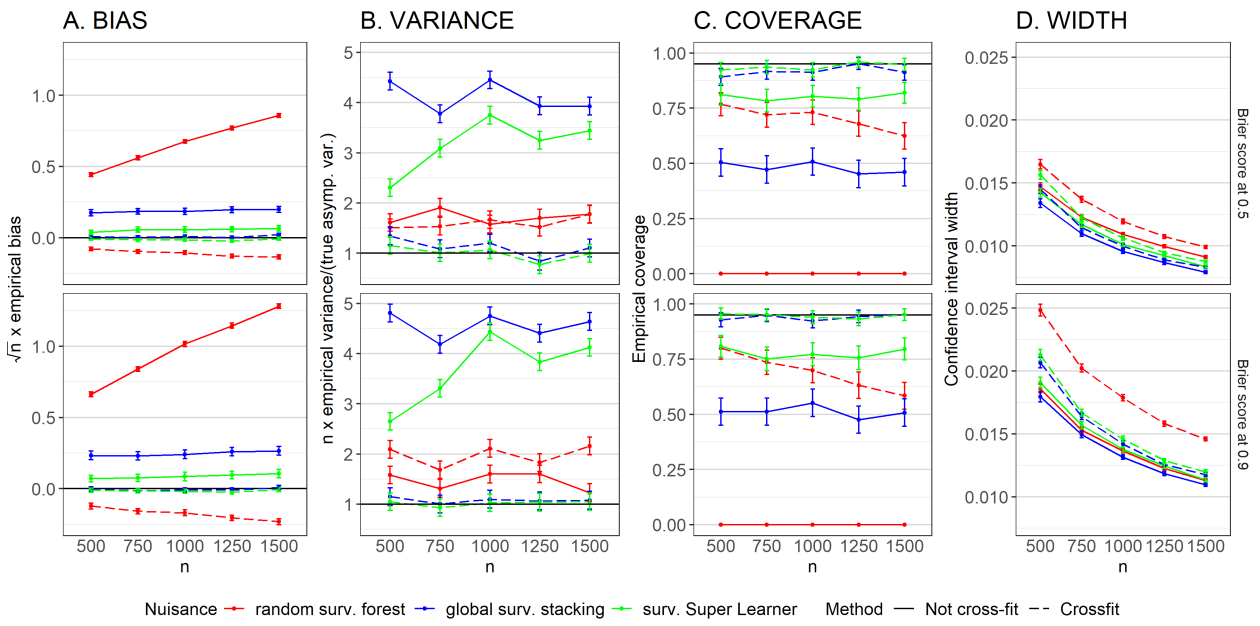

Figure A.3: Performance of the one-step VIM estimator for the importance of X_2 in Scenario 1 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

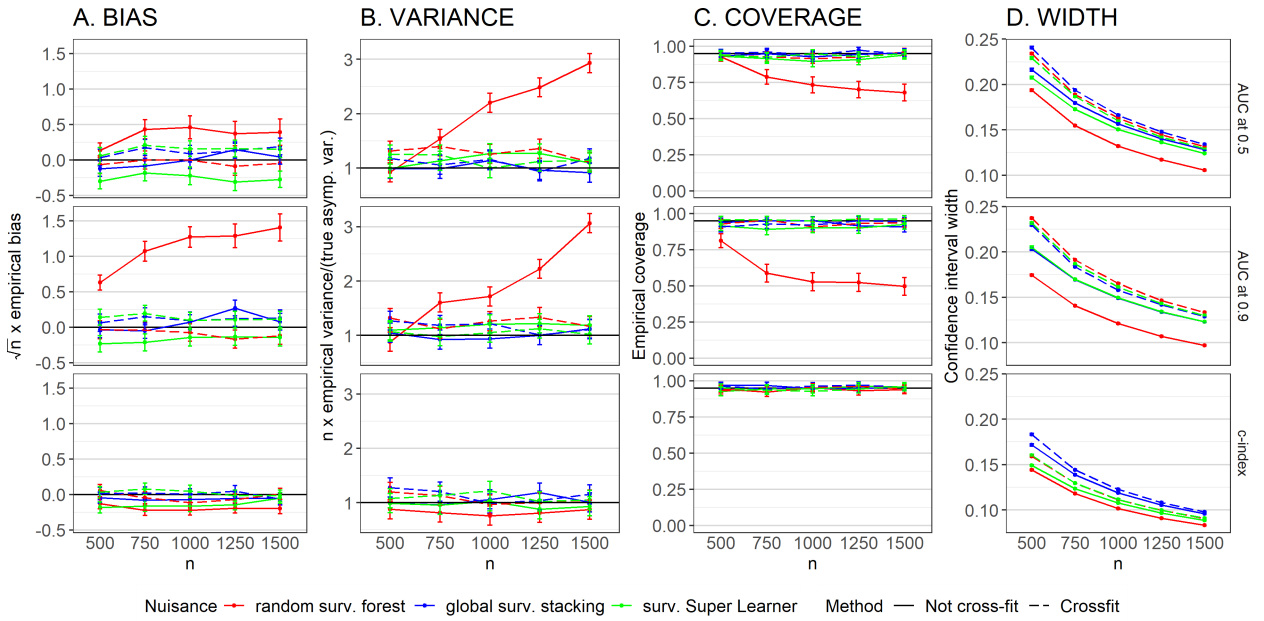

Figure A.4: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 2 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

vides valid inference under the null hypothesis of zero importance. Figure A.4 displays the AUC and c-index results for X_1, and Figures A.5 and A.6 display the results for Brier score VIM for X_1 and X_4, respectively. For X_4, rather than confidence interval width, we examine the empirical type I error rate. Similarly as in the main text, the non-cross-fitted estimators demonstrate increased bias, decreased confidence interval coverage, and increased type I error compared to their cross-fitted counterparts. The cross-fitted estimator using RSF performs well for assessing the Brier score VIM of X_4, but the associated confidence intervals are moderately anti-conservative in the case of X_1. This matches the results observed for AUC and c-index.

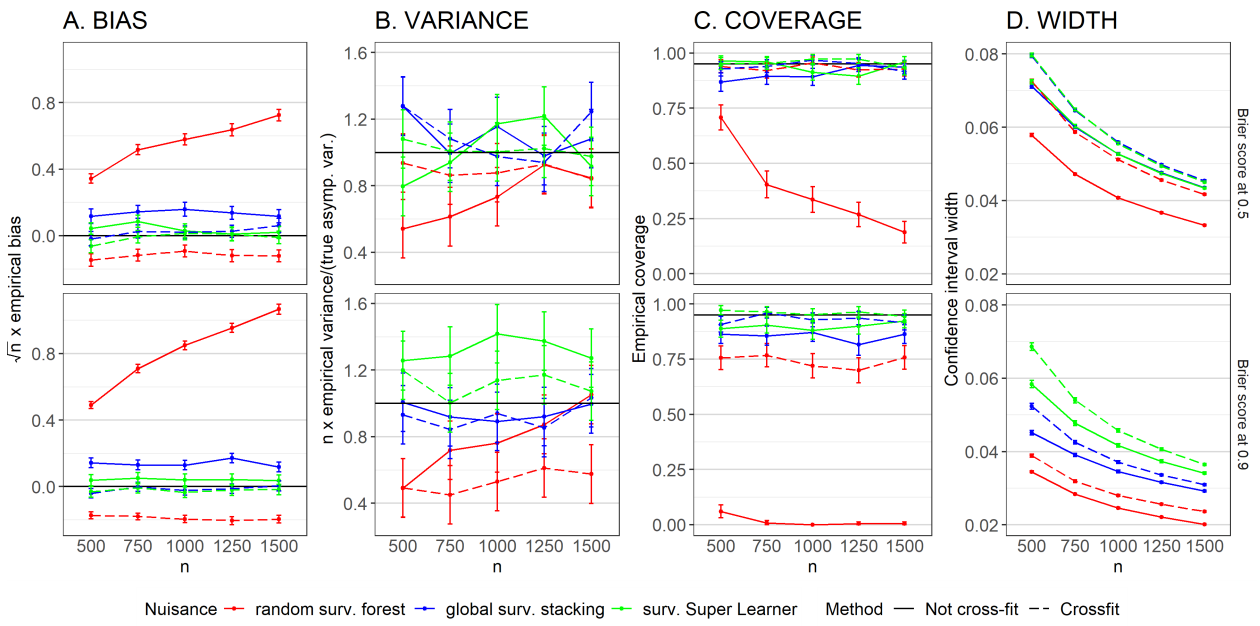

Figure A.5: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 2 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

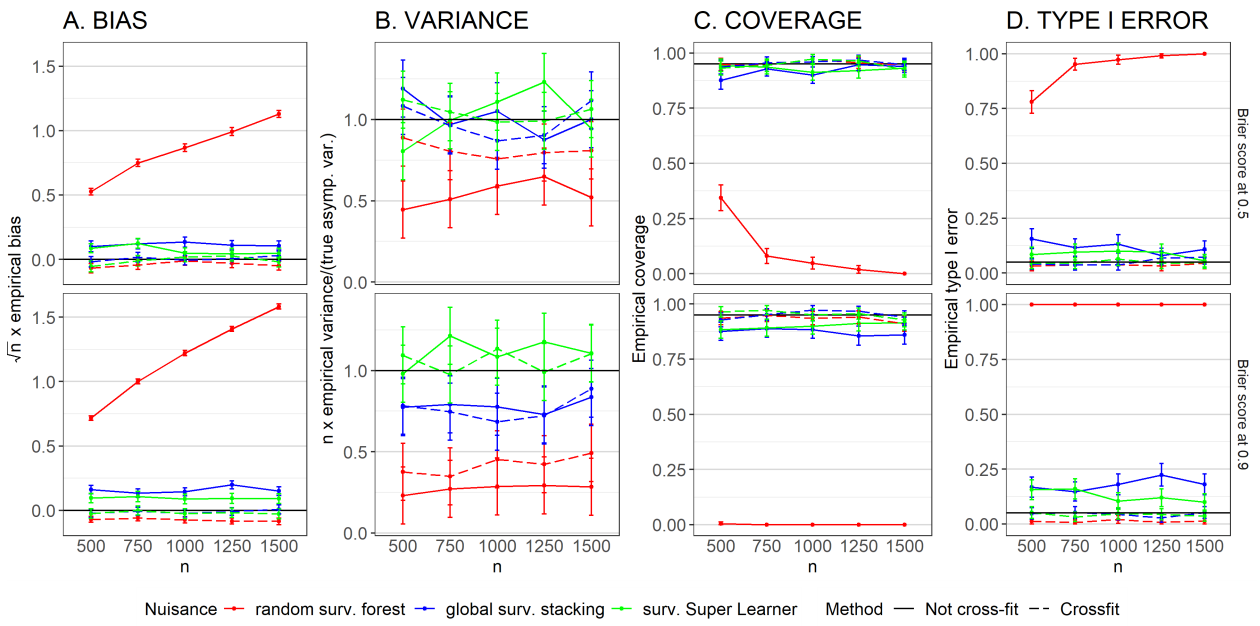

Figure A.6: Performance of the one-step VIM estimator for the (zero) importance of X_4 in Scenario 2 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

A.5.4 The effect of the censoring rate (Scenario 3)

In order to study the effect of censoring on our procedure, we performed a simulation study in Scenario 3, in which the censoring rate varied between 30% and 70%. The event times and covariates were generated as in Scenario 1, while $\beta_{0,C}$ was selected to achieve overall censoring rates in $\{30\%, 40\%, \ldots, 70\%\}$. For this scenario, we generated 500 random datasets of size 1000. We considered the importance of X_1 using AUC and Brier score at landmark times $\tau \in \{0.5, 0.9\}$ and c-index truncated at $\tau = 0.9$. We used Algorithm 2.2 to compute point and standard error estimates, from which we computed nominal 95% Wald-type confidence intervals. We evaluated performance using the empirical bias, the empirical variance, the empirical confidence interval coverage, and the average confidence interval width.

In Figure A.7 we display the results of this experiment. The bias of the cross-fitted global stacking and survival Super Learner estimators is largely unaffected by the censoring rate. The bias of the RSF estimator is modestly affected by increased censoring, even with cross-fitting. The variance of all estimators tends to increase with increasing censoring, most dramatically at the later landmark time $\tau = 0.9$. The cross-fitted global stacking and survival Super Learner estimators demonstrate nominal coverage. The coverage of the cross-fitted RSF estimator is slightly low when estimating AUC VIM, but seemingly unaffected by a higher censoring rate. As expected, confidence interval width increases with increased censoring. Overall, we see that the operating characteristics of the procedure are largely consistent across censoring levels. Unsurprisingly, the impact of censoring on estimator variance and confidence interval width is larger at the later landmark time.

A.5.5 Simulation results in Scenario 4

In Scenario 4, we set $p = 5$. The feature covariance matrix Σ was a 5×5 matrix with 1 on the diagonal. Off-diagonal elements were set to 0, except for $\Sigma 1, 4 = \Sigma_{4,1} = 0.7$ and $\Sigma_{2,3} = \Sigma_{3,2} = -0.3$. In this setting, therefore, while X_3 and X_4 do not directly affect the event time T, the true importance of X_1 and X_2 are altered due to their associations with

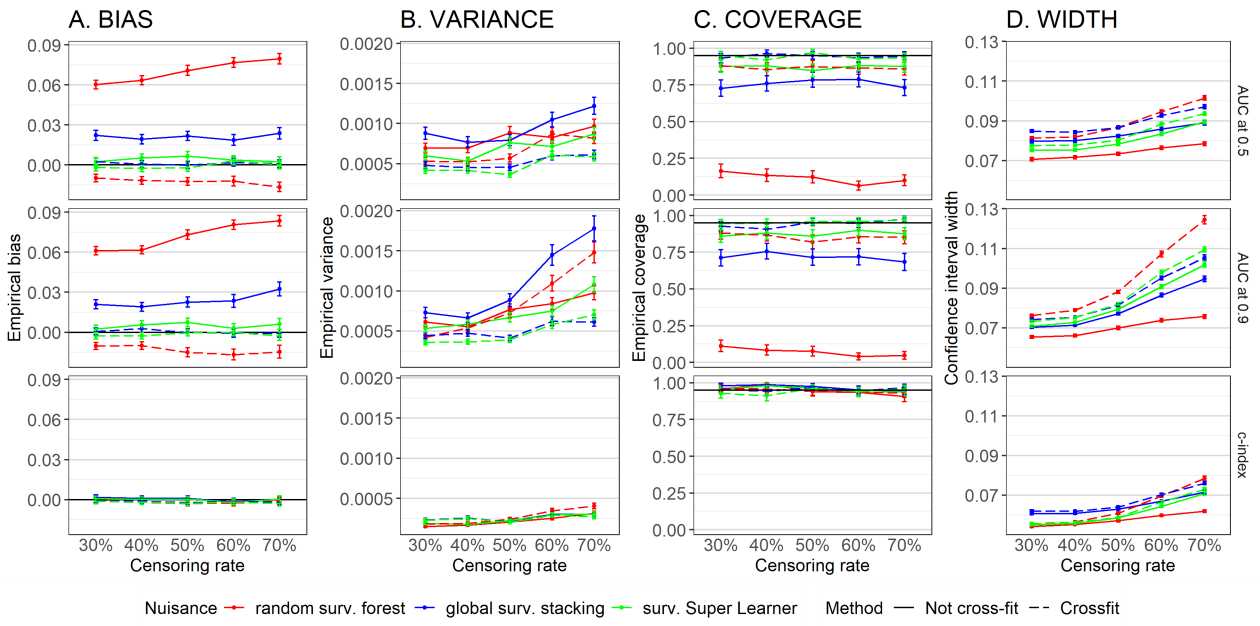

Figure A.7: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 3 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 1.5. (A) empirical bias; (B) empirical variance; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

