Statistical tools for immune correlates analysis of vaccine clinical trial data

Avi Kenny

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
Peter Gilbert
James Hughes

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical tools for immune correlates analysis
of vaccine clinical trial data

Avi Kenny

Chair of the Supervisory Committee:
Marco Carone
Department of Biostatistics

In vaccine research, it is important to identify biomarkers that can reliably predict vaccine efficacy relative to a clinical endpoint. Such biomarkers are known as immune correlates of protection (CoP) and can serve as surrogate endpoints in vaccine efficacy trials to accelerate the approval process. CoPs must be rigorously validated, and one method of doing so is through the controlled vaccine efficacy (CVE) curve, a function that represents the causal effect of the biomarker on population-level vaccine efficacy. In this dissertation, we propose and study two methods to estimate the CVE curve and construct pointwise confidence bands. The first method assumes a Cox proportional hazards model, allowing for possible nonlinearity in the additive predictor. The second method assumes monotonicity of the CVE curve and leverages modern tools from shape-constrained inference and nonparametric efficiency theory. We also develop a nonparametric test of the null hypothesis that the CVE curve is flat, and discuss extensions of these methods that lead to improved finite sample performance. Finally, we describe open-source software that we published to implement these methods, and apply the methods to data from several vaccine efficacy trials.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest gratitude and appreciation to all those who have supported and contributed to the completion of this dissertation. Without their guidance, encouragement, and assistance, this work would not have been possible.

First and foremost, I am profoundly indebted to my supervisor, Dr. Marco Carone, for his invaluable guidance throughout this research journey. His expertise, insightful feedback, and unwavering support have been instrumental in shaping this dissertation. I am truly grateful for his mentorship and the countless hours he dedicated to reviewing and refining my work.

I would also like to extend my sincere thanks to the members of my dissertation committee, Dr. Peter Gilbert, Dr. James Hughes, Dr. Patrick Heagerty, and Dr. Patricia Pavlinac. Their expertise, constructive criticism, and thought-provoking discussions greatly enriched the quality of this research. I am grateful for their time, dedication, and valuable suggestions, which have undoubtedly contributed to the enhancement of this dissertation.

I am grateful to the faculty and staff of the University of Washington for creating an enriching academic environment. The resources, facilities, and opportunities provided by the university have been essential in supporting my research endeavors.

My deepest gratitude goes to my mother Deborah, my late father Joel, my sisters Chava and Rachel, my partner Mallory, my grandparents, uncles, aunts, cousins, and the rest of my family for their unwavering love, encouragement, and understanding throughout this entire journey. Their support, belief in my abilities, and words of encouragement have been a constant source of motivation.

I would like to express my heartfelt appreciation to my friends and colleagues who provided

valuable insights, engaged in stimulating discussions, and offered their assistance whenever needed. Their friendship, encouragement, and collaboration have been instrumental in overcoming challenges and maintaining my motivation during this demanding process.

Thank you all for your unwavering support, guidance, and inspiration. This dissertation is a culmination of the collective efforts of many individuals, and I am humbled and honored to have had the privilege of working with each and every one of you.

Chapter 1

Introduction

The recent global outbreak of SARS-CoV-2, the coronavirus that causes COVID-19 disease, has highlighted the importance of developing, testing, and distributing safe and effective vaccines. In the context of a novel infectious pathogen, obtaining approval of a new vaccine typically requires conducting one or more large phase 3 randomized placebo-controlled trials with a clinically meaningful endpoint (e.g., moderate to severe COVID disease). However, placebo-controlled trials are typically resource-intensive and in some cases infeasible, and so it is desirable to have alternate modes of approval for the development of additional vaccines. In these situations, researchers have historically used immunological biomarkers as surrogate endpoints. To serve as a valid surrogate, a biomarker must be a correlate of protection (CoP), meaning that it can reliably be used to conduct accurate inference on vaccine efficacy against disease acquisition (Plotkin, 2010; Plotkin and Gilbert, 2012). CoPs are contrasted with correlates of risk (CoR), which are markers associated with but not necessarily predictive of the study endpoint. Aside from serving as a surrogate endpoint in a clinical trial, CoPs can be useful to assess efficacy of a vaccine for subpopulations that were underrepresented in or excluded from the original clinical trials (e.g., children) and to help approve modified versions of existing vaccines (e.g., a change in dosage or pathogen strain). However, to serve any of these purposes, a CoP must first be validated.

Historically, the validation of a CoP has involved the implementation of a number of statistical methods, including surrogate endpoint evaluation (Prentice, 1989; Molenberghs et al., 2008), prin-

cipal stratification vaccine efficacy moderation evaluation (Follmann, 2006; Moodie et al., 2018), mediation analysis (Cowling et al., 2019; Benkeser et al., 2021), and meta-analysis (Molenberghs et al., 2008). These methods are typically implemented across multiple phase 3 trials, and often multiple methods are used in a single analysis to add to the robustness of results. Recently, in Gilbert et al. (2022a), we proposed a novel method of CoP evaluation based on the controlled vaccine efficacy (CVE) curve, which represents the causal relationship between the biomarker and population-level vaccine efficacy. The calculation of the CVE curve involves a related function called the controlled risk (CR) curve, which represents the causal relationship between the biomarker and the population-level risk of experiencing the study endpoint by a certain time. This work builds off of previous research on controlled effects, dating back to Robins and Greenland (1992), and is closely related to the work of Joffe and Greene (2009), in which the authors consider the use of controlled effects for evaluating potential surrogate markers.

In Gilbert et al. (2022a), we suggested the use of a Cox proportional hazards model for estimation of the CR and CVE curves and a bootstrap-based approach for inference. However, while the Cox model may be a reasonable choice in many applications, there are several drawbacks of using the boostrap for inference. First, as is typical with resampling methods, it can be very computationally demanding. Second, the resampling procedure can vary between trials and can sometimes be quite complex, as it must account for the two-phase sampling structure (described in greater detail in section 2.1) often used in vaccine efficacy trials. Third, resampling with replacement systematically leads to ties in the event times, which is a known issue with the Cox partial likelihood (Lin and Zelterman, 2002); while a number of methods for dealing with ties are have been studied (e.g., Cox, 1972; Breslow, 1974; Efron, 1977), the decision of which method to use is somewhat subjective and can lead to bias (Xin et al., 2013).

Furthermore, incorporating the biomarker as a single variable in the Cox model linear predictor imposes a very strong assumption on the functional form of the CVE curve, and substantially restricts the set of shapes that this curve can take. It would be of practical use to extend the basic form of the model to allow for nonlinear relationships between the biomarker and the log-hazard function.

Beyond this, there are many settings in which the Cox model may not be appropriate, as the underlying linearity and proportional hazards assumptions are strong and do not necessarily hold in practice. To deal with these settings, we would like to explore nonparametric methods of estimation that allow for flexible estimation of various nuisance parameters, such as the conditional survival function of the outcome given the marker and baseline covariates. In particular, it is often reasonable to think that the relationship between the biomarker of interest and controlled vaccine efficacy is monotone; if this assumption holds, the methods of Westling et al. (2020) can be leveraged to construct an estimator that makes fewer assumptions about the underlying data-generating mechanism.

Finally, it is useful to quantify our confidence as to whether or not a marker is indeed a valid correlate of protection. In Gilbert et al. (2022a), a correlate of protection is defined as a marker (satisfying certain regularity conditions, as well as conditions related to causal inference) for which CVE increases as a function of the marker. As such, it is useful to derive a test of the null hypothesis that the CVE curve is flat.

The organization of this dissertation is as follows. In Chapter 2, we derive asymptotic results for the Cox model CR/CVE estimators, accounting for the use of estimated weights and two-phase sampling, and show how the basic form of the Cox model can be extended to handle a nonlinear relationship between the log hazard function and the biomarker of interest. We test the performance of the resulting models and estimators via simulation and illustrate the methods using data from the Coronavirus Efficacy (COVE) placebo-controlled phase 3 trial (NCT04470427) of the mRNA-1273 COVID-19 vaccine. In Chapter 3, we describe nonparametric CR/CVE estimators, derive a test statistic corresponding to null hypothesis of a flat CR curve, and derive the asymptotic properties of both the estimation procedure and the hypothesis test. We evaluate the performance of all methods via simulation and again apply them to the COVE phase 3 vaccine trial. In Chapter 4, we explore a possible solution to a bias issue related to the nonparametric estimator (and related to monotone-constrained estimators in general), conduct a simulation study to assess its benefits, and suggest possible routes for asymptotic analysis. In Chapter 5, we describe an open-source R software package that implements the methods developed in Chapters 2 and 3. Finally, in Chapter

6 we present select results from the application of our methods to three additional phase 3 clinical trials, to illustrate the versatility of these methods and to discuss interesting practical challenges that arose.

Chapter 2

Estimation and inference using a Cox proportional hazards model

2.1 Data structure and inferential goals

2.1.1 Data structure

Data come from a randomized controlled trial in which n population-representative individuals are assigned to the vaccine group and n' are assigned to the placebo group. Interest lies in summaries related to the time $T \in [0, \tau]$ of occurrence of an event of interest (e.g., the trial primary endpoint). However, due to right-censoring, investigators do not directly observe the event time T, but instead observe $Y \coloneqq \min\{T, C\}$, the minimum of the event time T and the right-censoring time $C \in [0, \tau]$, and $\Delta \coloneqq I(T \leq C)$, an indicator that the event occurred. Investigators also observe a vector X of baseline covariates for all individuals, and T is assumed to be conditionally independent of C given X. Of key interest is a univariate immunologic biomarker S, measured for a subsample of individuals in the vaccine group at a particular time post-vaccination and scaled to lie within $[0, 1]$. The biomarkers that are typically assessed have observed distributions that are influenced by the underlying measurement assays, which have associated lower limits of detection (LLOD) and upper limits of quantitation (ULOQ). Measurements below the LLOD cannot be distinguished from one another and have historically been assigned the value LLOD/2. Similarly, values above the ULOQ

are censored at the ULOQ value. Thus, the observed distribution of S is typically mixed, possibly with point masses at 0 (corresponding to the value LLOD/2) and 1 (corresponding to the ULOQ) and continuous in the open interval $(0, 1)$. For convenience, the notation $V \coloneqq (X, b(S))$ is used as shorthand for the entire set of covariates, where b is a user-specified basis function discussed in section 2.2.

A two-phase sampling structure is assumed in which investigators first observe data units $(X_1, Y_1, \Delta_1), ..., (X_n, Y_n, \Delta_n)$ in the vaccine group, known as the *phase-one sample*, and then construct a set of indicator variables $Z_1, ..., Z_n$ based on the phase-one data. S_i is observed for subject i if $Z_i = 1$; the set of individuals for whom $Z_i = 1$ is known as the *phase-two sample*. This sampling design is common in vaccine efficacy trials since the lab work required to obtain measurements of S is typically costly. It is assumed that the indicators $Z_1, ..., Z_n$ are sampled independently from a Bernoulli$\{\pi_0(X_i, Y_i, \Delta_i)\}$ distribution for a known deterministic function π_0 bounded below by some number $\epsilon_\pi > 0$. This yields a coarsened observed data structure that can be expressed as

$$O_1, ..., O_n \coloneqq (X_1, Y_1, \Delta_1, Z_1, Z_1 S_1), ..., (X_n, Y_n, \Delta_n, Z_n, Z_n S_n) \overset{iid}{\sim} P_0 \,. \tag{2.1}$$

In practice, a strategy known as *finite population stratified sampling* (FPSS) is often used instead of Bernoulli sampling (Breslow et al., 2009). With this strategy, the individuals in the phase-one sample are partitioned into a set of strata using a predetermined strategy that may depend on the observed data (Y_i, Δ_i, X_i). To represent this, let $C_i = c_0(X_i, Y_i, \Delta_i)$, where c_0 is a known deterministic function with range $\{1, 2, ..., d\}$ representing the strata indices, such that individuals i and j are in the same stratum if $C_i = C_j$. Next, a predetermined fraction of individuals are selected without replacement from each stratum, where the sampling fraction may depend on the stratum index, and again individual i is assigned $Z_i = 1$ if selected. Here, the notation $\pi_0(c) = \pi_0(c_0(x, y, \delta))$ is used to represent the probability that an individual is selected into the phase-two sample. Note that this breaks the iid structure of the data, since now Z_i and Z_j are dependent if $C_i = C_j$. To ease theoretical development, we proceed by assuming that (2.1) holds, such that the indicators Z are sampled independently from a Bernoulli distribution; in section 2.5, we justify the implications of this assumption.

2.1.2 Parameter of interest and causal identification

Using the potential outcomes framework of Rubin (2005), let $T(a)$ represent the potential outcome of T under assignment to group a, where $a = 1$ represents the vaccine group and $a = 0$ represents the placebo group. Similarly, let $T(a, s)$ represent the potential outcome under a hypothetical intervention that assigns individuals to both group a and biomarker value s. For a fixed time of interest t_0, the controlled risk is defined as $r_{C,0}(a, s) := P_0\{T(a, s) \leq t_0\}$, which represents the counterfactual probability of the event occurring by time t_0 under assignment to group a and biomarker level s. Note that the controlled risk depends on t_0, but for simplicity, this dependence is not reflected in the notation. The controlled risk curve is simply the function $s \mapsto r_{C,0}(1, s)$. In the special case in which assignment to the placebo group necessitates that $S = 0$, which often occurs for certain biomarkers when the population is naïve to the pathogen of interest, the CVE curve can be defined as

$$\text{CVE}_0 : s \mapsto 1 - \frac{r_{C,0}(1, s)}{r_{C,0}(0, 0)}. \tag{2.2}$$

Since in this case, we have $T(0) = T(0, 0)$ almost surely, it also holds that $r_{C,0}(0, 0) = P\{T(0) \leq t_0\}$. Essentially, $\text{CVE}_0(s)$ represents the population-level vaccine efficacy that would be observed under a hypothetical intervention that assigns the entire population both to the vaccine arm *and* to a specific biomarker level s. The setting in which the biomarker necessarily equals zero for the entire placebo arm is commonly encountered, such as in many COVID-19 and HIV-1 vaccine efficacy trials. For settings in which the biomarker can be nonzero in the placebo arm, the CVE curve can be generalized to a parameter called the CVE surface (Gilbert et al., 2022a), but consideration of this parameter is outside the scope of the current work.

Gilbert et al. (2022a) provide a set of conditions under which the controlled risk can be identified with the marginalized risk $r_{M,0}(a, s) := E_0\{P_0(T \leq t_0 \mid X, S = s, A = a)\}$ in the vaccine arm, where the expectation is over the marginal distribution of X. Briefly, for a fixed value s, suppose that the Stable Unit Treatment Value Assumption (SUTVA) holds, the covariates X sufficiently deconfound the causal effect of S on T, and the conditional density of S given $(X, A = 1)$ at $S = s$ is positive

almost surely (i.e., the positivity assumption holds). Then $r_{C,0}(1,s) = r_{M,0}(1,s)$. Given this result, we proceed by focusing attention on estimation and inference for $r_{M,0}(1,s)$; for convenience, the shorthand $r_{M,0}(s) \coloneqq r_{M,0}(1,s)$ is used. Our goal is to estimate the functions $s \mapsto r_{M,0}(s)$ and $s \mapsto \mathrm{CVE}_0(s)$ over $s \in [0,1]$ and form corresponding pointwise confidence bands.

2.2 Methods

2.2.1 Definition of proposed estimator

Following Gilbert et al. (2022a), it is assumed that the true full-data distribution follows a Cox model, and an inverse-probability-of-sampling (IPS) weighted Cox model (Prentice, 1986) is used to estimate the conditional survival function $Q_0(t \,|\, x,s) \coloneqq P_0(T > t \,|\, X = x, S = s, A = 1)$ in the vaccine arm. This involves maximizing the IPS-weighted sum of the Cox model partial log-likelihood contributions, restricted to the phase-two observations, yielding an estimator $\beta_n \coloneqq (\beta_{x,n}, \beta_{s,n})$ of the Cox model parameter vector $\beta_0 \coloneqq (\beta_{x,0}, \beta_{s,0})$. This in turn enables the construction of the estimator

$$Q_n(t \,|\, x,s) \coloneqq \exp\left[-\exp\left\{\beta'_{x,n}x + \beta'_{s,n}b(s)\right\}\Lambda_n(t)\right] .$$

Above, b is a user-specified function (either scalar-valued or vector-valued) that is typically chosen to be the identity function, as in the model considered in Gilbert et al., 2022a, but can take on other forms, such as one of the functions considered in section 2.2.2. Λ_n is the IPS-weighted version of the so-called Breslow estimator of the baseline cumulative hazard function Λ_0, defined as

$$\Lambda_n(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i \Delta_i I(Y_i \leq t)}{\pi_n(C_i) S_n(Y_i)} , \tag{2.3}$$

where

$$S_n(y) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i I(Y_i \geq y) e^{V_i'\beta_n}}{\pi_n(C_i)} ,$$

and where $\pi_n(C_i)$ is an estimator of the probability $\pi_0(C_i)$ of phase-two selection that will be defined later in this section. It is necessary to apply the inverse weights to (2.3), even though the sum does not involve S_i terms, to avoid terms of the form $1/0$ in the sum. The estimator we consider is defined as one minus the conditional survival function estimator marginalized over the observed covariate distribution in the vaccine arm, which can be written as

$$r_{M,n}(s) \coloneqq 1 - \frac{1}{n} \sum_{i=1}^{n} Q_n(t_0 \mid X_i, s) \,. \tag{2.4}$$

While the Cox model is estimated using information from the phase-two sample, the marginalization can be done over the covariate distribution from the entire phase-one sample for greater efficiency. Finally, combining (2.4) with an estimator $r_{M,n}(0,0)$ of the placebo group risk $r_{M,0}(0,0)$, an estimator of $CVE_0(s)$ can be defined as

$$\mathrm{CVE}_n(s) \coloneqq 1 - \frac{r_{M,n}(s)}{r_{M,n}(0,0)} \,.$$

Constructing an estimator of the placebo group risk is a well-studied problem, so we do not focus on it here. The only restriction placed on the estimator $r_{M,n}(0,0)$ is that it must be asymptotically linear and constructed from the placebo group data only; one example of such an estimator is a marginalized Cox model conditional risk function, similar in construction to (2.4) but not involving the biomarker.

Next, we describe how the IPS weights are constructed. Since the probability function π_0 is typically under the control of the investigators, one could in theory use the weights $Z_1/\pi_0(C_1), ..., Z_n/\pi_0(C_n)$. However, efficiency can be gained by using weights based on an estimator π_n of π_0 defined as

$$\pi_n(c) \coloneqq \frac{\sum_{j=1}^{n} I(c = C_j) Z_j}{\sum_{j=1}^{n} I(c = C_j)} \,. \tag{2.5}$$

The expression given in (2.5) represents the fraction of individuals in stratum C_i that were selected into phase-two sample. It is assumed that at least one person is always selected in each stratum and that the estimated weights are bounded below by ϵ_π (as was also the case for the true function π_0).

It may seem counterintuitive that estimating a known function could lead to increased efficiency, but in fact this is well-documented behavior (see, for example, Robins et al., 1994; Scott and Wild, 1997; Qi et al., 2005). In addition to using the IPS weights $Z_1/\pi_n(C_1), ..., Z_n/\pi_n(C_n)$ based on (2.5) to estimate the Cox model parameters, these weights will repeatedly be used to estimate terms of the form $E_0\{h(X, Y, \Delta, S)\}$ using the IPS-weighted estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i h(X_i, Y_i, \Delta_i, Z_i S_i)}{\pi_n(C_i)}\,, \tag{2.6}$$

where $h(x, y, \delta, s)$ is an arbitrary fixed function of an individual observation.

2.2.2 Allowing for nonlinearity in the linear predictor

We consider Cox models involving hazard functions of the form

$$\lambda_0(t \,|\, x, s) = \lambda_0(t) \exp\left\{\beta'_{x,0}x + \gamma_0(s)\right\},$$

where γ_0 is an unknown function. The usual choice is to assume that $\gamma_0(s) = \beta_{s,0}s$ for some unknown scalar parameter $\beta_{s,0}$, such that the log of the hazard function $\lambda_0(t \,|\, x, s)$ is a linear function of s; we refer to this as the *basic Cox model*. As this is fairly restrictive, it is worth exploring options that allow for this assumption to be relaxed. As a first attempt at this, we could consider adding polynomial terms of s to the linear predictor. For example, adding $d-1$ additional polynomial terms to the model would result in the form

$$\gamma_0(s) = \beta_{s,0,1}s + \beta_{s,0,2}s^2 + ... + \beta_{s,0,d}s^d\,, \tag{2.7}$$

where $\beta_{s,0} \coloneqq (\beta_{s,0,1}, \beta_{s,0,2}, ..., \beta_{s,0,d})$ is now a d-dimensional vector rather than a scalar. This method is advantageous in its simplicity, and in the ease of constructing the polynomial terms. However, a known issue with including polynomial terms in a regression linear predictor is that changes in the observed values of s at a fixed point s_0 can have a major influence on the shape of the curve for values of s that are not close to s_0.

An alternative solution that mitigates this issue is the use of regression splines. There are many different types of regression splines, each with their own advantages and disadvantages, but here, we consider the use of natural cubic splines; for an overview of regression splines, see Hastie et al. (2009). Briefly, for a fixed a set of real numbers $k_1 < ... < k_d$ called *knots*, a cubic spline is defined as a function that is equivalent to a cubic polynomial between any two adjacent knots (k_j, k_{j+1}). A natural cubic spline is further constrained such that it is continuous, twice continuously differentiable, and linear to the left of the first knot and to the right of the last knot. These constraints ensure a degree of smoothness of the resulting curve and decrease the number of corresponding model parameters. A natural cubic spline can be embedded within a Cox model through the construction of a *spline basis* – a set of functions $b_1, ..., b_d$, each of which is applied to to the variable of interest s, thus creating d additional variables in the dataset. This results in a model of the form

$$\gamma_0(s) = \beta'_{s,0} b(s) \coloneqq \beta_{s,0,1} b_1(s) + \beta_{s,0,2} b_2(s) + ... + \beta_{s,0,d} b_d(s)\,. \tag{2.8}$$

A natural cubic spline with $d+1$ knots can be represented with $d+1$ basis functions. However, for identifiability, and to maintain consistency of the value of $\lambda_0(t \,|\, s = 0, x)$ between models (2.7) and (2.8), we impose the constraint $b_1(0) + b_2(0) + ... + b_d(0) = 0$, such that the spline passes through the origin; this decreases the number of required basis functions by one for a fixed number of knots.

Because of the advantages of natural cubic splines over polynomials, we focus on model (2.8) moving forward. The basic Cox model is a submodel of (2.8), and so it is strictly more flexible. This model requires a choice for d, which can be seen as the number of degrees of freedom in the conditional hazard as a function of s. There are ways of obtaining an optimal choice for the value of d in a data-dependent manner, which typically involves the specification of a loss function and some form of sample splitting. However, this makes inference more cumbersome, and can reduce statistical power. Instead, we suggest pre-specifying a reasonable value of d; this is discussed further in section 2.5. If in addition, the knots are chosen to be equally-spaced across the unit interval then the functions $b_1, ..., b_d$ can be determined completely in advance. Given this strategy, we proceed by assuming that $\gamma_0(s) = \beta'_{s,0} b(s)$ for a known vector-valued basis function b, such that the conditional

hazard function can be written concisely as

$$\lambda_0(t \,|\, x, s) = \lambda_0(t) \exp\left(\beta'_{x,0} x + \beta'_{s,0} b(s)\right).$$

As mentioned in section 2.1.1, the observed distribution of S sometimes has a point mass at zero corresponding to one-half the marker LLOD. In these situations, it may be useful to further modify model (2.8) by adding the indicator variable $I(s = 0)$ to the sum of terms. This allows for more flexibility in terms of the estimation of CVE(0), which is particularly useful in applications in which there is a large gap in the observed marker distribution between the LLOD/2 value and the next largest observed value. This strategy may additionally allow for more accurate estimation of $\text{CVE}_0(0)$, which is often of direct interest in itself. One could also consider adding this indicator variable to the basic Cox model to make it slightly more flexible.

2.2.3 Asymptotic properties

Our strategy for estimating $\text{Var}\left\{r_{M,n}(s)\right\}$ is to derive the influence functions of β_n and $\Lambda_n(t_0)$, apply the delta method to find the influence function of $Q_n(t_0 \,|\, x, s)$ for fixed (x, s), and then account for the marginalization over the distribution of X to derive the influence function of $r_{M,n}(s)$. To begin, note that the estimator given in (2.6) is not asymptotically linear, since the $\pi_n(C_i)$ terms depend on the entire sample within a given stratum. Therefore, we make use of the first-order approximation given in Lemma 1:

Lemma 1. *For an arbitrary real-valued bounded fixed function $h(o) \coloneqq h(x, y, \delta, s)$ of an individual observation, if $E_0[\{h(O)\}^2 \,|\, Z = 1, C = c] < \infty$ for each stratum $c \in \{1, 2, ..., d\}$, then it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Z_i h(O_i)}{\pi_n(C_i)} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i h(O_i)}{\pi_0(C_i)} + \left\{1 - \frac{Z_i}{\pi_0(C_i)}\right\} E_0\left\{h(O) \,|\, Z = 1, C = C_i\right\}\right] + o_P(n^{-1/2}),$$

where $h(o_i) \coloneqq h(x_i, y_i, \delta_i, \delta_i s_i)$.

Next, note that the full-data influence function of the Cox model parameter vector estimator is given by $\tilde{\ell}_0 \coloneqq \mathcal{I}_0^{-1} \ell_0$, where $\mathcal{I}_0$ is the efficient information and ℓ_0 is the efficient score. To provide expressions for these quantities, it is useful to first define

$$
\begin{aligned}
S_0(y) &:= E_0\left\{e^{V'\beta_0} I(Y \geq y)\right\}, \\
\dot{S}_0(y) &:= E_0\left\{V e^{V'\beta_0} I(Y \geq y)\right\}, \\
\ddot{S}_0(y) &:= E_0\left\{V^{\otimes 2} e^{V'\beta_0} I(Y \geq y)\right\}, \\
m_0(y) &:= \dot{S}_0(y)/S_0(y),
\end{aligned}
\tag{2.9}
$$

where we recall that $V := (X, b(S))$ and where the notation $v^{\otimes 2} := vv'$ is used. The semiparametric efficient score and information are then defined as

$$
\begin{aligned}
\ell_0(o_i) &:= \delta_i\{v_i - m_0(y_i)\} - e^{v_i'\beta_0} \int_0^{y_i} \{v_i - m_0(x)\} d\Lambda_0(x), \\
\mathcal{I}_0 &:= E_0\left[e^{V'\beta_0} \int_0^{\tau} \{V - m_0(u)\}^{\otimes 2} I(Y \geq u) d\Lambda_0(u)\right].
\end{aligned}
$$

According to equation (19) of Breslow and Wellner (2007) it holds that

$$
\beta_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i}{\pi_n(C_i)} \tilde{\ell}_0(O_i) + o_P(n^{-1/2}).
$$

Applying Lemma 1 to this equation, it holds that

$$
\beta_n - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{Z_i}{\pi_0(C_i)} \tilde{\ell}_0(O_i) + \left\{1 - \frac{Z_i}{\pi_0(C_i)}\right\} E_0\left\{\tilde{\ell}_0(O) \,|C = C_i, Z = 1\right\}\right] + o_P(n^{-1/2}).
$$

Thus, the influence function of β_n under two-phase sampling with estimated weights is given by

$$
\varphi_{\beta,0} : o_i \mapsto \frac{z_i}{\pi_0(c_i)} \tilde{\ell}_0(o_i) + \left\{1 - \frac{z_i}{\pi_0(c_i)}\right\} E_0\left\{\tilde{\ell}_0(O) \,|C = c_i, Z = 1\right\}. \tag{2.10}
$$

Putting this aside, the following theorem is stated, which allows us to study the asymptotic behavior of the IPS-weighted Breslow estimator

Theorem 1. *Consider the estimator given in (2.3) and define*

$$\varphi_{\Lambda,0} : (o_i, t) \mapsto \frac{z_i \delta_i I(y_i \leq t)}{\pi_0(c_i) S_0(y_i)} + \mu_0(t)\varphi_{\beta,0}(o_i) + \nu_{2,0}(o_i, t) - \int_0^t \left\{ \frac{z_i I(y_i \geq x) e^{v_i' \beta_0} + \pi_0(c_i)\nu_{1,0}(o_i, x)}{\pi_0(c_i) S_0(x)} \right\} d\Lambda_0(x) \,, \tag{2.11}$$

where

$$\begin{aligned} \nu_{1,0}(o_i, y) &:= \left\{ 1 - \frac{z_i}{\pi_0(c_i)} \right\} E_0 \left\{ I(Y \geq y) e^{V'\beta_0} \,|C = c_i, Z = 1 \right\} , \\ \nu_{2,0}(o_i, t) &:= \left\{ 1 - \frac{z_i}{\pi_0(c_i)} \right\} E_0 \left\{ \frac{\Delta I(Y \leq t)}{S_0(Y)} \,|C = c_i, Z = 1 \right\} , \\ \mu_0(t) &:= \frac{\partial}{\partial \beta} \Lambda_0(t, \beta) \Big|_{\beta = \beta_0} = - \int_0^t m_0(u) d\Lambda_0(u) \,. \end{aligned}$$

and where S_0 and $\varphi_{\beta,0}$ are defined in (2.9) and (2.10), respectively. If there exists an $\epsilon > 0$ such that $\sup_{\beta : |\beta - \beta_0| < \epsilon} E(|V|^2 e^{2\beta' V} \,|\, Z = z, C = c) < \infty$ for each $(z, c) \in \{0, 1\} \times \{1, ..., d\}$, then for any fixed $t \in (0, \tau)$, it holds that

$$\Lambda_n(t) - \Lambda_0(t) = \frac{1}{n} \sum_{i=1}^n \varphi_{\Lambda,0}(O_i, t) + o_P(n^{-1/2}) \,.$$

Using several applications of the delta method on the influence functions given by equations (2.10) and (2.11), it holds that

$$Q_n(t_0 \,|\, v) - Q_0(t_0 \,|\, v) = \frac{1}{n} \sum_{i=1}^n \varphi_{Q,0}(O_i, v) + o_P(n^{-1/2}) \,,$$

where v is a fixed covariate vector and

$$\varphi_{Q,0} : (o_i, v) \mapsto -Q_0(t_0 \,|\, v) e^{v'\beta_0} \left\{ \Lambda_0(t_0) v' \varphi_{\beta,0}(o_i) + \varphi_{\Lambda,0}(o_i, t_0) \right\} .$$

Finally, building off of this result, the following theorem allows for asymptotic study of our estimator of interest.