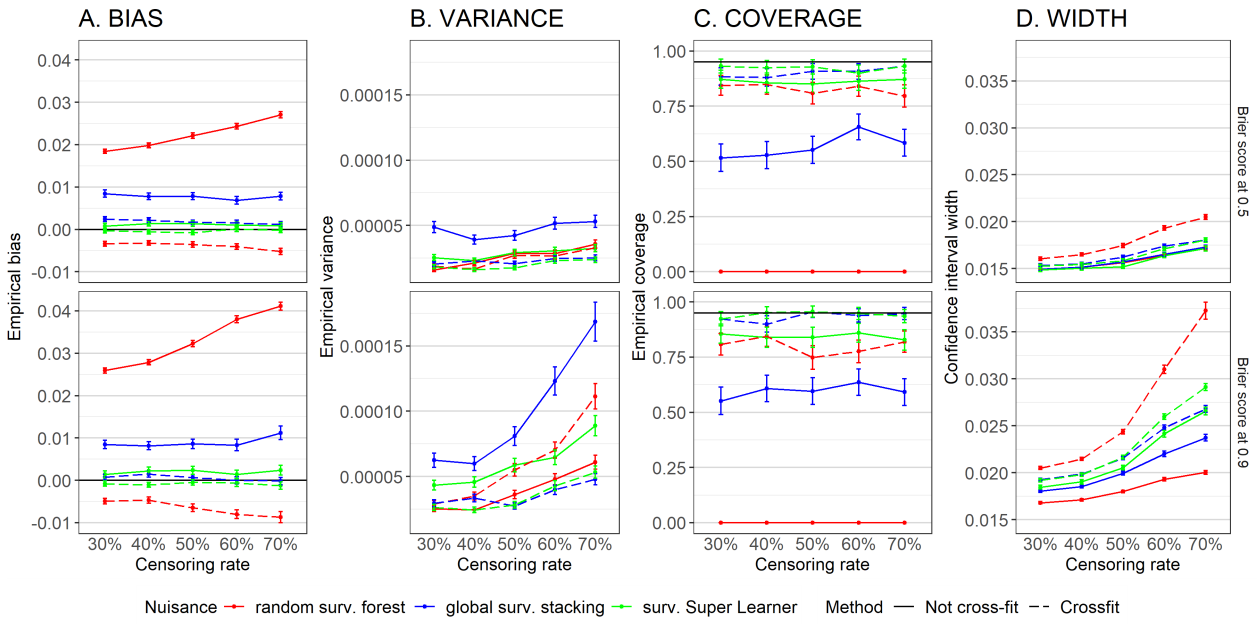

Figure A.8: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 3 in terms of Brier score. The two VIMs shown are the Brier score at times 0.5 and 0.9. (A) empirical bias; (B) empirical variance; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

X_4 and X_3, respectively. We generated 500 random datasets of size $n \in \{500, 750, \ldots, 1500\}$ and assessed the importance of X_1 and X_4 individually, as well as the joint importance of (X_1, X_4), using Algorithm 2.3. We estimated nuisance parameters as in other settings. We present results for empirical bias scaled by $n^{1/2}$, empirical variance scaled by n, empirical confidence interval coverage, average confidence interval width (for X_1 and (X_1, X_4)), and empirical type I error (for X_4).

Results for Scenario 4 are displayed in Figures A.9 – A.14. Generally, the proposed procedure performs similarly as in other simulation settings — the presence of correlated features seems to have little impact on operating characteristics. Furthermore, the procedure performs equally well for groups of features as for individual features. When features are expected to be correlated, considering groups of features may improve the interpretability of VIM analyses.

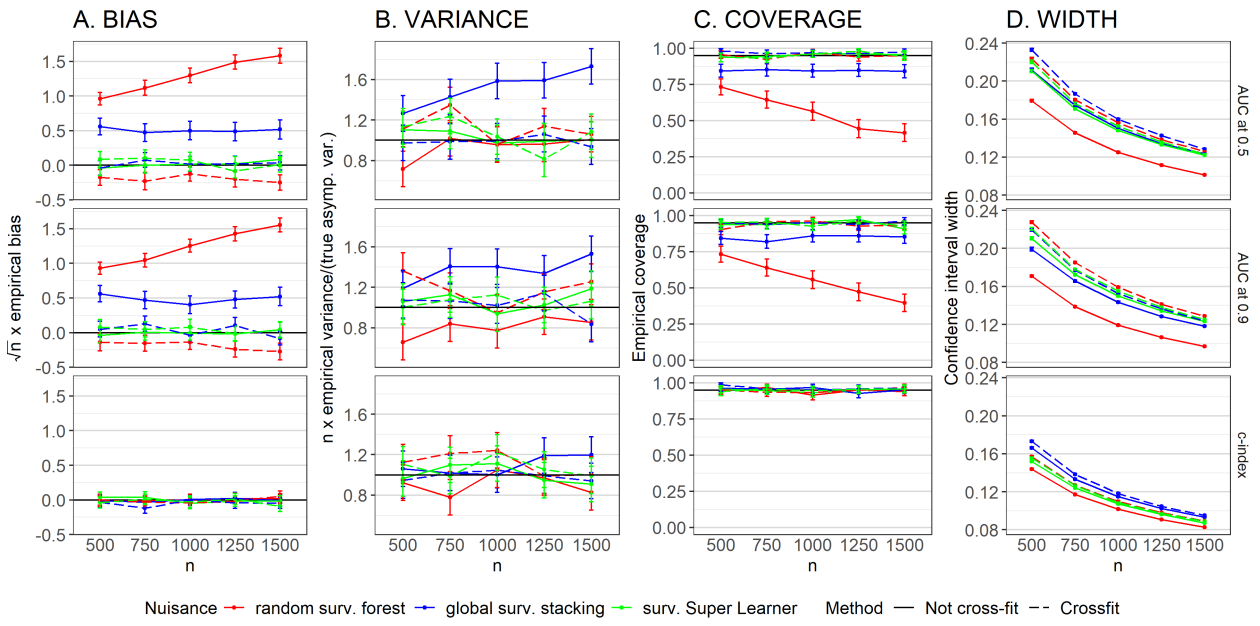

Figure A.9: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 4 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

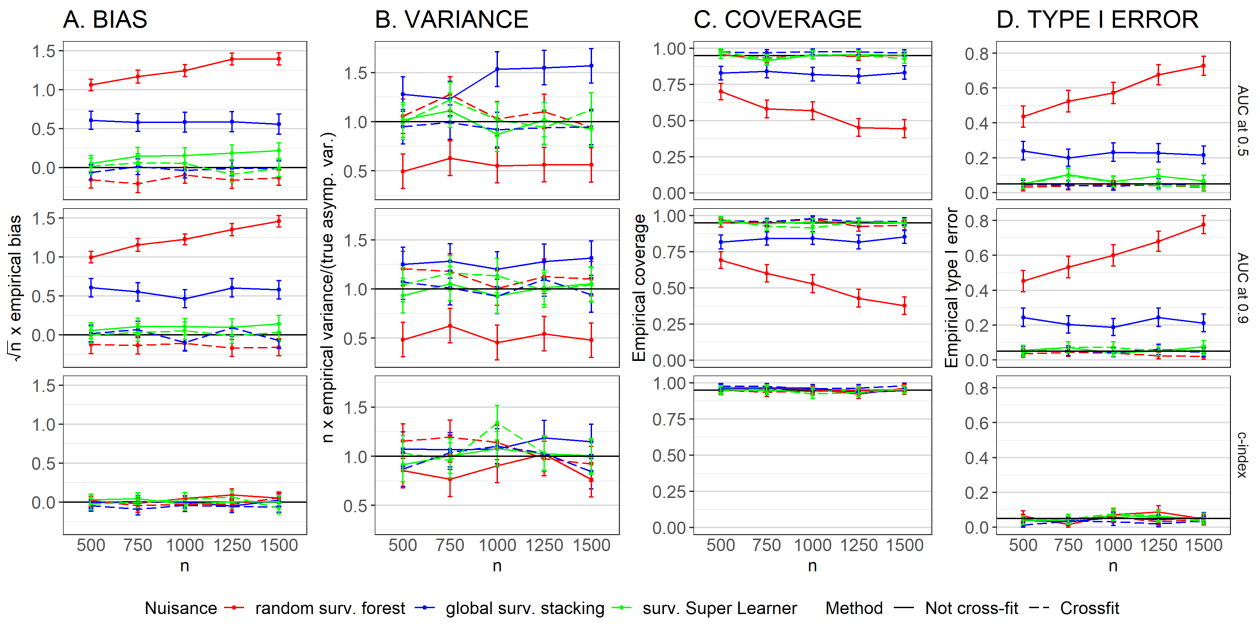

Figure A.10: Performance of the one-step VIM estimator for the (zero) importance of X_4 in Scenario 4 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

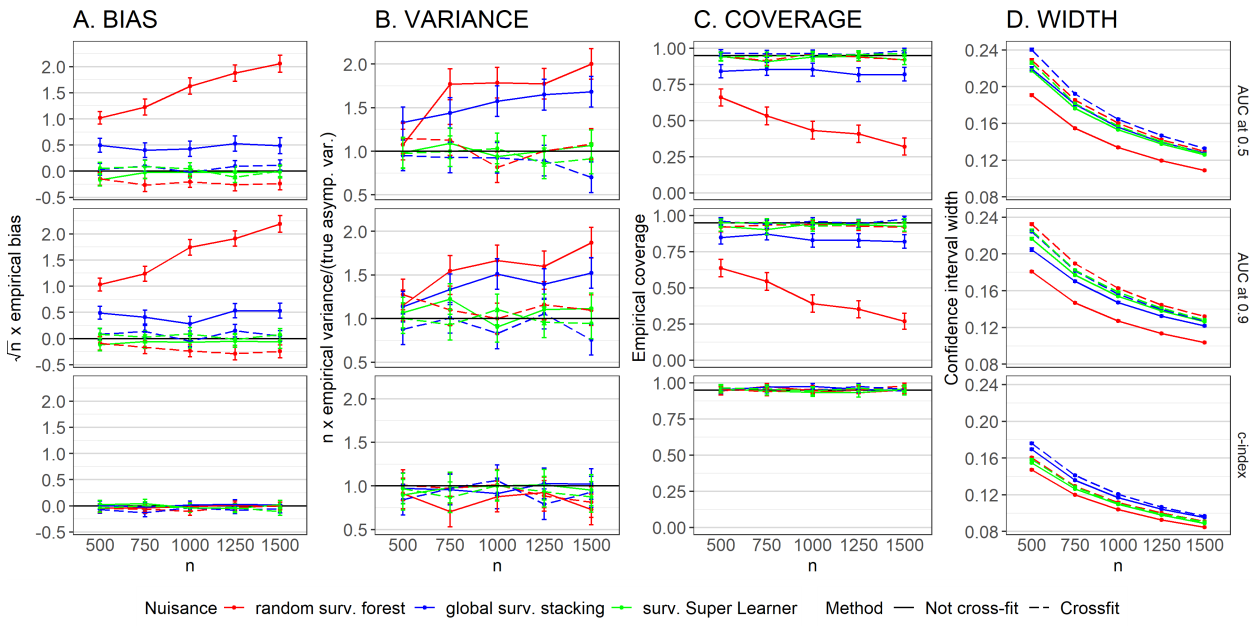

Figure A.11: Performance of the one-step VIM estimator for the joint importance of (X_1, X_4) in Scenario 4 in terms of AUC and c-index. The three VIMs shown are AUC at times 0.5 and 0.9 and the c-index truncated at time 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

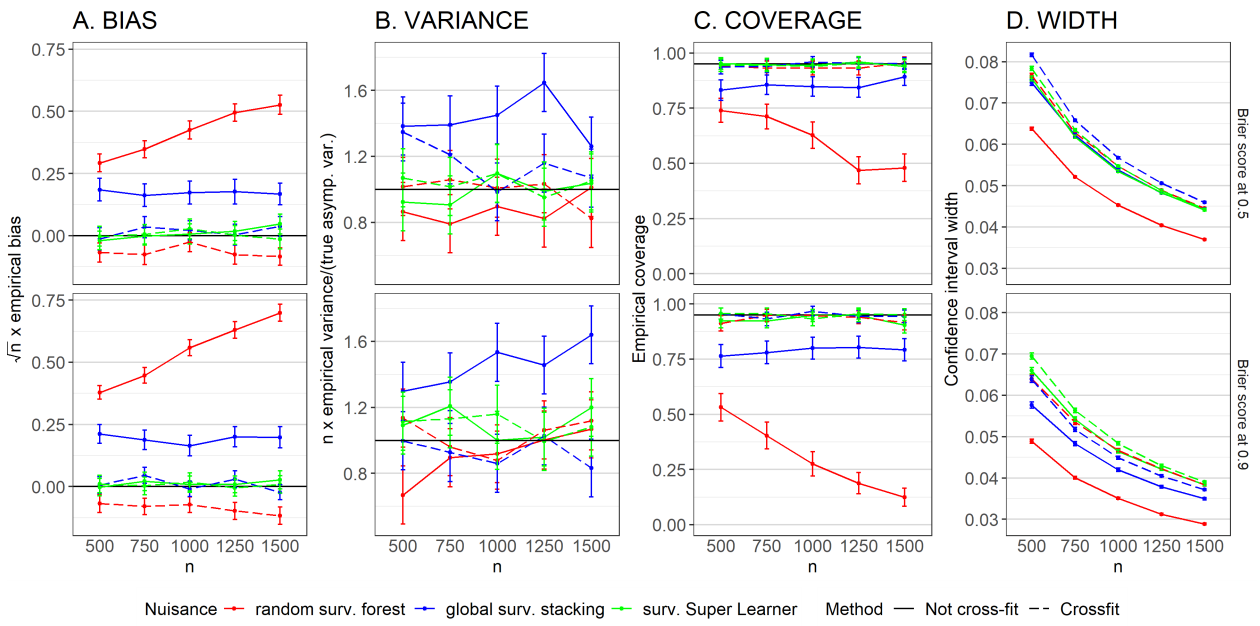

Figure A.12: Performance of the one-step VIM estimator for the importance of X_1 in Scenario 4 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

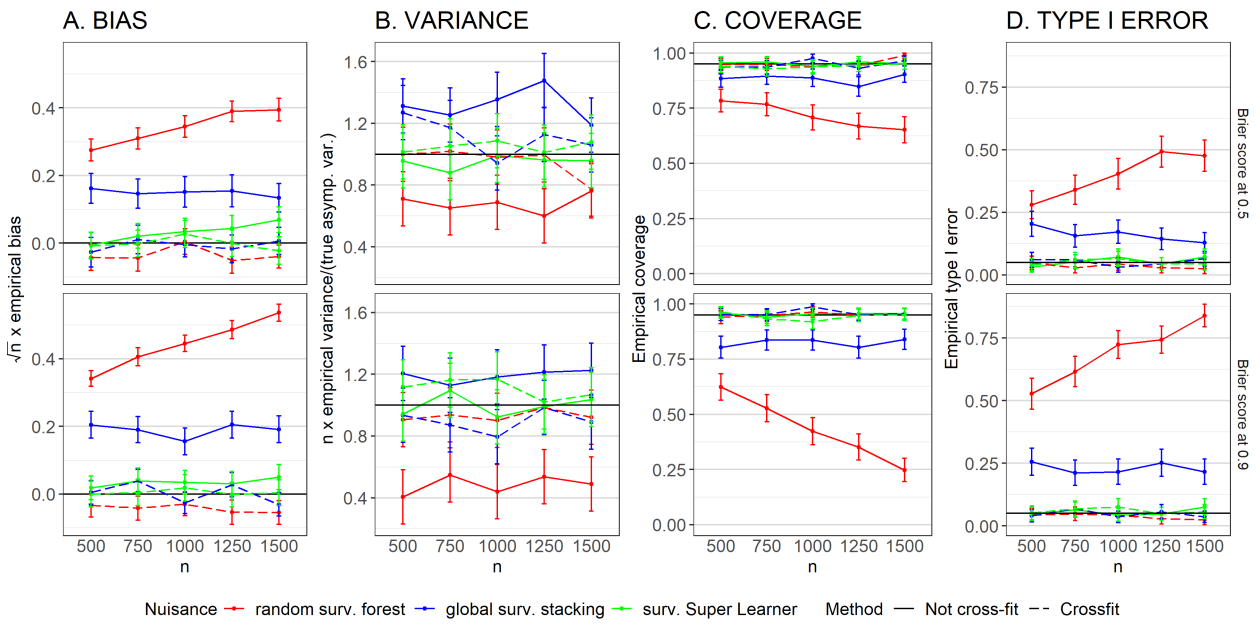

Figure A.13: Performance of the one-step VIM estimator for the (zero) importance of X_4 in Scenario 4 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) empirical type I error. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

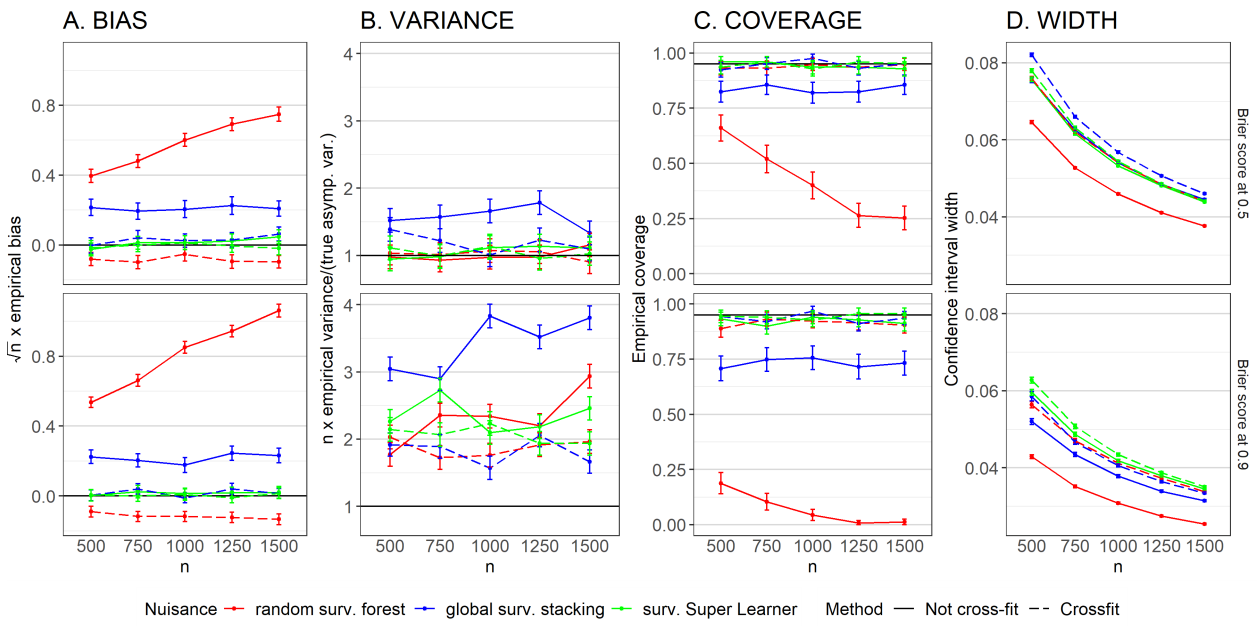

Figure A.14: Performance of the one-step VIM estimator for the joint importance of (X_1, X_4) in Scenario 4 in terms of Brier score. The two VIMs shown are Brier score at times 0.5 and 0.9. (A) empirical bias scaled by $n^{1/2}$; (B) empirical variance scaled by n/σ^2, where σ^2 is the theoretical asymptotic variance of the estimator; (C) empirical coverage of nominal 95% confidence intervals; (D) average confidence interval width. The colors denote different nuisance estimators, which were used to estimate both event and censoring distributions. Solid and dashed lines denote non-cross-fitted and cross-fitted estimators, respectively. Vertical bars represent 95% confidence intervals taking into account Monte Carlo error.

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Retrospective sampling

As in the prospective setting, we let W be the study entry time. We do not consider censoring in this setting, so T is observed for all participants. For notational consistency, we set $C = 0$ for all participants and define $Y := \max\{T, C\}$ and $\Delta := \mathbb{1}(T \geq C) = 1$. This implies that $Y = T$ for all participants, i.e., the observed follow-up times are equal to the event times. Under right truncation, an individual is sampled if $Y \leq W$. The observed data are $O := (X, Y, \Delta, W)$, and the sampling criterion is $W \geq Y$.

Identification of $S(t \,|\, x)$ in the retrospective setting follows from the prospective identification. Let τ denote a user-specified real number, and define the random variables $\bar{T} := \tau - T$, $\bar{C} := \tau - C$, $\bar{Y} := \tau - Y$, $\bar{W} := \tau - W$, and $\bar{\Delta} := \mathbb{1}(\bar{T} \leq \bar{C}) = 1$. In practice, we set τ as the maximum study entry time W, so that $\bar{T}$, $\bar{C}$, $\bar{Y}$ and $\bar{W}$ are non-negative. If T has bounded support, the upper bound of that support would be another natural choice for τ. (In principle, τ could be any real number, including 0, in which case the transformed data could take negative values. For the sake of applying our prospective results to the retrospective setting, we assume the transformed data are nonnegative.) We suppose that Assumption C holds.

We note that $\bar{T}$ is subject to conditionally independent left truncation by $\bar{W}$. Denoting by $\bar{\Lambda}(t \,|\, x)$ and $\bar{S}(t \,|\, x)$ the conditional cumulative hazard and survival functions of $\bar{T}$ given X at t, we can directly use the prospective setting results to identity $\bar{\Lambda}(\cdot \,|\, x)$ at generic time point t by

$$\int_0^t \frac{\bar{\pi}(x)\bar{F}_1(du \,|\, x)}{\bar{G}_1(u \,|\, x)\bar{\pi}(x)\left\{1 - \bar{F}_1(u^- \,|\, x)\right\} + \bar{G}_0(u \,|\, x)\left\{1 - \bar{\pi}(x)\right\}\left\{1 - \bar{F}_0(u^- \,|\, x)\right\}},$$

where $\bar{F}_1, \bar{F}_0$, $\bar{G}_1$, and $\bar{G}_0$ are defined analogously as in the prospective setting, and $\bar{\pi}(x) :=$

$P(\bar{\Delta} = 1 \,|\, X = x)$. Because there is no censoring in this setting, $\bar{\pi}(x) = 1$, and the above identification can be written in the form

$$\bar{\Lambda}^{\text{obs}}(t \,|\, x) := \int_0^t \left[\bar{G}_1(u \,|\, x) \left\{1 - \bar{F}_1(u^- \,|\, x)\right\}\right]^{-1} \bar{F}_1(du \,|\, x) \ .$$

Finally, we note that $S(t \,|\, x)$ can be written as $1 - \bar{S}(\tau - t \,|\, x)$, and so it suffices to use the above identification of $\bar{\Lambda}(\cdot \,|\, x)$ in order to estimate $S(\cdot \,|\, x)$. This result demonstrates that estimating the conditional hazard of T given X under right truncation can be accomplished by simply estimating the conditional hazard of $\tau - T$ given X under left truncation.

In the retrospective setting, estimation proceeds by simply (i) transforming the data to reverse time, taking $\bar{Y}_i = \tau - Y_i$, $\bar{\Delta}_i = 1$, $\bar{W}_i = \tau - W_i$, and $\bar{t} = \tau - t$; (ii) following Steps 1, 2, and 3 in Section 3.3.3 of the main text to produce an estimate $\bar{S}_n(\bar{t} \,|\, x)$ of $\bar{S}(\bar{t} \,|\, x)$, with $\bar{S}_n$ being either the product integral or exponential form; and (iii) computing $S_n(t \,|\, x) = 1 - \bar{S}_n(\bar{t} \,|\, x)$.

B.2 Details of identification result

Let $F_{T,C,W}$ and $F_{C,W}$ denote the conditional distribution functions of (T, C, W) given X and (C, W) given X, respectively. We begin by using standard probability rules to write

$$\begin{aligned}
\pi(x)F_1(u \,|\, x) &= P(\Delta = 1 \,|\, X = x, W \leq Y) P(Y \leq u \,|\, \Delta = 1, X = x, W \leq Y) \\
&= \frac{P(\Delta = 1, Y \leq u, W \leq Y \,|\, X = x)}{P(W \leq Y \,|\, X = x)} \\
&= \frac{P(T \leq C, T \leq u, W \leq T \,|\, X = x)}{P(W \leq Y \,|\, X = x)} \\
&= \frac{\iiint \mathbb{1}(t \leq u, c \geq t, w \leq t) F_{T,C,W}(dt, dc, dw \,|\, x)}{P(W \leq Y \,|\, X = x)} \\
&\overset{\text{(a)}}{=} \frac{\iint \mathbb{1}(t \leq u, c \geq t, w \leq t) F_{C,W}(dc, dw \,|\, x) F(dt \,|\, x)}{P(W \leq Y \,|\, X = x)} \\
&= \frac{\int_0^u P(C \geq t, W \leq t \,|\, X = x) F(dt \,|\, x)}{P(W \leq Y \,|\, X = x)} \ ,
\end{aligned}$$

where (a) follows from Assumption C. The differential of this function with respect to u is

$$\pi(x)F_1(du \,|\, x) = \frac{P(W \leq u, C \geq u \,|\, X = x) F(du \,|\, x)}{P(W \leq Y \,|\, X = x)} \ .$$

The denominator of Λ^{obs} is

$$\begin{aligned} & G_1(u \,|\, x)\pi(x)\left\{1 - F_1(u^- \,|\, x)\right\} + G_0(u \,|\, x)\left\{1 - \pi(x)\right\}\left\{1 - F_0(u^- \,|\, x)\right\} \\ &= P(W \le u, Y \ge u \,|\, W \le Y, X = x) \ , \end{aligned}$$

where we have applied the law of total probability. Continuing from this expression we have

$$\begin{aligned} P(W \le u, Y \ge u \,|\, W \le Y, X = x) &= \frac{P(W \le u \le Y, W \le Y \,|\, X = x)}{P(W \le Y \,|\, X = x)} \\ &= \frac{P(W \le u \le Y \,|\, X = x)}{P(W \le Y \,|\, X = x)} \\ &= \frac{P(W \le u, C \ge u, T \ge u \,|\, X = x)}{P(W \le Y \,|\, X = x)} \\ &\overset{\text{(b)}}{=} \frac{P(W \le u, C \ge u \,|\, X = x)S(u^- \,|\, x)}{P(W \le Y \,|\, X = x)} \ , \end{aligned}$$

where (b) follows from Assumption C. Combining this with the numerator, we have

$$\begin{aligned} \Lambda^{\text{obs}}(t \,|\, x) &= \int_0^t \frac{\pi(x)F_1(du \,|\, x)}{G_1(u \,|\, x)\pi(x)\left\{1 - F_1(u^- \,|\, x)\right\} + G_0(u \,|\, x)\left\{1 - \pi(x)\right\}\left\{1 - F_0(u^- \,|\, x)\right\}} \\ &= \int_0^t \left(\frac{P(W \le u, C \ge u \,|\, X = x)}{P(W \le Y \,|\, X = x)}\right) \Bigg/ \left(\frac{P(W \le u, C \ge u \,|\, X = x)S(u^- \,|\, x)}{P(W \le Y \,|\, X = x)}\right) F(du \,|\, x) \\ &= \int_0^t \frac{F(du \,|\, x)}{S(u^- \,|\, x)} \\ &= \Lambda(t \,|\, x) \ . \end{aligned}$$

B.3 Identification under alternative assumption

In this section, we consider an alternative identifying assumption for use in contexts in which censoring can only occur in individuals who satisfy the sampling criterion. Assumption D is given in three parts as:

Assumption D1: $W < C$ almost surely;

Assumption D2: T and W are conditionally independent given X;

Assumption D3: T and C are conditionally independent given (X, W) and $W \le T$.

Let F_W denote the conditional distribution function of W given X. Let $H_{T,C,W}$ and H_W denote respectively the conditional distribution functions of (T, C, W) and W given both X and $W \leq T$. Let $H_{T,C\,|\,W}$, $H_{C\,|\,W}$, and $H_{T\,|\,W}$ denote respectively the conditional distribution functions of (T, C), C, and T given both (W, X) and $W \leq T$.

We note that Assumption D2 allows us to write

$$\begin{aligned} H_{T\,|\,W}(t\,|\,w,x) &= P(T \leq t\,|\,X = x, W = w, W \leq T) \\ &= P(T \leq t\,|\,X = x, W = w, T \geq w) \\ &= \frac{P(T \leq t, T \geq w\,|\,X = x, W = w)}{P(T \geq w\,|\,X = x, W = w)} \\ &= \frac{\mathbb{1}(w \leq t)P(w \leq T \leq t\,|\,X = x, W = w)}{P(T \geq w\,|\,X = x, W = w)} \\ &\overset{(a)}{=} \frac{\mathbb{1}(w \leq t)P(w \leq T \leq t\,|\,X = x)}{P(T \geq w\,|\,X = x)}\,, \end{aligned} \tag{B.1}$$

where (a) follows from Assumption D2. We then use standard probability rules to write

$$\begin{aligned} \pi(x)F_1(u\,|\,x) &= P(\Delta = 1\,|\,X = x, W \leq Y)P(Y \leq u\,|\,\Delta = 1, X = x, W \leq Y) \\ &= P(\Delta = 1, Y \leq u\,|\,W \leq Y, X = x) \\ &\overset{(b)}{=} P(\Delta = 1, Y \leq u\,|\,W \leq T, X = x) \\ &= P(T \leq C, T \leq u\,|\,W \leq T, X = x) \\ &= \iiint \mathbb{1}(t \leq u, c \geq t)H_{T,C,W}(dt, dc, dw\,|\,x) \\ &= \iiint \mathbb{1}(t \leq u, c \geq t)H_{T,C|W}(dt, dc\,|\,w, x)H_W(dw\,|\,x) \\ &\overset{(c)}{=} \iiint \mathbb{1}(t \leq u, c \geq t)H_{C|W}(dc\,|\,w, x)H_{T|W}(dt\,|\,w, x)H_W(dw\,|\,x) \\ &\overset{(d)}{=} \iiint \frac{\mathbb{1}(t \leq u, c \geq t, w \leq t)H_{C|W}(dc\,|\,w, x)F(dt\,|\,x)H_W(dw\,|\,x)}{P(T \geq w\,|\,X = x)}\,, \end{aligned}$$

where (b) follows from Assumption D1, (c) from Assumption D3, and (d) from equation (B.1). The differential of this function with respect to u is

$$\pi(x)F_1(du\,|\,x) = \iint \frac{\mathbb{1}(c \geq u, w \leq u)H_{C|W}(dc\,|\,w, x)H_W(dw\,|\,x)F(du\,|\,x)}{P(T \geq w\,|\,X = x)}\,.$$