Theorem 2. *Consider the estimator given in (2.4) and define*

$$\varphi_{r,0} : (o_i, s) \mapsto E_0 \left\{ Q_0(t_0 \,|\, X, s) \right\} - Q_0(t_0 \,|\, x_i, s) - E_0 \left[\varphi_{Q,0} \left\{ o_i, (X, s) \right\} \right] . \tag{2.12}$$

If the conditions of Theorem 1 hold and the support of V is contained in some compact set, then for any fixed $s \in [0, 1]$, it holds that

$$r_{M,n}(s) - r_{M,0}(s) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{r,0}(O_i, s) + o_P(n^{-1/2}) .$$

2.2.4 Variance estimation

To estimate $\mathrm{Var}\{r_{M,n}(s)\}$, an estimator of the influence function given in (2.12) must be constructed. First, several component function estimators are defined:

$$\begin{aligned}
\dot{S}_n(y) &\coloneqq \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i V_i I(Y_i \geq y) e^{V_i' \beta_n}}{\pi_n(C_i)} , \\
\ddot{S}_n(y) &\coloneqq \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i V_i^{\otimes 2} I(Y_i \geq y) e^{V_i' \beta_n}}{\pi_n(C_i)} , \\
m_n(y) &\coloneqq \dot{S}_n(y) / S_n(y) .
\end{aligned}$$

An estimator $\tilde{\ell}_n \coloneqq \mathcal{I}_n^{-1} \ell_n$ of $\tilde{\ell}_0$ can then be defined, where $\mathcal{I}_n$ and ℓ_n are defined as

$$\begin{aligned}
\mathcal{I}_n &\coloneqq \frac{1}{n} \sum_{\{i:\Delta_i = 1\}} \frac{Z_i}{\pi_n(C_i)} \left[\frac{\ddot{S}_n(Y_i)}{S_n(Y_i)} - \{m_n(Y_i)\}^{\otimes 2} \right] , \\
\ell_n(o_i) &\coloneqq \delta_i \left\{ v_i - m_n(y_i) \right\} - \frac{1}{n} \sum_{\{j:\Delta_j = 1\}} \frac{Z_j e^{v_i' \beta_n} I(Y_j \leq y_i) \left\{ v_i - m_n(Y_j) \right\}}{\pi_n(C_j) S_n(Y_j)} .
\end{aligned}$$

The influence function $\varphi_{\beta,0}$ given in (2.10) can then be estimated using

$$\varphi_{\beta,n}(o_i) \coloneqq \frac{z_i}{\pi_n(c_i)} \tilde{\ell}_n(o_i) + k_n(o_i) \sum_{\{j:C_j = c_i, Z_j = 1\}} \tilde{\ell}_n(O_j) ,$$

where

$$k_n(o_i) \coloneqq \left\{1 - \frac{z_i}{\pi_n(c_i)}\right\} \left\{\sum_{j=1}^{n} I(C_j = c_i) Z_j\right\}^{-1} .$$

The influence function $\varphi_{\Lambda,0}$ given in (2.11) can be estimated using

$$\varphi_{\Lambda,n}(o_i, t) \coloneqq \frac{z_i \delta_i I(y_i \leq t)}{\pi_n(c_i) S_n(y_i)} + \mu_n(t) \varphi_{\beta,n}(o_i) + \nu_{2,n}(o_i, t) \\ - \frac{1}{n} \sum_{\{j:\Delta_j=1\}} \frac{Z_j I(Y_j \leq t) \left\{ z_i I(y_i \geq Y_j) e^{v_i' \beta_n} + \pi_n(c_i) \nu_{1,n}(o_i, Y_j) \right\}}{\pi_n(c_i) \pi_n(C_j) \left\{ S_n(Y_j) \right\}^2} ,$$

which involves the nuisance estimators

$$\nu_{1,n}(o_i, y_j) \coloneqq k_n(o_i) \sum_{\{k:C_k=c_i, Z_k=1\}} I(Y_k \geq y_j) e^{V_k' \beta_n} ,$$
$$\nu_{2,n}(o_i, t) \coloneqq k_n(o_i) \sum_{\{k:C_k=c_i, Z_k=1\}} \frac{\Delta_k I(Y_k \leq t)}{S_n(Y_k)} ,$$
$$\mu_n(t) \coloneqq -\frac{1}{n} \sum_{\{j:\Delta_j=1\}} \frac{Z_j I(Y_j \leq t) m_n(Y_j)}{\pi_n(C_j) S_n(Y_j)} .$$

The influence function $\varphi_{Q,0}$ can be estimated using

$$\varphi_{Q,n}(o_i, v) \coloneqq -Q_n(t_0 \,|\, v) e^{v' \beta_n} \left\{ \Lambda_n(t_0) v' \varphi_{\beta,n}(o_i) + \varphi_{\Lambda,n}(o_i, t_0) \right\} .$$

Finally, the influence function $\varphi_{r,0}$ can be estimated using

$$\varphi_{r,n}(o_i, s) \coloneqq Q_n(t_0 \,|\, x_i, s) + \frac{1}{n} \sum_{j=1}^{n} \left[\varphi_{Q,n} \left\{ o_i, (X_j, s) \right\} - Q_n(t_0 \,|\, X_j, s) \right] .$$

The variance of $r_{M,n}(s)$ can be estimated by estimating the second moment of its influence function (2.12), as usual. However, directly evaluating this can be computationally intense, as it involves a double-sum over the entire dataset, which must be done separately for each value of s. To ease computation, define

$$K_{1,n}(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} Q_n(t_0 \,|\, X_i, s)\,,$$
$$K_{2,n}(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} Q_n(t_0 \,|\, X_i, s) e^{(X_i,s)'\beta_n}\,,$$
$$K_{3,n}(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} Q_n(t_0 \,|\, X_i, s) e^{(X_i,s)'\beta_n}(X_i, s)\,.$$

Then it holds that

$$\begin{aligned}\sigma^2_{r,n}(s) &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\left\{\varphi_{r,n}(O_i, s)\right\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[Q_n(t_0 \,|\, X_i, s) - \Lambda_n(t_0)\left\{K_{3,n}(s)\right\}'\varphi_{\beta,n}(O_i) - K_{2,n}(s)\varphi(O_i, t_0) - K_{1,n}(s)\right]^2 .\end{aligned}$$

While this still requires separate calculation for each value of s, this form eliminates the double-sum, therefore dramatically reducing computation time.

Since $r_{M,0}(s) \in [0,1]$, it may be useful to form $(1-\alpha)$-level confidence intervals on the $\text{logit}\{r_{M,0}(s)\}$ scale and then transform the limits back to the $r_{M,0}(s)$ scale. This yields estimated confidence interval limits of the form

$$\xi^{-1}\left[\xi\{r_{M,n}(s)\} \pm n^{-1/2} z_{\alpha/2}\dot{\xi}\left\{r_{M,n}(s)\right\}\sigma_{r,n}(s)\right],$$

where $\xi(x) \coloneqq \text{logit}(x)$, $\dot{\xi}(x) \coloneqq \frac{d}{dx}\text{logit}(x)$, $\xi^{-1}(x) \coloneqq \text{expit}(x)$, and $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard Normal distribution.

To form confidence intervals for $\text{CVE}_n(s)$, first note that the estimators $r_{M,n}(s)$ and $r_{M,n}(0,0)$ are independent, since the former is constructed using only the vaccine group data and the latter is constructed using only the placebo group data. Suppose that our estimator $r_{M,n}(0,0)$ of the placebo group risk is asymptotically normal with limiting variance $\sigma^2_{P,0}$, estimated by $\sigma^2_{P,n}$. Since $\text{CVE}_0(s) \in (-\infty, 1)$ for any s, it is desirable to form $(1-\alpha)$-level confidence intervals on the $\log\{1 - \text{CVE}_0(s)\}$ scale and then transforming back, leading to limits of the form

$$1 - \exp\left[\log\left\{\frac{r_{M,n}(s)}{r_{M,n}(0,0)}\right\} \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{\sigma^2_{r,n}(s)}{\{r_{M,n}(s)\}^2} + \frac{\sigma^2_{P,n}(s)}{\{r_{M,n}(0,0)\}^2}}\right] .$$

2.3 Simulation study

2.3.1 Simulation study methods

To assess the performance of the estimation and inference procedure described in section 2.2, we conducted a simulation study. Two baseline covariates (X_1, X_2) were sampled, $X_1 \sim$ Bernoulli(0.5) and

$X_2 \sim$ Uniform$\{0.0, 0.1, ..., 1.0\}$. The biomarker was sampled from one of two distributions, either a standard uniform distribution or a Normal distribution with mean 0.5 and standard deviation 0.2 (truncated to lie within $[0, 1]$). The survival time T was generated via the method of Bender et al. (2005) according to a Cox model, with the baseline hazard function following a Weibull distribution with scale parameter $\lambda_1 = 2 \times 10^{-4}$ and shape parameter $v_1 = 1.5$. The linear predictor of the Cox model followed one of three different forms, which differ in terms of the shape of the relationship between the biomarker S and the resulting log hazard function, leading to conditional survival functions that had one of the following forms:

$$\begin{aligned}
Q_0^{(1)}(t \,|\, x, s) &\coloneqq \exp\left\{-\lambda_1 t^{v_1} \exp(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 s)\right\} , \\
Q_0^{(2)}(t \,|\, x, s) &\coloneqq \exp\left\{-\lambda_1 t^{v_1} \exp(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 s^3)\right\} , \\
Q_0^{(3)}(t \,|\, x, s) &\coloneqq \exp\left\{-\lambda_1 t^{v_1} \exp(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \text{expit}(20s - 10))\right\} ,
\end{aligned}$$

where $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, -2)$. The censoring variable C followed a Cox model involving a Weibull baseline conditional hazard function with parameters $\lambda_2 = 5 \times 10^{-5}$ and $v_2 = 1.5$; the full conditional censoring function is given by

$$Q_0^C(t \,|\, x, s) \coloneqq \exp\left\{-\lambda_2 t^{v_2} \exp(\alpha_1 x_1 + \alpha_2 x_2)\right\} .$$

Using the variables T and C, we generated the observed survival variables Y and Δ. We then used

the function

$$(x, y, \delta) \mapsto \delta I(y \leq t_0) + \{1 - \delta I(y \leq t_0)\}\text{expit}(x_1 + x_2 - 1)$$

to generate the two-phase sampling indicator Z. The strata variable was defined by the 22 unique combinations of the two baseline covariates, allowing us to calculate estimated weights for each stratum. The size of the phase-one sample was 1,000 individuals and the time of interest was $t_0 = 200$.

In all simulations, the marginalized risk was estimated using three estimators, a basic Cox model (i.e., with the marker modeled as a linear term), a Cox model including an embedded natural cubic spline basis with four degrees of freedom (as per the specification given in section 2.2.2), and a similar spline model but using a basis with eight degrees of freedom. Knots were chosen to be equally spaced with respect to the observed quantiles of the distribution, with the boundary knots at the 5% and 95% quantiles. Although technically this is a data-driven choice, we expect the detrimental effects on inference to be negligible as long as the spline is fully pre-specified; see section 2.5. Performance of the three Cox model variants was compared in terms of bias, variance, and 95% confidence interval coverage for estimating the function $s \mapsto r_{M,0}(s)$. Simulations were written using the R programming language and structured using the *SimEngine* simulation framework (Kenny and Wolock, 2021). We ran 1,000 simulation replicates for every level combination.

2.3.2 Simulation study results

To get a sense of the behavior of the three estimators, it is useful to visualize sample paths of the estimated marginalized risk functions plotted against the true functions. This is provided in Figure 2.1.

Based on these plots and on our understanding of parametric models, we expect to see a bias-variance trade-off as our models increase in flexibility, from the basic Cox model to the spline model with eight degrees of freedom. In Figure 2.2, the bias of the three estimators is displayed for all data-generating mechanisms. As expected, all estimators are unbiased when the true relationship

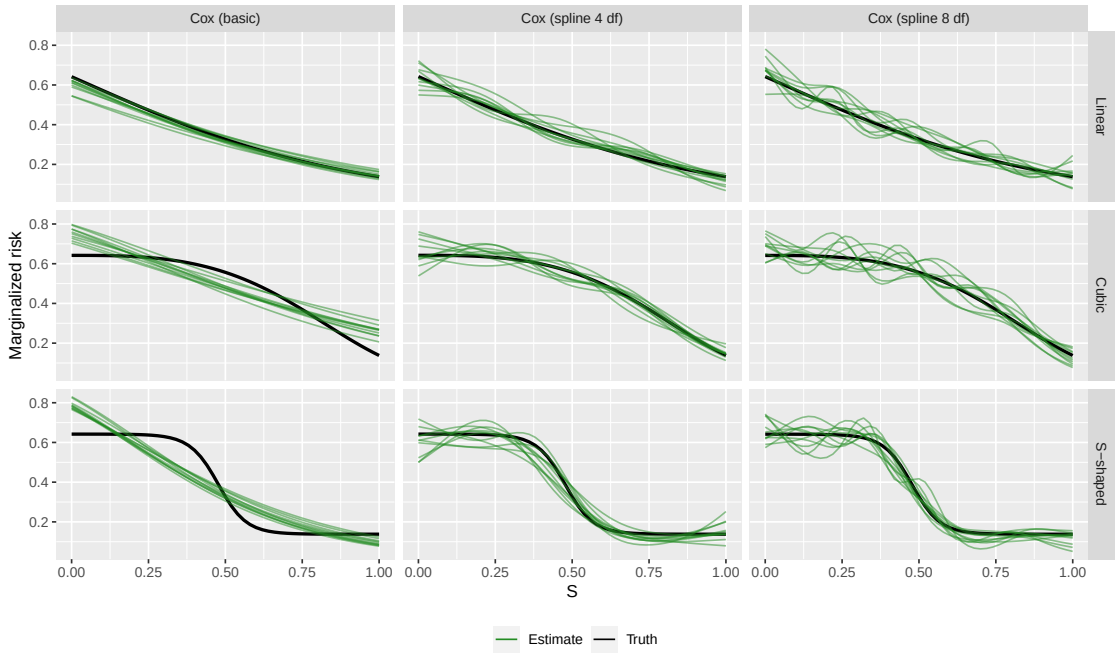


Figure 2.1: Sample paths of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (spline 8 df)" = Cox model with a spline with 8 degrees of freedom), displayed for the standard uniform marker distribution and three conditional survival functions ("Linear" = $Q_0^{(1)}$, "Cubic" = $Q_0^{(2)}$, "S-shaped" = $Q_0^{(3)}$). For each scenario, ten sample paths are displayed for each estimator. The estimated marginalized risk functions are plotted in green and the true functions are plotted in black.

between the biomarker and the log hazard function is linear. When the relationship is cubic, there is bias in the basic Cox model estimator but little bias for both spline-based estimators. When the relationship is S-shaped, the basic Cox model has substantial bias, the spline model with 4 degrees of freedom has a moderate amount of bias, and the spline model with 8 degrees of freedom has very little bias. This is all consistent with our expectations, as the spline models are specifically designed to detect nonlinear relationships.

Figure 2.3 displays 95% confidence interval coverage of the three estimators. As expected, for the data-generating mechanisms with a linear relationship, coverage is close to 95% for all estimators. When the basic Cox model is incorrect, coverage falls as low as 0% for some values of s in both the cubic mechanism and the S-shaped mechanism. The degree of undercoverage also depends on the marginal distribution of the biomarker. The spline model with 8 degrees of freedom achieves

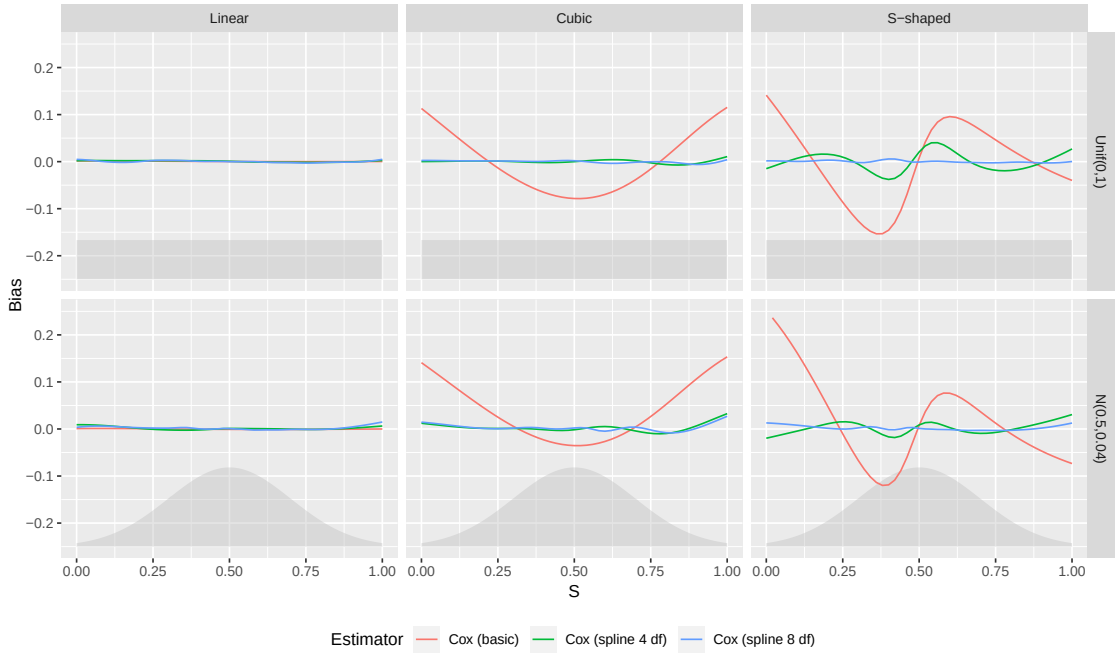


Figure 2.2: Bias of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (spline 8 df)" = Cox model with a spline with 8 degrees of freedom), displayed for three conditional survival functions ("Linear" = $Q_0^{(1)}$, "Cubic" = $Q_0^{(2)}$, "S-shaped" = $Q_0^{(3)}$) and two distributions of S.

excellent coverage across all scenarios, and the spline model with 4 degrees of freedom performs fairly well overall, although there are some areas with undercoverage in the S-shaped mechanism.

Figure 2.4 displays the standard deviations of all three estimators. As expected, as the model increases in complexity from the basic Cox model to the spline model with 8 degrees of freedom, the standard deviation of the estimator increases. This is consistent with the expected bias-variance trade-off. The standard deviation of both spline models also increases substantially for values of s close to 0 and 1.

Figure 2.5 shows that variance estimation appears highly accurate for sufficiently large sample sizes ($n \geq 800$) and that variance is underestimated for low sample sizes ($n = 100$ and $n = 200$). Surprisingly, variance is accurately estimated using the basic Cox model even for the Cubic and S-shaped conditional distributions (in which this model is incorrectly specified), although this may just be an artifact of the particular simulation setup considered. The plot corresponds to the point $s = 0.2$, but results are similar for other points considered, in that standard deviation is accurately

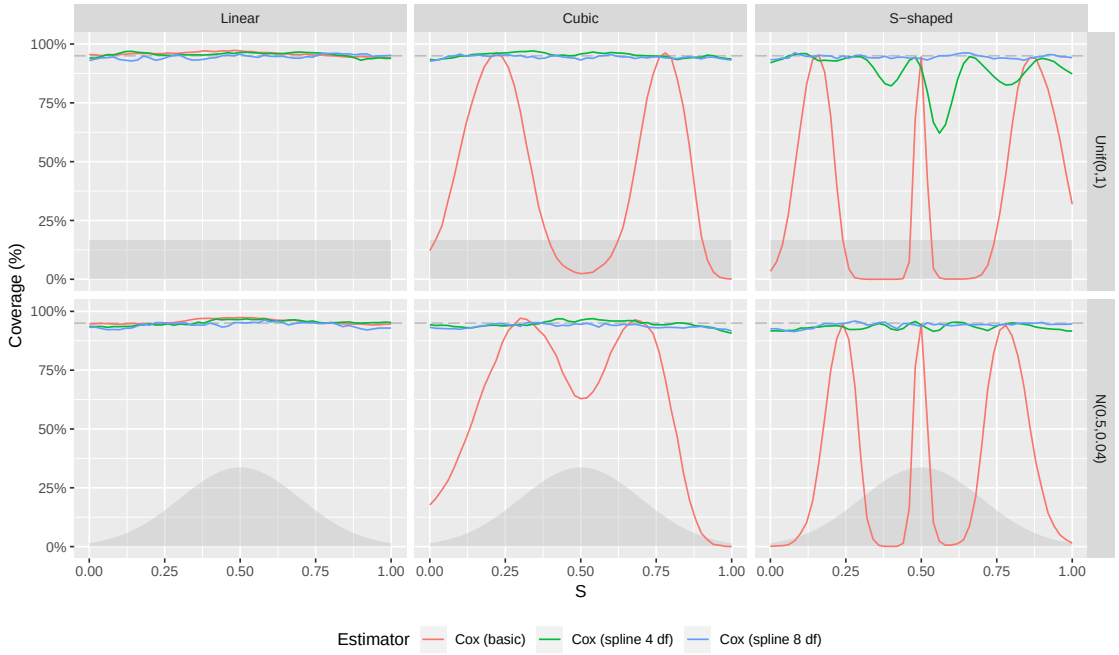


Figure 2.3: 95% confidence interval coverage of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (spline 8 df)" = Cox model with a spline with 8 degrees of freedom), displayed for three conditional survival functions ("Linear" = $Q_0^{(1)}$, "Cubic" = $Q_0^{(2)}$, "S-shaped" = $Q_0^{(3)}$) and two distributions of S.

estimated for larger sample sizes and underestimated for the lowest sample sizes. Given these results, it may be desirable to develop a small-sample correction to the variance estimator, but this is not pursued further in the current work.

In section 2.7, additional simulation results are presented related to the use of an indicator function corresponding to the marker LLOD, as described at the end of section 2.2.2.

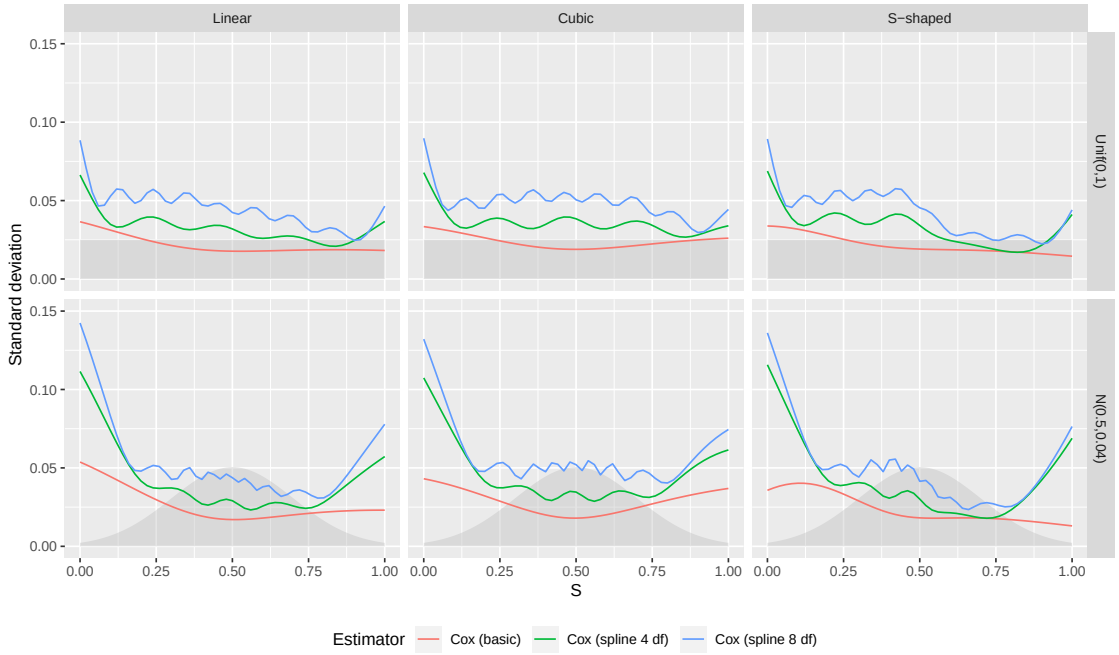


Figure 2.4: Standard deviation of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (spline 8 df)" = Cox model with a spline with 8 degrees of freedom), displayed for three conditional survival functions ("Linear" = $Q_0^{(1)}$, "Cubic" = $Q_0^{(2)}$, "S-shaped" = $Q_0^{(3)}$) and two distributions of S.

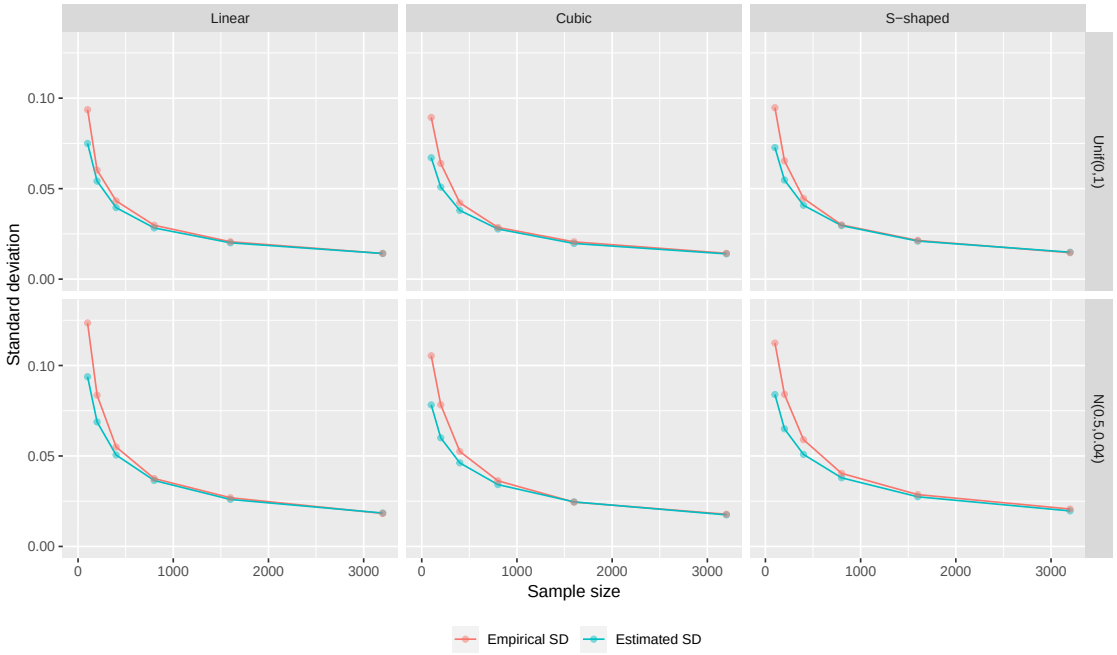


Figure 2.5: Empirical standard deviation ("Empirical SD") and estimated standard deviation ("Estimated SD") for the basic Cox model, evaluated at $s = 0.2$, displayed for three conditional survival functions ("Linear" $= Q_0^{(1)}$, "Cubic" $= Q_0^{(2)}$, "S-shaped" $= Q_0^{(3)}$), two distributions of S, and several sample sizes.

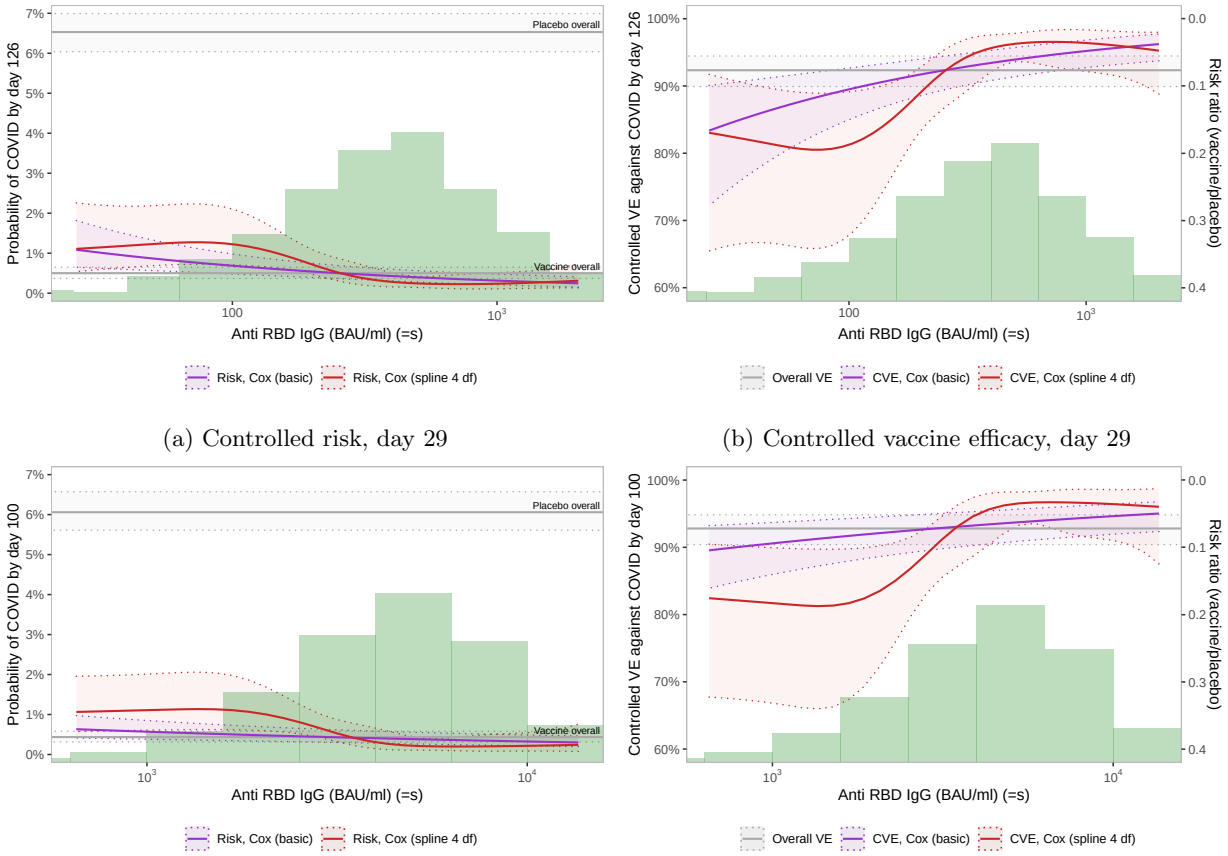
2.4 Immune correlates analysis of the Coronavirus Efficacy (COVE) trial of the mRNA-1273 COVID-19 vaccine

In this section, we illustrate the use of three variants of the Cox model on data from the Coronavirus Efficacy (COVE) placebo-controlled phase 3 trial (NCT04470427) of the mRNA-1273 COVID-19 vaccine. This trial involved randomizing 30,420 adults with high risk of SARS-CoV-2 infection or severe COVID-19 disease at a 1:1 ratio to the vaccine arm or the placebo arm across 99 sites in the United States. Vaccine efficacy was estimated at 94.1% (95%CI: 89.3% to 96.8%) (Baden et al., 2020), with 185 symptomatic COVID-19 disease events in the placebo arm and 11 events in the vaccine arm.

In Gilbert et al. (2022b), the authors considered four immune markers to be potential correlates of protection, including IgG bAbs to the spike protein, IgG bAbs to spike receptor binding domain (RBD), 50% inhibitory dilution (ID50) nAb titer, and 80% inhibitory dilution (ID80) nAb titer; each marker was measured twice, at day 29 and day 57 post-vaccination. A variety of methods were used for this analysis, including the marginalized Cox proportional hazards model considered in this work(Gilbert et al., 2022a), nonparametric targeted minimum loss-based threshold regression (van der Laan et al., 2022), and mediation analysis (Benkeser et al., 2021).

Here, results for two markers are presented, RBD and ID50. Results for the other two markers given in section 2.8. Figure 2.6 shows CR and CVE curves for the IgG bAbs to spike receptor binding domain (RBD) marker, estimated using two models, a basic Cox proportional hazards model (with the biomarker included as a single term in the linear predictor) and a Cox model including an embedded natural cubic spline basis with four degrees of freedom (as per the specification given in section 2.2.2). For each figure, pointwise 95% confidence intervals are plotted as well. Subfigures (2.6a) and (2.6b) correspond to the day 29 marker measurement whereas subfigures (2.6c) and (2.6d) correspond to the day 57 measurement.

Figure 2.7 shows analogous results for the 50% inhibitory dilution (ID50) nAb titer marker.



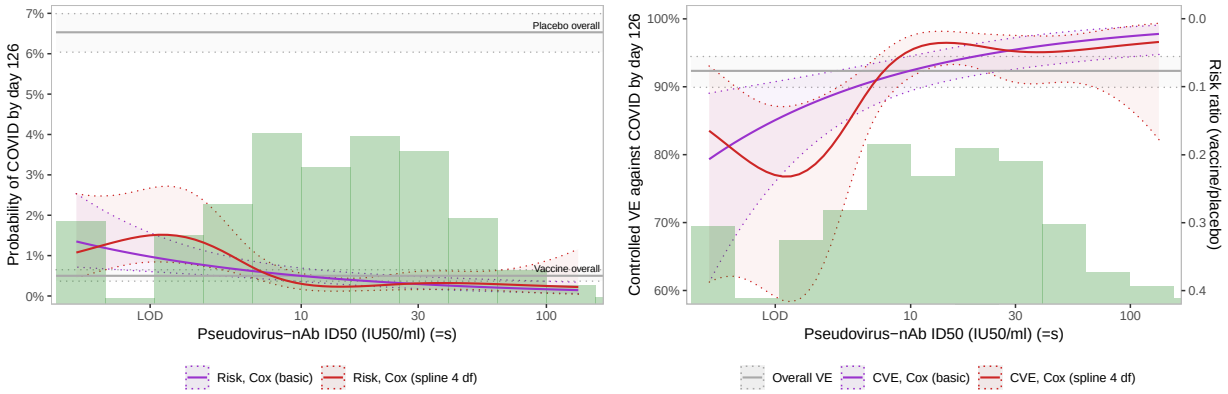
(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29

(c) Controlled risk, day 57

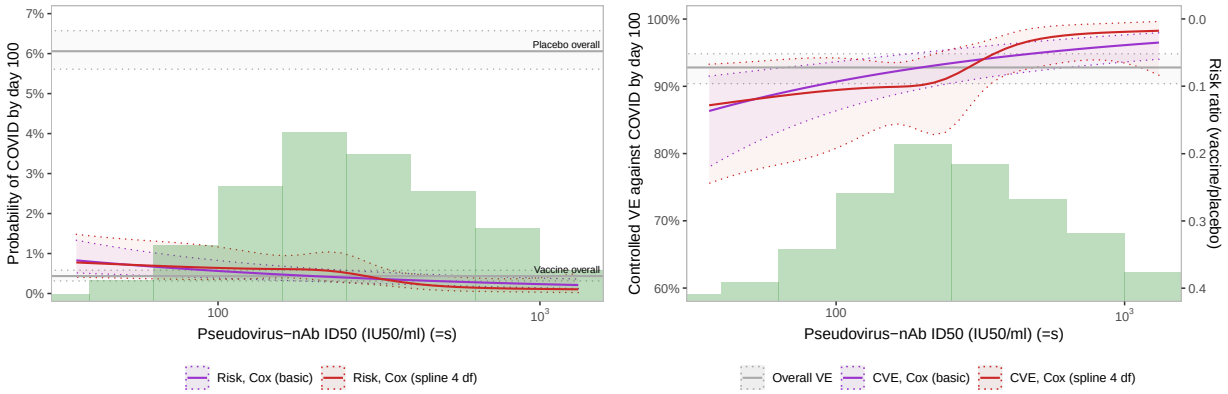
(d) Controlled vaccine efficacy, day 57

Figure 2.6: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the IgG bAbs to spike receptor binding domain (RBD) marker, measured at days 29 and 57. Curves are estimated using a basic Cox model (purple) and a Cox model with a spline with 4 degrees for freedom (red). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.



(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29



(c) Controlled risk, day 57

(d) Controlled vaccine efficacy, day 57

Figure 2.7: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the 50% inhibitory dilution (ID50) nAb titer marker, measured at days 29 and 57. Curves are estimated using a basic Cox model (purple) and a Cox model with a spline with 4 degrees for freedom (red). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

2.5 Discussion

In this chapter, we derived analytic variance estimators for the CR and CVE parameters, showed how these can be used to construct pointwise confidence bands, outlined several ways that the basic Cox model can be extended to account for nonlinearity in the linear predictor, and tested these methods in simulation and real data. This work elaborates upon and extends the methods of Gilbert et al. (2022a) in ways we believe will have practical value in future immune correlates analyses.

The natural cubic spline model represents a practical and flexible way to account for potential nonlinearity in the relationship between the biomarker and the log-hazard function. It includes the basic Cox model as a submodel and results in a bias-variance tradeoff that may be desirable in many applications. One feature of this model is that the resulting CVE curves will not necessarily be monotonic; whether this is an advantage or a disadvantage depends on the particular application. A downside of this model is that it requires the user to specify the number and placement of knots in advance to obtain valid inference. A commonly-used strategy is to place the knots based on the observed quantiles of the data, with the boundary knots placed near (but not at) the minimum and maximum data values and the remaining knots placed at equally-spaced quantiles between these. However, selecting knot locations based on the data is a form of double-dipping, and so we do not encourage this. If any information is known in advance about the expected marker distribution (e.g., from a phase 1 trial pilot study), quantiles can be estimated from these data. Otherwise, we suggest using quantiles of the standard uniform distribution. In terms of the number of knots, we suggest using five knots (i.e., four degrees of freedom) as a rule of thumb, as this meaningfully extends the basic Cox model without adding too much extra variance. When combined with a quantile-based strategy for knot selection, this results in knots being placed at the 5%, 27.5%, 50%, 72.5%, and 95% quantiles, which is precisely the recommendation of Harrell et al. (2001). Of course, any rule of thumb is somewhat arbitrary. Alternatively, if the spline model is being used as a sensitivity analysis, researchers may choose to use several models involving different choices for the number and/or placement of knots. A useful direction for future research is to determine a data-driven strategy for selecting the number and placement of knots that preserves inferential

properties, possibly through the use of sample splitting techniques. Another possible direction for future research is to assess the performance of monotone-constrained regression splines (Wang and Small, 2015).

The models that include an indicator variable corresponding to the value $S = 0$ (one-half the marker LLOD) were studied via simulation in section 2.7. While somewhat specialized, these models may be useful if specific interest lies in the value $\text{CVE}_0(0)$, which in the terms of mediation analysis can be interpreted as one minus the natural direct effect of the vaccine (Benkeser et al., 2021). Adding this indicator variable to either the basic Cox model or the spline model results in increased flexibility, but at the cost of higher variance. In particular, we do not see much benefit of adding this variable to the spline model, since it can result in erratic model behavior just above $S = 0$, and since the spline model does a fairly good job at estimating $\text{CVE}_0(0)$ without the addition of the indicator. Adding the indicator to the basic Cox model may be a good choice if a large point mass is expected at the LLOD, but as noted in section 2.8, it can result in curves that are non-monotonic.

To enable asymptotic study, it was assumed that the indicator variables denoting selection into the phase-two sample were generated from a Bernoulli distribution, even though in reality they are typically sampled using a FPSS mechanism. Although in finite samples, this may lead to confidence intervals that are somewhat conservative, this assumption should make no difference in terms of the asymptotic results, since the estimated weights described in section 2.2.1 were used, as noted in section 4.3 of Breslow et al. (2015). Future research could examine whether efficiency could be gained from adapting finite sample correction procedures (e.g., Kott, 1988) to the FPSS mechanism.

An important limitation of this work is that results are only valid if the underlying assumptions hold, including those related to the Cox model (e.g., the proportional hazards assumption) and the causal assumptions that allow for the marginalized risk to be equated with the controlled risk; these assumptions may not be met in practice and the plausibility of each assumption should be carefully considered in each data analysis. We are currently working in parallel on a nonparametric method of estimation and inference for the CR and CVE curves that relaxes the assumptions of the Cox model by allowing for estimation of the conditional survival function using flexible machine

learning tools. Furthermore, while many methods exist for extending linear models to account for potential nonlinearities (including other types of splines), we only considered the use of natural cubic splines in this chapter; future research could look at other possible modeling choices.

2.6 Proofs

Proofs of all theorems and lemmas are given below. Throughout, empirical process notation is used, in which $P_0 f \coloneqq E_0\{f(O)\}$ and $P_n f \coloneqq n^{-1}\sum_{i=1}^{n} f(O_i)$. Furthermore, we repeatedly make use of decompositions of the form

$$P_n f_n - P_0 f_0 = (P_n - P_0)f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)\,, \tag{2.13}$$

where f_0 is an arbitrary function depending on P_0 and f_n is an estimator of f_0. Theorem 19.24 of Van der Vaart (2000) tells us that if (i) $P_0(f_n - f_0)^2 = o_P(1)$ and (ii) f_n falls in a P_0-Donsker class with probability tending to one, then $(P_n - P_0)(f_n - f_0) = o_P(n^{-1/2})$. Furthermore, since $P_0(f_n - f_0)^2 = o_P(1)$ implies that $P_0(f_n - f_0) = o_P(1)$, conditions (i) and (ii) jointly imply that $P_n f_n - P_0 f_0 = o_P(1)$, since $(P_n - P_0)f_0 = o_P(1)$ by the central limit theorem. Note that in the proofs below, f_n and f_0 are repeatedly redefined, as to avoid excessive notation.

2.6.1 Proof of Lemma 1

Applying the delta method to (2.5), it holds that

$$\frac{1}{\pi_n(c_i)} - \frac{1}{\pi_0(c_i)} = \frac{1}{n}\sum_{j=1}^{n} \frac{I(C_j = c_i)\{p_0^1(c_i) - p_0(c_i)Z_j\}}{\{p_0^1(c_i)\}^2} + R_n^{(1)}(o_i)\,, \tag{2.14}$$

where $p_0(c) \coloneqq P_0(C = c)$, $p_0^1(c) \coloneqq P_0(C = c, Z = 1)$, and

$$R_n^{(1)}(o_i) \coloneqq \frac{1}{\pi_n(c_i)} - \frac{1}{\pi_0(c_i)} - \frac{1}{n}\sum_{j=1}^{n} \frac{I(C_j = c_i)\{p_0^1(c_i) - p_0(c_i)Z_j\}}{\{p_0^1(c_i)\}^2} = o_P(n^{-1/2})\,.$$

For an arbitrary real-valued fixed function $h(O) \coloneqq h(X, Y, \Delta, S)$ of an individual observation, define $f_n(o_i) \coloneqq z_i h(o_i)/\pi_n(c_i)$ and $f_0(o_i) \coloneqq z_i h(o_i)/\pi_0(c_i)$ and note that

$$P_n f_n - P_0 f_0 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i}{\pi_n(C_i)}h(O_i) - E_0\{h(O)\}\right]. \tag{2.15}$$

The asymptotic behavior of (2.15) can be studied through the decomposition given in (2.13). The term $(P_n - P_0)f_0$ is already linear. Using (2.14), the term $P_0(f_n - f_0)$ can be written as

$$\begin{aligned} P_0(f_n - f_0) &= E_0\left[\left\{\frac{1}{\pi_n(C)} - \frac{1}{\pi_0(C)}\right\} Zh(O)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E_0\left[\frac{I(C = C_i)\{p_0^1(C) - p_0(C)Z_i\}Zh(O)}{\{p_0^1(C)\}^2}\right] + R_n^{(2)}, \end{aligned} \tag{2.16}$$

where $R_n^{(2)} \coloneqq E_0\{Zh(O)R_n^{(1)}(O)\}$. To show that $R_n^{(2)} = o_P(n^{-1/2})$, begin by noting that

$$\begin{aligned} R_n^{(1)}(o_i) = {} & \frac{\sum_{j=1}^n I(C_j = c_i)}{\{p_0^1(c_i)\}^2 \sum_{j=1}^n I(C_j = c_i)Z_j}\left\{\frac{1}{n}\sum_{j=1}^{n} I(C_j = c_i)Z_j - p_0^1(c_i)\right\}^2 \\ & + \frac{1}{\{p_0^1(c_i)\}^2}\left\{\frac{1}{n}\sum_{j=1}^{n} I(C_j = c_i) - p_0(c_i)\right\}\left\{p_0^1(c_i) - \frac{1}{n}\sum_{j=1}^{n} I(C_j = c_i)Z_j\right\}. \end{aligned}$$

This allows us to write $R_n^{(2)} = R_n^{(2,1)} + R_n^{(2,2)}$, where

$$\begin{aligned} R_n^{(2,1)} &= E_0\left[Zh(O)\frac{\sum_{j=1}^n I(C_j = C)}{\{p_0^1(C)\}^2 \sum_{j=1}^n I(C_j = C)Z_j}\left\{\frac{1}{n}\sum_{j=1}^{n} I(C_j = C)Z_j - p_0^1(C)\right\}^2\right], \\ R_n^{(2,2)} &= E_0\left[Zh(O)\frac{1}{\{p_0^1(C)\}^2}\left\{\frac{1}{n}\sum_{j=1}^{n} I(C_j = C) - p_0(C)\right\}\left\{p_0^1(C) - \frac{1}{n}\sum_{j=1}^{n} I(C_j = C)Z_j\right\}\right]. \end{aligned}$$

Iterating the expectation and noting that the random variable C is discrete, the term $R_n^{(2,1)}$ can be written as

$$\begin{aligned} R_n^{(2,1)} = \frac{1}{d}\sum_{k=1}^{d}\Bigg[& E_0\{Zh(O) \mid C = c_k\}\frac{\sum_{j=1}^n I(C_j = c_k)}{\{p_0^1(c_k)\}^2 \sum_{j=1}^n I(C_j = c_k)Z_j} \\ & \times \left\{\frac{1}{n}\sum_{j=1}^{n} I(C_j = c_k)Z_j - p_0^1(c_k)\right\}^2 P_0(C = c_k)\Bigg], \end{aligned}$$

where d is the total number of strata. Since, in the statement of the lemma, it is assumed that $E_0[\{h(O)\}^2 \,|\, Z = 1, C = c] < \infty$ for $c \in \{1, 2, ..., d\}$ (i.e., for each stratum), it also holds that $E_0[Zh(O) \,|\, C = c] < \infty$. Since the squared term is $O_P(n^{-1})$, it holds that $R_n^{(2,1)} = o_P(n^{-1/2})$ by the continuous mapping theorem. An analogous argument can be used to show that $R_n^{(2,2)} = o_P(n^{-1/2})$, and so $R_n^{(2)} = o_P(n^{-1/2})$ as well.

Finally, we must show that $(P_n - P_0)(f_n - f_0) = o_P(n^{-1/2})$. Using Theorem 19.24 of Van der Vaart (2000), this will be the case if (i) $P_0(f_n - f_0)^2 = o_P(1)$ and (ii) f_n falls in a P_0-Donsker class with probability tending to one. For the first condition, it holds that

$$
\begin{aligned}
P_0(f_n - f_0)^2 &= E_0 \left\{ \frac{Zh(O)}{\pi_n(C)} - \frac{Zh(O)}{\pi_0(C)} \right\}^2 \\
&= E_0 \left[E_0[Z\{h(O)\}^2 \,|\, C] \left\{ \frac{\sum_{j=1}^n I(C = C_j)}{\sum_{j=1}^n I(C = C_j) Z_j} - \frac{p_0(C)}{p_0^1(C)} \right\}^2 \right] \\
&= c_1 E_0 \left\{ \frac{\sum_{j=1}^n I(C = C_j)}{\sum_{j=1}^n I(C = C_j) Z_j} - \frac{p_0(C)}{p_0^1(C)} \right\}^2 \\
&= o_P(1) \,,
\end{aligned}
$$

for some constant c_1, where again the assumption that $E_0[\{h(O)\}^2 \,|\, Z = 1, C = c_k] < \infty$ for $k \in \{1, ..., d\}$ is used, as well as the continuous mapping theorem. To satisfy the Donsker condition, note that the function $o \mapsto z/\pi_n(c)$ falls in the Donsker class of functions of bounded variation, since the argument z takes only binary values and the argument c takes only values from a fixed finite set of integers (i.e., the stratum indices). Since h is a fixed bounded function, the class f_n will be Donsker as well (see van der Vaart and Wellner, 1996, example 2.10.10).

Since $R_n^{(2)} = o_P(n^{-1/2})$ and $(P_n - P_0)(f_n - f_0) = o_P(n^{-1/2})$, equations (2.13) and (2.16) jointly imply that

$$
\begin{aligned}
P_n f_n &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i h(O_i)}{\pi_0(C_i)} + E_0\left[\frac{I(C=C_i)\{p_0^1(C)-p_0(C)Z_i\}Zh(O)}{\{p_0^1(C)\}^2}\right]\right) + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i h(O_i)}{\pi_0(C_i)} + E_0\left[\frac{\{p_0^1(C_i)-p_0(C_i)Z_i\}h(O)}{\{p_0^1(C_i)\}^2}\,\middle|\, Z=1, C=C_i\right] p_0^1(C_i)\right) + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i h(O_i)}{\pi_0(C_i)} + \left\{1-\frac{Z_i}{\pi_0(C_i)}\right\} E_0\left\{h(O)\,|\,Z=1, C=C_i\right\}\right] + o_P(n^{-1/2})\,. \qquad \square
\end{aligned}
$$

2.6.2 Proof of Theorem 1

To study the IPS-weighted Breslow estimator given by (2.3), first define

$$
\begin{aligned}
S_0(u,\beta) &:= \int e^{\beta' v} I(y \geq u) dP_0(y,v) = E_0\left\{e^{V'\beta} I(Y \geq u)\right\}, \\
S_n(u,\beta) &:= \int e^{\beta' v} I(y \geq u) d\tilde{P}_n(y,v) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i e^{\beta' V_i} I(Y_i \geq u)}{\pi_n(C_i)}, \\
\Lambda_0(t,\beta) &:= \int \frac{\delta I(y \leq t)}{S_0(y,\beta)} dP_0(\delta,y) = E_0\left\{\frac{\Delta I(Y \leq t)}{S_0(Y,\beta)}\right\}, \\
\Lambda_n(t,\beta) &:= \int \frac{\delta I(y \leq t)}{S_n(y,\beta)} d\tilde{P}_n(\delta,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i \Delta_i I(Y_i \leq t)}{\pi_n(C_i) S_n(Y_i,\beta)},
\end{aligned}
$$

where $\tilde{P}_n$ is the IPS-weighted empirical measure that places mass $Z_i/\{n\pi_n(C_i)\}$ at each observation. Note that $\Lambda_n(t,\beta_n) = \Lambda_n(t)$, $\Lambda_0(t,\beta_0) = \Lambda_0(t)$, $S_n(u,\beta_n) = S_n(u)$, and $S_0(u,\beta_0) = S_0(u)$. Consider the decomposition given by

$$
\Lambda_n(t) - \Lambda_0(t) = \overbrace{\Lambda_n(t,\beta_n) - \Lambda_n(t,\beta_0)}^{\text{Term 1}} + \overbrace{\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0)}^{\text{Term 2}}.
$$

In what follows, we derive the influence function contributions of Terms 1 and 2, and prove that all remainder terms are $o_P(n^{-1/2})$. We focus first on Term 1. A Taylor expansion of $\beta \mapsto \Lambda_n(\beta,t)$ yields

$$
\Lambda_n(t,\beta_n) - \Lambda_n(t,\beta_n) = \left\{\mu_n(t,\beta_0)\right\}'(\beta_n - \beta_0) + \frac{1}{2}(\beta_n - \beta_0)'\left\{H_n(t,\beta_n^*)\right\}(\beta_n - \beta_0), \tag{2.17}
$$

where $\beta_n \in [\beta_0, \beta_n]$, and the gradient $\mu_n(t,\beta_0)$ and Hessian $H_n(t,\beta_0)$ are given by

$$\mu_n(t,\beta) = \frac{\partial}{\partial\beta}\Lambda_n(t,\beta) = -\int \frac{\delta I(y \le t)\dot{S}_n(y,\beta)}{\left\{S_n(y,\beta)\right\}^2} d\tilde{P}_n(y,\delta)$$
$$H_n(t,\beta) = \frac{\partial^2}{\partial\beta^2}\Lambda_n(t,\beta) = \int \delta I(y \le t) \left[\frac{2\left\{\dot{S}_n(y,\beta)\right\}^{\otimes 2} - S_n(y,\beta)\ddot{S}_n(y,\beta)}{\left\{S_n(y,\beta)\right\}^3}\right] d\tilde{P}_n(y,\delta)\,,$$

where $\dot{S}_n(t,\beta)$ and $\ddot{S}_n(t,\beta)$ are defined analogously to $S_n(t,\beta)$, such that $\dot{S}_n(t,\beta_n) = \dot{S}_n(t)$ and $\ddot{S}_n(t,\beta_n) = \ddot{S}_n(t)$. Adding and subtracting $\{\mu_0(t)\}'(\beta_n - \beta_0)$ from (2.17), it holds that

$$\begin{aligned}\Lambda_n(t,\beta_n) - \Lambda_n(t,\beta_0) &= \left\{\mu_0(t)\right\}'(\beta_n - \beta_0) + R_n^{(3)} + R_n^{(4)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mu_0(t)\right\}'\varphi_{\beta,0}(O_i) + R_n^{(3)} + R_n^{(4)} + o_P(n^{-1/2})\,,\end{aligned}$$

where

$$\begin{aligned}R_n^{(3)} &\coloneqq \left\{\mu_n(t,\beta_0) - \mu_0(t)\right\}'(\beta_n - \beta_0)\,, \\ R_n^{(4)} &\coloneqq \frac{1}{2}(\beta_n - \beta_0)' H_n(t,\beta_n^*)(\beta_n - \beta_0)\,.\end{aligned}$$

Next, consider Term 2. Adding and subtracting terms and using the relationship $\delta\{S_0(y)\}^{-1}dP_0(y,\delta) = d\Lambda_0(y)$, this can be expressed as:

$$\begin{aligned}\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0) &= \int \frac{\delta I(y \le t)}{S_n(y,\beta_0)} d\tilde{P}_n(y,\delta) - \int \frac{\delta I(y \le t)}{S_0(y,\beta_0)} dP_0(y,\delta) \\ &= \int \frac{\delta I(y \le t)}{S_0(y,\beta_0)} d\tilde{P}_n(y,\delta) - \int \frac{\delta I(y \le t) S_n(y,\beta_0)}{\{S_0(y,\beta_0)\}^2} dP_0(y,\delta) + R_n^{(5)} + R_n^{(6)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i\Delta_i I(Y_i \le t)}{\pi_n(C_i)S_0(Y_i,\beta_0)} - \frac{Z_i}{\pi_n(C_i)}\int_0^{t\wedge Y_i}\frac{e^{V_i'\beta_0}d\Lambda_0(y)}{S_0(y,\beta_0)}\right\} + R_n^{(5)} + R_n^{(6)}\,,\end{aligned} \tag{2.18}$$

where

$$R_n^{(5)} := \int \delta I(y \le t) \left\{ \frac{1}{S_n(y,\beta_0)} - \frac{1}{S_0(y,\beta_0)} \right\} d(\tilde{P}_n - P_0)(y,\delta)\,,$$
$$R_n^{(6)} := \int \frac{\delta I(y \le t) \left\{ S_n(y,\beta_0) - S_0(y,\beta_0) \right\}^2}{S_n(y,\beta_0) \left\{ S_0(y,\beta_0) \right\}^2} dP_0(y,\delta)\,.$$

Next, it holds that for all $c \in \{1, 2, ..., d\}$,

$$E_0 \left[\left\{ \int_0^{t\wedge Y} \frac{e^{V'\beta_0}}{S_0(y)} d\Lambda_0(y) \right\}^2 \Big| Z = 1, C = c \right] \le \left\{ \frac{\Lambda_0(t)}{S_0(t)} \right\}^2 E_0 \left(e^{2V'\beta_0} \Big| Z = 1, C = c \right) < \infty\,.$$

Therefore, Lemma 1 can be applied to write

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \frac{Z_i}{\pi_n(C_i)} \int_0^{t\wedge Y_i} \frac{e^{V_i'\beta_0} d\Lambda_0(y)}{S_0(y,\beta_0)} \\
&\quad = \frac{1}{n} \sum_{i=1}^n \left[\frac{Z_i}{\pi_0(C_i)} \int_0^{t\wedge Y_i} \frac{e^{V_i'\beta_0} d\Lambda_0(y)}{S_0(y,\beta_0)} + \left\{ 1 - \frac{Z_i}{\pi_0(C_i)} \right\} E_0 \left\{ \int_0^{t\wedge Y} \frac{e^{V'\beta_0} d\Lambda_0(y)}{S_0(y,\beta_0)} \Big| Z = 1, C = C_i \right\} \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^n \left[\frac{Z_i}{\pi_0(C_i)} \int_0^{t\wedge Y_i} \frac{e^{V_i'\beta_0} d\Lambda_0(y)}{S_0(y,\beta_0)} + \int_0^t \frac{\nu_{1,0}(O_i,y)}{S_0(y,\beta_0)} d\Lambda_0(y) \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^n \left[\int_0^t \left\{ \frac{Z_i I(y \le Y_i) e^{V_i'\beta_0} + \pi_0(C_i)\nu_{1,0}(O_i,y)}{\pi_0(C_i) S_0(y,\beta_0)} \right\} d\Lambda_0(y) \right]\,.
\end{aligned}$$

Combining this with (2.18), it holds that

$$\begin{aligned}
\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0) = \frac{1}{n} \sum_{i=1}^n \Bigg[&\frac{Z_i \Delta_i I(Y_i \le t)}{\pi_n(C_i) S_0(Y_i)} - \int_0^t \left\{ \frac{Z_i I(y \le Y_i) e^{V_i'\beta_0} + \pi_0(C_i)\nu_{1,0}(O_i,y)}{\pi_0(C_i) S_0(y)} \right\} d\Lambda_0(y) \Bigg] \\
&+ R_n^{(5)} + R_n^{(6)} + o_P(n^{-1/2})\,.
\end{aligned} \tag{2.19}$$

Similarly applying Lemma 1 to the first term inside the summation of (2.19), it holds that

$$\begin{aligned}
\Lambda_n(t,\beta_0) - \Lambda_0(t,\beta_0) = \frac{1}{n} \sum_{i=1}^n \Bigg[&\frac{Z_i \Delta_i I(Y_i \le t)}{\pi_0(C_i) S_0(Y_i)} - \int_0^t \left\{ \frac{Z_i I(y \le Y_i) e^{V_i'\beta_0} + \pi_0(C_i)\nu_{1,0}(O_i,y)}{\pi_0(C_i) S_0(y)} \right\} d\Lambda_0(y) \\
&+ \nu_{2,0}(O_i,t) \Bigg] + R_n^{(5)} + R_n^{(6)} + o_P(n^{-1/2})\,.
\end{aligned}$$

If all remainder terms are $o_P(n^{-1/2})$, then combining the influence function contributions from Term 1 and Term 2, it holds that

$$\Lambda_n(t) - \Lambda_0(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i\Delta_i I(Y_i \le t)}{\pi_0(C_i)S_0(Y_i)} - \int_0^t \left\{\frac{Z_i I(y \le Y_i)e^{V_i'\beta_0} + \pi_0(C_i)\nu_{1,0}(O_i, y)}{\pi_0(C_i)S_0(y)}\right\} d\Lambda_0(y) \right.$$
$$\left. + \nu_{2,0}(O_i, t) + \{\mu_0(t)\}'\varphi_{\beta,0}(O_i)\right] + o_P(n^{-1/2}),$$

which is precisely the result of the theorem. We first analyze $R_n^{(3)}$. Since $\beta_n - \beta_0 = O_p(n^{-1/2})$, it holds that $R_n^{(3)} = o_P(n^{-1/2})$ if $|\mu_n(t, \beta_0) - \mu_0(t)| = o_P(1)$. Define

$$f_n(o) := \frac{z\delta I(y \le t)\dot{S}_n(y, \beta_0)}{\pi_n(c)\left\{S_n(y, \beta_0)\right\}^2},$$
$$f_0(o) := \frac{z\delta I(y \le t)\dot{S}_0(y, \beta_0)}{\pi_0(c)\left\{S_0(y, \beta_0)\right\}^2}.$$

Then it holds that $\mu_n(t, \beta_0) - \mu_0(t) = P_0 f_0 - P_n f_n$. The decomposition given in (2.13) and Theorem 19.24 of Van der Vaart (2000) can be used to show that $P_n f_n - P_0 f_0 = o_P(1)$. To show that $P_0(f_n - f_0)^2 = o_P(1)$, first note that

$$\left|\dot{S}_n(y, \beta_0)\right| \le \epsilon_\pi^{-1}\frac{1}{n}\sum_{i=1}^{n}\left|V_i e^{V_i'\beta_0}\right| \xrightarrow{p} \epsilon_\pi^{-1} E_0\left(|V|\, e^{V'\beta_0}\right) < \infty.$$

Therefore, there exists a number C_1 such that each component of $|\dot{S}_n(y, \beta_0)|$ is bounded above by C_1 uniformly over all possible y and n with probability tending to one. Next, note that

$$\begin{aligned} P_0(f_n - f_0)^2 &= E_0[\{f_n(O)\}^2] - 2E_0\{f_n(O)f_0(O)\} + E_0[\{f_0(O)\}^2] \\ &= E_0\left[\frac{Z\Delta I(Y \le t)\{\dot{S}_n(Y, \beta_0)\}^2}{\{\pi_n(C)\}^2\{S_n(Y, \beta_0)\}^4}\right] \\ &\quad - 2E_0\left[\frac{Z\Delta I(Y \le t)\dot{S}_0(Y, \beta_0)\dot{S}_n(Y, \beta_0)}{\pi_n(C)\pi_0(C)\{S_n(Y, \beta_0)S_0(Y, \beta_0)\}^2}\right] \\ &\quad + E_0\left[\frac{Z\Delta I(Y \le t)\{\dot{S}_0(Y, \beta_0)\}^2}{\{\pi_0(C)\}^2\{S_0(Y, \beta_0)\}^4}\right]. \end{aligned}$$

Considering the term inside the first expectation, since $y \leq t$ with probability one, it holds that

$$\left| \frac{z\delta I(y \leq t)\{\dot{S}_n(y, \beta_0)\}^2}{\{\pi_n(c)\}^2\{S_n(y, \beta_0)\}^4} \right| \leq \epsilon_\pi^{-2} \left| \frac{\{\dot{S}_n(y, \beta_0)\}^2}{\{S_n(t, \beta_0)\}^4} \right| .$$

Since $S_n(t, \beta_0) \xrightarrow{p} S_0(t, \beta_0)$, there exists a number $\epsilon_s > 0$ such that $S_n(t, \beta_0) > \epsilon_s$ for sufficiently large n. Since $|\dot{S}_n(y, \beta_0)|$ is uniformly bounded over all possible y, the term inside the first expectation can be uniformly bounded, and so by the dominated convergence theorem, it holds that

$$E_0 \left[\frac{Z\Delta I(Y \leq t)\{\dot{S}_n(Y, \beta_0)\}^2}{\{\pi_n(C)\}^2\{S_n(Y, \beta_0)\}^4} \right] \xrightarrow{p} E_0 \left[\frac{Z\Delta I(Y \leq t)\{\dot{S}_0(Y, \beta_0)\}^2}{\{\pi_0(C)\}^2\{S_0(Y, \beta_0)\}^4} \right] .$$

The second expectation will similarly converge, and so it follows that $P_0(f_n - f_0)^2 = o_P(1)$. To show that f_n falls in a P_0-Donsker class with probability tending to one, note that since (1) the set of functions $y \mapsto \{S_n(y, \beta_0)\}^{-2}$ indexed by n (for $n > n^*$ for sufficiently large n^*) is uniformly bounded and monotone, (2) the domain and range of π_n are finite sets, and (3) the set of functions $y \mapsto \dot{S}_n(y, \beta_0)$ indexed by n is uniformly bounded with finite variation norm (for $n > n^*$ for sufficiently large n^*) is, then the set of functions f_n falls in a P_0-Donsker class. Thus it holds that $|\mu_n(t, \beta_0) - \mu_0(t)| = o_P(1)$, and so $R_n^{(3)} = o_P(n^{-1/2})$.