Let $R(u,x) := \iint \frac{\mathbb{1}(c \geq u, w \leq u) H_{C|W}(dc \,|\, w,x) H_W(dw \,|\, x)}{P(T \geq w \,|\, X = x)}$. As in Section B.2, the denominator of Λ^{obs} can be written as

$$\begin{aligned}
P(W \leq u, Y \geq u \,|\, W \leq Y, X = x) &\overset{\text{(e)}}{=} P(W \leq u, Y \geq u \,|\, W \leq T, X = x) \\
&= P(W \leq u, C \geq u, T \geq u \,|\, W \leq T, X = x) \\
&= \iiint \mathbb{1}(t \geq u, c \geq u, w \leq u) H_{T,C,W}(dt, dc, dw \,|\, x) \\
&= \iiint \mathbb{1}(t \geq u, c \geq u, w \leq u) H_{T,C|W}(dt, dc \,|\, w, x) H_W(dw \,|\, x) \\
&\overset{\text{(f)}}{=} \iiint \mathbb{1}(t \geq u, c \geq u, w \leq u) H_{C|W}(dc \,|\, w, x) H_{T|W}(dt \,|\, w, x) H_W(dw \,|\, x) \\
&\overset{\text{(g)}}{=} \iiint \frac{\mathbb{1}(t \geq u, c \geq u, w \leq u, w \leq t) H_{C|W}(dc \,|\, w, x) F_T(dt \,|\, x) H_W(dw \,|\, x)}{P(T \geq w \,|\, X = x)} \\
&= \iiint \frac{\mathbb{1}(t \geq u, c \geq u, w \leq u) H_{C|W}(dc \,|\, w, x) F_T(dt \,|\, x) H_W(dw \,|\, x)}{P(T \geq w \,|\, X = x)} \\
&= S(u^- \,|\, x) R(u,x) \ ,
\end{aligned}$$

where (e) follows from Assumption D1, (f) from Assumption D3, and (g) from equation (B.1). Combining this with the numerator, we find, as claimed, that

$$\begin{aligned}
\Lambda^{\text{obs}}(t \,|\, x) &= \int_0^t \frac{\pi(x) F_1(du \,|\, x)}{G_1(u \,|\, x)\pi(x)\{1 - F_1(u^- \,|\, x)\} + G_0(u \,|\, x)\{1 - \pi(x)\}\{1 - F_0(u^- \,|\, x)\}} \\
&= \int_0^t \frac{R(u,x)}{R(u,x) S(u^- \,|\, x)} F(du \,|\, x) = \int_0^t \frac{F(du \,|\, x)}{S(u^- \,|\, x)} = \Lambda(t \,|\, x) \ .
\end{aligned}$$

B.4 Simulation details

B.4.1 Additional details on estimators and data-generating mechanism

Here, we describe the estimators included in the simulation studies. The `R` package implementation is given in parentheses.

1. Global survival stacking (`survML`): We estimated F_1, F_0, G_1, and G_0 using pooled binary regression with Super Learner, as implemented in the `SuperLearner` software package. The algorithm library consisted of the marginal mean, logistic regression

with all pairwise interactions, generalized additive models, multivariate adaptive regression splines, random forests, and gradient-boosted trees. We estimated $\pi(x)$ using the same Super Learner library. We used five-fold cross-validation and the built-in nonnegative least-squares method to determine the optimal convex combination of these algorithms. We considered three time grids $\mathcal{C}$ for the pooled regression: a grid made up of every observed follow-up time and grids of 10 or 40 cutpoints evenly spaced on the quantile scale of observed follow-up times. We used the exponential form $S_{n,e}(t \mid x)$. The approximation time grid $\mathcal{B}$ was set to the observed follow-up times. The predictions across times in the approximation grid were isotonized using the pool adjacent violators algorithm, as implemented in the `Iso` software package. Table B.1 details the Super Learner library used for estimating binary regressions in global and local survival stacking.

2. Local survival stacking (`survML`): We used Super Learner as the binary classifier in local survival stacking, using the same algorithm library as for global survival stacking. Tuning was performed in the same manner as described above, and the same time grids were included, based on observed event times $\mathcal{R}_n$. Algorithm B.1 details the procedure to construct the local survival stacking algorithm.

3. Survival Super Learner (`survSuperLearner`): We used the same library of algorithms for both the censoring and event time distributions, including the marginal Kaplan-Meier estimator, the Cox proportional hazards model with Breslow baseline hazard estimator, exponential regression, Weibull regression, log-logistic regression, a generalized additive proportional hazards model, and random survival forest. We did not evaluate this method in any settings with truncation since it is not designed to handle truncation. Table B.2 details the Super Learner library used for `survSuperLearner`.

4. Random forests (`LTRCforests`): We used conditional inference forests for left-truncated, right-censored data. We set the `mtry` parameter equal to the square root of

the number of predictors, rounded up.

5. Linear Cox proportional hazards regression (`survival`): We used a main-terms linear Cox proportional hazards model with Breslow baseline hazard estimator.

6. Generalized additive Cox proportional hazards regression (`mgcv`): We used a main-terms generalized additive Cox proportional hazards model with Breslow baseline hazard estimator.

Below, we include additional information on the simulation data-generating mechanisms.

- **Figure B.1:** Example densities for the time-to-event variable T under the two data-generating mechanisms used in simulations.

- **Table B.3:** Average truncation rates in numerical experiments.

Algorithm B.1 Local survival stacking

Choose grid of time-points $\mathcal{C} := \{t_1^*, t_2^*, \ldots, t_k^*\}$ on which to discretize. Set $t_{k+1}^* = \infty$.
Choose how to include time in model (continuous, dummy variable, etc.).
for $t_j^* \in \mathcal{C}$ **do**
    Including only participants with $Y \geq t_j^*$ and $W \leq t_j^*$, construct dataset $D_{t_j^*}$ consisting of participant baseline covariates, outcomes $\mathbb{1}(t_j^* \leq Y < t_{j+1}^*)$, and time using chosen basis.
end for
Construct full stacked dataset by combining $\{D_{t_1^*}, D_{t_2^*}, \ldots, D_{t_k^*}\}$.
Fit binary regression or classification algorithm of choice.
Generate hazard predictions $\{\lambda_n(t_1^* \,|\, x), \lambda_n(t_2^* \,|\, x), \ldots, \lambda_n(t_k^* \,|\, x)\}$ from fitted model.
Compute estimate $S_n(t \,|\, x) = \prod_{t_j^* \in \mathcal{C}: t_j^* \leq t} \{1 - \lambda_n(t_j^* \,|\, x)\}$.

Algorithm name	Algorithm description	Tuning parameters
`SL.mean`	Marginal mean	NA
`SL.glm.interaction`	Logistic regression with pairwise interactions	NA
`SL.gam`	Generalized additive model	default
`SL.earth`	Multivariate adaptive regression splines	default
`SL.ranger`	Random forest	default
`SL.xgboost`	Gradient-boosted trees	`ntrees` $\in \{250, 500, 1000\}$ `max_depth` $\in \{1, 2\}$

Table B.1: Algorithms included in the Super Learner for global and local survival stacking. All tuning parameters besides those for `SL.xgboost` were set to default values. In particular, `gam` was implemented with `degree = 2`; `earth` with `degree = 2, penalty = 3, nk` = number of predictors plus 1, `endspan = 0`, `minspan = 0`; and `ranger` with `num.trees = 500, mtry` = the square root of the number of predictors, `min.node.size = 1, sample.fraction = 1` with replacement. For `SL.xgboost`, `shrinkage` was set to 0.01, `minobspernode` was set to 1, and each combination of `ntrees` and `max_depth` was included in the Super Learner library.

Algorithm name	Algorithm description
`survSL.km`	Kaplan-Meier estimator
`survSL.expreg`	Survival regression assuming event and censoring times follow an exponential distribution conditional on covariates
`survSL.weibreg`	Survival regression assuming event and censoring times follow a Weibull distribution conditional on covariates
`survSL.loglogreg`	Survival regression assuming event and censoring times follow a log-logistic distribution conditional on covariates
`survSL.gam`	Main-terms generalized additive Cox proportional hazards estimator as implemented in the `mgcv` package
`survSL.coxph`	Main-terms Cox proportional hazards estimator with Breslow baseline cumulative hazard
`survSL.rfsrc`	Random survival forests as implemented in the package `randomForestSRC`

Table B.2: Algorithms included in the survival Super Learner. All tuning parameters were set to default values. In particular, `gam` was implemented with `degree = 1`; and `rfsrc` with `ntree = 500`, `mtry =` the square root of the number of predictors, `nodesize = 15`, `splitrule = "logrank"`, `sampsize = 1` with replacement.

Study design	Skew	Setting	Truncation rate
Prospective	Right	Non-proportional hazards	70%
		Proportional hazards	66%
		Discrete	70%
	Left	Non-proportional hazards	46%
		Proportional hazards	51%
		Discrete	46%
Retrospective	Right	Non-proportional hazards	35%
	Left	Non-proportional hazards	65%

Table B.3: Average truncation rates across simulations.

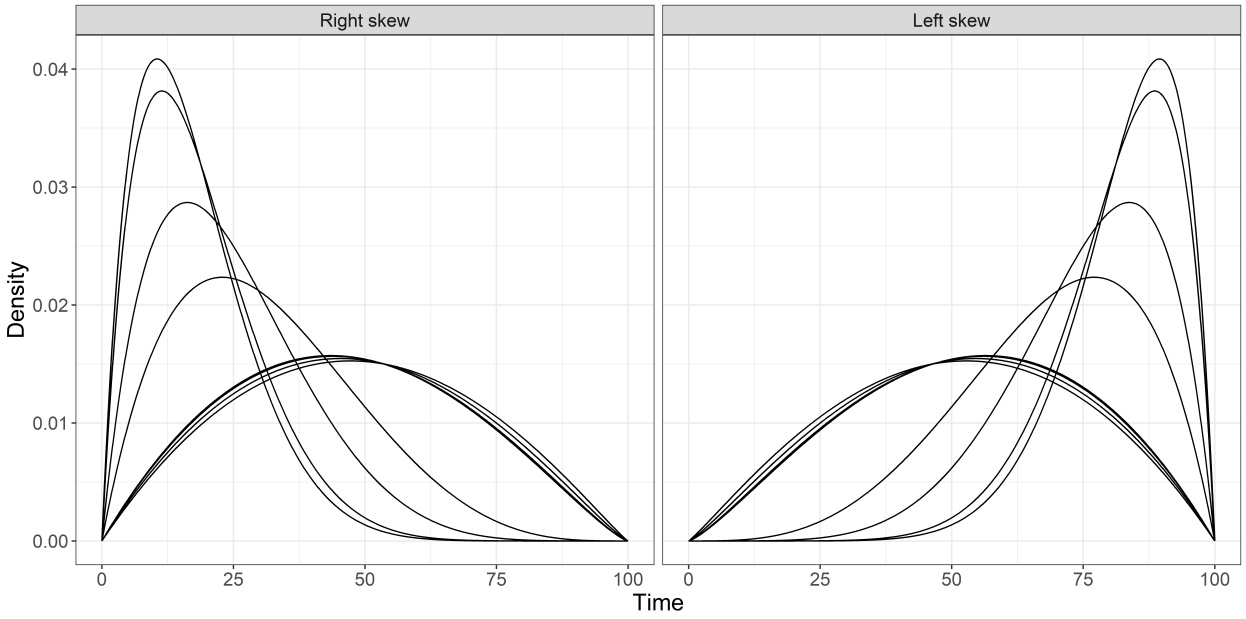

Figure B.1: Example densities for the time-to-event variable T under the two data-generating mechanisms used in simulations. Each plot shows the conditional density of T given X for ten random draws from the distribution of X.

B.5 Additional numerical results

B.5.1 Performance under prospective sampling with non-proportional hazards (Scenarios 1 and 2)

These simulations were performed under Scenarios 1 and 2, as described in Section 3.4.1 of the main text. Figures B.2 and B.3 display the full results, including mean integrated squared error (MISE) over the interval $[0, 100]$ and MSE at landmark times corresponding to the 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times. Global survival stacking performs well overall and is generally not sensitive to the choice of time grid $\mathcal{C}$ used for the pooled binary regression.

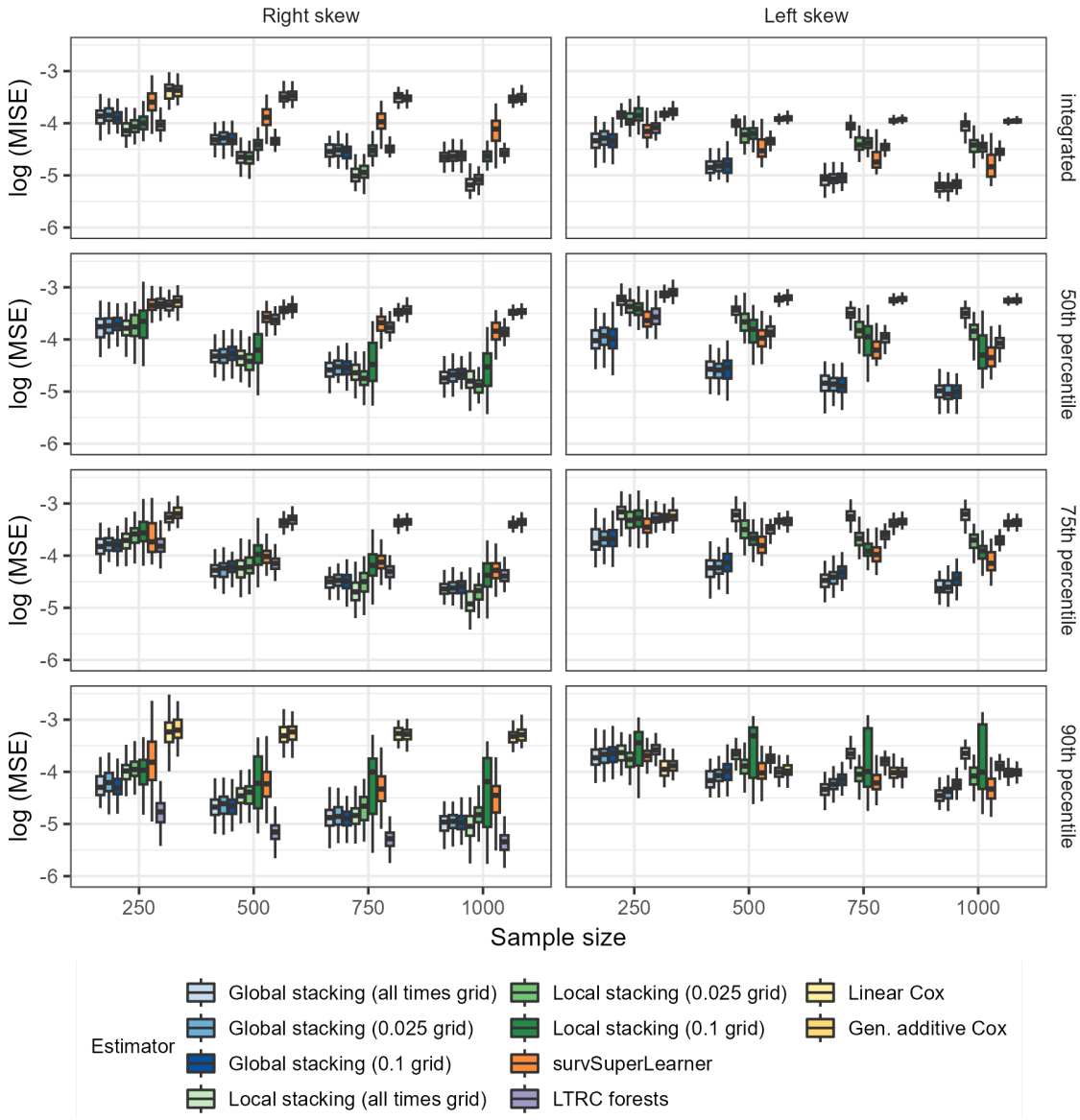

Figure B.2: Performance of conditional survival estimators with right-censored data (Scenario 1). The methods compared were global survival stacking, local survival stacking, survival Super Learner, random forests, a main-terms linear Cox model with Breslow baseline hazard estimator, and a main-terms generalized additive Cox model with Breslow baseline hazard estimator. Time grids are based on quantiles of observed follow-up times (global stacking) or observed event times (local stacking). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