To analyze $R_n^{(4)}$, first note that

$$\left| \dot{S}_n(y, \beta_n^*) \right| \leq \epsilon_\pi^{-1} \frac{1}{n} \sum_{i=1}^{n} |V_i| e^{V_i' \beta_n^*} .$$

To show that the term on the right is bounded, set $f_n(V_i) \coloneqq |V_i| e^{V_i' \beta_n^*}$ and $f_0(V_i) \coloneqq |V_i| e^{V_i' \beta_0}$. Then it holds that

$$\frac{1}{n} \sum_{i=1}^{n} |V_i| e^{V_i' \beta_n^*} - E_0 \left(|V| e^{V' \beta_0} \right) = (P_n - P_0) f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0) .$$

The first term on the right-hand side is linear. For the second term, since $\beta_n^* \xrightarrow{p} \beta_0$, it holds that $e^{V' \beta_n^*} \xrightarrow{p} e^{V' \beta_0}$. Also, for sufficiently large n, $|\beta_n - \beta_0| < \epsilon$ for any $\epsilon > 0$, which implies that $|\beta_n^* - \beta_0| < \epsilon$. By assumption, $\sup_{\beta : |\beta - \beta_0| < \epsilon} E(|V| e^{V' \beta}) < \infty$, and so by the dominated convergence

theorem, it holds that

$$P_0(f_n - f_0) = E_0\left(|V|e^{V'\beta_n^*}\right) - E_0\left(|V|e^{V'\beta_0}\right).$$

For the third term, similar arguments using the dominated convergence theorem can be used to show that

$$P_0(f_n - f_0)^2 = E_0\left(|V|e^{2V'\beta_n^*}\right) - 2E_0\left\{|V|e^{V'(\beta_n^* + \beta_0)}\right\} + E_0\left(|V|e^{2V'\beta_0}\right) = o_P(1).$$

Since each f_n falls in a Donsker class, $(P_n - P_0)(f_n - f_0) = o_P(n^{-1/2})$ by Theorem 19.24 of Van der Vaart (2000). Therefore, it holds that

$$\left|\dot{S}_n(y, \beta_n^*)\right| \leq \epsilon_\pi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}|V_i|e^{V_i'\beta_n^*}\right) \xrightarrow{p} \epsilon_\pi^{-1}E_0\left(|V|e^{V'\beta_0}\right) < \infty.$$

Therefore, each component of $|\dot{S}_n(y, \beta_n^*)|$ is bounded by C_1 with probability tending to one. Analogous arguments can be used to show that there exists a constant C_2 such that each component of $\left|\ddot{S}_n(y, \beta_n^*)\right|$ is bounded by C_2. Also, we previously showed that for sufficiently large n, there exists a number ϵ_s such that $S_n(t, \beta_0) > \epsilon_s$; analogous arguments can be used to show that $S_n(t, \beta_n^*) > \epsilon_s$ as well, which allows us to write

$$\begin{aligned}\left|\left\{H\Lambda_n(t, \beta_n^*)\right\}_{ij}\right| &\leq 2\left|\int \frac{\delta I(y \leq t)\{\dot{S}_n(y, \beta_n^*)\}_i\{\dot{S}_n(y, \beta_n^*)\}_j}{\{S_n(y, \beta_n^*)\}^3} d\tilde{P}_n(y, \delta)\right| \\ &\quad + \left|\int \frac{\delta I(y \leq t)\{\ddot{S}_n(y, \beta_n^*)\}_{ij}}{\{S_n(y, \beta_n^*)\}^2} d\tilde{P}_n(y, \delta)\right| \\ &\leq 2\epsilon_s^{-3}C_1^2 + \epsilon_s^{-2}C_2.\end{aligned}$$

for sufficiently large n, where the notation $\{A\}_i$ represents the i^{th} element of vector A and $\{B\}_{ij}$ represents the element in row i and column j of matrix B. Therefore, it holds that

$$\left|R_n^{(4)}\right| \leq \left(C_1^2\epsilon_s^{-3} + C_2\epsilon_s^{-2}/2\right)(\beta_n - \beta_0)'(\beta_n - \beta_0).$$

The term $R_n^{(5)}$ will be $o_P(n^{-1/2})$ by arguments analogous to those used above using Theorem 19.24 of Van der Vaart (2000), since

$$\int \frac{z\delta I(y \leq t)}{\{\pi_0(c)\}^2} \left\{ \frac{1}{S_n(y, \beta_0)} - \frac{1}{S_0(y, \beta_0)} \right\}^2 dP_0(y, \delta) = o_P(1) \,.$$

To analyze $R_n^{(6)}$, note that

$$\left| R_n^{(6)} \right| \leq \left\{ S_n(t, \beta_0) \right\}^{-1} \left\{ S_0(t, \beta_0) \right\}^{-2} \int \left\{ S_n(y, \beta_0) - S_0(y, \beta_0) \right\}^2 dP_0(y) = o_P(n^{-1/2}) \,. \qquad \square$$

2.6.3 Proof of Theorem 2

Using the notation $PQ_n \coloneqq E_P[Q_n(t_0 \,|\, X, s)]$ and $PQ_0 \coloneqq E_P[Q_0(t_0 \,|\, X, s)]$, it holds that

$$r_{M,n}(s) - r_{M,0}(s) = (P_0 - P_n)Q_0 + P_0(Q_0 - Q_n) + (P_n - P_0)(Q_0 - Q_n) \,.$$

The first term is linear and can be written as

$$(P_0 - P_n)Q_0 = \frac{1}{n} \sum_{i=1}^{n} \left[E_0 \left\{ Q_0(t_0 \,|\, X, s) \right\} - Q_0(t_0 \,|\, X_i, s) \right] .$$

The second term can be represented as

$$\begin{aligned} P_0(Q_0 - Q_n) &= E_0 \left\{ Q_n(t_0 \,|\, X, s) - Q_0(t_0 \,|\, X, s) \right\} \\ &= -\frac{1}{n} \sum_{i=1}^{n} E_0 \left[\varphi_{Q,0} \left\{ O_i, (X, s) \right\} \right] + E_0(R_n^{(7)}(X)) \,, \end{aligned}$$

where

$$R_n^{(7)}(x) \coloneqq Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s) - \frac{1}{n} \sum_{i=1}^{n} \varphi_{Q,0}\{O_i, (x, s)\} = o_P(n^{-1/2}) \,.$$

To prove that $E_0\{R_n^{(7)}(X)\} = o_P(n^{-1/2})$, note that

$$\left|R_n^{(7)}(x)\right| \leq \left|Q_n(t_0 \,|\, x,s)\right| + \left|Q_0(t_0 \,|\, x,s)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} \varphi_{Q,0}\{O_i,(x,s)\}\right|$$
$$\leq 2 + e^{(x,s)'\beta_0} \sup_o \left|\Lambda_0(t_0)(x,s)'\varphi_{\beta,0}(o) + \varphi_{\Lambda,0}(o,t_0)\right| .$$

Therefore, since the support of V is contained within a compact set, $R_n^{(7)}(x)$ will be uniformly bounded if the influence functions $\varphi_{\beta,0}$ and $\varphi_{\Lambda,0}$ are bounded. To bound $\varphi_{\beta,0}$, note that

$$\sup_o \left|\varphi_{\beta,0}(o)\right| = \sup_o \left|\frac{z}{\pi_0(c)}\tilde{\ell}_0(o) + \left\{1 - \frac{z}{\pi_0(c)}\right\} E_0\left\{\tilde{\ell}_0(O) \,|\, C=c, Z=1\right\}\right|$$
$$\leq \mathcal{I}_0^{-1}\left(1 + 2\epsilon_\pi^{-1}\right) \sup_o \left|\ell_0(o)\right|$$
$$\leq \mathcal{I}_0^{-1}\left(1 + 2\epsilon_\pi^{-1}\right) \sup_{(\delta,v,y)} \left|\delta\{v - m_0(y)\} - e^{v'\beta_0}\int_0^y \{v - m_0(x)\} d\Lambda_0(x)\right| .$$

Since the support of V is contained in some compact set and m_0 is bounded, $|v - m_0(x)| < (C_3, ..., C_3)$ for any $x \in [0,\tau]$ for some constant $C_3 > 0$. Therefore, it holds that

$$\sup_o \left|\varphi_{\beta,0}(o)\right| \leq \mathcal{I}_0^{-1}\left(1 + 2\epsilon_\pi^{-1}\right)(C_3, ..., C_3)\left\{1 + \Lambda_0(\tau) \sup_v \left|e^{v'\beta_0}\right|\right\} < \infty .$$

Given this result, it is straightforward to show that $\varphi_{\Lambda,0}$ is bounded as well. Therefore, it holds that

$$P_0(Q_0 - Q_n) = -\frac{1}{n}\sum_{i=1}^{n} E_0\left[\varphi_{Q,0}\left\{O_i,(X,s)\right\}\right] + o_P(n^{-1/2}) .$$

For the third term, Theorem 19.24 of Van der Vaart (2000) can again be leveraged. It trivially holds that $P_0(Q_n - Q_0)^2 = o_P(1)$, since survival functions are bounded above by 1. The Donsker condition will also hold since for each n, $x \mapsto Q_n(t_0 \,|\, x, s)$ is a continuous transformation of a bounded linear function of x. Combining these results, it holds that

$$r_{M,n}(s) - r_{M,0}(s) = \frac{1}{n}\sum_{i=1}^{n}\left(E_0\left\{Q_0(t_0 \,|\, X, s)\right\} - Q_0(t_0 \,|\, X_i, s) - E_0\left[\varphi_{Q,0}\left\{O_i,(X,s)\right\}\right]\right) + o_P(n^{-1/2}) .$$

□

2.7 Additional simulation results

At the end of section 2.2.2, we mentioned that in situations in which the observed distribution of S has a point mass at zero, corresponding to one-half the marker LLOD, it could be useful to include an indicator variable $I(s = 0)$ in the Cox model linear predictor. In this section, we augmented two of the three models considered in section 2.3 by adding this indicator variable to the linear predictor and tested these models via simulation. The two models augmented were the basic Cox model (with the marker modeled as a linear term), and the Cox model that includes an embedded natural cubic spline basis with four degrees of freedom.

This set of simulations described here is analogous to the set described in section 2.3, but with two differences. First, the distribution of S included a point mass of roughly 10% at zero. Second, data were generated from a Cox model that included a step function in the linear predictor corresponding to the biomarker; this resulted in a conditional survival function of the form

$$Q_0^{(4)}(t \,|\, x, s) \coloneqq \exp\left\{-\lambda_1 t^{v_1} \exp(\alpha_1 x_1 + \alpha_2 x_2 + 0.5\alpha_3 I(s > 0))\right\} .$$

We generated data corresponding to $Q_0^{(4)}$ above, as well as $Q_0^{(1)}$, and focused on the case where the marginal distribution is Normal (with the added point mass). As in section 2.3, performance was evaluated via bias, confidence interval coverage, and standard deviation.

Figure 2.8 shows that, as expected, all four estimators are roughly unbiased when the true model is linear. When the true model has a step function in the linear predictor corresponding to the marker, the models that incorporate the edge indicator variable are roughly unbiased, whereas the basic Cox model and the spline model have substantial bias. The basic Cox model is biased for nearly the entire unit interval, whereas the spline model is minimally biased at the edge itself and over roughly $[0.25, 1]$, but badly biased over $(0, 0.25]$. The only model that has substantial bias in estimating $\text{CVE}_0(0)$ is the basic Cox model.

In Figure 2.9, it can be seen that all estimators roughly maintain nominal coverage when the true

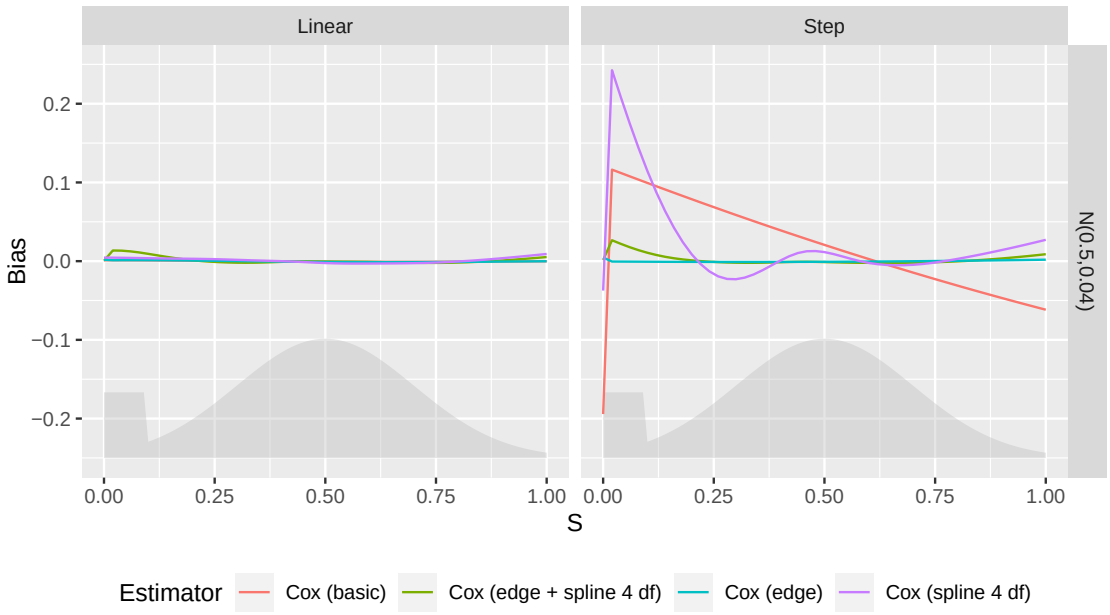


Figure 2.8: Bias of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (edge)" = Basic Cox model augmented with an indicator for S=0), "Cox (edge + spline 4df)" = Cox spline model augmented with an indicator for S=0), displayed for two conditional survival functions ("Linear" $= Q_0^{(1)}$, "Step" $= Q_0^{(4)}$) and a $N(0.5, 0.04)$ distribution for S.

model is linear, but the models that do not include the edge indicator result in poor coverage near the edge.

Figure 2.10 shows that, as expected, the basic Cox model has the lowest standard deviation whereas the spline model that includes the edge indicator has the highest standard deviation. In particular, the standard deviation of the spline model that includes the edge indicator is huge for values just above zero. This can result in highly erratic sample paths for this model, and for this reason, we do not recommend its use. However, the basic Cox model augmented with an indicator variable at the edge may be useful if a point mass is expected at the edge.

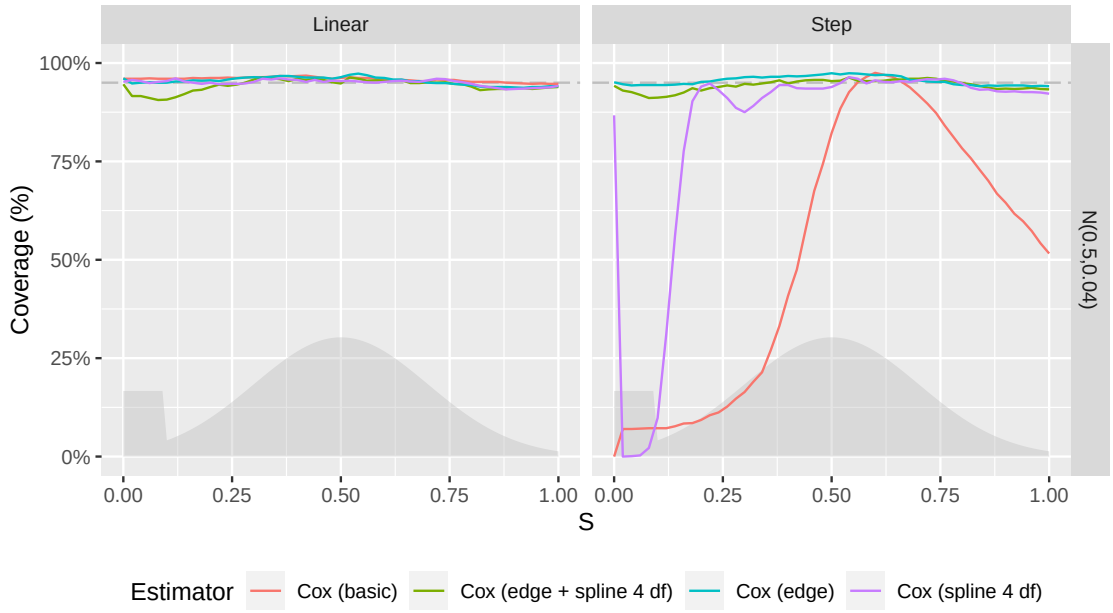


Figure 2.9: 95% confidence interval coverage of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (edge)" = Basic Cox model augmented with an indicator for S=0), "Cox (edge + spline 4df)" = Cox spline model augmented with an indicator for S=0), displayed for two conditional survival functions ("Linear" = $Q_0^{(1)}$, "Step" = $Q_0^{(4)}$) and a $N(0.5, 0.04)$ distribution for S.

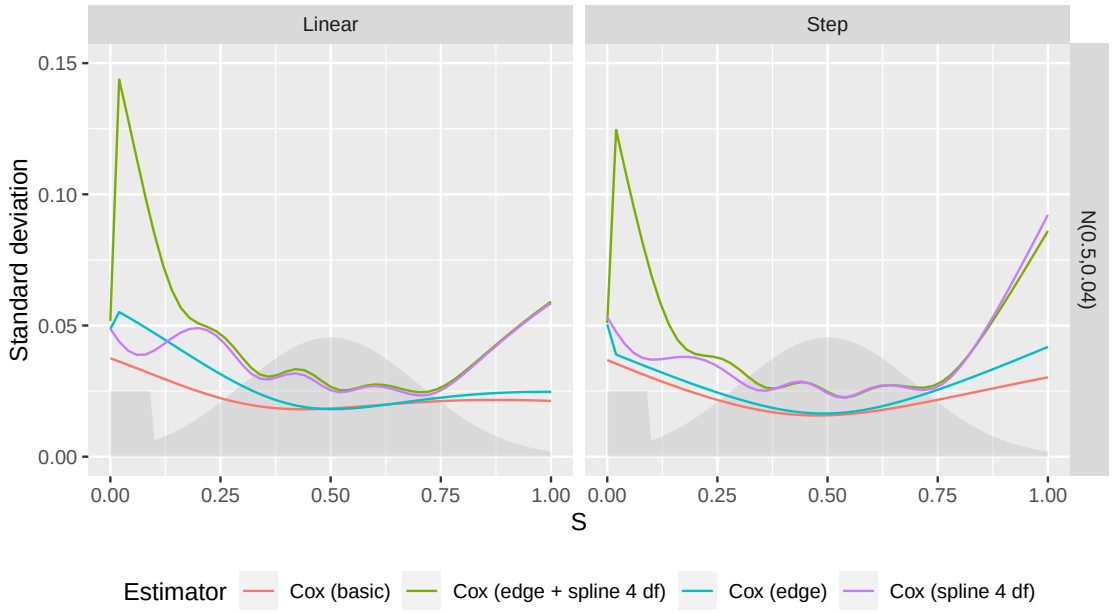


Figure 2.10: Standard deviation of three estimators ("Cox (basic)" = Cox model with a linear term for the marker, "Cox (spline 4 df)" = Cox model with a spline with 4 degrees of freedom, "Cox (edge)" = Basic Cox model augmented with an indicator for S=0), "Cox (edge + spline 4df)" = Cox spline model augmented with an indicator for S=0), displayed for two conditional survival functions ("Linear" = $Q_0^{(1)}$, "Step" = $Q_0^{(4)}$) and a $N(0.5, 0.04)$ distribution for S.

2.8 Additional results from the immune correlates analysis of the Coronavirus Efficacy (COVE) trial

2.8.1 Additional biomarkers

In section 2.4, results were given for two of the four biomarkers considered by Gilbert et al. (2022b) as part of their immune correlates analysis. Here, results are presented for the other two markers, IgG bAbs to the spike protein and 80% inhibitory dilution (ID80) nAb titer. Figure 2.11 shows CR and CVE curves for the IgG bAbs to spike receptor binding domain (RBD) marker, again estimated using two models, a basic Cox proportional hazards model and a Cox model including an embedded natural cubic spline basis with four degrees of freedom. Subfigures (2.11a) and (2.11b) correspond to the day 29 marker measurement whereas subfigures (2.11c) and (2.11d) correspond to the day 57 measurement.

Figure 2.12 shows analogous results for the 80% inhibitory dilution (ID80) nAb titer marker.

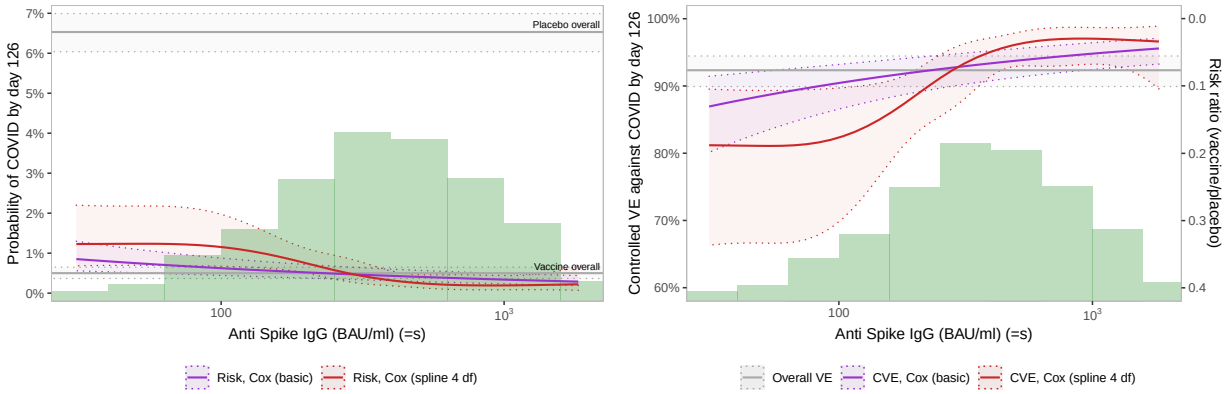
2.8.2 Use of LLOD indicator function

In section 2.4, we provided data analysis results for two representative biomarkers using two models, a basic Cox proportional hazards model (with the biomarker included as a single term in the linear predictor) and a Cox model including an embedded natural cubic spline basis with four degrees of freedom. In this section, we demonstrate the use of the augmented basic Cox model that includes an indicator variable at the left edge, corresponding to $S = 0$; this model was described in section 2.2.2 and tested via simulation in section 2.7. Select results are displayed to illustrate several practical properties of this marker.

Figure 2.13 displays CVE curves for the 80% inhibitory dilution (ID80) nAb titer marker, measured at day 29. The estimated curve from the edge model appears roughly similar to that of the spline model in that the estimated value of $\text{CVE}_0(0)$ is exactly the same and the curve flattens out at about 95% shortly thereafter. These two models differ in that the edge model allows for an immediate jump after the edge value, whereas the spline model results in a smoother curve. On the other hand, the basic Cox model results in a smooth, gradual estimated curve, which is more

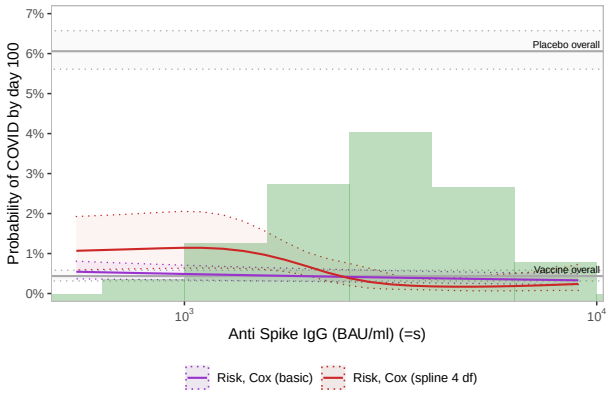
likely driven by the constraints of the model rather than the actual shape of the curve.

Figure 2.14 displays CVE curves for the 50% inhibitory dilution (ID80) nAb titer marker, measured at day 29. This figure demonstrates a potential drawback of the edge model, which is that it can result in estimated curves that are non-monotonic and may appear implausible.

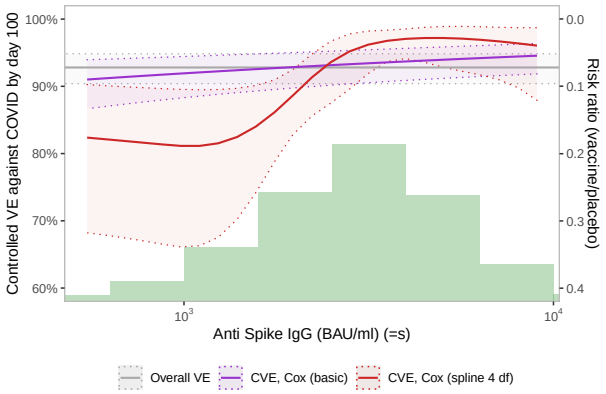


(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29

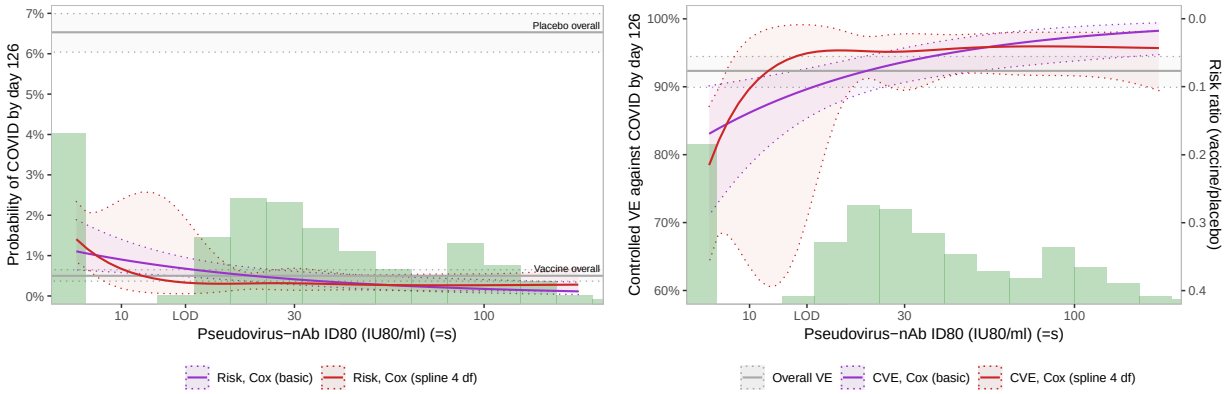


(c) Controlled risk, day 57



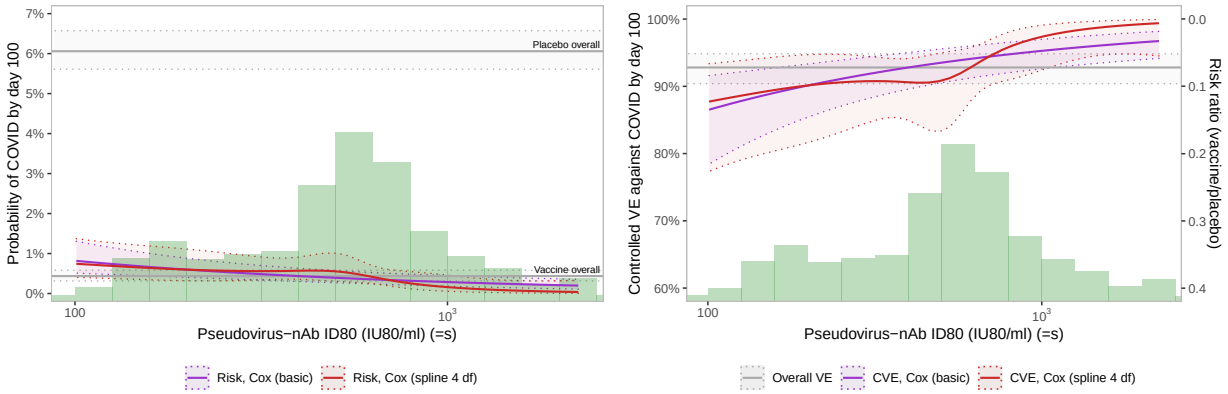
(d) Controlled vaccine efficacy, day 57

Figure 2.11: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the IgG bAbs to the spike protein marker, measured at days 29 and 57. Curves are estimated using a basic Cox model (purple) and a Cox model with a spline with 4 degrees for freedom (red). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.



(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29



(c) Controlled risk, day 57

(d) Controlled vaccine efficacy, day 57

Figure 2.12: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the 80% inhibitory dilution (ID80) nAb titer marker, measured at days 29 and 57. Curves are estimated using a basic Cox model (purple) and a Cox model with a spline with 4 degrees for freedom (red). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

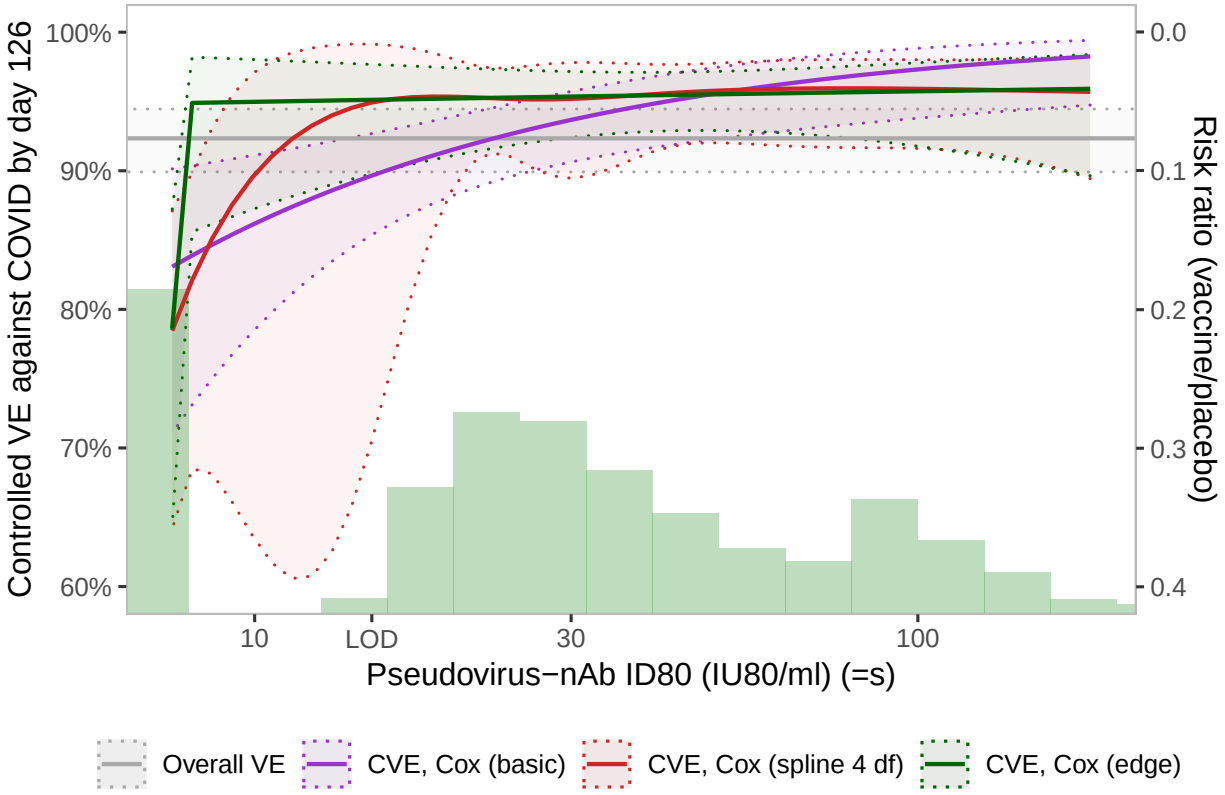


Figure 2.13: Controlled vaccine efficacy (CVE) curve for the 80% inhibitory dilution (ID80) nAb titer marker, measured at day 29. Curves are estimated using a basic Cox model (purple), a Cox model with a spline with 4 degrees for freedom (red), and a basic Cox model augmented with an edge indicator variable (green). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

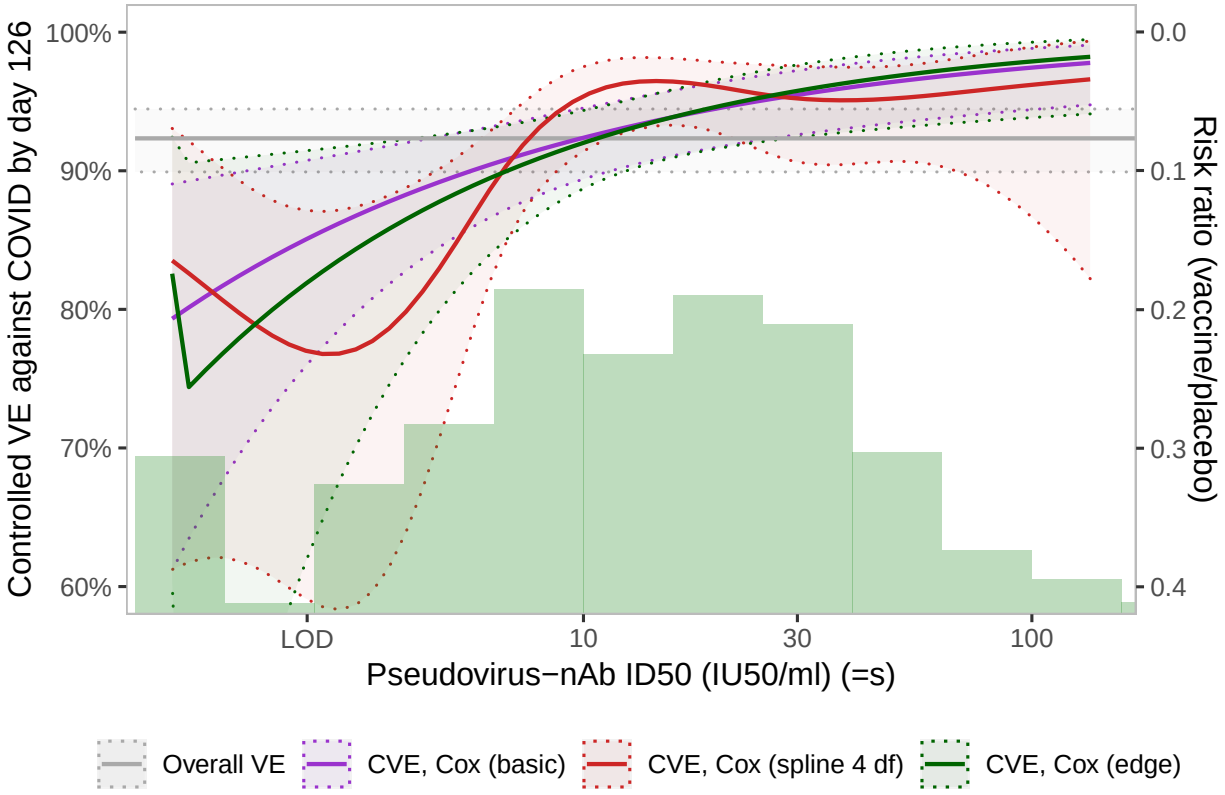


Figure 2.14: Controlled vaccine efficacy (CVE) curve for the 50% inhibitory dilution (ID50) nAb titer marker, measured at day 29. Curves are estimated using a basic Cox model (purple), a Cox model with a spline with 4 degrees for freedom (red), and a basic Cox model augmented with an edge indicator variable (green). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

Chapter 3

Nonparametric estimation and inference for the CR and CVE curves

Throughout this chapter, the data structure is exactly the same as that described in section 2.1. Our primary objective is to develop and evaluate nonparametric procedures that allow us to perform the following two inferential tasks:

1. Estimate the functions $s \mapsto r_{M,0}(s)$ and $s \mapsto \text{CVE}_0(s)$ over $s \in [0, 1]$ and construct corresponding pointwise confidence bands.

2. Test the null hypothesis that the function $s \mapsto \text{CVE}_0(s)$ is constant against the alternative hypothesis that $s \mapsto \text{CVE}_0(s)$ is nonconstant and nonincreasing.

For the first objective, since we have assumed monotonicity of $r_{M,0}(s)$, we can leverage the general results of Westling and Carone (2020) on nonparametric inference for monotone functions. The resulting procedure can be seen as a generalization of causal isotonic regression (Westling et al., 2020) to settings involving a survival data structure and two-phase sampling. In section 3.1.8, we show how simple extensions of this method allows for estimation and inference for the CVE curve. For the second objective, we develop a novel directional hypothesis test that leverages the monotonicity assumption.

3.1 Methods: estimation

3.1.1 Review of isotonic regression and generalized Grenander-type estimators

We first briefly review the classical form of isotonic regression (Barlow and Brunk, 1972), in which the observed data consist of iid pairs $(S_1, Y_1), ..., (S_n, Y_n)$ drawn from some bivariate distribution and the quantity of interest is a univariable regression function $\theta_0 : s \mapsto E_0[Y \mid S = s]$. The *isotonic regression estimator* of θ_0 is defined as

$$\theta_n \coloneqq \underset{r \in \mathcal{R}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left\{ Y_i - r(S_i) \right\}^2 \right],$$

the minimizer in r of the residual sum of squares over the space $\mathcal{R}$ of all non-decreasing functions. The estimator θ_n takes the form of a step function that can be found by using the pool adjacent violators algorithm (PAVA) (Groeneboom and Jongbloed, 2014). Of course, this estimator is only sensible if the true regression θ_0 is itself nondecreasing. Although the PAVA approach perhaps gives the best intuition for the behavior of θ_n, we consider an alternative representation in terms of the greatest convex minorant (GCM) operator, as this representation facilitates a connection between isotonic regression and our method. The GCM is defined for a bounded function f on a closed interval as the supremum over all convex functions g defined on that interval that satisfy $g \leq f$. Let F_n^S denote the empirical distribution function of S based on $(S_1, ..., S_n)$. The *cumulative sum diagram* is defined as the linear interpolation of the set of points $(0, 0) \cup \{(F_n^S(S_i), \Gamma_n(S_i)) : i \in \{1, ..., n\}\}$, where $\Gamma_n(u) \coloneqq \frac{1}{n} \sum_{i=1}^{n} Y_i I(S_i \leq u)$. Denoting the GCM of the cumulative sum diagram over $[0, 1]$ as $s \mapsto \bar{\Psi}_n(s)$, then for any point u, the isotonic regression estimator can be expressed as

$$\theta_n(u) \coloneqq \frac{d}{dt} \bar{\Psi}_n(t) \Big|_{t = F_n^S(u)},$$

where here d/dt represents a left derivative. Since the GCM is by definition convex, its derivative is nondecreasing, and so θ_n is guaranteed to be nondecreasing.

To build intuition for why this procedure works, one can study the simple case in which $S_1, ..., S_n$ are standard uniform random variables. Then, for a large sample, F_n^S roughly equals the identity

function, and so θ_n roughly equals the left derivative of $\bar{\Psi}_n$. Furthermore, each y-coordinate $\Gamma_n(u)$ of the cumulative sum diagram can be seen as an estimator of the primitive $\Gamma_0(u) \coloneqq E_0\{YI(S \leq u)\} = \int_0^u \theta_0(t)dt$. Since θ_0 equals the derivative of Γ_0, this procedure can be seen as finding a convex function $\bar{\Psi}_n$ that approximates the estimated primitive function Γ_n and then differentiating the result. When S is not uniformly distributed, $\Gamma_0(u) = \int_0^u \theta_0(t)dF_0^S(t)$ instead, where F_0^S is the distribution function of S. It may seem circuitous that, instead of estimating the regression function θ_0 directly, the primitive is estimated and then differentiated. The reason why this strategy is sensible is that the primitive $\Gamma_0(u)$, when expressed as a functional of a distribution within a fully nonparametric model, is pathwise differentiable at the true distribution P_0 and thus estimable at root-n rate under regularity conditions, whereas this is not the case for $\theta_0(u)$.

Both the classical isotonic regression and the causal isotonic regression estimator of Westling et al. (2020) fall in the class of so-called *generalized Grenander-type* estimators. This class of estimators was studied by Westling and Carone (2020), who derived general results on nonparametric inference for monotone functions. We will leverage their results to construct a generalized Grenander-type nonparametric estimator of the controlled risk curve, which will in turn be used to construct an estimator of the CVE curve.

The motivation for this class of estimators can be described as follows. First, define $\Phi_0 : [0,1] \to [0,1]$ to be an arbitrary nondecreasing, càdlàg (right-continuous with left-hand limits) function, and $\Phi_0^- : x \mapsto \inf\{u \in [0,1] : \Phi_0(u) \geq x\}$ to be its generalized inverse. Also define $\Psi_0(s) \coloneqq \int_0^s \theta_0\{\Phi_0^-(t)\}dt$ and let $\bar{\Psi}_0$ denote the greatest convex minorant of Ψ_0. For any value $s \in [0,1]$ such that $\theta_0(s)$ is left-continuous and $\Phi_0(u) < \Phi_0(s)$ for all $u < s$, it can be shown that $\theta_0(s)$ equals the left derivative of $\bar{\Psi}_0$ evaluated at $\Phi_0(s)$. Defining $\Gamma_0 \coloneqq \Psi_0 \circ \Phi_0$, note that $\Psi_0 = \Gamma_0 \circ \Phi_0^-$ and $\Gamma_0(u) = \int_0^u \theta_0(t)d\Phi_0(t)$. Westling and Carone (2020) proceed by studying estimators of the form

$$\theta_n(s) \coloneqq \frac{d}{du}\bar{\Psi}_n(u)\bigg|_{u=\Phi_n(s)}, \tag{3.1}$$

where Γ_n and Φ_n^- are estimators of Γ_0 and Φ_0^-, respectively, $\Psi_n \coloneqq \Gamma_n \circ \Phi_n^-$, and $\bar{\Psi}_n$ is the GCM of Ψ_n. Estimators of this form are referred to as generalized Grenander-type estimators. The

function $\bar{\Psi}_n$ can equivalently be expressed as the GCM of the linear interpolation of the set of points $(0,0) \cup \{(\Phi_n(S_i), \Gamma_n(S_i)) : i \in \{1, ..., n\}\}$. The need to estimate Ψ_0 is clear, since its definition contains the unknown function θ_0. However, the purpose of the function Φ_0 is more subtle and context-dependent. In general, this function allows the resulting estimators to include a data-dependent domain transformation, which can be advantageous in many settings; see Westling and Carone (2020) for further discussion.

If we have $\theta_0(s) = E_0[Y|S = s]$ and set $\Phi_0 \coloneqq F_0^S$, then $\Gamma_0(u) = \int_0^u \theta_0(t) dF_0^S(s) = E_0\{YI(S \leq u)\}$. Plugging in the estimators $\Gamma_n(u) \coloneqq \frac{1}{n}\sum_{i=1}^n Y_i I(S_i \leq u)$ and F_n^S yields the classical isotonic regression estimator.

3.1.2 Definition of proposed estimator

In our setting, $\theta_0(s) \coloneqq r_{M,0}(s)$. For a fixed point $u \in [0,1]$, the primitive $\Gamma_0(u)$ can be expressed as

$$\Gamma_0(u) = \int_0^u r_{M,0}(s) dF_0^S(s) = \int\int I(s \leq u) \left\{1 - Q_0(t_0 \,|\, x, s)\right\} dF_0^S(s) dF_0^X(x)\,, \tag{3.2}$$

where $Q_0(t_0 \,|\, x, s) \coloneqq P_0(T \geq t_0 \,|\, X = x, S = s)$ is a conditional survival function. The primitive $\Gamma_0(u)$, when considered as a functional of a distribution P evaluated at $P = P_0$, is a pathwise differentiable parameter in a nonparametric model; throughout this work, when we refer to population quantities as parameters, we are implicitly considering them as functionals of a distribution, evaluated at P_0. We proceed by deriving the nonparametric efficient influence function of Γ_0 at P_0 and constructing a one-step estimator Γ_n. First, it can be shown that under the full-data structure (in which S is observed for every individual), the nonparametric efficient influence function of $\Gamma_0(u)$ at P_0 is given by

$$\tilde{\varphi}_{\Gamma,0} : (x, y, \delta, s, u) \mapsto I(s \leq u) \left\{ \frac{\omega_0(x, y, \delta, s)}{g_0(x, s)} + r_{M,0}(s) \right\} + \eta_0(x, u) - 2\Gamma_0(u)\,.$$

This influence function involves the nuisances

$$\omega_0(x,y,\delta,s) \coloneqq Q_0(t_0 \,|\, x,s)\left\{\frac{\delta I(y \le t_0)}{Q_0(y \,|\, x,s)Q_0^C(y \,|\, x,s)} - \int_0^{t_0 \wedge y} \frac{\Lambda_0(dt \,|\, x,s)}{Q_0(t \,|\, x,s)Q_0^C(t \,|\, x,s)}\right\},$$
$$g_0(x,s) \coloneqq f_0^{S|X}(s \,|\, x)/f_0^S(s)\,,$$
$$\eta_0(x,u) \coloneqq \int I(s \le u)\left\{1 - Q_0(t_0 \,|\, x,s)\right\} dF_0^S(s)\,,$$

where $Q_0^C(t_0 \,|\, x,s) \coloneqq P_0(C \ge t_0 \,|\, X = x, S = s)$ is a conditional survival function of the censoring variable C and $t \mapsto \Lambda_0(t \,|\, x,s) \coloneqq -\log Q_0(t|x,s)$ is the conditional cumulative hazard function of T. However, as described in section 2.1.1, S is not observed for every individual. To proceed, we apply a result from Rose and van der Laan (2011) that allows us to determine the form of the influence function under two-phase sampling given the influence function in the full-data setting. Suppose that the missing at random (MAR) assumption holds for S; that is, $\pi_0(x,y,\delta) \coloneqq P_0(Z = 1 \,|\, X = x, Y = y, \Delta = \delta) = P_0(Z = 1 \,|\, X = x, Y = y, \Delta = \delta, S = s)$ for any (x,y,δ,s). Then the efficient influence function of $\Gamma_0(u)$ can be written as

$$\varphi_{\Gamma,0} : (o,u) \mapsto \frac{z}{\pi_0(o)}\tilde{\varphi}_{\Gamma,0}(o,u) + \left\{1 - \frac{z}{\pi_0(o)}\right\} E_0\left\{\tilde{\varphi}_{\Gamma,0}(O,u)\big|\, Z = 1, X = x, Y = y, \Delta = \delta\right\}\,, \quad (3.3)$$

where as in section 2.1.1 $o \coloneqq (x,y,\delta,z,zs)$ is shorthand for the observed data unit, $\tilde{\varphi}_{\Gamma,0}(o,u) \coloneqq \tilde{\varphi}_{\Gamma,0}(x,y,\delta,zs,u)$, and $\pi_0(o) \coloneqq \pi_0(x,y,\delta)$. The efficiency of (3.3) is relative to a model that includes the MAR assumption for S, as well as the assumptions related to the survival data structure outlined in section 2.1.1. Next, define the nuisance function q_0 as

$$q_0(x,y,\delta,u) \coloneqq E_0\left[q_0^*(X,Y,\Delta,S,u) \,|\, Z = 1, X = x, Y = y, \Delta = \delta\right]\,,$$

where for convenience, we use the shorthand $q_0(o,u) \coloneqq q_0(x,y,\delta,u)$, and where

$$q_0^*(x,y,\delta,s,u) \coloneqq I(s \le u)\left\{\frac{\omega_0(x,y,\delta,s)}{g_0(x,s)} + r_{M,0}(s)\right\}\,.$$

This allows us to write the influence function given in (3.3) as

$$\varphi_{\Gamma,0} : (o,u) \mapsto \frac{zI(s \leq u)}{\pi_0(o)} \left\{ \frac{\omega_0(o)}{g_0(x,s)} + r_{M,0}(s) \right\} + \left\{ 1 - \frac{z}{\pi_0(o)} \right\} q_0(o,u) + \eta_0(x,u) - 2\Gamma_0(u)\,. \quad (3.4)$$

The difficulty in constructing a one-step estimator $\Gamma_n(u)$ of $\Gamma_0(u)$ is that we must be able to specify a complete set of nuisance estimators that are consistent with one another in the sense that the estimated influence function $\varphi_{\Gamma,n}$ can be expressed as a functional of a well-defined distribution $\hat{P}_n$; in section 3.1.4, we outline one strategy for specifying a set of nuisances. With these nuisance estimators in hand, define the one-step estimator $\Gamma_n(u)$ as

$$\Gamma_n(u) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Z_i I(S_i \leq u)}{\pi_0(O_i)} \left\{ \frac{\omega_n(O_i)}{g_n(X_i,S_i)} + \tilde{r}_{M,n}(S_i) \right\} + \left\{ 1 - \frac{Z_i}{\pi_0(O_i)} \right\} q_n(O_i,u) \right].$$

Finally, applying the general specification given in (3.1), our estimator is defined as

$$r_{M,n}(s) \coloneqq -\frac{d}{du} \bar{\Psi}_n(u) \Big|_{u=F_n^S(s)}\,, \quad (3.5)$$

where F_n^S is an estimator of the distribution function F_0^S (described in section 3.1.4), F_n^{S-} is the generalized inverse of F_n^S, $\Psi_n \coloneqq -\Gamma_n \circ F_n^{S-}$, and $\bar{\Psi}_n$ is the GCM of Ψ_n; note that the negative signs are present since $r_{M,0}$ is a decreasing function.

3.1.3 Asymptotic properties

Since $r_{M,n}(s)$ is a generalized Grenander-type estimator, the results of Westling and Carone (2020) can be applied to study its asymptotic behavior. In the following theorem, sufficient conditions are provided to guarantee consistency of $r_{M,n}(s)$ for any $s \in (0,1)$ and obtain a distributional limit.

Theorem 3. *Fix a point $s \in (0,1)$ and suppose that $r_{M,0}$ is differentiable and strictly decreasing, F_0^S is continuously differentiable with a positive derivative, and that the regularity conditions (A4) – (A11) given in Appendix 3.6.1 hold. Then $r_{M,n}(s) \xrightarrow{p} r_{M,0}(s)$ and*

$$n^{1/3} \left\{ r_{M,n}(s) - r_{M,0}(s) \right\} \rightsquigarrow \tau_0(s)\mathbb{Z}\,, \quad (3.6)$$

where $\mathbb{Z}$ follows the Chernoff distribution and where

$$\begin{aligned}
\dot{r}_{M,0}(s) &\coloneqq \frac{d}{du} r_{M,0}(u)\Big|_{u=s}, \\
\gamma_0(x,s) &\coloneqq E_0\left[\left\{\frac{\omega_0(X,Y,\Delta,S)}{\pi_0(X,Y,\Delta)}\right\}^2 \middle| X=x, S=s, Z=1\right], \\
g_{z,0}(x,s) &\coloneqq P_0(Z=1 \,|\, X=x, S=s), \\
\tau_0(s) &\coloneqq \left\{4\dot{r}_{M,0}(s)\int \frac{\gamma_0(x,s)g_{z,0}(x,s)}{f_0^{S|X}(s\,|\,x)} dF_0^X(x)\right\}^{1/3}.
\end{aligned}$$

The distributional limit given in (3.6) can be used to form a pointwise confidence band for the function $s \mapsto r_{M,0}(s)$. Also, since $r_{M,0}(s) \in [0,1]$ for any s, it may be desirable to form confidence intervals on the logit$\{r_{M,n}(s)\}$ scale to ensure that the resulting limits are also contained within $[0,1]$. Consistent estimation of the scale factor $\tau_0(s)$ follows from consistent estimation of the component nuisance estimators, discussed further in section 3.1.4. Doubly-robust estimation of $\tau_0(s)$ may be possible but is not considered in this work.

Note that the validity of pointwise confidence intervals constructed using (3.6) requires the assumption in Theorem 3 that the marginalized risk curve is strictly decreasing. If instead, for a fixed value s_0, the true curve is flat in a neighborhood around s_0 (i.e., for some $\epsilon > 0$, $s \mapsto r_{M,0}(s)$ is constant for all $s \in [s_0 - \epsilon, s_0 + \epsilon]$), then the asymptotic results of Theorem 3 will not apply to $r_{M,n}(s_0)$. This is a notable limitation of this method; the construction of pointwise confidence intervals that are valid in both the strictly decreasing case and the nonincreasing case represents a worthwhile direction for future research.

3.1.4 Estimation of nuisance parameters

As discussed in section 3.1.2, the construction of a one-step estimator $\Gamma_n(u)$ of $\Gamma_0(u)$ requires that a complete set of nuisance estimators can be specified that are consistent with one another. For the distribution function F_0^S, we can use the inverse-probability-of-sampling (IPS) weighted empirical distribution function given by

$$F_n^S(u) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i I(S_i \leq u)}{\pi_0(O_i)}\,. \tag{3.7}$$

This strategy of estimating an expectation (involving the biomarker variable S) using an IPS-weighted empirical mean estimator will be used repeatedly. Next, the conditional survival functions Q_0 and Q_0^C can be estimated using any regression technique that allows for the incorporation of censoring information and observation weights, provided that the estimators satisfy the regularity conditions of Appendix 3.6.1. We recommend either the approach of Wolock et al. (2022) or the approach of Westling et al. (2021), both of which allow for the use of flexible machine learning estimators with tuning parameters selected using cross-validation. Once estimators Q_n and Q_n^C are obtained, define the conditional cumulative hazard function estimator as $\Lambda_n(t \,|\, x, s) \coloneqq -\log Q_n(t \,|\, x, s)$ and define ω_n as

$$\omega_n(x, y, \delta, s) \coloneqq Q_n(t_0 \,|\, x, s) \left\{ \frac{\delta I(y \leq t_0)}{Q_n(y \,|\, x, s) Q_n^C(y \,|\, x, s)} - \int_0^{t_0 \wedge y} \frac{\Lambda_n(dt \,|\, x, s)}{Q_n(t \,|\, x, s) Q_n^C(t \,|\, x, s)} \right\},$$

where the integral is approximated numerically. Similarly, the estimators Q_n, Q_n^C, and F_n^S are used to define the nuisance estimators

$$\begin{aligned}
\eta_n(x, u) &\coloneqq \int I(s \leq u) \left\{1 - Q_n(t_0 \,|\, x, s)\right\} dF_n^S(s) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i I(S_i \leq u) \left\{1 - Q_n(t_0 \,|\, x, S_i)\right\}}{\pi_0(O_i)}\,, \\
\tilde{r}_{M,n}(s) &\coloneqq 1 - \frac{1}{n}\sum_{i=1}^{n} Q_n(t_0 \,|\, X_i, s)\,, \\
\tilde{\Gamma}_n(u) &\coloneqq \int I(s \leq u) \tilde{r}_{M,n}(s) dF_n^S(s) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i I(S_i \leq u) \tilde{r}_{M,n}(S_i)}{\pi_0(O_i)}\,.
\end{aligned}$$

To construct an estimator g_n of the density ratio g_0, the conditional density $f_n^{S|X}(s \,|\, x)$ can first be constructed using any conditional density estimator that allows for the incorporation of observation weights. Then, since $f_0^S(s) = \int f_0^{S|X}(s \,|\, x) dF_0^X(x)$, estimators of f_0^S and g_0 can be constructed as

$$f_n^S(s) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_n^{S|X}(s \mid X_i)\,,$$
$$g_n(x,s) \coloneqq f_n^{S|X}(s \mid x)/f_n^S(s)\,.$$

Next, an estimator q_n of the regression q_0 must be constructed that is consistent with the other nuisance estimators in the sense discussed at the end of section 3.1.2. Since q_0 is an integral with respect to the conditional density of S given (Z, X, Y, Δ), first note that this density can be decomposed as

$$f_0^{S|Z,X,Y,\Delta}(s \mid z,x,y,\delta) = \frac{f_0^{Y,\Delta|X,S}(y,\delta \mid x,s)g_0(x,s)f_0^S(s)}{\int f_0^{Y,\Delta|X,S}(y,\delta \mid x,t)g_0(x,t)dF_0^S(t)}\,.$$

This result allows us to express q_0 as

$$\begin{aligned} q_0(o,u) &= \int q_0^*(x,y,\delta,s,u) f_0^{S|Z,X,Y,\Delta}(s \mid 1,x,y,\delta)ds \\ &= \left\{\int f_0^{Y,\Delta|X,S}(y,\delta \mid x,s)g_0(x,s)dF_0^S(s)\right\}^{-1} \\ &\quad \times \int q_0^*(x,y,\delta,s,u) f_0^{Y,\Delta|X,S}(y,\delta \mid x,s)g_0(x,s)dF_0^S(s)\,. \end{aligned}$$

Next, using the estimated conditional survival functions of T and C, an estimator of $f_0^{Y,\Delta|X,S}$ is given by

$$f_n^{Y,\Delta|X,S}(y,\delta \mid x,s) \coloneqq -I(\delta=1)Q_n^C(y \mid x,s)\dot{Q}_n(y \mid x,s) - I(\delta=0)Q_n(y \mid x,s)\dot{Q}_n^C(y \mid x,s)\,, \quad (3.8)$$

where $\dot{Q}_n$ and $\dot{Q}_n^C$ are the numerically-approximated derivatives of Q_n and Q_n^C, respectively. If the estimated conditional survival and censoring functions are not differentiable, they can be smoothed (e.g., using kernel smoothing); however, this is not strictly necessary, as a numerical derivative based on function evaluations over a grid will be well-defined even if the underlying function is not differentiable. Finally, defining q_n^* to equal q_0^* but with the nuisances ω_0, g_0, $r_{M,0}$, and Γ_0 replaced by ω_n, g_n, $\tilde{r}_{M,n}$, and $\tilde{\Gamma}_n$, respectively, q_n is defined as

$$q_n(o,u) \coloneqq \left\{ \sum_{i=1}^{n} \frac{Z_i f_n^{Y,\Delta|X,S}(y,\delta \,|\, x, S_i) g_n(x, S_i)}{\pi_0(O_i)} \right\}^{-1} \times \left\{ \sum_{i=1}^{n} \frac{Z_i q_n^*(S_i, o, u) f_n^{Y,\Delta|X,S}(y,\delta \,|\, x, S_i) g_n(x, S_i)}{\pi_0(O_i)} \right\}.$$

Finally, we describe a strategy for constructing a consistent estimator of the Chernoff scale factor $\tau_0(s)$ for fixed s. We start by describing one set of possible estimators of the underlying nuisance functions $\dot{r}_{M,0}(s)$, $\gamma_0(x,s)$, and $g_{z,0}(x,s)$.

Estimation of $\dot{r}_{M,0}(s)$ is complicated by the fact that $s \mapsto r_{M,n}(s)$ is a step function by construction. As noted by Westling et al. (2020), one approach is to construct a smoothed version of $r_{M,n}$ and then differentiate the result. They suggest using local quadratic kernel smoothing of the linear spline defined by the set of points $\mathcal{S} \coloneqq \{(s^*, r_{M,n}(s^*)) : s^* \in \mathcal{S}^*\}$, where $\mathcal{S}^*$ is the set of midpoints of the jump points of $r_{M,n}$. Another strategy is to just numerically differentiate the linear spline itself; as with the conditional survival and censoring functions, although the spline is technically nondifferentiable at the knots, a numerical derivative based on function evaluations over a grid is well-defined everywhere. Another strategy that we have found to work well in practice is to apply the monotone Hermite spline method of Fritsch and Carlson (1980) to the set of points $\mathcal{S}$ (where we choose to also include the endpoints 0 and 1 in the set $\mathcal{S}^*$) and then differentiate the result. All three methods described guarantee that the resulting derivative estimator $\dot{r}_{M,n}(s)$ is positive for $s \in [0,1]$, provided that $r_{M,n}$ has at least one jump point.

For $\gamma_0(x,s)$, an estimator $\gamma_n(x,s)$ can be constructed by regressing the pseudo-outcomes $\{\omega_n(O_i)/\pi_0(O_i)\}^2$ on (X_i, S_i) among the subset of the data for which $\Delta_i = 1$ using any suitable regression technique. To estimate $g_{z,0}(x,s) = P_0(Z = 1 \,|\, X = x, S = s)$, an application of Bayes' rule yields $g_{z,0}(x,s) = f_0^{S|X,Z}(s \,|\, x, 1)\tilde{g}_{z,0}(x) / f_0^{S|X}(s \,|\, x)$, where $f_0^{S|X,Z}(s \,|\, x, z)$ is the conditional density of S given (X, Z) and $\tilde{g}_{z,0}(x) \coloneqq P_0(Z = 1 \,|\, X = x)$. An estimator of $f_n^{S|X,Z}$ of $f_0^{S|X,Z}$ can be constructed using the same conditional density estimator, but restricted to the subset of the data for which $Z = 1$, since it does not need to be evaluated for $z = 0$. Since the function $\pi_0(x, y, \delta) \coloneqq P_0(Z = 1 \,|\, X = x, Y = y, \Delta = \delta)$ is known and since $\tilde{g}_{z,0}(x) = E_0\{\pi_0(x, Y, \Delta)\}$ holds,

the estimator

$$\tilde{g}_{z,n}(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \pi_0(x, Y_i, \Delta_i)$$

can be used. Alternatively (or if π_0 is unknown), $\tilde{g}_{z,0}$ can be estimated using any suitable technique for regression of binary outcomes. Putting these together, we can define

$$g_{z,n}(x, s) \coloneqq \frac{f_n^{S|X,Z}(s \mid x, 1)}{f_n^{S|X}(s \mid x)} \tilde{g}_{z,n}(x)\,.$$

Finally, with these component nuisance estimators in hand, define $\tau_n(s)$ as

$$\tau_n(s) \coloneqq \left\{ \frac{4\dot{r}_{M,n}(s)}{n f_n^S(s)} \sum_{i=1}^{n} \frac{\gamma_n(X_i, s) g_{z,n}(X_i, s)}{g_n(X_i, s)} \right\}^{1/3}.$$

By the continuous mapping theorem, $\tau_n(s)$ is consistent for $\tau_0(s)$ if each of the component estimators $\dot{r}_{M,n}(s)$, $\gamma_n(x, s)$, and $g_{z,n}(x, s)$ is consistent.

3.1.5 Use of estimated IPS weights to gain precision

Thus far, we have assumed that the two-phase sampling probabilities, governed by the function π_0, are known exactly, and that the sampling indicator Z is selected independently for each individual via Bernoulli sampling. However, in practice this might not be the case. First, even if the researcher dictates the form of π_0, real-world complications such as loss-to-follow-up and missingness may lead to the actual two-phase sampling probabilities being different than the design probabilities. Second, other sampling strategies are sometimes used in practice, most notably finite population stratified sampling (FPSS) (Breslow and Lumley, 2013). With FPSS, the population is partitioned into a fixed number of strata based on available variables, and a predetermined fraction or number of individuals are sampled without replacement from each stratum. Suppose that each individual i is assigned to the stratum indexed by C_i, where C_i deterministically depends on (X_i, Y_i, Δ_i). The corresponding estimated FPSS sampling weight is then calculated as the fraction of individuals in stratum C_i who were selected into the phase-two sample, given by

$$\pi_n^*(O_i) := \frac{\sum_{j=1}^n I(C_i = C_j) Z_j}{\sum_{j=1}^n I(C_i = C_j)}, \tag{3.9}$$

where for consistency of notation, we consider C_i to be part of the vector O_i. One property of these estimated weights is that they always sum to one. This leads to desirable finite-sample properties, such as guaranteeing that the IPS-weighted empirical distribution function estimator given in (3.7), with the true weights replaced by the estimated weights, equals one when it is evaluated at the largest observed data point. Note that even if Bernoulli sampling is used, these weights can still be calculated, provided that (i) the range of the function π_0 is a finite set, effectively partitioning the observations into strata such that individuals i and j are in the same stratum if and only if $\pi_0(Y_i, \Delta_i, W_i) = \pi_0(Y_j, \Delta_j, W_j)$, and (ii) $\pi_n^*(O_i) \neq 0$ for all $i \in \{1, ..., n\}$. Additionally, even if the Bernoulli sampling probabilities are known exactly, there is often a benefit in terms of precision from using the estimated IPS weights given in (3.9) rather than the true IPS weights; this counterintuitive phenomenon has been documented in a number of settings (e.g., Scott and Wild, 1997; Deville and Särndal, 1992; Qi et al., 2005).

A complication is that true FPSS breaks the iid structure of the data-generating mechanism, since the phase-two selection indicators $Z_1, ..., Z_n$ are not independent, and so derivation of theoretical results under this sampling scheme is difficult. However, since the variance of an arbitrary estimator is typically smaller under FPSS relative to Bernoulli sampling, we can proceed by treating the data as coming from a Bernoulli sampling model, using the estimated weights, and noting that the resulting confidence intervals may be conservative. Thankfully, the use of estimated weights rather than known weights does not change the influence function given in (3.4) of Γ_n (Rose and van der Laan, 2011), nor does it change the asymptotic behavior of the resulting estimator $r_{M,n}(s)$. Additionally, although the influence function of F_n^S does change when the estimated weights are used, the asymptotic results given in Theorem 3 do not change, as discussed in Westling and Carone (2020).

3.1.6 Cross-fitting to avoid empirical process conditions

Regularity condition (A6) of Theorem 3 necessitates that as sample size increases, the nuisance estimators used to calculate the estimated influence function $\varphi_{\Gamma,n}$ (almost surely) fall in sufficiently small function classes. This condition is necessary to guarantee that the remainder term from the one-step estimator $\Gamma_n(u)$ is asymptotically negligible. The use of machine learning techniques to construct these nuisance estimators is appealing, as discussed in section 3.1.4, but certain classes of machine learning estimators may fail to satisfy this condition, and for other classes it may be difficult to determine whether this condition holds. However, many authors have noted that similar empirical process conditions can be avoided through the use of *cross-fitting*, a form of sample splitting (e.g., Bickel, 1982; Robins et al., 2008).