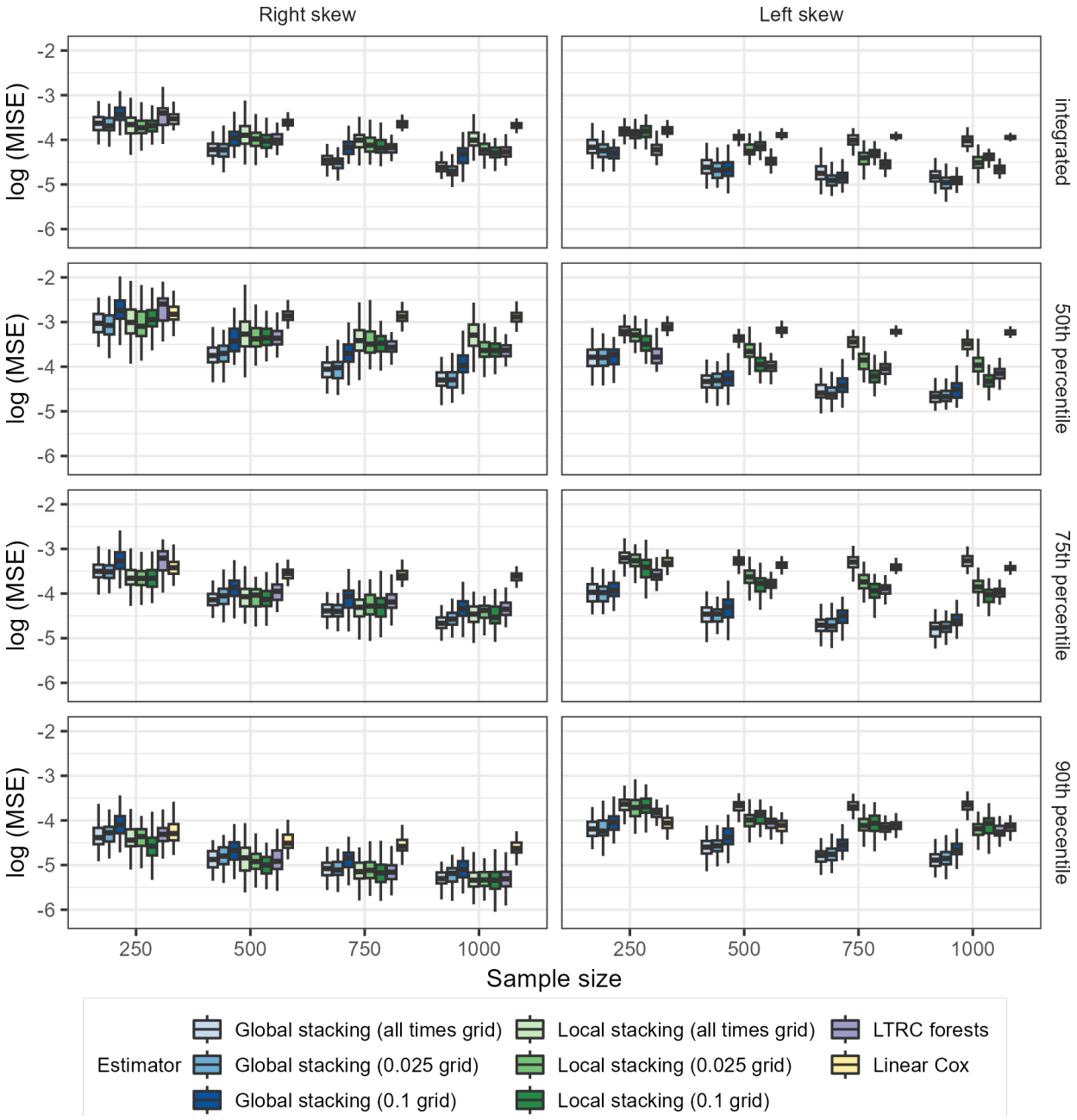

Figure B.3: Performance of conditional survival estimators with left-truncated, right-censored data (Scenario 2). The methods compared were global survival stacking, local survival stacking, random forests, and a main-terms linear Cox model with Breslow baseline hazard estimator. Time grids are based on quantiles of observed follow-up times (global stacking) or observed event times (local stacking). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

B.5.2 Performance under retrospective sampling (Scenario 3)

For the retrospective simulation study, data were generated as described in Section 3.4.1 of the main text. Only observations with $Y \leq W$ were sampled. There was no censoring in the retrospective study design. Figure B.4 shows the results of the retrospective simulation study. The results are similar as those of the prospective study with left truncation, with global survival stacking demonstrating consistent overall performance.

B.5.3 Estimator performance under proportional hazards (Scenario 4)

We evaluated the performance of global and local survival stacking when the data satisfied the proportional hazards assumption. The covariate vector X, censoring variable C, and study entry variable W were generated in the same manner as in the primary numerical experiments described in Section 3.4 of the main text. Given covariate vector $X = x$, we simulated the event time T to be distributed as $100c(x)Z_3$, where $c(x) = \exp\left\{\frac{1}{2}(x_1 + x_2 + x_3 + x_4 + x_5)\right\}$ was the hazard ratio. In the right-skewed setting, Z_3 was a Beta(2,3) random variable, and in the left-skewed setting Z_3 was a Beta(3,2) random variable. We considered the prospective setting with left truncation and 25% censoring rate. We evaluated performance in the same manner as described in Section 3.4 of the main text. We compared global survival stacking, local survival stacking, and the Cox model with Breslow baseline hazard estimator.

We display the results for the proportional hazards simulation in Figure B.5. The Cox model, which in this case was correctly specified, yields the best overall performance across all metrics. Among the machine learning approaches, local stacking on a 40 cutpoint grid performs the best by a modest margin, and global survival stacking demonstrates good performance as well. As in the primary empirical results in Section 3.4 of the main text, local stacking is more sensitive to grid size choice. Local stacking on a grid of all observed event times tends to show increasing estimation error beyond a sample size of 500.

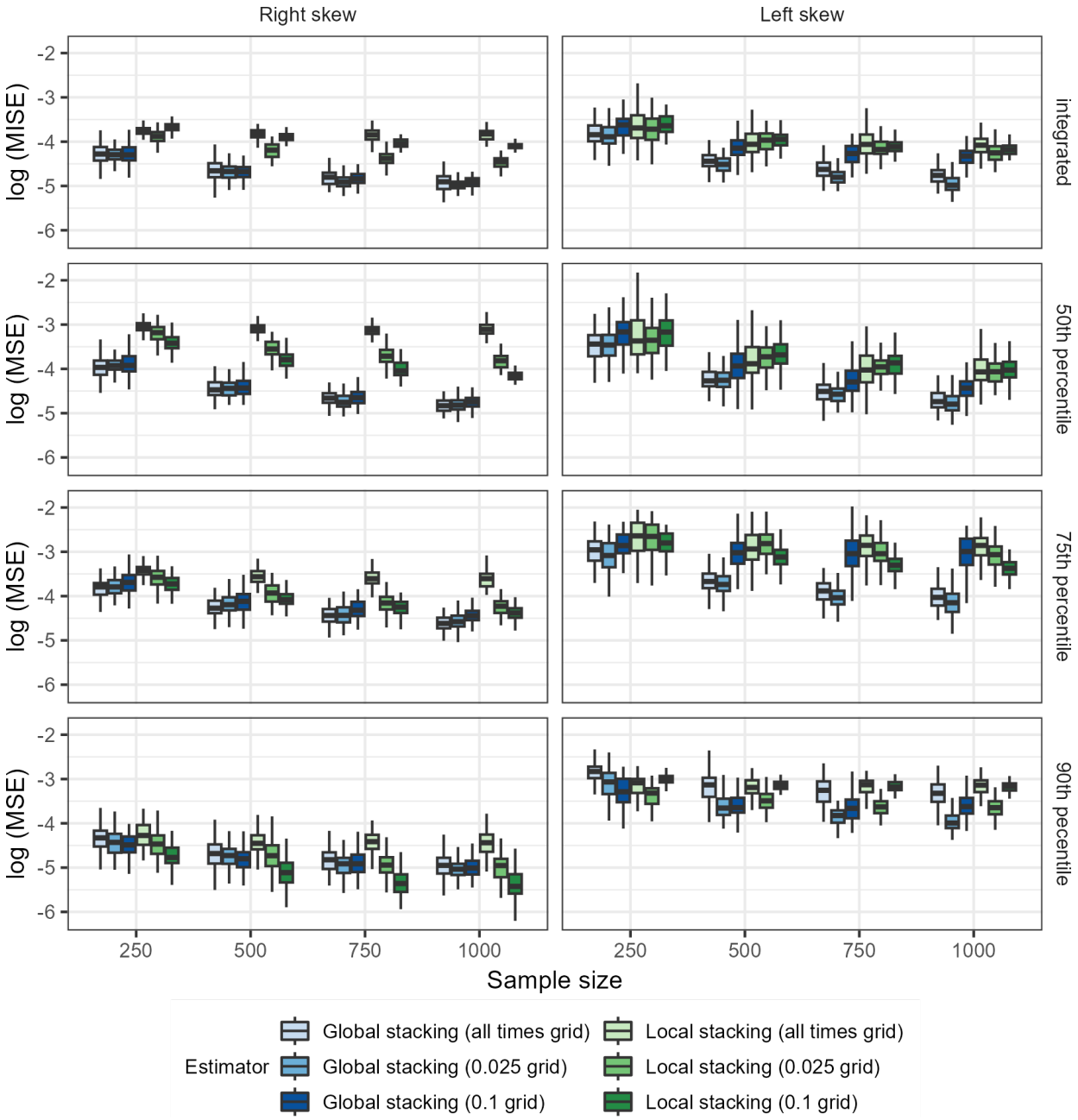

Figure B.4: Performance of conditional survival estimators with right-truncated data (Scenario 3). The methods compared were global survival stacking and local survival stacking. Time grids are based on quantiles of observed follow-up times (global stacking) or observed event times (local stacking). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

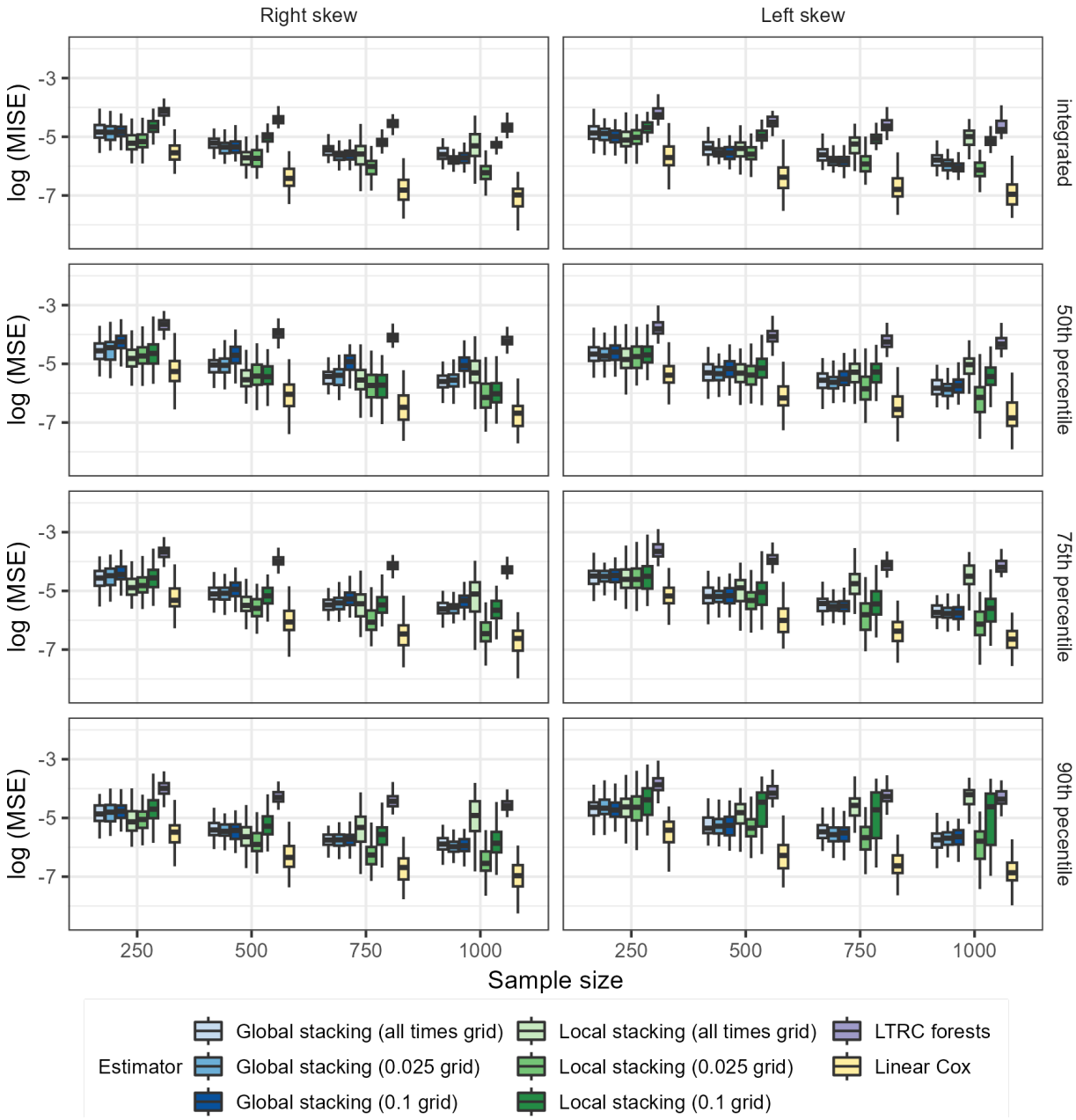

Figure B.5: Performance of conditional survival estimators with right-censored, left-truncated data generated under a proportional hazards model (Scenario 4). The methods compared were global survival stacking, local survival stacking, random forests, and the main-terms Cox model with Breslow baseline hazard estimator. Time grids are based on quantiles of observed follow-up times (global stacking) or observed event times (local stacking). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

B.5.4 Estimator performance when events are observed at discrete times (Scenario 5)

For the discrete-time numerical experiments, we generated X, T, C, and W in the same manner as in the primary numerical experiments described in Section 3.4 of the main text. For m the desired number of times in the discrete-time grid, we divided the interval $[0, 100]$ into m equally sized intervals $I_1, I_2, \ldots, I_m$. For all Y falling in I_j, we set $\tilde{Y}$ equal to the right endpoint of interval I_j and used $\tilde{Y}$ as the observed follow-up time. In this way, while the distribution of T was continuous, $\tilde{Y}$ was observed on a discrete time scale. Likewise, for all W falling in I_j, we set $\tilde{W}$ equal to the left endpoint of interval I_j. We considered the prospective setting with left truncation and 25% censoring rate. We evaluated performance in the same manner as described in Section 3.4 of the main text and compared the performance of global and local stacking on grids of all observed follow-up and event times, respectively. We used the product integral form for global stacking. We included the main-terms Cox model as a comparator.

We display the results for the discrete-time experiment with 10 intervals in Figure B.6, with 20 intervals in Figure B.7, and with 50 intervals in Figure B.8. With 10 and 20 intervals, the overall performance of global and local stacking are similar. With 50 intervals, global survival stacking generally outperforms local survival stacking in the left-skewed setting, and the two perform similarly in the right-skewed setting. The MSE and MISE for global stacking are similar in the 50 interval setting as in the continuous-time setting.

B.5.5 Computational considerations

In order to benchmark computational burden, we simulated samples of size 500 in the prospective study design without left truncation under the left-skewed data-generating mechanism. We fit each estimator as described above and generated conditional survival function estimates for a test data set of size 100 on an evenly spaced grid of times from $t = 0.1$ to $t = 100$. The computational benchmarking simulations were run on an Amazon Web Services EC2 `r6a.large` instance with 2 vCPUs and 16GB memory. There were 100 simulation

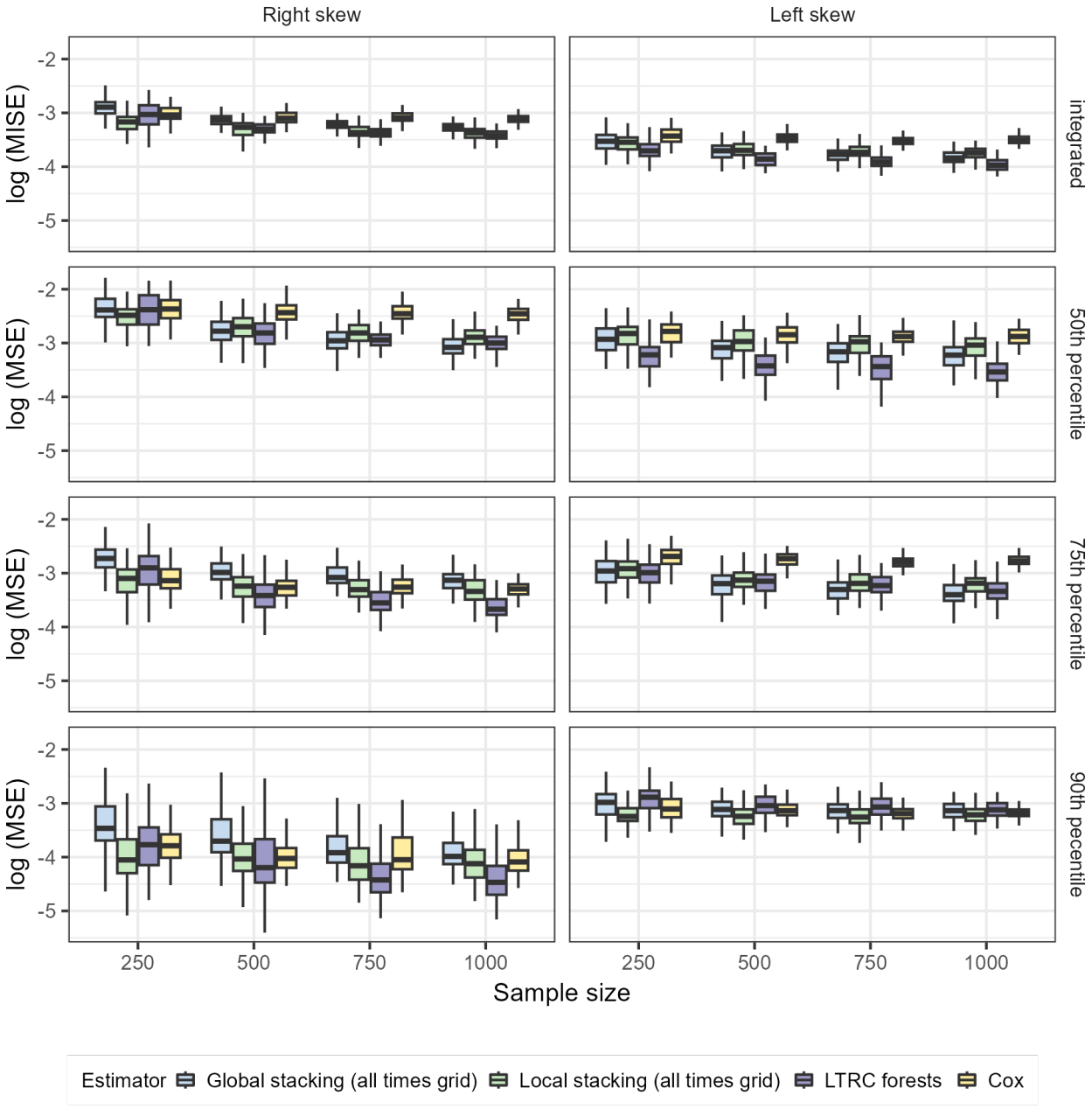

Figure B.6: Performance of conditional survival estimators with right-censored, left-truncated data observed on a discrete grid of 10 time-points (Scenario 5). The methods compared were global survival stacking, local survival stacking, random forests, and the main-terms Cox model with Breslow baseline hazard estimator. Global and local survival stacking were implemented using a grid of every observed follow-up time (global) or every observed event time (local). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

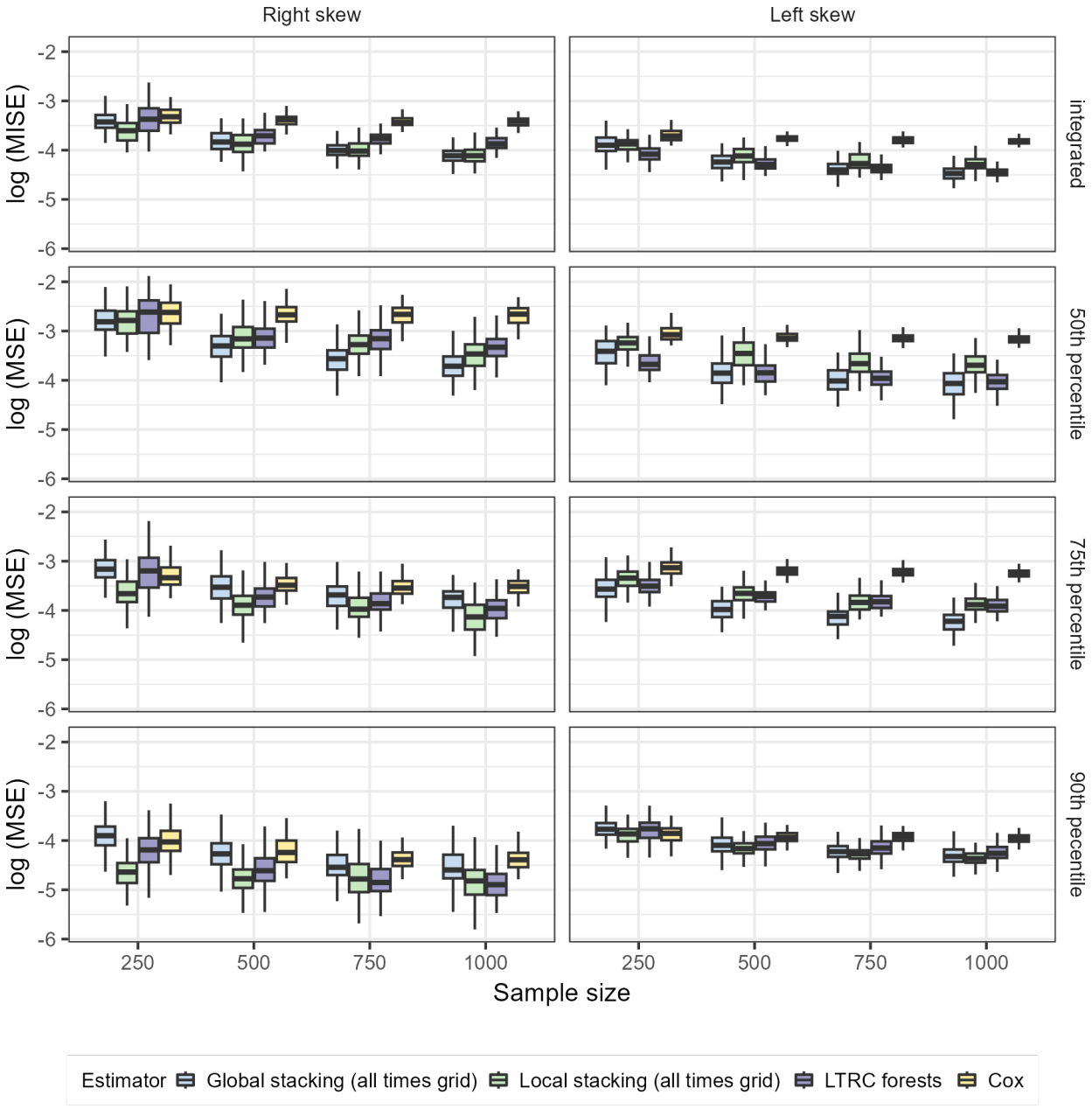

Figure B.7: Performance of conditional survival estimators with right-censored, left-truncated data observed on a discrete grid of 20 time-points (Scenario 5). The methods compared were global survival stacking, local survival stacking, random forests, and the main-terms Cox model with Breslow baseline hazard estimator. Global and local survival stacking were implemented using a grid of every observed follow-up time (global) or every observed event time (local). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

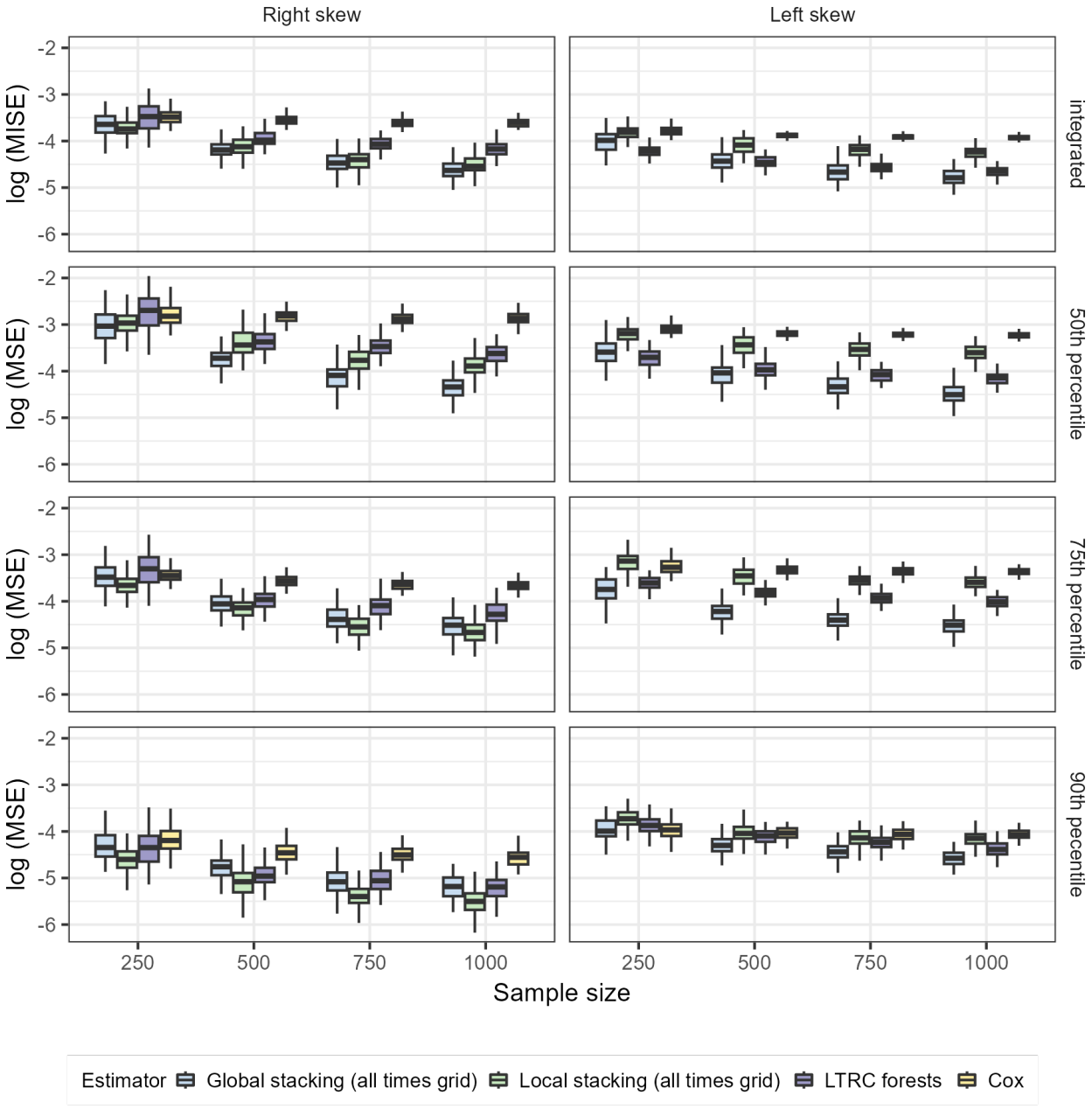

Figure B.8: Performance of conditional survival estimators with right-censored, left-truncated data observed on a discrete grid of 50 time-points (Scenario 5). The methods compared were global survival stacking, local survival stacking, random forests, and the main-terms Cox model with Breslow baseline hazard estimator. Global and local survival stacking were implemented using a grid of every observed follow-up time (global) or every observed event time (local). From top to bottom, rows correspond to MISE and to MSE at 50^{th}, 75^{th}, and 90^{th} percentiles of observed event times.

Estimator	Mean runtime (s)	Std. dev. runtime (s)
Global stacking (all times grid)	998	36.0
Global stacking (40 cutpoint grid)	222	43.2
Global stacking (10 cutpoint grid)	129	1.1
Local stacking (all times grid)	493	21.7
Local stacking (40 cutpoint grid)	52	0.8
Local stacking (10 cutpoint grid)	20	0.3
survSuperLearner	61	1.8
LTRC forests	60	1.9
Linear Cox	0.03	0.002
Gen. additive Cox	4.6	0.14

Table B.4: Computation time for conditional survival estimators from numerical experiments.

replicates for each estimator.

Table B.4 displays the results of this experiment. Global survival stacking was slower than alternative methods, and its speed is highly dependent on the size of the grid used in the pooled binary regression.

B.5.6 Comparison of survival function mappings in global survival stacking

When the product integral is discretized, the differential of the cumulative hazard is a probability and must lie in $[0, 1]$. Our method may yield an estimated differential that lies outside of $[0, 1]$, leading to survival function estimates that are negative, particularly in the tails of the distribution of Y. The exponential form protects against this potential issue and is analogous to exponentiating the negative Nelson-Aalen cumulative hazard estimate (Fleming and Harrington, 1984). We note that in settings without truncation, $S_{n,p}$ naturally respects the $[0, 1]$ bounds. When the distribution function of T is continuous, we expect minimal differences in performance between the two forms of the survival function estimator. However, because the exponential mapping from hazard to survival function only holds mathematically if T has a continuous distribution, it is not clear if it should perform as well as the product

Estimator	Training sample size	Percent of estimates outside $[0, 1]$
Exponential	250	0
	500	0
	750	0
	1000	0
Product integral	250	0.6%
	500	1.3%
	750	1.5%
	1000	1.4%

Table B.5: Percentage of estimated survival probabilities falling outside $[0, 1]$ using two forms of the global survival stacking estimator in the prospective study design with left truncation and right censoring.

integral form when the hazard is evaluated on a grid of times.