In our context, to form the cross-fitted version Γ_n° of the one-step estimator Γ_n, first partition the dataset into K subsets, each of size n/K (where for ease of exposition n/K is assumed to be an integer), indexed by the sets $\mathcal{V}_{n1}, ..., \mathcal{V}_{nK}$. Next, define $\mathcal{T}_{nk} \coloneqq \{i : i \notin \mathcal{V}_{nk}\}$ for $k \in \{1, ..., K\}$ and for each k, use the observations indexed by $\mathcal{T}_{nk}$ to form the nuisance estimators ω_{nk}, g_{nk}, and so on. The cross-fitted estimator $\Gamma_n^\circ(u)$ of $\Gamma_0(u)$ can then be defined as

$$\Gamma_n^\circ(u) \coloneqq \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_{nk}}\left[\frac{Z_i I(S_i \le u)}{\pi_0(O_i)}\left\{\frac{\omega_{nk}(O_i)}{g_{nk}(X_i, S_i)} + \tilde{r}_{M,nk}(S_i)\right\}\right.$$
$$\left. + \left\{1 - \frac{Z_i}{\pi_0(O_i)}\right\} q_{nk}(O_i, u) + \eta_{nk}(X_i, u) - \tilde{\Gamma}_{nk}(u)\right].$$

3.1.7 Handling bias at the edge

In general, monotone function estimators are known to have asymptotic bias when evaluated at the endpoints of the domain of the function of interest. Additionally, although values in the interior of the domain can be estimated consistently and without asymptotic bias, we have observed in simulation (see section 3.3) that finite sample bias occurs; typically the magnitude of the bias increases as the distance between the edge and the point at which the function is evaluated decreases. For a decreasing function, the estimator will, on average, overestimate the true function towards the lower endpoint and underestimate the function towards the upper endpoint. This is concerning,

since this bias is anti-conservative in the sense that the overall change in the value of the function over its domain will be overestimated. Here, we discuss two approaches to compensating for this shortcoming.

First, when reporting results, the display of the estimated function $s \mapsto r_{M,n}(s)$ can simply be truncated; for example, one may choose to only display estimates and pointwise confidence intervals for values of s between the 5% and 95% quantiles of the estimated marginal distribution of S. This is a sensible thing to do in general, given that the bias increases (sometimes dramatically) towards the edges. However, the obvious question is how to choose the cutoff points; this represents a potential direction for future research. One possible approach is to simulate data that has similar properties to the actual data (in terms of sample size, number of events, distribution of the marker, etc.), choose an acceptable level of bias, and select cutoff points such that in the simulated dataset, the estimated bias (averaged over a sufficiently large number of dataset replicates) does not exceed this level within the interval defined by the cutoff points.

A second approach leverages the fact that the observed distribution of the marker S often has a point mass at zero, corresponding to one-half the LLOD, as discussed in section 2.1.1. This provides us with the opportunity to construct estimators of the parameters $r_{M,0}(0)$ and $\text{CVE}_0(0)$ that converge at root-n rate. Benkeser et al. (2021) note that the parameter $\text{CVE}_0(0)$ can be interpreted as a natural direct effect (in the language of mediation analysis); that is, this is the portion of the vaccine effect that occurs through mechanisms other than the marker of interest. Given this, we could choose to use the natural direct effect estimator given in Benkeser et al. (2021) as the edge estimator. However, while their method accounts for right-censoring, it treats the endpoint as binary (rather than as a time-to-event variable) and thus ignores some of the available information by constructing an estimator from a coarsened data structure. Instead, we adapt that counterfactual survival curve method of Westling et al. (2021) to our setting. The only required modification is that the method needs to be able to handle a two-phase sampling structure. The form of this cross-fitted one-step estimator is given by

$$r^{\circ}_{M,n} \coloneqq 1 + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_{nk}}\left(\frac{[Z_i I(S_i=0) + g_{s,nk}(O_i)\{\pi_0(O_i) - Z_i\}]\omega_{nk}(X_i, Y_i, \Delta_i, 0)}{\pi_0(O_i)\tilde{g}_{s,nk}(X_i)} - Q_{nk}(t_0 \,|\, X_i, 0)\right), \tag{3.10}$$

where $\tilde{g}_{s,n}(x) \coloneqq g_n(x,0)(1-p_n)$, $\mathcal{V}_{nk}$, Q_{nk}, ω_{nk}, and so on are defined as in section 3.1.6, and

$$g_{s,n}(x,y,\delta) \coloneqq \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i f_n^{Y,\Delta|X,S}(y,\delta \,|\, x, S_i) g_n(x,S_i)}{\pi_0(O_i)}\right\}^{-1}\left\{f_n^{Y,\Delta|X,S}(y,\delta \,|\, x, 0) g_n(x,0)(1-p_n)\right\}.$$

A non-cross-fitted estimator can be formed by substituting Q_n for Q_{nk}, ω_n for ω_{nk}, and so on in (3.10), in which case the double sum reduces to a sum over $i \in \{1, ..., n\}$. Under a set of regularity conditions (see Westling et al., 2021), the influence function $\varphi_{r,0}$ of $r^{\circ}_{M,n}$ is given by

$$o \mapsto 1 - Q_0(t_0 \,|\, x, 0) + \frac{[zI(s=0) + g_{s,0}(o)\{\pi_0(o) - z\}]\omega_0(x,y,\delta,0)}{\pi_0(o)\tilde{g}_{s,0}(x)} - \tilde{r}_{M,0}(0).$$

There are now two possible estimators of $r_{M,0}(0)$, the generalized Grenander-type estimator $r_{M,n}(0)$ given by (3.5) and the edge estimator $r^{\circ}_{M,n}$. If the goal is estimating $r_{M,0}(0)$, using $r_{M,n}(0)$ will generally not yield reliable estimates because of the bias issue discussed above, and so we instead recommend using $r^{\circ}_{M,n}$. However, investigators will usually be interested in estimating the entire curve, and it is possible to use $r^{\circ}_{M,n}$ to adjust values of the curve close to (but not equal to) zero. Given the monotonicity assumption, since $r_{M,0}(s) < r_{M,0}(0)$ for $s > 0$, we propose the edge-corrected estimator $r^{*}_{M,n}$ given by

$$r^{*}_{M,n}(s) \coloneqq I(s=0) r^{\circ}_{M,n} + I(s>0)\min\left\{r^{\circ}_{M,n}, r_{M,n}(s)\right\}.$$

We choose to define $r^{*}_{M,n}$ using the indicator function $I(s=0)$ so that it is unbiased at the edge; if it was instead defined as the minimum of the two component estimators, there would be a small finite sample bias. Since $r_{M,0}(s)$ is strictly decreasing in s and $r_{M,n}(s)$ is consistent for $s > 0$, it is the case that $\left|r^{*}_{M,n}(s) - r_{M,n}(s)\right| = o_P(1)$ for each fixed $s \in (0,1]$. That is, $r^{*}_{M,n}(s)$ asymptotically behaves like $r_{M,n}(s)$ everywhere except at $s = 0$, where it behaves like $r^{\circ}_{M,n}$. Given this, a valid

approach to constructing pointwise confidence intervals is to use the confidence interval for $r^{\circ}_{M,n}$ at $s = 0$ and use the confidence intervals for $r_{M,n}$ for $s > 0$. However, even though this approach is asymptotically valid, it can lead to undesirable behavior in finite samples, such as the point estimate lying outside the confidence interval. One could also consider constructing each pointwise confidence interval as the union of the two intervals of the component estimators, but this will often be overly conservative. Instead, we propose a hybrid of these two approaches in which the confidence interval for $r^{\circ}_{M,n}$ is used at $s = 0$ and the confidence intervals for $r_{M,n}$ are used for $s > 0$, *unless* s lies in the subinterval of $(0, 1]$ for which the edge correction is in effect (i.e., the subinterval for which $r^{\circ}_{M,n} \leq r_{M,n}(s)$). If s lies in this subinterval, the pointwise minimum of the two lower confidence limits is used (which is more conservative than using the pointwise lower limits for $r_{M,n}$) and the pointwise minimum of the two upper limits (which is less conservative than using the pointwise maximum or the limits for $r_{M,n}$, but is justified by the fact that $r_{M,0}$ is decreasing). More precisely, for a fixed point s, denote the confidence intervals for $r_{M,n}(s)$ and $r^{\circ}_{M,n}$ by $(l(s), u(s))$ and (l°, u°), respectively. Then the confidence interval $(l^*(s), u^*(s))$ for $r^*_{M,n}(s)$ is defined as

$$l^*(s) \coloneqq I(s = 0)l^{\circ} + I(s > 0)\left[I\left\{r^{\circ}_{M,n} \leq r_{M,n}(s)\right\}\min\{l^{\circ}, l(s)\} + I\left\{r^{\circ}_{M,n} > r_{M,n}(s)\right\}l(s)\right],$$
$$u^*(s) \coloneqq I(s = 0)u^{\circ} + I(s > 0)\left[I\left\{r^{\circ}_{M,n} \leq r_{M,n}(s)\right\}\min\{u^{\circ}, u(s)\} + I\left\{r^{\circ}_{M,n} > r_{M,n}(s)\right\}u(s)\right].$$

Note that asymptotically, the subinterval for which $r^{\circ}_{M,n} \leq r_{M,n}(s)$ will shrink to length zero and this approach will be equivalent to using the confidence interval for $r^{\circ}_{M,n}$ at $s = 0$ and using the confidence intervals for $r_{M,n}$ for $s > 0$.

Another finite sample issue with the estimator $r^*_{M,n}(s)$ is that if the proportion of observations for which $S = 0$ is small, the variance of $r^{\circ}_{M,n}$ will be high, which will in turn destabilize $r^*_{M,n}(s)$ for $s > 0$. It may be possible to construct an adaptive estimator that uses, for example, a weighted combination of the estimators $r^*_{M,n}(s)$ and $r_{M,n}(s)$ with data-dependent weights, but that is outside the scope of the current work. In lieu of this, a decision rule can be created for when to use one estimator versus the other. For example, one may decide to use $r^*_{M,n}(s)$ if at least 10% of the

observations have $S = 0$. Like the truncation of the display of the function $s \mapsto r_{M,n}(s)$ discussed above, one could determine an appropriate cutoff value for a given decision rule via simulation.

Finally, note that similar corrections can be done if there is a point mass at the ULOQ. The arguments and expressions are analogous.

3.1.8 Estimating controlled vaccine efficacy

So far, we have focused estimation and inference on the controlled risk parameter $r_{M,0}(s)$. However, as discussed in section 2.1.2, the controlled vaccine efficacy parameter $\text{CVE}_0(s)$ defined in (2.2) is often of primary interest. Estimation of $\text{CVE}_0(s)$ necessitates estimation of the counterfactual quantity $P\{T(0) \leq t_0\}$, which can be identified with $r_{M,0}(0,0)$ if the assumptions discussed in section 2.1.2 hold. This quantity is simply the risk (i.e., the cumulative incidence) under assignment to the placebo group; estimation of this quantity is a well-studied problem and many approaches can be used, including Kaplan-Meier curves (Kaplan and Meier, 1958), marginalized Cox models (Self and Prentice, 1988), and machine learning techniques (Wang et al., 2019). Assume that an approach is chosen that yields an asymptotically linear estimator $r_{M,n}(0,0)$ of $r_{M,0}(0,0)$. The natural estimator to consider is then given by

$$\text{CVE}_n(s) \coloneqq 1 - \frac{r_{M,n}(s)}{r_{M,n}(0,0)}\,. \tag{3.11}$$

This estimator is sensible because monotonicity of $r_{M,0}(s)$ implies monotonicity of $\text{CVE}_0(s)$. Since $r_{M,n}(0,0)$ is asymptotically linear, it converges at a rate of $n^{1/2}$, whereas $r_{M,n}(s)$ converges at a rate of $n^{1/3}$. Therefore, the limiting distribution of $\text{CVE}_n(s)$ is unaffected by the variance of $r_{M,n}(0,0)$ and so the variance of $\text{CVE}_n(s)$ can be estimated as the variance of $r_{M,n}(s)$ divided by $\{r_{M,n}(0,0)\}^2$. However, estimation of this variance in finite samples may be improved by a correction that incorporates the excess variance due to estimation of $r_{M,0}(0,0)$. Suppose that the estimator $r_{M,n}(0,0)$ has a Normal limiting distribution with variance $\sigma^2_{P,0}$ and only uses information from the placebo group such that it is independent of $r_{M,n}(s)$ for all $s \in [0,1]$. Then using the delta method, it holds that

$$\frac{r_{M,n}(s)}{r_{M,n}(0,0)} - \frac{r_{M,0}(s)}{r_{M,0}(0,0)} = \frac{r_{M,n}(s) - r_{M,0}(s)}{r_{M,n}(0,0)} - \frac{r_{M,0}(s)}{\{r_{M,0}(0,0)\}^2}\left\{r_{M,n}(0,0) - r_{M,0}(0,0)\right\} + R(s)\,,$$

where the remainder term $R(s)$ can be expressed in second order as

$$R(s) \coloneqq \frac{r_{M,n}(s)}{r_{M,n}(0,0)} - \frac{r_{M,0}(s)}{r_{M,0}(0,0)} + \frac{r_{M,0}(s)r_{M,n}(0,0)}{\{r_{M,0}(0,0)\}^2} - \frac{r_{M,0}(s)}{r_{M,0}(0,0)} = O_P(n^{-5/6})$$

This allows us to write

$$\begin{aligned} n^{1/3}\left\{\frac{r_{M,n}(s)}{r_{M,n}(0,0)} - \frac{r_{M,0}(s)}{r_{M,0}(0,0)}\right\} &= \frac{n^{1/3}\{r_{M,n}(s) - r_{M,0}(s)\}}{r_{M,0}(0,0)} \\ &\quad - \frac{n^{-1/6}r_{M,0}(s)}{\{r_{M,0}(0,0)\}^2} n^{1/2}\left\{r_{M,n}(0,0) - r_{M,0}(0,0)\right\} + o_P(1)\,, \end{aligned}$$

And so the distribution of $\mathrm{CVE}_n(s)$ can be approximated as

$$\mathrm{CVE}_n(s) - \mathrm{CVE}_0(s) \overset{d}{\approx} \frac{r_{M,0}(s)}{n^{1/2}\{r_{M,0}(0,0)\}^2} N\big(0, \sigma^2_{P,0}\big) - \frac{\tau_0(s)}{n^{1/3} r_{M,0}(0,0)}\mathbb{Z}\,,$$

where as in (3.6), $\mathbb{Z}$ follows the Chernoff distribution, and where $N\big(0, \sigma^2_{P,0}\big) \perp\!\!\!\perp \mathbb{Z}$. This allows us to write

$$\mathrm{Var}\left\{\mathrm{CVE}_n(s)\right\} \approx \frac{\{r_{M,n}(s)\}^2}{n\{r_{M,n}(0,0)\}^4}\sigma^2_{P,n} + \left\{\frac{\tau_n(s)}{n^{1/3} r_{M,n}(0,0)}\right\}^2 \mathrm{Var}(\mathbb{Z})\,.$$

Next, if the edge correction is performed, there is more work to do, since the edge-corrected estimator $r^{\circ}_{M,n}$ of $r_{M,0}(0)$ converges at root-n rate. Let $\tilde{\varphi}_{r,0}$ denote the influence function of $r_{M,n}(0,0)$ and recall that $\varphi_{r,0}$ denotes the influence function of $r^{\circ}_{M,n}$. Define $\mathrm{CVE}^*_n(s)$ as the edge-corrected version of (3.11) with $r_{M,n}(s)$ replaced by $r^*_{M,n}(s)$. Then by the delta method, the influence function of the edge value $\mathrm{CVE}^*_n(0)$ is given by

$$o \mapsto \frac{r^*_{M,0}(0)\tilde{\varphi}_{r,0}(o)}{\left\{r_{M,0}(0,0)\right\}^2} - \frac{\varphi_{r,0}(o)}{r_{M,0}(0,0)}\,.$$

This influence function can be used to form confidence intervals in the same way as described in

section 3.1.7. Also, since $\text{CVE}_0(s) \in (-\infty, 1)$ for any s, it may be desirable to form confidence intervals on the $\log\{1 - \text{CVE}_0(s)\}$ scale to ensure that the resulting limits are also contained within $(-\infty, 1)$.

3.2 Methods: hypothesis testing

3.2.1 Derivation of test statistic

Interest lies in testing the null hypothesis that the function $s \mapsto \text{CVE}_0(s)$ is constant against the alternative hypothesis that $s \mapsto \text{CVE}_0(s)$ is nonconstant and nonincreasing. Since $\text{CVE}_0(s)$ is constant if and only if $r_{M,0}(s)$ is constant, the null hypothesis can also be stated in terms of the CR curve. Given that the alternative hypothesis assumes that the direction of monotonicity is known, the test will be directional. However, it is straightforward to extend the test to the case in which the direction of monotonicity is not known. Fix a univariate distribution P_0^* such that the support of P_0^* coincides with the support of S, and let $U \sim P_0^*$. The slope of the closest linear approximation to the function $r_{M,0}$ over $[0, 1]$ in a P_0^*-least-squares sense is given by

$$\frac{\text{Cov}_0^*\{U, r_{M,0}(U)\}}{\text{Var}_0^*(U)} .$$

Although it is not true under arbitrary alternatives, under monotonicity, it can be shown that this parameter equals zero if and only if $r_{M,0}$ is constant on $[0, 1]$. This may suggest consideration of the test statistic

$$\frac{1}{n} \sum_{i=1}^{n} \left(U_i - \bar{U}_n\right) r_{M,n}(U_i) \ ,$$

where $U_1, ..., U_n$ is a sample from P_0^* and $r_{M,n}$ is defined as in (3.5). However, this statistic will not converge at a $n^{1/2}$ rate since $r_{M,n}$ converges at a $n^{1/3}$ rate, and so $n^{1/2}$-rate inference is not possible with this statistic. Analogous test statistics that use other choices for estimating $r_{M,0}$ either lead to similar issues or require additional assumptions about $r_{M,0}$. However, note that the projection of $r_{M,0}$ onto the space of linear functions is analogous to the projection of the primitive

$\Theta_0 : u \mapsto \int_0^u r_{M,0}(t)dt$ onto the space of quadratic functions (without a constant term), and that the quadratic term coefficient in the latter corresponds to half the linear term coefficient in the former. The population quadratic coefficient value from this projection is given by

$$\frac{E_0^*\left\{\left(\lambda_{2,*}U^2 - \lambda_{3,*}U\right)\Theta_0(U)\right\}}{\lambda_{2,*}\lambda_{4,*} - \lambda_{3,*}^2}, \tag{3.12}$$

where $\lambda_{k,*} \coloneqq E_0^*[U^k]$. This suggests using the numerator of (3.12) as the parameter on which a test statistic can be based. However, for reasons explained in section 3.2.2, it is beneficial to use the corresponding parameter from the space of quadratic functions *with* a constant term. This coefficient is given by

$$\frac{E_0^*\left[\{(\lambda_{1,*}\lambda_{2,*} - \lambda_{3,*})(U - \lambda_{1,*}) + (\lambda_{2,*} - \lambda_{1,*}^2)(U^2 - \lambda_{2,*})\}\Theta_0(U)\right]}{(\lambda_{2,*}^2 - \lambda_{4,*})(\lambda_{1,*}^2 - \lambda_{2,*}) - (\lambda_{3,*} - \lambda_{1,*}\lambda_{2,*})^2}. \tag{3.13}$$

Thus, we base our test on the numerator of (3.13), the use of which is justified in Theorem 4.

Theorem 4. *Suppose that the distribution P_0^* is continuous with positive support on $[0,1]$ and is chosen such that the denominator of (3.13) is nonzero. Also suppose that the function $s \mapsto r_{M,0}(s)$ is nonincreasing. Then the quantity*

$$\beta_0 \coloneqq E_0^*\left[\{(\lambda_{1,*}\lambda_{2,*} - \lambda_{3,*})(U - \lambda_{1,*}) + (\lambda_{2,*} - \lambda_{1,*}^2)(U^2 - \lambda_{2,*})\}\Theta_0(U)\right] \tag{3.14}$$

equals zero if and only if $s \mapsto r_{M,0}(s)$ is constant on $(0,1)$.

It may seem peculiar that the open interval $(0,1)$ is specified in the statement of the theorem rather than the closed interval $[0,1]$, but this is in fact an important limitation. Specifically, the null hypothesis cannot be rejected by a test based on β_0 if the true controlled risk function is $r_{M,0}(s) = k_1 + k_2 I(s > 0)$ for some pair of constants (k_1, k_2), representing a decrease in risk immediately after the LLOD and a flat function thereafter. A solution to this shortcoming is discussed in section 3.2.4.

Since Theorem 4 holds for any distribution P_0^* with support on $[0,1]$ for which the denominator of (3.13) is positive, a simple way to construct a test statistic based on β_0 is to take P_0^* to be the

standard uniform distribution and estimate β_0 (up to a constant) with the estimator

$$\begin{aligned}\beta_n &\coloneqq \int_0^1 \left(u^2 - u + \frac{1}{6}\right) \Theta_n(u) du \\ &\propto E_0^* \left[\{(\lambda_{1,*}\lambda_{2,*} - \lambda_{3,*})(U - \lambda_{1,*}) + (\lambda_{2,*} - \lambda_{1,*}^2)(U^2 - \lambda_{2,*})\} \Theta_n(U) \right] ,\end{aligned}$$

where the integral is numerically approximated and Θ_n is an estimator of Θ_0 that will be described in the next section.

3.2.2 Estimating Θ_0

The primitive function Θ_0 is similar in construction to the function Γ_0 used in section 3.1, and as such, a similar strategy to estimation can be taken. For a fixed value u, the parameter $\Theta_0(u)$ can be written as

$$\Theta_0(u) = \int_0^u r_{M,0}(s) ds = \int \int I(s \leq u) \left\{1 - Q_0(t_0 \,|\, x, s)\right\} dF_0^X(x) ds \,.$$

Note that this is equivalent to (3.2) if Φ_0 is taken to be the identity function. Again applying the Rose and van der Laan (2011) result used in section 3.1.2, it can be shown that the efficient influence function of Θ_0 at P_0 relative to a model that includes the MAR assumption for S and the assumptions related to the survival data structure outlined in section 2.1.1 is given by

$$\varphi_{\Theta,0} : (o, u) \mapsto \frac{z I(s \leq u) \omega_0(o)}{\pi_0(o) f_0^{S|X}(s \,|\, x)} + \left\{1 - \frac{z}{\pi_0(o)}\right\} \tilde{q}_0(o, u) + \eta_0^*(x, u) - \Theta_0(u) \,. \tag{3.15}$$

which involves the nuisances

$$\begin{aligned}\eta_0^*(x, u) &\coloneqq \int I(s \leq u) \left\{1 - Q_0(t_0 \,|\, x, s)\right\} ds \,, \\ \tilde{q}_0(o, u) &\coloneqq E_0 \left\{ \frac{I(S \leq u) \omega_0(O)}{f_0^{S|X}(S \,|\, X)} \middle| Z = 1, X = x, Y = y, \Delta = \delta \right\} .\end{aligned}$$

This influence function can be used to construct a one-step estimator Θ_n of Θ_0, given by

$$\Theta_n(u) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i I(S_i \leq u)\omega_n(O_i)}{\pi_0(O_i) f_n^{S|X}(S_i \,|\, X_i)} + \left\{1 - \frac{Z_i}{\pi_0(O_i)}\right\}\tilde{q}_n(O_i, u) + \eta_n^*(X_i, u)\right].$$

This specification involves the same nuisance estimators ω_n and $f_n^{S|X}$ described in section 3.1.4, and as well as

$$\eta_n^*(x,u) \coloneqq u - \int I(s \leq u) Q_n(t_0 \,|\, x, s) ds\,,$$
$$\tilde{q}_n(o,u) \coloneqq \left\{\int f_n^{Y,\Delta|X,S}(y,\delta \,|\, x,s) f_n^{S|X}(s \,|\, x) ds\right\}^{-1} \int I(s \leq u)\omega_n(x,y,\delta,s) f_n^{Y,\Delta|X,S}(y,\delta \,|\, x,s) ds\,,$$

where the integrals are numerically approximated. Above, the estimator $\tilde{q}_n$ is constructed in a manner analogous to that of q_n and involves the same conditional density estimator $f_n^{Y,\Delta|X,S}$ defined in (3.8). We can also plug these nuisances into (3.15) to construct an estimator $\varphi_{\Theta,n}$ of the influence function $\varphi_{\Theta,0}$; this additionally requires the plug-in estimator

$$\tilde{\Theta}_n(u) \coloneqq \int\int I(s \leq u)\left\{1 - Q_n(t_0 \,|\, x, s)\right\} dF_n^X(x) ds\,,$$

of $\Theta_0(u)$, where again the integral is numerically approximated. As mentioned in section 3.2.1, we chose to base the test statistic β_n on the parameter corresponding to a quadratic regression with an estimated intercept, even though the true intercept is known to be zero, since $\Theta_0(0) = 0$ by definition. The reason for choosing to estimate this value is that, by construction, $\Theta_n(0)$ will generally not equal zero in finite samples, even though it will tend to zero in probability. Therefore, a quadratic regression with an estimated intercept will yield a better fit to the estimated function $u \mapsto \Theta_n(u)$, which will in turn improve the performance of the hypothesis test.

3.2.3 Asymptotic distribution of β_n

First, a set of conditions is provided under which $\Theta_n(u)$ is an asymptotically linear estimator of $\Theta_0(u)$ for fixed u.

Theorem 5. *Suppose that regularity conditions (A4), (A2.i), (A1.iii), (A5.ii), and (B1) – (B4) given in section 3.6 hold. Then for fixed $u \in [0,1]$, it holds that*

$$\Theta_n(u) - \Theta_0(u) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\Theta,0}(O_i, u) + r_{\Theta,n}(u)\,,$$

where $r_{\Theta,n}(u) = o_P(n^{-1/2})$.

If the conditions of Theorem 5 hold, then

$$\begin{aligned}\beta_n - \beta_0 &= \int_0^1 \left(u^2 - u + \frac{1}{6}\right)\left\{\Theta_n(u) - \Theta_0(u)\right\} du \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^1 \left(u^2 - u + \frac{1}{6}\right)\varphi_{\Theta,0}(O_i, u)du\right\} + \int_0^1 \left(u^2 - u + \frac{1}{6}\right) r_{\Theta,n}(u)du\,,\end{aligned}$$

where $r_{\Theta,n}$ is the remainder term corresponding to Θ_n. Therefore, if $\int_0^1 (u^2 - u + 1/6) r_{\Theta,n}(u)du = o_P(n^{-1/2})$, then β_n is asymptotically linear with influence function

$$\varphi_{\beta,0} : o \mapsto \int_0^1 \left(u^2 - u + \frac{1}{6}\right)\varphi_{\Theta,0}(o, u)du\,.$$

Given this result, the test statistic

$$T_{\beta,n} \coloneqq n\left[\sum_{i=1}^{n}\{\varphi_{\beta,n}(O_i)\}^2\right]^{-1/2}\beta_n\,,$$

is used as the basis for our hypothesis test, where $\varphi_{\beta,n}$ equals $\varphi_{\beta,0}$ with the estimated influence function $\varphi_{\Theta,n}$ plugged in for $\varphi_{\Theta,0}$. Under the null hypothesis, this test statistic converges in distribution to a standard Normal random variable, and since a known direction of monotonicity is assumed, a directional test can be constructed.

3.2.4 Modification of test statistic to allow for detection of an edge jump

In section 3.2.1, we discussed how a hypothesis test based on the parameter β_0 would not be able to detect a decreasing controlled risk function if the form of that function is $r_{M,0}(s) = k_1 + k_2 I(s > 0)$ for some pair of constants (k_1, k_2); we refer to the difference $r_{M,0}(\epsilon) - r_{M,0}(0)$ (where 0 corresponds to the value LLOD/2 and ϵ corresponds to the value LLOD) as the *edge jump*. Furthermore, even if

the function $s \mapsto r_{M,0}(s)$ has an edge jump *and* is decreasing over some subinterval of $(0, 1]$, a test based on β_0 would be expected to have low power if the magnitude of the change over $(0, 1]$ is low relative to the magnitude of the edge jump. Given this, we propose a correction to the test for cases in which there is mass at the LLOD. Our strategy is to derive a second test statistic specifically designed to detect the edge jump and then combine this with the statistic $T_{\beta,n}$ to form a modified test statistic. The parameter on which our edge jump test statistic is based is $\beta_{e,0} \coloneqq \Theta_0(1) - r_{M,0}(0)$. It is the case that $\beta_{e,0} = 0$ if and only if the null hypothesis is true, since $r_{M,0}$ represents the value at the edge and $\Theta_0(1)$ represents the average value over $(0, 1]$, which is necessarily less than $r_{M,0}(0)$ under the monotonicity assumption. This is true regardless of the functional form of $s \mapsto r_{M,0}(s)$; given this, one might wonder why we do not simply base a hypothesis test on $\beta_{e,0}$. The two reasons why this would not be sensible are (i) the parameter $r_{M,0}(0)$ is pathwise differentiable only if there is a point mass at the LLOD (as discussed in section 3.1.7), and (ii) all of the information about the change in the function value over $(0, 1]$ is effectively ignored by the construction of $\beta_{e,0}$. Thus, this test statistic is only well-suited to detect an edge jump.

Given that we have already described estimators $\Theta_n(1)$ and $r^{\circ}_{M,n}$ of $\Theta_0(1)$ and $r_{M,0}(0)$, respectively, there is no additional work required to construct an estimator of $\beta_{e,0}$, and the estimator $\beta_{e,n} \coloneqq \Theta_n(1) - r^{\circ}_{M,n}$ can be used. This estimator is a linear function of the two component estimators, and so its influence function is simply the difference of the two component influence functions, given by $\varphi_{\beta,e,0} : o \mapsto \varphi_{\Theta,0}(o, 1) - \varphi_{r,0}(o)$. Although tests based on the parameters β_0 and $\beta_{e,0}$ could be conducted separately, it is desirable to construct a single test statistic to avoid the loss of power that would result from a multiple comparison correction. Given that the null hypothesis is true if and only if $\beta_0 = \beta_{e,0} = 0$, and that both β_0 and $\beta_{e,0}$ are of the same sign under any (monotone) alternative hypothesis, a new test statistic could be formed based on the sum $\beta_n + \beta_{e,n}$. The drawback of this approach is that if one statistic has much higher variance than the other, the signal could be washed out by the noise, so to speak, if the statistic with lower variance is the one with a stronger signal. Therefore, we choose to divide each statistic by its estimated standard deviation before taking the sum. The modified test statistic is given by

$$T^*_{\beta,n} \coloneqq \sqrt{\frac{n}{2+2\rho_n}} \left(\frac{\beta_n}{\sigma_{\beta,n}} + \frac{\beta_{e,n}}{\sigma_{\beta,e,n}} \right), \tag{3.16}$$

where

$$\begin{aligned}
\sigma_{\beta,n} &\coloneqq \left(\frac{1}{n} \sum_{i=1}^{n} \{\varphi_{\beta,n}(O_i)\}^2 \right)^{1/2}, \\
\sigma_{\beta,e,n} &\coloneqq \left(\frac{1}{n} \sum_{i=1}^{n} \{\varphi_{\beta,e,n}(O_i)\}^2 \right)^{1/2}, \\
\rho_n &\coloneqq \frac{1}{n\sigma_{\beta,n}\sigma_{\beta,e,n}} \sum_{i=1}^{n} \varphi_{\beta,n}(O_i)\varphi_{\beta,e,n}(O_i).
\end{aligned}$$

Under the null hypothesis, the test statistic $T^*_{\beta,n}$ converges in distribution to a standard Normal random variable, and as with $T_{\beta,n}$, this statistic can be used to construct a directional test. The only remaining issue is that the estimator $r^{\circ}_{M,n}$ is impractical when there is not a point mass at the LLOD, and thus $T^*_{\beta,n}$ may give misleading results. One approach is to take an analogous strategy to the one taken in section 3.1.7, using $T^*_{\beta,n}$ for the hypothesis test if a certain percentage of the observations have $S = 0$ and using β_n otherwise.

3.3 Simulation study

3.3.1 Estimation

To assess the operating characteristics of our proposed estimator, we conducted a simulation study. We generated data from six mechanisms corresponding to combinations of three conditional distributions $f_0^{S|X}$ of the marker and two conditional survival functions Q_0. Two baseline covariates (X_1, X_2) were drawn, with $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim \text{Uniform}\{0.0, 0.1, ..., 1.0\}$. The three conditional distributions were the standard uniform distribution, a Normal distribution with mean 0.5 and variance 0.2 (truncated to lie within $[0, 1]$), and a Normal distribution with mean $0.3 + 0.4X_2$ and variance 0.3 (also truncated to lie within $[0, 1]$). The two conditional survival distributions both followed proportional hazards models, with one following a Cox model and the other having a more complex nonlinear form. For both, the baseline hazard function of the survival variable

followed a Weibull distribution with scale parameter $\lambda_1 = 2 \times 10^{-4}$ and shape parameter $v_1 = 1.5$. The baseline conditional censoring function also followed a Weibull distribution, with parameters $\lambda_2 = 5 \times 10^{-5}$ and $v_2 = 1.5$. This led to the following forms for the two conditional survival functions (denoted $Q_0^{(1)}$ and $Q_0^{(2)}$) and the conditional censoring function Q_0^C:

$$
\begin{aligned}
Q_0^{(1)}(t \mid x, s) &\coloneqq \exp\left\{-\lambda_1 t^{v_1} \exp(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 s)\right\} , \\
Q_0^{(2)}(t \mid x, s) &\coloneqq \exp\left[-\lambda_1 t^{v_1} \exp\{\alpha_3 \text{expit}(20s - 10) + \alpha_2 x_1 x_2\}\right] , \\
Q_0^{C}(t \mid x, s) &\coloneqq \exp\left\{-\lambda_2 t^{v_2} \exp(\alpha_1 x_1 + \alpha_2 x_2)\right\} ,
\end{aligned}
$$

where $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, -2)$. Using these functions, we generated survival and censoring times using the method of Bender et al. (2005). We used the function

$$
\pi_0(x, y, \delta) \coloneqq \delta I(y \leq t_0) + \{1 - \delta I(y \leq t_0)\}\text{expit}(x_1 + x_2 - 1)
$$

to generate the two-phase sampling indicator Z. With this function, all individuals who experienced the event prior to the time of interest t_0 and a fraction of the remaining individuals were selected into the phase-two sample. This formulation allows for the use of estimated weights, since the set of possible covariate combinations is discrete. The time of interest used was $t_0 = 200$ and the size of each phase-1 sample was 1,000 individuals.

We used IPS-weighted Cox proportional hazards models to construct an estimator Q_n of the conditional survival function Q_0 and an estimator Q_n^C of the conditional censoring function Q_0^C. Q_n is consistent when $Q_0 = Q_0^{(1)}$ but inconsistent when $Q_0 = Q_0^{(2)}$, and Q_n^C is always consistent. We used a correct parametric model to estimate the conditional density $f_0^{S|X}$. All other nuisance estimators are constructed according to the specifications given in section 3.1.4.

In all simulations, we compared the performance of our proposed estimator against the estimator derived from the marginalized Cox model described by Gilbert et al. (2022a) in terms of bias, variance, and 95% confidence interval coverage for the estimation of $r_{M,0}(s)$ for $s \in [0, 1]$. Simulations were conducted using the R programming language and structured using the *SimEngine* simulation framework (Kenny and Wolock, 2021). For each level combination, we ran 1,000 simulation

replicates.

Figure 3.1, displays the bias of both estimators across all six data-generating mechanisms considered. The Cox estimator was unbiased across the entire unit interval when it is correctly specified, as expected; when it is incorrect it was badly biased. The nonparametric estimator $r_{M,n}(s)$ showed little to no bias towards the interior of the unit interval, but has a positive bias that increases as $s \to 0$ and a negative bias that increases as $s \to 1$. This pattern of bias is a characteristic feature of generalized Grenander-type estimators, and of many monotone-constrained function estimators in general.

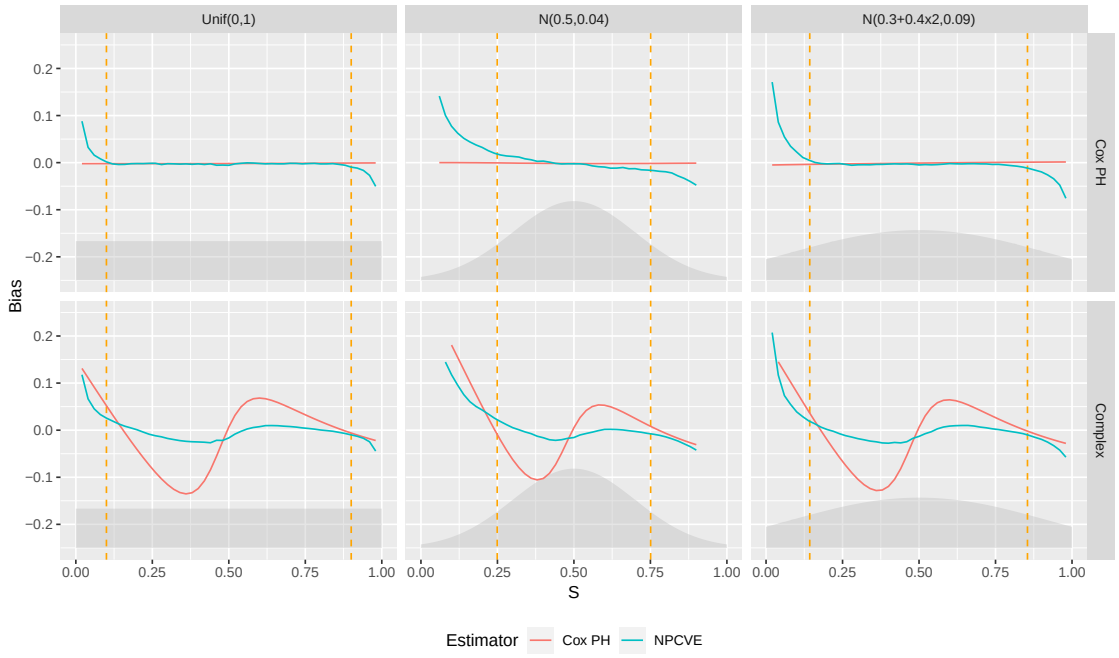


Figure 3.1: Bias of two estimators ("NPCVE" = nonparametric CVE, "Cox PH" = marginalized Cox proportional hazards), displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

Figure 3.2 displays the 95% confidence interval coverage of both estimators across the same set of data-generating mechanisms. The Cox estimator achieved nominal coverage when correctly specified, but when it was incorrect, coverage is unacceptably low, with 0% coverage for some values of s. Thus, even when the proportional hazards assumption holds, the Cox estimator may yield incorrect inference. The nonparametric estimator achieved nominal coverage over most of the unit

interval, although some areas of undercoverage and overcoverage are observed.

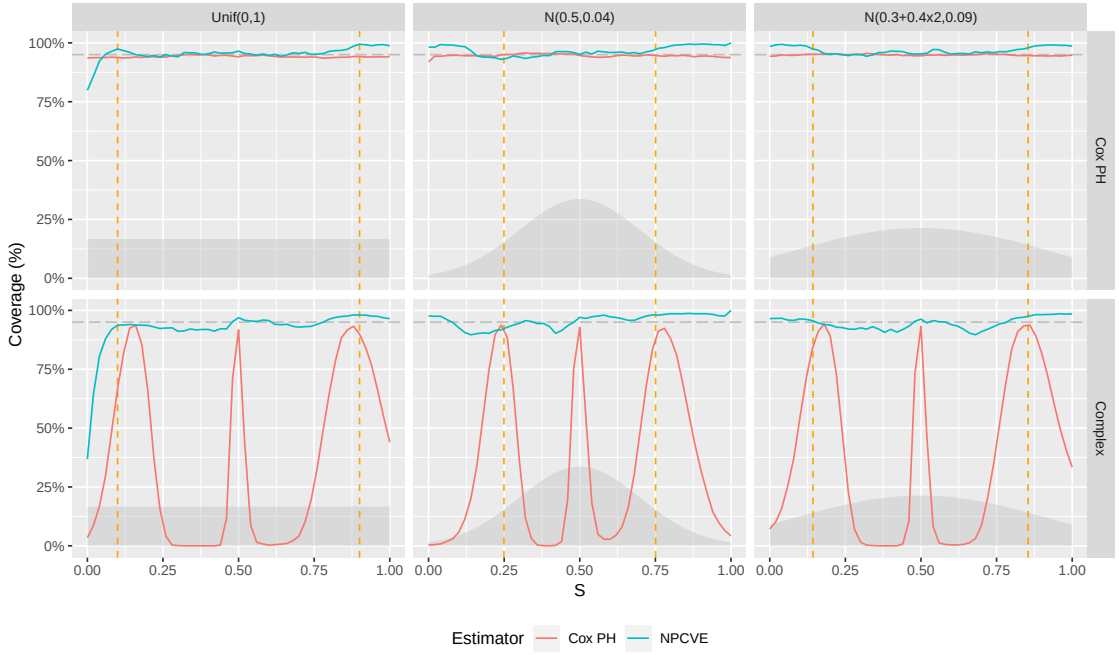


Figure 3.2: 95% confidence interval coverage of two estimators ("NPCVE" = nonparametric CVE, "Cox PH" = marginalized Cox proportional hazards), displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

Figure 3.3 displays the standard deviations of both estimators. In all scenarios, the standard deviation of the nonparametric estimator was greater than that of the Cox estimator, with the ratio of standard deviations increasing towards the endpoints.

Additional simulation results are presented in section 3.7.

3.3.2 Testing

In the second part of our simulation study, we evaluated the operating characteristics of the hypothesis test described in section 3.2. We used the same data-generating mechanism that was used previously in section 3.3.1, but with several differences. The parameter α_3 was varied between 0 and -0.5. The distribution of S used was a mixture of a standard uniform distribution and a point mass at zero, where the propensity score $\tilde{g}_{s,0}(x) := P_0(S = 0 \,|\, X = x)$ took one of the following three forms:

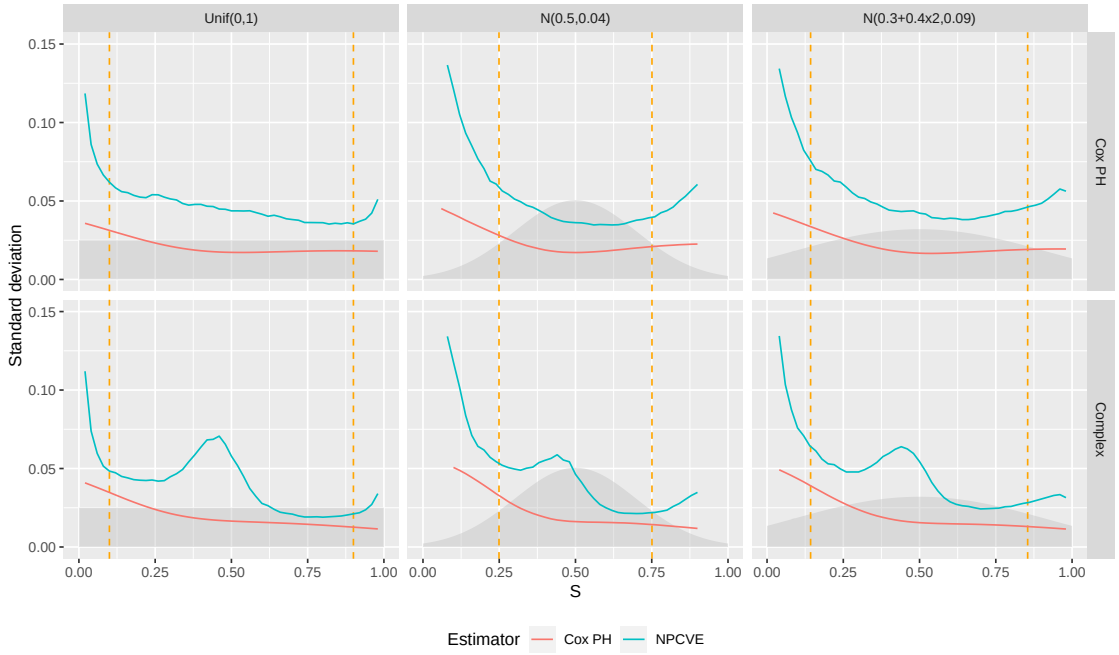


Figure 3.3: Standard deviation of two estimators ("NPCVE" = nonparametric CVE, "Cox PH" = marginalized Cox proportional hazards), displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

$$\begin{aligned}
\tilde{g}_{s,0}^{(1)}(x) &:= 0\,, \\
\tilde{g}_{s,0}^{(2)}(x) &:= \text{expit}(x_1 + x_2 - 2.5)\,, \\
\tilde{g}_{s,0}^{(3)}(x) &:= \text{expit}(x_1 + x_2 - 1.4)\,.
\end{aligned}$$

The function $\tilde{g}_{s,0}^{(2)}$ resulted in a point mass of roughly 10% at the edge and the function $\tilde{g}_{s,0}^{(3)}$ resulted in a point mass of roughly 40% at the edge. We tested two conditional survival functions, including the function $Q_0^{(1)}$ specified in section 3.3.1 and the function

$$Q_0^{(3)}(t \,|\, x, s) = \exp\left\{-\lambda_1 t^{v_1} \exp(\alpha_1 x_1 + \alpha_2 x_2 + 0.5\alpha_3 I(s = 0))\right\}\,,$$

which is constant as a function of s other than an edge jump, as described in section 3.2.4. The same set of nuisance estimators that were used in the previous simulations were used here. We

studied the two versions of the hypothesis test described in section 3.2, the standard version of the test and the version that is adapted to detect the edge jump, assessing performance through type I error rates and power. Results are shown in Figure 3.4.

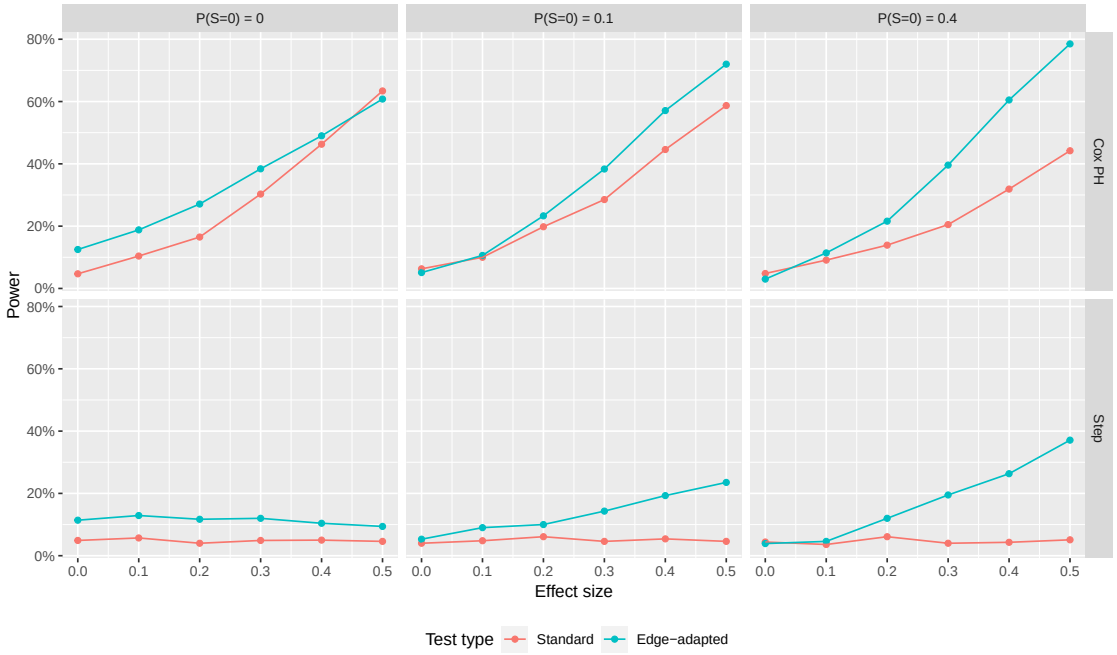


Figure 3.4: Power of two versions of hypothesis test, where the marker S follows a mixture of a $U(0,1)$ distribution and a point mass at zero, displayed for three propensity scores $\tilde{g}_{s,0}$ ("P(S=0)=0" $= \tilde{g}^{(1)}_{s,0}$, "P(S=0)=0.1" $= \tilde{g}^{(2)}_{s,0}$, "P(S=0)=0.4" $= \tilde{g}^{(3)}_{s,0}$) and two conditional survival functions ("Cox PH" $= Q^{(1)}_0$, "Step" $= Q^{(3)}_0$).

When the true conditional survival function follows a Cox model, the power of the edge-adapted test increases relative to the power of the standard test as the percentage of observations for which $S = 0$ increases. We also observe that the edge-adapted test has higher power when there is no mass at the edge, but this is because it is invalid in this scenario, as evidenced by the inflated type I error rates. When the true conditional survival function is an edge jump, the standard version of the test has no power regardless of effect size, as expected. On the other hand, the edge-adapted test has power that increases as the effect size increases and as the percentage of observations for which $S = 0$ increases.

In section 3.7, we provide analogous results corresponding to the two other choices for $f_0^{S|X}$ that were used in section 3.3 (also with and without the addition of point masses at the edge). The

results are qualitatively similar to those in Figure 3.4.

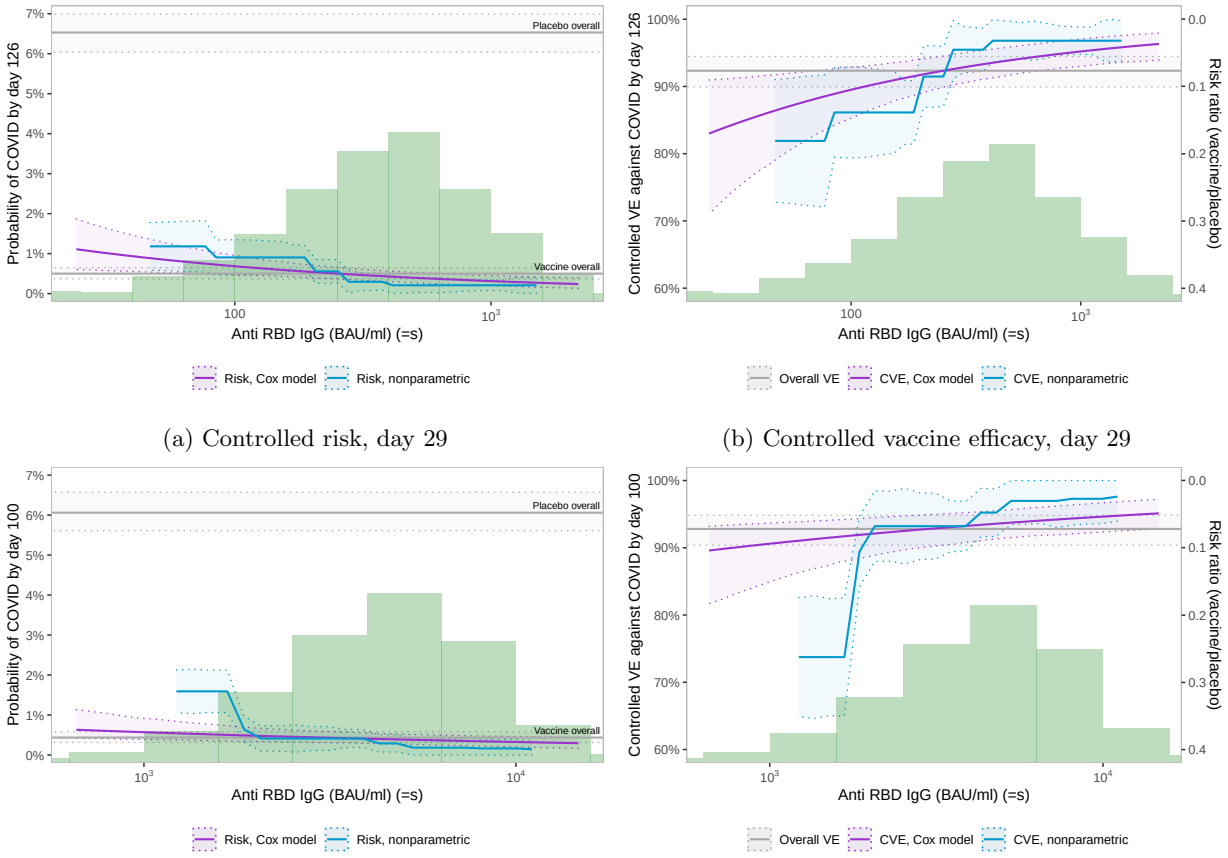
3.4 Immune correlates analysis of the Coronavirus Efficacy (COVE) trial of the mRNA-1273 COVID-19 vaccine

In this section, we illustrate the use of our estimation and hypothesis testing methods using the dataset from the Coronavirus Efficacy (COVE) placebo-controlled phase 3 trial (NCT04470427) of the mRNA-1273 COVID-19 vaccine, conducted across 99 sites in the United States starting in July 2020. In this trial, 30,420 adults aged 18 or older with high risk of SARS-CoV-2 infection or severe COVID-19 disease were randomized at a 1:1 ratio to the vaccine arm or the placebo arm. Overall vaccine efficacy was estimated to be 94.1% (95%CI: 89.3% to 96.8%) (Baden et al., 2020), with 185 instances of symptomatic COVID-19 disease in the placebo arm and 11 instances in the vaccine arm, leading to the United States Food and Drug Administration (FDA) giving an Emergency Use Authorization for the vaccine.

Gilbert et al. (2022b) assessed four immune markers as potential correlates of protection, including IgG bAbs to the spike protein, IgG bAbs to spike receptor binding domain (RBD), 50% inhibitory dilution (ID50) nAb titer, and 80% inhibitory dilution (ID80) nAb titer. Measurements were taken at both day 29 and day 57 post-vaccination. The authors used a variety of methods, including marginalized Cox proportional hazards models (Gilbert et al., 2022a), nonparametric targeted minimum loss-based threshold regression (van der Laan et al., 2022), and mediation analysis (Benkeser et al., 2021), concluding that all four markers were associated with vaccine efficacy.

Here, results for two representative markers, RBD and ID50, are presented, with results for the other two markers given in section 3.8. Figure 3.5 shows CR and CVE curves for the IgG bAbs to spike receptor binding domain (RBD) marker, estimated using both a Cox proportional hazards model and the nonparametric estimator, along with pointwise 95% confidence intervals. Subfigures (3.5a) and (3.5b) correspond to the day 29 marker measurement whereas subfigures (3.5c) and (3.5d) correspond to the day 57 measurement.

Figure 3.6 shows analogous results for the 50% inhibitory dilution (ID50) nAb titer marker. The



(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29

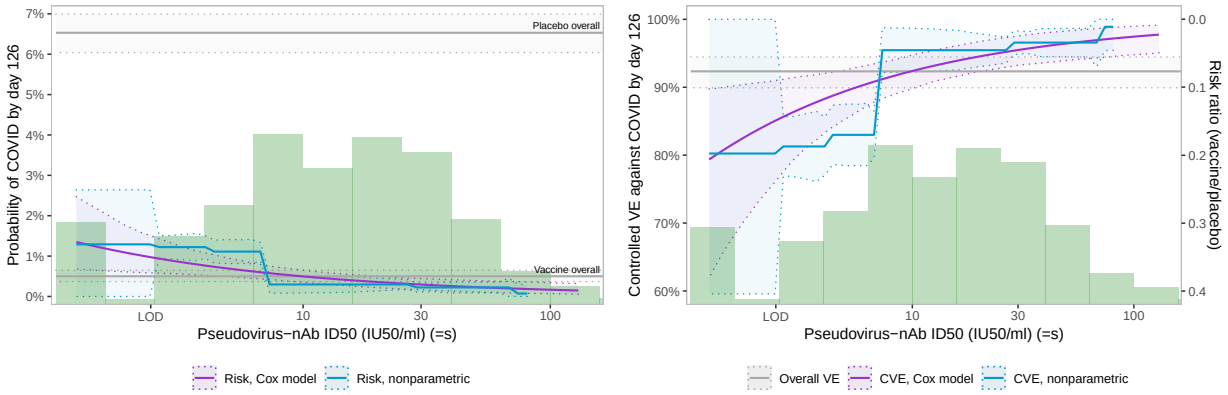
(c) Controlled risk, day 57

(d) Controlled vaccine efficacy, day 57

Figure 3.5: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the IgG bAbs to spike receptor binding domain (RBD) marker, measured at days 29 and 57. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

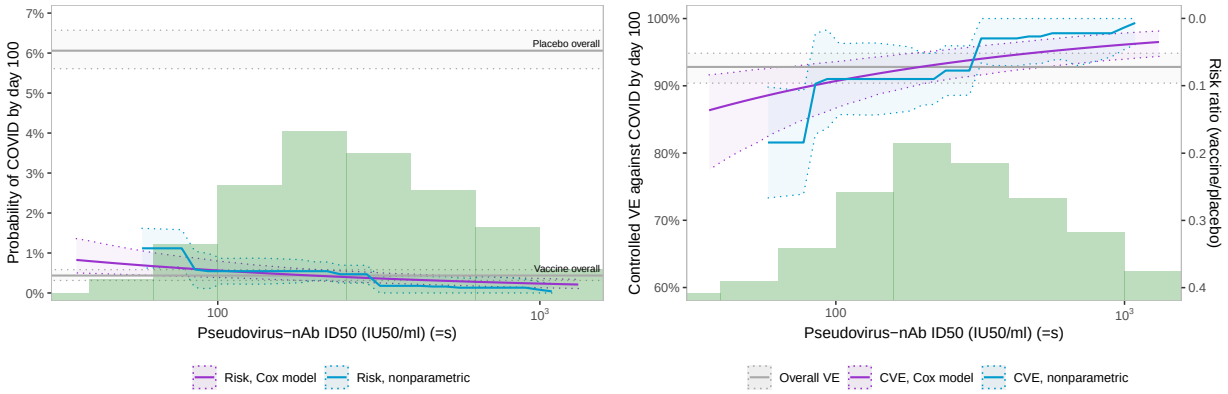
edge correction described in section 3.1.7 was used for the day 29 ID50 measurements (subfigures 3.6a and 3.6b).

Finally, in Table 3.1, results are displayed for the hypothesis test described in section 3.2 for the two markers considered above. P-values corresponding to the ID-50 marker at both day 29 and day 57 are significant at the $\alpha = 0.05$ level, although this does not account for multiple testing. The standard version of the test (rather than the edge-adapted version) was used, as the estimated marginal distribution of the marker has less than 10% mass at the edge for both markers at both



(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29



(c) Controlled risk, day 57

(d) Controlled vaccine efficacy, day 57

Figure 3.6: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the 50% inhibitory dilution (ID50) nAb titer marker, measured at days 29 and 57. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

time points; this decision rule was explained in section 3.2.4.

Marker	Day	P-value
IgG bAbs to spike receptor binding domain (RBD)	29	0.1384
IgG bAbs to spike receptor binding domain (RBD)	57	0.9342
50% inhibitory dilution (ID50) nAb titer	29	0.0007
50% inhibitory dilution (ID50) nAb titer	57	0.0425

Table 3.1: Hypothesis testing results for RBD and ID50 markers, measured at days 29 and 57.

Interestingly, although all of the CVE curves depicted in Figures 3.5 and 3.6 appear to be

nonconstant, the results in Table 3.1 do not support rejecting the null hypothesis at the $\alpha = 0.05$ level for the IgG bAbs to spike receptor binding domain (RBD) marker. The apparent contradiction between the estimation and testing results for this marker (when measured at day 57) may be due to the edge bias issue of the estimation procedure leading to the change in the marker being overestimated, as the hypothesis test does not have this issue.

3.5 Discussion

In this chapter, we described and studied a method that allows for nonparametric inference on the CR and CVE curves under the assumption that these curves are monotone. Our method can be seen as a generalization of the work of Westling et al. (2020), which combines ideas from the literature on shape-constrained nonparametric inference and on causal inference. Westling et al. (2020) noted that a natural extension of their work would be to extend the methods to settings in which a coarsened version of the full data are observed, and in this work we performed this generalization when the outcome is right-censored and the biomarker is observed subject to two-phase sampling. We chose to focus on the types of coarsening commonly encountered in vaccine efficacy trials, and further research could focus on similar extensions to other types of coarsening, such as left-truncation or missingness in the covariates. Additionally, theory could be developed to allow for the hypothesis testing approach taken in section 3.2 to be generalized to settings in which interest lies in testing the constancy of any arbitrary function assumed to be monotone. As part of this work, it would be useful to see if choices for P_0^* other than the standard uniform distribution lead to increased power, and to study the behavior of such tests in the context of local alternatives.

Several improvements may be possible to the estimation and inference procedures studied in this work. First, as discussed in section 3.1.3, the validity of the pointwise confidence intervals around $r_{M,n}$ depends on the assumption that the true function $r_{M,0}$ is strictly decreasing (rather than just nonincreasing). This assumption may not be the case in practice, and if the estimated curve has flat or nearly-flat regions, the estimated derivative $\dot{r}_{M,n}$ may be very small, resulting in unrealistically narrow confidence intervals in finite samples. Theoretical work that allows for pointwise confidence intervals to be constructed that are valid under the assumption that $r_{M,0}$ is nonincreasing would

be valuable. Second, it is somewhat odd that our method (as well as the methods of Westling and Carone (2020)) assume that the true function is differentiable, yet the estimator is always a nondifferentiable step function. Given this, it may be worth considering a procedure that allows for smoothing of the estimated function, ideally in a way that is asymptotically negligible. Such a smoothing procedure may result in better finite sample performance and an estimator that is more realistic and intuitively appealing to researchers. Third, it would be desirable to have a principled way to choose the cutoff points defining the range of the marker for which the estimated function is displayed, as described in section 3.1.7. Fourth, while the NPCVE estimator is constrained by construction to be monotone, the pointwise confidence limits are not, in the sense that the functions $s \mapsto l(s)$ and $s \mapsto u(s)$, where $l(s)$ and $u(s)$ are the confidence limits of $r_{M,n}$, may not be monotone. Intuitively, this would be a desirable property; further theoretical work may be necessary to achieve this. Fifth, it would be useful to extend the approach taken in this work to settings in which the biomarker can be nonzero in the placebo arm, for which the CVE surface (Gilbert et al., 2022a) is of interest.

It is also worth noting that in the construction of our estimator, we chose to construct the empirical distribution function F_n^X using data from the vaccine group only, rather than both the vaccine and placebo groups, even though the latter approach would lead to greater precision. We chose not to do so because in practice, researchers often use machine learning techniques to construct a so-called *baseline risk score*, a single continuous variable that predicts the risk of experiencing the outcome by time t_0, using baseline covariate data in the placebo arm Fong et al. (2022). This model is then used to construct the baseline risk score in the vaccine group, which is then considered as part of the covariate vector X and used for inference. Therefore, using the placebo covariate data to also construct F_n^X would constitute a form of double-dipping (in this case, creating dependent covariates but treating them as independent, thereby invalidating inference).

Because the CVE curve was defined recently, the only other method that has been studied for estimation is the Cox model approach of Gilbert et al. (2022a). As our simulations demonstrate, this method leads to substantial bias and undercoverage if one or more of the assumptions of the Cox model are incorrect in a given setting. The main advantages of our method relative

to the Cox model approach are that (i) it does not assume proportional hazards or a log-linear form of the conditional survival function predictor, and (ii) the conditional survival function Q_0, conditional censoring function Q_0^C, and density ratio g_0 can be estimated using flexible machine learning techniques. Several disadvantages of our method relative to the Cox model are (i) if the Cox model is correct, our estimator has higher variance, (ii) the Cox model does not have the edge bias issues of our method, and (iii) while the Cox model implicitly assumes monotonicity of $r_{M,0}$, it does not assume the direction of monotonicity and also allows for valid inference in the case when $r_{M,0}$ is constant. In the future, it may be worth considering nonparametric estimators that make different assumptions, such as assuming a degree of smoothness of $r_{M,0}$ rather than monotonicity.

This work focuses on a setting in which a single biomarker is of interest, but in practice, there are often many possible biomarkers available. As such, it would be desirable to have some way of quantitatively differentiating two or more biomarkers based on their usefulness as individual correlates of protection. Additionally, it will be useful to see if there is a principled way to combine multiple biomarkers into a single marker that serves as a better correlate of protection than any of the biomarkers do individually.

3.6 Proofs

3.6.1 Proof of Theorem 3

Theorem 3 is an adaptation of Theorem 4 of Westling and Carone (2020) to our setting. Therefore, our strategy will be to derive conditions under which the estimator $\Gamma_n(u)$ is asymptotically linear (for fixed u) and then to show that conditions A4-A5 and B1-B5 of Westling and Carone (2020) (which are referred to as WC.A4, WC.A5, etc.) are satisfied. This will require the introduction of regularity conditions (A4) – (A11) below. Throughout, we make use of empirical process notation, denoting $P_0 f \coloneqq E_0\{f(O)\}$ and $P_n f \coloneqq n^{-1}\sum_{i=1}^n f(O_i)$.