We performed a simulation study to compare the two forms (product integral and exponential) of our estimator in the prospective setting with left truncation and right censoring. Both estimators used a grid of 40 cutpoints evenly spaced on the quantile scale. Data were generated as in the other prospective settings, and performance was again evaluated using MISE and MSE at three landmark times. In addition to assessing performance, we also recorded the proportion of estimated survival probabilities in the test data that fell outside the interval $[0, 1]$.

The overall performance of global survival stacking appears insensitive to the choice of survival function mapping (Figure B.9). For the product integral form, between 0.6% and 1.5% of estimated survival probabilities fell outside the unit interval, depending on the training data sample size (Table B.5). For the exponential form, none of the survival function estimates fell outside the unit interval. When the distribution of T is continuous, we recommend using the exponential form to protect again potential issues arising in a particular sample.

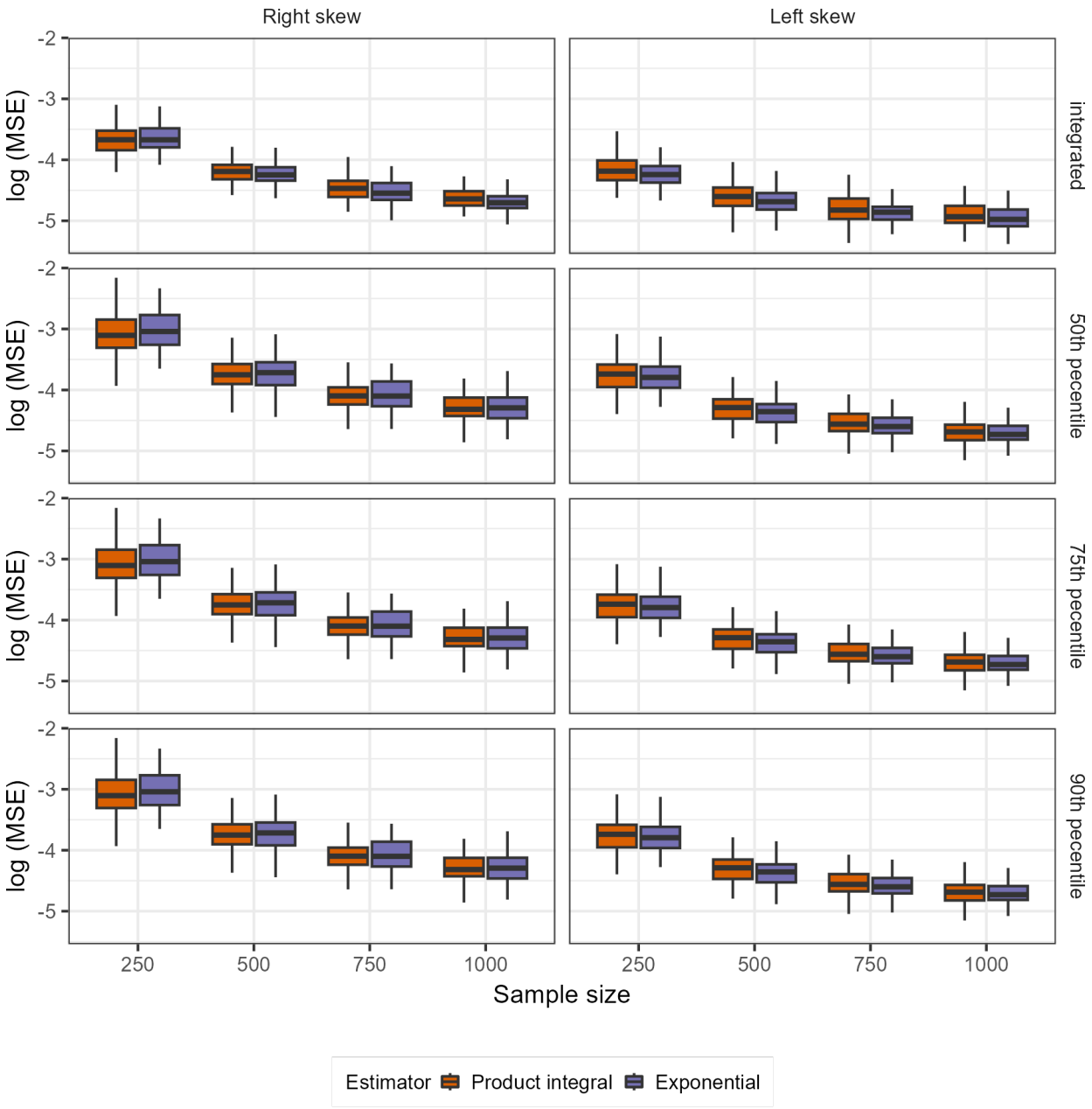

Figure B.9: Performance of different forms of the global survival stacking estimator in the prospective study design with left truncation and right censoring. The two forms are based on the mappings from hazard to survival function (product integral and exponential).

B.6 Details on publicly available datasets

We describe the publicly available survival datasets analyzed in Section 3.4.3 of the main text.

FLCHAIN: The Assay of Serum-Free Light Chain study investigated the relationship between serum-free light chain and mortality in residents of Olmstead County (Kyle et al., 2006). We used eight features for prediction: age, sex, calendar year of sample collection, serum-free light chain kappa portion, serum-free light chain lambda portion, free light chain group, serum creatinine, and an indicator of monoclonal gammopathy diagnosis. After removal of individuals with missing data, the dataset consisted of 6542 individuals. This dataset is available in the `survival` package (Therneau, 2022).

GBSG: The German Breast Cancer Study Group data is derived from a 1984-1989 trial of patients with node-positive breast cancer (Schumacher et al., 1994). The outcome of interest was recurrence-free survival time, with seven features of interest: hormone therapy, age, menopausal status, tumor size, tumor grade, number of positive nodes, progesterone receptor positivity, and estrogen receptor positivity. We used dummy variables for tumor grade, which consists of three categories. After removal of individuals with missing data, the dataset consisted of 684 individuals. It is available in the `survival` package (Therneau, 2022).

METABRIC: This dataset was produced by the Molecular Taxonomy of Breast Cancer International Consortium (Curtis et al., 2012). The outcome of interest was mortality, and the features of interest included expression of four different genes (MKI67, EGFR, PGR, and ERBB2), as well as five clinical features (hormone treatment, radiotherapy, chemotherapy, estrogen receptor positivity, and age at diagnosis). This dataset consisted of 1904 individuals, after removal of individuals with missing data. It is available in the `DeepSurv` software package (Katzman et al., 2018).

NWTCO: The National Wilms' Tumor Study investigated the relationship between time to tumor relapse and several prognostic variables, including two types of histology (D'Angio

et al., 1976). We included five features: local histology, central histology, age, disease stage, and an indicator of whether the individual was a participant in NWTS 3 or 4. We used dummy variables for disease stage, which consists of four categories. This dataset consisted of 4028 individuals and is available in the `survival` package (Therneau, 2022).

SUPPORT: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments investigated the relationship between clinical outcomes among seriously ill hospitalized adults (Knaus et al., 1995). For our analysis, the outcome of interest was mortality, with 14 features of interest: sex, age, race, number of comorbidities, blood pressure, heart rate, respiration, white blood cell count, temperature, serum creatinine, serum sodium, dementia diagnosis, diabetes diagnosis, and cancer diagnosis. Dummy variables were used for race (five categories) and cancer (three categories). After removal of individuals with missing data, the dataset consisted of 8873 individuals. This dataset is available on the Vanderbilt Biostatistics website.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1 Proof of Theorem 4

Proof of Theorem 4. We analyze the behavior of the estimator under H_0 and under H_1.

Case 1: $\psi_0 = 0$: We begin by studying the linear term in (4.5). First, we note that the collection of random variables $\left\{2\left[S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right]\right\}_{i=1}^n$ is i.i.d. over i and has mean zero, with variance σ^2_{0,B_n} and third absolute moment denoted by ρ_{0,B_n}.

We let F_n denote the cdf of the random variable

$$(n\sigma^2_{0,B_n})^{-1/2}\sum_{i=1}^n 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\},$$

and let F_0 the distribution function of a standard Gaussian random variable. An application of the Berry-Esseen theorem yields that

$$\sup_{u\in\mathbb{R}}|F_n(u) - F_0(u)| \leq U_{n,B_n} := \frac{C\rho_{0,B_n}}{n^{1/2}\sigma^3_{0,B_n}} \tag{C.1}$$

When $\psi_0 = 0$, we have that $\sigma^2_{0,B_n} = 4B_n^{-1}P_0\phi_0^2$, and so

$$U_{n,B_n} = \frac{C\rho_{0,B_n}B_n^{3/2}}{2n^{1/2}(P_0\phi_0^2)}\,.$$

Furthermore, we have that

$$2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} = 2(2S_i^{(B_n)} - 1)\phi_0(Z_i) = 2\phi_0(Z_i)\frac{1}{B_n}\sum_{b=1}^{B_n}(2S_{ib} - 1)\,.$$

We let $S^*_{ib} := 2S_{ib} - 1$ and note that this is a Rademacher random variable. Using the

independence of Z and S_b for all b, we then write

$$\rho_{0,B_n} = \frac{8}{B_n^3}\mathbb{E}_0\left[\left|\phi_0(Z)\sum_{b=1}^{B_n} S_b^*\right|^3\right] \leq \frac{8}{B_n^3}\mathbb{E}_0\left[|\phi_0(Z)|^3\left|\sum_{b=1}^{B_n} S_b^*\right|^3\right]$$
$$= \frac{8}{B_n^3}\mathbb{E}_0\left[|\phi_0(Z)|^3\right]\mathbb{E}_0\left[\left|\sum_{b=1}^{B_n} S_b^*\right|^3\right].$$

Applying the Marcinkiewicz–Zygmund inequality, we have that

$$\mathbb{E}_0\left[\left|\sum_{b=1}^{B_n} S_b^*\right|^3\right] \lesssim \mathbb{E}_0\left[\left(\sum_{b=1}^{B_n}|S_b^*|^2\right)^{3/2}\right] = B_n^{3/2}\ .$$

We can therefore upper bound ρ_{0,B_n} by a constant times $B_n^{-3/2}$, yielding that $U_{n,B_n} \lesssim n^{-1/2}$. This upper bound tends to 0 as $n \to \infty$, irrespective of the rate of B_n, implying that

$$\left(n\sigma_{0,B_n}^2\right)^{-1/2}\sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1-S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} \rightsquigarrow N(0,1)\ .$$

and therefore that $\frac{1}{n}\sum_{i=1}^{n}\left\{S_i^{(B_n)}\phi_0(Z_i) - (1-S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} = O_P\left(\left(nB_n\right)^{-1/2}\right)$.

We now consider the remainder term in (4.5). By assumption, $\sup_{b\in\{1,\ldots,B_n\}} r_{nb} = O_P\left(n^{-\alpha}\right)$. That is to say, for any $\epsilon > 0$, there exists $M > 0$ and $N > 0$ such that for all $n > N$

$$P_0\left(\left|\frac{\sup_{b\in\{1,\ldots,B_n\}} r_{nb}}{n^{-\alpha}}\right| > M\right) < \epsilon\ .$$

Now, fix $\epsilon > 0$ and choose M and N from above. We see that for all $n > N$

$$P_0\left(\left|\frac{\frac{1}{B_n}\sum_{b=1}^{B_n} r_{nb}}{n^{-\alpha}}\right| > M\right) \leq P_0\left(\left|\frac{\sup_{b\in\{1,\ldots,B_n\}} r_{nb}}{n^{-\alpha}}\right| > M\right) < \epsilon\ ,$$

and so $r_n^{(B_n)} = O_P\left(n^{-\alpha}\right)$ as well.

In total, then, we have

$$\begin{aligned}
&\left(\frac{n}{\sigma^2_{0,B_n}}\right)^{1/2} \left(\psi_n^{(B_n)} - \psi_0\right) \\
&\qquad = \left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} + \left(\frac{n}{\sigma^2_{0,B_n}}\right)^{1/2} r_n^{(B_n)} \\
&\qquad = \left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} + \left(\frac{nB_n}{\tau_0^2}\right)^{1/2} O_P(n^{-\alpha}) \\
&\qquad = \left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} + O_P(n^{\frac{1+\delta}{2}})O_P(n^{-\alpha}) \\
&\qquad = \left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} + O_P(n^{\frac{1+\delta}{2}})O_P(n^{-\alpha}) \\
&\qquad = \left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} + o_P(1) \rightsquigarrow N(0,1) \ .
\end{aligned}$$

Case 2: $\psi_0 \neq 0$: In this case, the plug-in estimator constructed without sample-splitting admits the representation

$$\left(n\nu_0^2\right)^{-1/2} \left(\psi_n^* - \psi_0\right) = (n\nu_0^2)^{-1/2} \sum_{i=1}^{n} \left\{\phi_0(Z_i) - \phi_{0,s}(Z_i)\right\} + o_P(1) \ .$$

This implies that ψ_n^*, when suitably scaled and centered, converges weakly to a standard Gaussian random variable. As $n \to \infty$, clearly $\sigma^2_{0,B_n} \to \nu_0^2$ and for each i, $S_i^{(B_n)} - \frac{1}{2} = o_P(1)$ by the Weak Law of Large Numbers. Thus,

$$\begin{aligned}
&\left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} - (n\nu_0^2)^{-1/2} \sum_{i=1}^{n} \left\{\phi_0(Z_i) - \phi_{0,s}(Z_i)\right\} \\
&\qquad = \left\{n\left(\sigma^2_{0,B_n} - \nu_0^2\right)\right\}^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} \\
&\qquad\qquad + (n\nu_0^2)^{-1/2} \sum_{i=1}^{n} 2\left\{\left(S_i^{(B_n)} - \frac{1}{2}\right)\phi_0(Z_i) - \left(\frac{1}{2} - S_i^{(B_n)}\right)\phi_{0,s}(Z_i)\right\} \\
&\qquad = o_P(1)O_P(1) + o_P(1) = o_P(1).
\end{aligned}$$

Therefore, we can write

$$\begin{aligned}
&\left(\frac{n}{\sigma^2_{0,B_n}}\right)^{1/2} \left(\psi_n^{(B_n)} - \psi_0\right) \\
&\qquad = \left(n\sigma^2_{0,B_n}\right)^{-1/2} \sum_{i=1}^{n} 2\left\{S_i^{(B_n)}\phi_0(Z_i) - (1 - S_i^{(B_n)})\phi_{0,s}(Z_i)\right\} + \left(\frac{n}{\sigma^2_{0,B_n}}\right)^{1/2} r_n^{(B_n)} \\
&\qquad = (n\nu_0^2)^{-1/2} \sum_{i=1}^{n} \{\phi_0(Z_i) - \phi_{0,s}(Z_i)\} + \left(\frac{n}{\nu_0^2}\right)^{1/2} r_n^{(B_n)} + o_P(1) \\
&\qquad = (n\nu_0^2)^{-1/2} \sum_{i=1}^{n} \{\phi_0(Z_i) - \phi_{0,s}(Z_i)\} + o_P(1) \rightsquigarrow N(0,1) \ .
\end{aligned}$$

□

C.2 Additional simulation results

In this section, we provide simulation results not shown in the main text. Figures C.1 and C.2 show results for estimating binary classification accuracy VIM under Scenarios 1 and 2, respectively. Figures C.3 and C.4 show results for estimating R^2 VIM under Scenarios 1 and 2, repsectively.

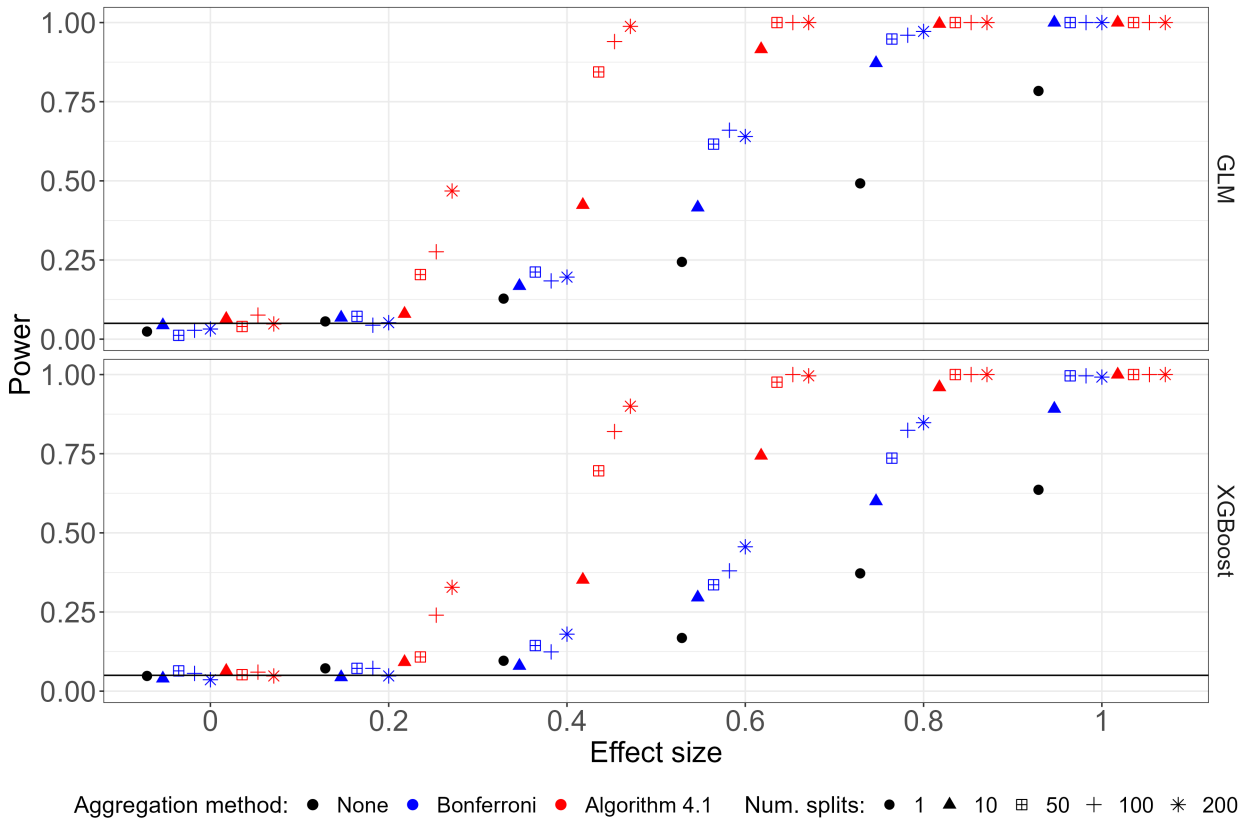

Figure C.1: Performance of sample splitting approaches for testing the hypothesis of zero importance using accuracy predictiveness in Scenario 1 ($p = 5$). When the effect size is zero, the null hypothesis holds. Oracle prediction functions were estimated using either probit regression (top row) or gradient-boosted trees (bottom row). The number of splitting iterations ranged from 1 to 200. Where more than one splitting iteration was used, results from multiple iterations were combined using either a Bonferroni correction (blue) or our proposed method, Algorithm 4.1 (red).

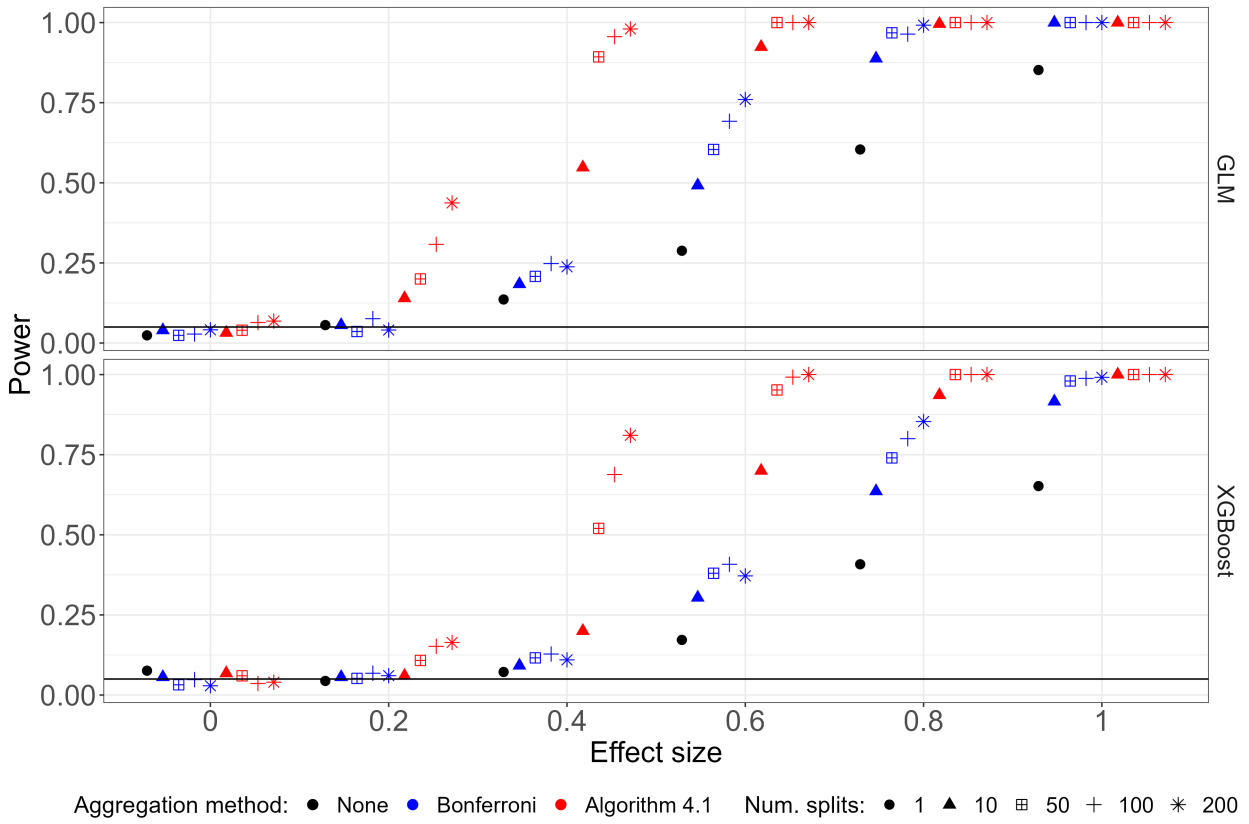

Figure C.2: Performance of sample splitting approaches for testing the hypothesis of zero importance using accuracy predictiveness in Scenario 2 ($p = 100$). When the effect size is zero, the null hypothesis holds. Oracle prediction functions were estimated using either probit regression (top row) or gradient-boosted trees (bottom row). The number of splitting iterations ranged from 1 to 200. Where more than one splitting iteration was used, results from multiple iterations were combined using either a Bonferroni correction (blue) or our proposed method, Algorithm 4.1 (red).

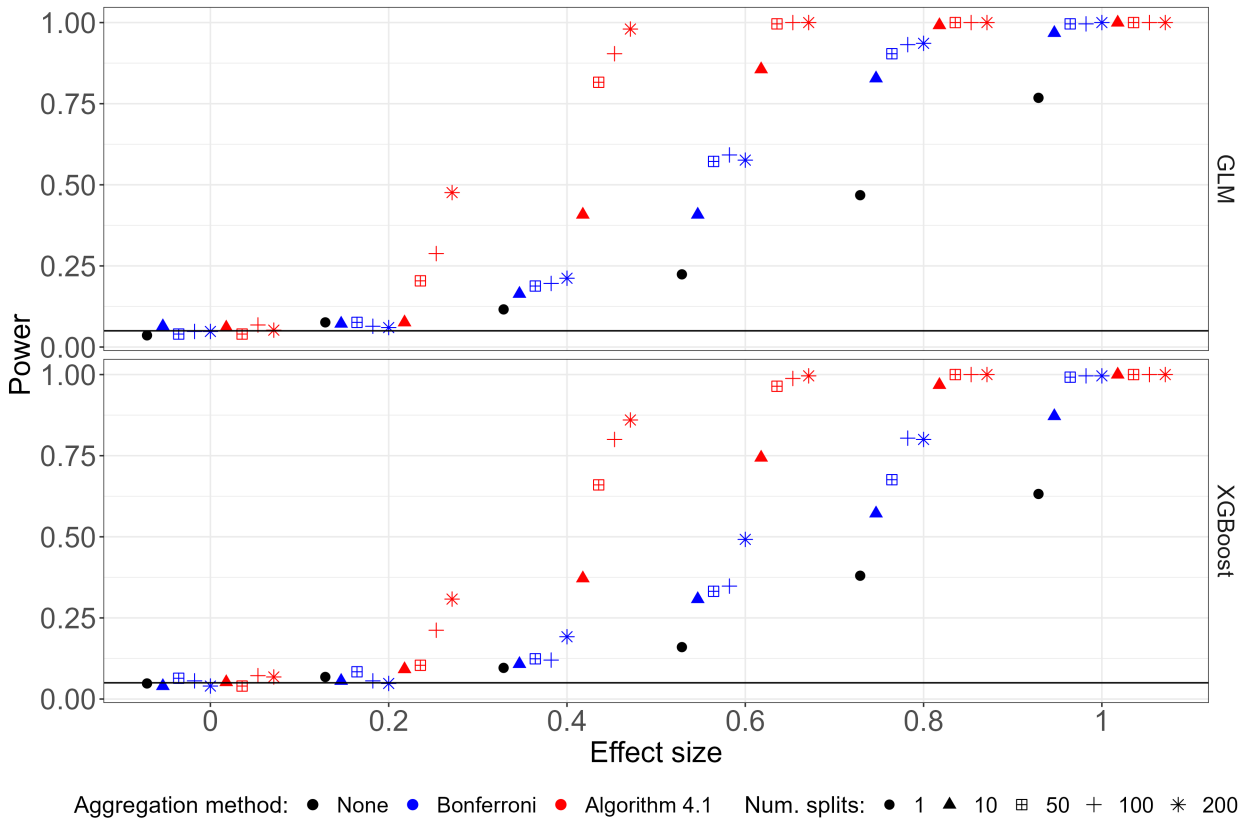

Figure C.3: Performance of sample splitting approaches for testing the hypothesis of zero importance using R^2 predictiveness in Scenario 1 ($p = 5$). When the effect size is zero, the null hypothesis holds. Oracle prediction functions were estimated using either probit regression (top row) or gradient-boosted trees (bottom row). The number of splitting iterations ranged from 1 to 200. Where more than one splitting iteration was used, results from multiple iterations were combined using either a Bonferroni correction (blue) or our proposed method, Algorithm 4.1 (red).

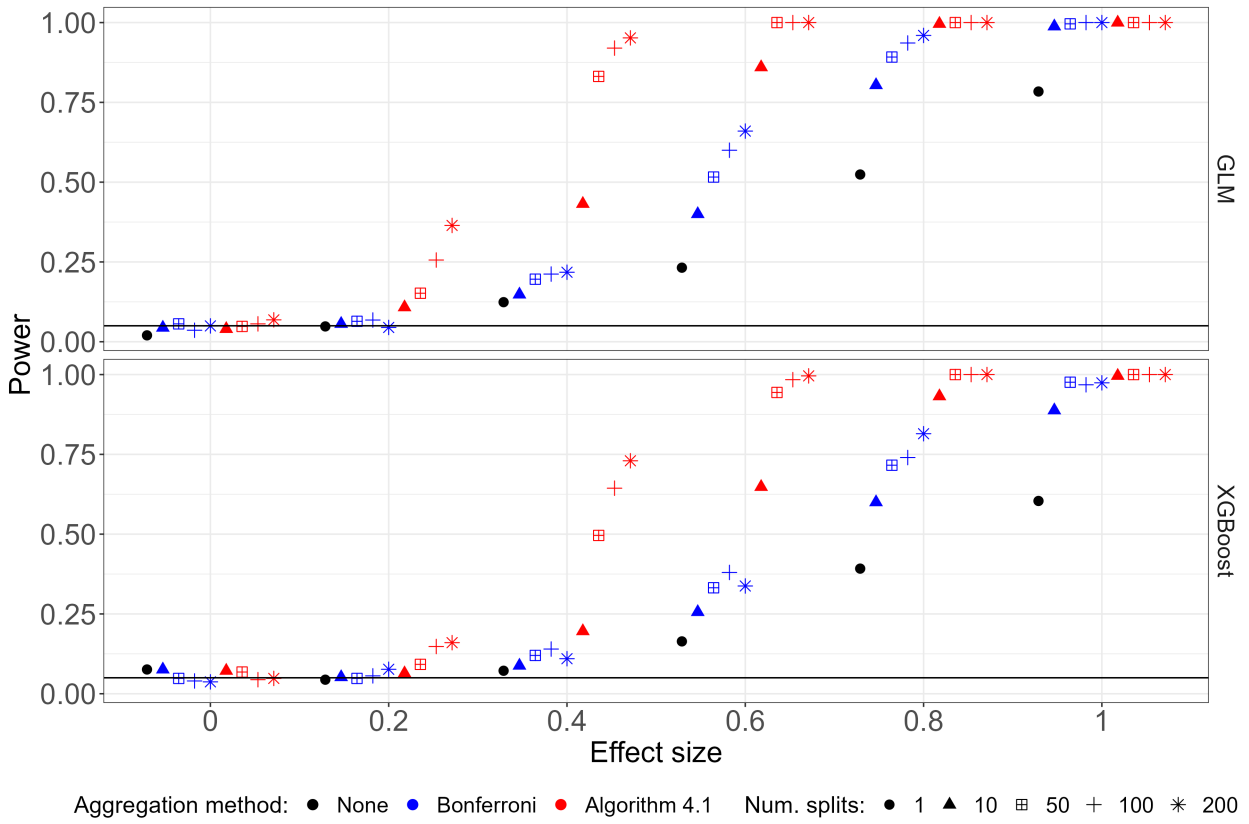

Figure C.4: Performance of sample splitting approaches for testing the hypothesis of zero importance using R^2 predictiveness in Scenario 2 ($p = 100$). When the effect size is zero, the null hypothesis holds. Oracle prediction functions were estimated using either probit regression (top row) or gradient-boosted trees (bottom row). The number of splitting iterations ranged from 1 to 200. Where more than one splitting iteration was used, results from multiple iterations were combined using either a Bonferroni correction (blue) or our proposed method, Algorithm 4.1 (red).