Asymptotic linearity of $\Gamma_n(u)$

Here, conditions are provided under which the estimator $\Gamma_n(u)$ is asymptotically linear for fixed u. Define the remainder term

$$R_{\Gamma,n}(u) \coloneqq \tilde{\Gamma}_n(u) - \Gamma_0(u) + E_0\left\{\varphi_{\Gamma,n}(O,u)\right\},$$

where $\varphi_{\Gamma,n}$ is an estimator of the influence function $\varphi_{\Gamma,0}$ of $\Gamma_n(u)$ given in (3.4). $\Gamma_n(u)$ is asymptotically linear if the following three conditions hold:

1. $R_{\Gamma,n}(u) = o_P(n^{-1/2})$,
2. $E_0\left\{\varphi_{\Gamma,n}(O,u) - \varphi_{\Gamma,0}(O,u)\right\}^2 = o_P(1)$
3. $o \mapsto \varphi_{\Gamma,n}(o,u) - \varphi_{\Gamma,0}(o,u)$ falls in some Donsker class with probability tending to one.

First, conditions are provided under which $R_{\Gamma,n}(u) = o_P(n^{-1/2})$. The remainder term can be expressed as

$$\begin{aligned}
R_{\Gamma,n}(u) &\coloneqq \tilde{\Gamma}_n(u) - \Gamma_0(u) + E_0\left\{\varphi_{\Gamma,n}(O,u)\right\} \\
&= \tilde{\Gamma}_n(u) - \Gamma_0(u) + E_0\left[\frac{ZI(S\le u)}{\pi_0(O)}\left\{\frac{\omega_n(O)}{g_n(X,S)} + \tilde{r}_{M,n}(S)\right\} + \left\{1 - \frac{Z}{\pi_0(O)}\right\} q_n(O,u)\right. \\
&\qquad \left. + \eta_n(X,u)\right] - 2\tilde{\Gamma}_n(u) \\
&= E_0\left\{\frac{I(S\le u)\omega_n(O)}{g_n(X,S)}\right\} + E_0\left\{I(S\le u)\tilde{r}_{M,n}(S)\right\} + E_0\{\eta_n(X,u)\} - \tilde{\Gamma}_n(u) - \Gamma_0(u) \\
&= E_0\left\{\frac{I(S\le u)\omega_n(O)}{g_n(X,S)}\right\} + \int\int I(s\le u)\left\{1 - Q_n(t_0 \mid x,s)\right\} dF_0^S(s) dF_n^X(x) \\
&\qquad + \int\int I(s\le u)\left\{1 - Q_n(t_0 \mid x,s)\right\} dF_n^S(s) dF_0^X(x) - \tilde{\Gamma}_n(u) - \Gamma_0(u)
\end{aligned}$$

Adding and subtracting terms, this can be expressed as

$$R_{\Gamma,n}(u) = E_0\left[\frac{I(S\le u)\left\{Q_0(t_0 \mid X,S) - Q_n(t_0 \mid X,S)\right\}}{g_0(X,S)}\right] + E_0\left\{\frac{I(S\le u)\omega_n(O)}{g_n(X,S)}\right\} + r_n^{(1)}, \quad (3.17)$$

where

$$r_n^{(1)} \coloneqq \int \int I(s \leq u) \left\{1 - Q_n(t_0 \,|\, x, s)\right\} d(F_0^S - F_n^S)(s) d(F_n^X - F_0^X)(x) \,.$$

Putting this aside, it holds that:

$$\begin{aligned}
&E_0 \left\{ \frac{I(S \leq u)\}\omega_n(O)}{g_n(X,S)} \right\} \\
&= E_0 \left[\frac{I(S \leq u)\}}{g_n(X,S)} Q_n(t_0 \,|\, X, S) \left\{ \frac{\Delta I(Y \leq t_0)}{Q_n(Y \,|\, X, S) Q_n^C(Y \,|\, X, S)} - \int_0^{t_0 \wedge Y} \frac{\Lambda_n(du \,|\, X, S)}{Q_n(u \,|\, X, S) Q_n^C(u \,|\, X, S)} \right\} \right] \\
&= E_0 \left(\frac{I(S \leq u)}{g_n(X,S)} Q_n(t_0 \,|\, X, S) \left[\int \int \frac{\delta I(y \leq t_0) f_0^{Y,\Delta|X,S}(y, \delta \,|\, X, S)}{Q_n(y \,|\, X, S) Q_n^C(y \,|\, X, S)} d\delta dy \right. \right. \\
&\qquad \left. \left. - E_0 \left\{ \int_0^{t_0} \frac{I(Y > u) \Lambda_n(du \,|\, X, S)}{Q_n(u \,|\, X, S) Q_n^C(u \,|\, X, S)} \,|\, X, S \right\} \right] \right) \\
&= E_0 \left[\frac{I(S \leq u)}{g_n(X,S)} Q_n(t_0 \,|\, X, S) \left\{ \int \frac{I(y \leq t_0) f_0^{Y,\Delta|X,S}(y, 1 \,|\, X, S)}{Q_n(y \,|\, X, S) Q_n^C(y \,|\, X, S)} dy \right. \right. \\
&\qquad \left. \left. - \int_0^{t_0} \frac{P_0(Y > u \,|\, X, S) \Lambda_n(du \,|\, X, S)}{Q_n(u \,|\, X, S) Q_n^C(u \,|\, X, S)} \right\} \right] \\
&= E_0 \left\{ \frac{I(S \leq u)}{g_n(X,S)} Q_n(t_0 \,|\, X, S) \int_0^{t_0} \frac{Q_0(u \,|\, X, S) Q_0^C(u \,|\, X, S)}{Q_n(u \,|\, X, S) Q_n^C(u \,|\, X, S)} (\Lambda_0 - \Lambda_n)(du \,|\, X, S) \right\}
\end{aligned}$$

Adding and subtracting terms, it follows that

$$\begin{aligned}
&E_0 \left\{ \frac{I(S \leq u) \omega_n(O)}{g_n(X,S)} \right\} \\
&= E_0 \left\{ \frac{I(S \leq u)}{g_n(X,S)} Q_n(t_0 \,|\, X, S) \int_0^{t_0} \frac{Q_0(u \,|\, X, S)}{Q_n(u \,|\, X, S)} (\Lambda_0 - \Lambda_n)(du \,|\, X, S) \right\} + r_n^{(2)} \,,
\end{aligned} \tag{3.18}$$

where

$$r_n^{(2)} \coloneqq E_0 \Bigg[\frac{I(S \leq u)}{g_n(X,S)} Q_n(t_0 \,|\, X, S) \\ \times \int_0^{t_0} \frac{Q_0(u \,|\, X, S) Q_0^C(u \,|\, X, S)}{Q_n(u \,|\, X, S)} \left\{ \frac{1}{Q_n^C(u \,|\, X, S)} - \frac{1}{Q_0^C(u \,|\, X, S)} \right\} (\Lambda_0 - \Lambda_n)(du \,|\, X, S) \Bigg] .$$

Next, using the Duhamel equation of Gill and Johansen (1990), it holds that

$$Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s) = Q_n(t_0 \,|\, x, s) \int_0^{t_0} \frac{Q_0(u \,|\, x, s)}{Q_n(u \,|\, x, s)} (\Lambda_0 - \Lambda_n)(du \,|\, x, s) .$$

Plugging this into (3.18), it follows that

$$E_0 \left\{ \frac{I(S \leq u) \omega_n(O)}{g_n(X,S)} \right\} = E_0 \left[\frac{I(S \leq u) \{ Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S) \}}{g_n(X,S)} \right] + r_n^{(2)} ,$$

Finally, plugging this into (3.17), it follows that

$$R_{\Gamma,n}(u) = r_n^{(1)} + r_n^{(2)} + r_n^{(3)} ,$$

where

$$r_n^{(3)} \coloneqq E_0 \left[I(S \leq u) \left\{ Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s) \right\} \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\} \right] .$$

Next, we will show that each of the three components of $R_{\Gamma,n}(u)$ is $o_P(n^{-1/2})$. Doing so requires the following assumptions:

(A1) The following terms all converge to zero in probability:

(i) $\int \left[\int I(s \leq u) Q_n(t_0 \,|\, x, s) d(F_0^S - F_n^S)(s) \right]^2 dF_0^X(x) = o_P(1)$;

(ii) $\int \left[\int I(s \leq u) \left\{ Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s) \right\} d(F_0^S - F_n^S)(s) \right]^2 dF_0^X(x) = o_P(1)$;

(iii) $E_0 \left[\left\{ \sup_{t \in [0, t_0]} \left| \frac{Q_n(t_0 \,|\, X, S)}{Q_n(t \,|\, X, S)} - \frac{Q_0(t_0 \,|\, X, S)}{Q_0(t \,|\, X, S)} \right| \right\}^2 \Big| I(S = 0) = s \right] = o_P(1)$ for $s \in \{0, 1\}$;

(A2) The following rate conditions hold:

(i) $E_0 \left| \int_0^{t_0} \frac{Q_0(u \mid X, S)}{Q_n(u \mid X, S)} \left\{ \frac{Q_0^C(u \mid X, S)}{Q_n^C(u \mid X, S)} - 1 \right\} (\Lambda_0 - \Lambda_n)(du \mid X, S) \right| = o_P(n^{-1/2})$;

(ii) $\left[E_0 \left\{ Q_n(t_0 \mid X, S) - Q_0(t_0 \mid X, S) \right\}^2 \right]^{1/2} \left[E_0 \left\{ \frac{1}{g_n(X, S)} - \frac{1}{g_0(X, S)} \right\}^2 \right]^{1/2} = o_P(n^{-1/2})$.

Condition (A1) requires that certain second-order remainder terms do not contribute to the limiting distribution of $\Gamma_n(u)$. Roughly speaking, (A2) requires that the estimators of the conditional survival function Q_0 (and thus the conditional cumulative hazard function Λ_0), the conditional censoring function Q_0^C, and the density ratio function g_0 converge quickly enough.

To begin, first note that $r_n^{(1)}$ can be written as

$$r_n^{(1,1)} = \int h_n(x) d(F_n^X - F_0^X)(x) \,,$$

where

$$h_n(x) \coloneqq \int I(s \leq u) \left\{ 1 - Q_n(t_0 \mid x, s) \right\} d(F_0^S - F_n^S)(s) \tag{3.19}$$

By (A1.i), it holds that $P_0 h_n^2 = o_P(1)$. Next, assume that

(A3) h_n falls in a P_0-Donsker class with probability tending to one.

Using Theorem 19.24 of Van der Vaart (2000), it follows that $r_n^{(1)} = o_P(n^{-1/2})$. For the second component of the remainder term (and elsewhere in the proof), it is useful to assume that certain nuisance quantities are uniformly bounded away from zero. Specifically, assume that

(A4) There exist positive numbers ϵ_π, ϵ_g, and ϵ_c such that P_0-almost everywhere, it holds that $\pi_0(o) > \epsilon_\pi^{-1}$, $g_0(x, s) > \epsilon_g^{-1}$, and $Q_0^C(t_0 \mid x, s) > \epsilon_c^{-1}$.

Then (A2.i) and (A4) jointly imply that

$$|r_n^{(2)}| \leq \epsilon_g E_0 \left| \int_0^{t_0} \frac{Q_0(u \mid X, S)}{Q_n(u \mid X, S)} \left\{ \frac{Q_0^C(u \mid X, S)}{Q_n^C(u \mid X, S)} - 1 \right\} (\Lambda_0 - \Lambda_n)(du \mid X, S) \right| = o_P(n^{-1/2}) \,.$$

Roughly speaking, condition (A2.i) requires that the estimator of the conditional survival function Q_0 (and thus the conditional cumulative hazard function Λ_0) and the estimator of the conditional censoring function Q_0^C converge quickly enough. For the third component of the remainder, the Cauchy-Schwartz inequality and (A2.ii) imply that

$$\begin{aligned}
\left|r_n^{(3)}\right| &= \left| E_0 \left[I(S \leq u) \left\{ Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S) \right\} \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\} \right] \right| \\
&\leq \left(E_0 \left\{ Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S) \right\}^2 \right)^{1/2} \left(E_0 \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\}^2 \right)^{1/2} \\
&= o_P(1)
\end{aligned}$$

Given the results above, if conditions (A4) – (A3) are assumed, then it holds that $R_{\Gamma,n}(u) = o_P(n^{-1/2})$. Next, conditions are provided to guarantee that $E_0 \left\{ \varphi_{\Gamma,n}(O,u) - \varphi_{\Gamma,0}(O,u) \right\}^2 = o_P(1)$. First, assume that

(A5) The following terms all converge to zero in probability:

(i) $E_0 \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\}^2 = o_P(1)$;

(ii) $E_0 \left\{ \sup_{t \in [0,t_0]} \left| \frac{1}{Q_n^C(t \,|\, X, S)} - \frac{1}{Q_0^C(t \,|\, X, S)} \right| \right\}^2 = o_P(1)$;

(iii) $E_0 \left\{ \int \{ Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S) \} d(F_n^X - F_0^X)(x) \right\}^2 = o_P(1)$;

(iv) $\int \{ Q_n(t_0 \,|\, x, 0) - Q_0(t_0 \,|\, x, 0) \} d(F_n^X - F_0^X)(x) = o_P(1)$;

(v) $E_0 \{ q_n(O,u) - q_0(O,u) \}^2 = o_P(1)$.

This assumption requires that certain second-order remainder terms tend to zero. Next, note that

$$\varphi_{\Gamma,n}(o,u) - \varphi_{\Gamma,0}(o,u) = \sum_{j=1}^{5} D_{j,u,n}(o) \,,$$

where

$$
\begin{aligned}
D_{1,u,n}(o) &\coloneqq \frac{zI(s \le u)}{\pi_0(o)} \left\{ \frac{\omega_n(o)}{g_n(x,s)} - \frac{\omega_0(o)}{g_0(x,s)} \right\}, \\
D_{2,u,n}(o) &\coloneqq \frac{zI(s \le u)}{\pi_0(o)} \left\{ \tilde{r}_{M,n}(s) - r_{M,0}(s) \right\}, \\
D_{4,u,n}(o) &\coloneqq \left\{ 1 - \frac{z}{\pi_0(o)} \right\} \left\{ q_n(o,u) - q_0(o,u) \right\}, \\
D_{5,u,n}(o) &\coloneqq \eta_n(x,u) - \eta_0(x,u), \\
D_{6,u,n}(o) &\coloneqq 2 \left\{ \Gamma_0(u) - \tilde{\Gamma}_n(u) \right\}.
\end{aligned}
$$

By the triangle inequality, it holds that

$$
E_0 \left\{ \varphi_{\Gamma,n}(O,u) - \varphi_{\Gamma,0}(O,u) \right\}^2 \le \left\{ \sum_{j=1}^{5} \left(P_0 D_{j,u,n}^2 \right)^{1/2} \right\}^2.
$$

We proceed by bounding each term $P_0 D_{j,u,n}^2$ individually. To bound $P_0 D_{1,u,n}^2$, note that

$$
P_0 D_{1,u,n}^2 \le \epsilon_\pi E_0 \left\{ \frac{\omega_n(O)}{g_n(X,S)} - \frac{\omega_0(O)}{g_0(X,S)} \right\}^2.
$$

$$
P_0 D_{1,u,n}^2 \le \epsilon_\pi E_0 \left\{ \frac{\omega_n(O)}{g_n(X,S)} - \frac{\omega_0(O)}{g_0(X,S)} \right\}^2 = \epsilon_\pi E_0 \left\{ \sum_{j=1}^{5} \Omega_{j,n}(O) \right\}^2 \le \epsilon_\pi \left\{ \sum_{j=1}^{5} \left(P_0 \Omega_{j,n}^2 \right)^{1/2} \right\}^2,
$$

where

$$
\begin{aligned}
\Omega_{1,n}(o) &\coloneqq \left\{ \frac{1}{g_n(x,s)} - \frac{1}{g_0(x,s)} \right\} \left\{ \frac{Q_0(t_0 \,|\, x,s) \delta I(y \le t_0)}{Q_0(y \,|\, x,s) Q_0^C(y \,|\, x,s)} - \int_0^{t_0 \wedge y} \frac{Q_0(t_0 \,|\, x,s) \Lambda_0(dt \,|\, x,s)}{Q_0(t \,|\, x,s) Q_0^C(t \,|\, x,s)} \right\}, \\
\Omega_{2,n}(o) &\coloneqq \frac{\delta I(y \le t_0)}{g_n(x,s) Q_0^C(y \,|\, x,s)} \left\{ \frac{Q_n(t_0 \,|\, x,s)}{Q_n(y \,|\, x,s)} - \frac{Q_0(t_0 \,|\, x,s)}{Q_0(y \,|\, x,s)} \right\}, \\
\Omega_{3,n}(o) &\coloneqq \frac{\delta I(y \le t_0) Q_n(t_0 \,|\, x,s)}{g_n(x,s) Q_n(y \,|\, x,s)} \left\{ \frac{1}{Q_n^C(y \,|\, x,s)} - \frac{1}{Q_0^C(y \,|\, x,s)} \right\}, \\
\Omega_{4,n}(o) &\coloneqq \frac{1}{g_n(x,s)} \int_0^{t_0 \wedge y} \frac{Q_0(t_0 \,|\, x,s)}{Q_0(t \,|\, x,s)} \left\{ \frac{1}{Q_0^C(t \,|\, x,s)} - \frac{1}{Q_n^C(t \,|\, x,s)} \right\} \Lambda_0(dt \,|\, x,s), \\
\Omega_{5,n}(o) &\coloneqq \frac{1}{g_n(x,s)} \int_0^{t_0 \wedge y} \left\{ \frac{Q_0(t_0 \,|\, x,s) \Lambda_0(dt \,|\, x,s)}{Q_0(t \,|\, x,s) Q_n^C(t \,|\, x,s)} - \frac{Q_n(t_0 \,|\, x,s) \Lambda_n(dt \,|\, x,s)}{Q_n(t \,|\, x,s) Q_n^C(t \,|\, x,s)} \right\}.
\end{aligned}
$$

We proceed by bounding each term $P_0\Omega_{1,n}^2, ..., P_0\Omega_{5,n}^2$ individually. To bound $P_0\Omega_{1,n}^2$, first note that

$$\begin{aligned} P_0\Omega_{1,n}^2 \leq \epsilon_c^2 E_0 \Bigg[& \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\}^2 \\ & \left\{ \left| \frac{\Delta I(Y \leq t_0) Q_0(t_0 \,|\, X, S)}{Q_0(Y \,|\, X, S)} \right| + \left| \int_0^{t_0 \wedge Y} \frac{Q_0(t_0 \,|\, X, S) Q_0^C(Y \,|\, X, S) \Lambda_0(dt \,|\, X, S)}{Q_0(t \,|\, X, S) Q_0^C(t \,|\, X, S)} \right| \right\}^2 \Bigg] \\ \leq \epsilon_c^2 E_0 \Bigg[& \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\}^2 \left\{ 1 + \left| \int_0^{t_0 \wedge Y} \frac{Q_0(t_0 \,|\, X, S) \Lambda_0(dt \,|\, X, S)}{Q_0(t \,|\, X, S)} \right| \right\}^2 \Bigg] . \end{aligned}$$

Using the identity $\int_0^t Q(t)\{Q(u)\}^{-1}\Lambda(du) = 1 - Q(t)$, which holds for any survival function (and corresponding cumulative hazard function), it holds that

$$\begin{aligned} P_0\Omega_{1,n}^2 & \leq \epsilon_c^2 E_0 \left[\left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\}^2 \left\{ 1 + \left| 1 - Q_0(t_0 \wedge Y \,|\, X, S) \right| \right\}^2 \right] \\ & \leq 4\epsilon_c^2 E_0 \left\{ \frac{1}{g_n(X,S)} - \frac{1}{g_0(X,S)} \right\}^2 . \end{aligned}$$

By (A5.i), it holds that $P_0\Omega_{1,n}^2 = o_P(1)$. To bound $P_0\Omega_{2,n}^2$, note that (A1.iii) implies

$$P_0\Omega_{2,n}^2 \leq \epsilon_g^2 \epsilon_c^2 E_0 \left\{ \frac{Q_n(t_0 \,|\, X, S)}{Q_n(Y \,|\, X, S)} - \frac{Q_0(t_0 \,|\, X, S)}{Q_0(Y \,|\, X, S)} \right\} = o_P(1) ,$$

and similarly, for $P_0\Omega_{3,n}^2$, (A5.ii) implies

$$P_0\Omega_{3,n}^2 \leq \epsilon_g^2 E_0 \left\{ \frac{1}{Q_n^C(Y \,|\, X, S)} - \frac{1}{Q_0^C(Y \,|\, X, S)} \right\} = o_P(1) .$$

To bound $P_0\Omega_{4,n}^2$, note that

$$\begin{aligned} P_0\Omega_{4,n}^2 & \leq \epsilon_g^2 E_0 \left\{ \sup_{t \in [0,t_0]} \left| \frac{1}{Q_n^C(t \,|\, X, S)} - \frac{1}{Q_0^C(t \,|\, X, S)} \right| \int_0^{t_0 \wedge Y} \frac{Q_0(t_0 \,|\, X, S)}{Q_0(t \,|\, X, S)} \Lambda_0(dt \,|\, X, S) \right\}^2 \\ & = \epsilon_g^2 E_0 \left\{ \sup_{t \in [0,t_0]} \left| \frac{1}{Q_n^C(t \,|\, X, S)} - \frac{1}{Q_0^C(t \,|\, X, S)} \right| \{ 1 - Q_0(t \,|\, X, S) \} \right\}^2 \\ & = \epsilon_g^2 E_0 \left\{ \sup_{t \in [0,t_0]} \left| \frac{1}{Q_n^C(t \,|\, X, S)} - \frac{1}{Q_0^C(t \,|\, X, S)} \right| \right\}^2 . \end{aligned}$$

Then (A5.ii) implies that $P_0\Omega_{4,n}^2 = o_P(1)$. To bound $P_0\Omega_{5,n}^2$, an argument identical to one used in the proof of Lemma 3 in Westling et al. (2021), omitted for brevity, can be used to state that

$$\begin{aligned} P_0\Omega_{5,n}^2 &\leq \epsilon_g^2 E_0 \left[\int_0^{t_0 \wedge y} \left\{ \frac{Q_n(t_0 \,|\, x, s)\Lambda_n(dt \,|\, x, s)}{Q_n(t \,|\, x, s)Q_n^C(t \,|\, x, s)} - \frac{Q_0(t_0 \,|\, x, s)\Lambda_0(dt \,|\, x, s)}{Q_0(t \,|\, x, s)Q_n^C(t \,|\, x, s)} \right\} \right]^2 \\ &\leq 3\epsilon_g^2 \epsilon_c^4 E_0 \left\{ \sup_{t \in [0, t_0]} \left| \frac{Q_n(t_0 \,|\, x, s)}{Q_n(Y \,|\, x, s)} - \frac{Q_0(t_0 \,|\, x, s)}{Q_0(Y \,|\, x, s)} \right| \right\}^2 . \end{aligned}$$

Condition (A5.ii) implies that $P_0\Omega_{5,n}^2 = o_P(1)$. Combining these results, it follows that $P_0 D_{1,n}^2 = o_P(1)$. Next, we bound $P_0 D_{2,n}^2$. Given that

$$P_0 D_{2,u,n}^2 \leq \epsilon_\pi E_0 \left\{ \tilde{r}_{M,n}(S) - r_{M,0}(S) \right\}^2 ,$$

it suffices to bound $P_0(\tilde{r}_{M,n} - r_{M,0})^2$. It holds that

$$\begin{aligned} P_0(\tilde{r}_{M,n} - r_{M,0})^2 &= E_0 \left\{ \int \{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)\} dF_0^X(x) + \int Q_n(t_0 \,|\, x, S) d(F_n^X - F_0^X)(x) \right\}^2 \\ &\leq \left\{ \left(E_0 \left\{ \int \{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)\} dF_0^X(x) \right\}^2 \right)^{1/2} \right. \\ &\quad \left. + \left(E_0 \left\{ \int Q_n(t_0 \,|\, x, S) d(F_n^X - F_0^X)(x) \right\}^2 \right)^{1/2} \right\}^2 . \end{aligned}$$

For the first term above, Jensen's inequality and (A1.iii) imply that

$$\begin{aligned} E_0 \left\{ \int \{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)\} dF_0^X(x) \right\}^2 &= E_0 \left\{ \int \frac{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)}{g_0(x, S)} dF_0^{X|S}(x \,|\, S) \right\}^2 \\ &\leq \epsilon_g^2 E_0 \left\{ \int \{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)\} dF_0^{X|S}(x \,|\, S) \right\}^2 \\ &\leq \epsilon_g^2 E_0 \left\{ Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S) \right\}^2 \\ &= o_P(1) , \end{aligned}$$

For the piece on the right, first note that

$$
\begin{aligned}
&E_0\left\{\int Q_0(t_0 \,|\, x, S) d(F_n^X - F_0^X)(x)\right\}^2 \\
&\quad = \int\int\int Q_0(t_0 \,|\, x_1, s) Q_0(t_0 \,|\, x_2, s) dF_0^S(s) d(F_n^X - F_0^X)(x_1) d(F_n^X - F_0^X)(x_2) \\
&\quad = o_P(1)\,.
\end{aligned}
$$

Using (A5.iii), it holds that

$$
\begin{aligned}
&E_0\left\{\int Q_n(t_0 \,|\, x, S) d(F_n^X - F_0^X)(x)\right\}^2 \\
&\quad = E_0\left\{\int Q_0(t_0 \,|\, x, S) d(F_n^X - F_0^X)(x) + \int \{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)\} d(F_n^X - F_0^X)(x)\right\}^2 \\
&\quad \leq \left\{\left(E_0\left\{\int Q_0(t_0 \,|\, x, S) d(F_n^X - F_0^X)(x)\right\}^2\right)^{1/2}\right. \\
&\qquad \left. + \left(E_0\left\{\int \{Q_n(t_0 \,|\, x, S) - Q_0(t_0 \,|\, x, S)\} d(F_n^X - F_0^X)(x)\right\}^2\right)^{1/2}\right\}^2 \\
&\quad = o_P(1)\,.
\end{aligned}
$$

Therefore, it is the case that $P_0 D_{2,u,n}^2 = o_P(1)$. Next, to bound $P_0 D_{4,u,n}^2$, note that (A5.v) implies that

$$
P_0 D_{4,u,n}^2 = E_0\left[\left\{1 - \frac{z}{\pi_0(o)}\right\}\{q_n(o,u) - q_0(o,u)\}\right]^2 \leq (\epsilon_\pi - 1)^2 E_0\left\{q_n(o,u) - q_0(o,u)\right\}^2 = o_P(1)\,.
$$

Next, we bound $P_0 D_{5,u,n}^2$. It is the case that

$$
\begin{aligned}
P_0 D_{5,u,n}^2 &= \int\left\{\int I(s \leq u)\{1 - Q_n(t_0 \,|\, x, s)\} d(F_n^S - F_0^S)(s)\right. \\
&\qquad \left. + \int I(s \leq u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\} dF_0^S(s)\right\}^2 dF_0^X(x) \\
&= \left\{(P_0 h_n^2)^{1/2} + \left(\int\left\{\int I(s \leq u)\{Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s)\} dF_0^S(s)\right\}^2 dF_0^X(x)\right)^{1/2}\right\}^2 .
\end{aligned}
$$

It was already shown that $P_0 h_n^2 = o_P(1)$. To bound the term on the right, note that Jensen's

inequality and (A1.iii) imply that

$$\begin{aligned}
&\int \left\{ \int I(s \leq u)\{Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s)\} dF_0^S(s) \right\}^2 dF_0^X(x) \\
&\quad \leq \int \int I(s \leq u)\{Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s)\}^2 dF_0^S(s) dF_0^X(x) \\
&\quad \leq \int \int \{Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s)\}^2 dF_0^S(s) dF_0^X(x) \\
&\quad = o_P(1)\,.
\end{aligned}$$

Thus, it is the case that $P_0 D_{5,u,n}^2 = o_P(1)$. Next, to bound $P_0 D_{6,u,n}^2$, it is sufficient to show that $\tilde{\Gamma}_n(u) - \Gamma_0(u) = o_P(1)$. First, note that condition (A5.iii) implies that

$$\int \int \{Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s)\} dF_0^S(s) d(F_n^X - F_0^X)(x) = o_P(1)\,.$$

On the other hand, since the function h_n, as defined in (3.19) belongs to a Donsker class with probability tending to one, it holds that

$$P_n h_n = \int \int I(s \leq u)\{1 - Q_n(t_0 \,|\, x, s)\} d(F_n^S - F_0^S)(s) dF_n^X(x) = o_P(1)\,. \tag{3.20}$$

Adding and subtracting $\Gamma_0(u)$, it follows that

$$\begin{aligned}
\left\{ \tilde{\Gamma}_n(u) - \Gamma_0(u) \right\} &+ \int \int I(s \leq u)\{Q_n(t_0 \,|\, x, s) - Q_0(t_0 \,|\, x, s)\} dF_0^S(s) dF_0^X(x) \\
&+ \int \int I(s \leq u)\{1 - Q_0(t_0 \,|\, x, s)\} dF_0^S(s) d(F_0^X - F_n^X)(x) = o_P(1)\,.
\end{aligned}$$

The second term in the sum above is $o_P(1)$ by (A1.iii) and the third term is $o_P(1)$ by the weak law of large numbers. Therefore, it follows that $\tilde{\Gamma}_n(u) - \Gamma_0(u) = o_P(1)$, which in turn implies that $P_0 D_{6,u,n}^2 = o_P(1)$. Combining the results above, it holds that
$E_0\{\varphi_{\Gamma,n}(O,u) - \varphi_{\Gamma,0}(O,u)\}^2 = o_P(1)$. Finally, assume that

(A6) $o \mapsto \varphi_{\Gamma,n}(o,u) - \varphi_{\Gamma,0}(o,u)$ falls in some Donsker class with probability tending to one.

Then conditions (A4) – (A6) jointly guarantee that $\Gamma_n(u)$ is an asymptotically linear estimator of

$\Gamma_0(u)$ for each $u \in [0,1]$. As discussed in section 3.1.6, condition (A6) can be avoided through the use of cross-fitting.

Satisfying conditions WC.A4 and WC.A5

First, note that the class of functions $\{o \mapsto z/\pi_0(o) : u \in [0,1]\}$ has a single element and is uniformly bounded. Since the class of functions $\{o \mapsto I\{s \leq u : u \in [0,1]\}\}$ is P_0-Donsker, the class of functions equal to the pointwise products of elements of these two classes

$$\left\{o \mapsto \varphi_{\S,0}(o,u) : u \in [0,1]\right\},$$

where

$$\varphi_{\S,0}(o,u) \coloneqq \frac{zI(s \leq u)}{\pi_0(o)},$$

is P_0-Donsker as well. Therefore, it follows that

$$n^{1/2} \sup_{u \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \varphi_{\S,0}(O_i,u) - F_0^S(u) \right| = O_P(1),$$

which in turn implies

$$n^{1/3} \sup_{u \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \varphi_{\S,0}(O_i,u) - F_0^S(u) \right| = o_P(1).$$

Thus, condition WC.A4 is satisfied. Next, assume that

(A7) $||(\Gamma_n - \Gamma_0) - r_{M,0}(u)(F_n^S - F_0^S)||_{\infty,[0,1]} = o_P(1)$.

This is equivalent to condition WC.A5.

Satisfying conditions WC.B1 and WC.B2

We start by introducing additional notation and definitions. The bracketing number $N_{[]}(\epsilon, \mathcal{G}, L_2(P))$ is the smallest number of ϵ-brackets needed to cover the class $\mathcal{G}$ with respect to the $L_2(P)$ norm and

the covering number $N(\epsilon, \mathcal{G}, L_2(\mathcal{Q}))$ is the minimum number of ϵ-balls in $L_2(\mathcal{Q})$ required to cover $\mathcal{G}$ with respect to the $L_2(\mathcal{Q})$ norm (van der Vaart and Wellner, 2013). Next, define the following:

$$
\begin{aligned}
g_{u,t} &\coloneqq o \mapsto \varphi_{\Gamma,0}(o, u+t) - \varphi_{\Gamma,0}(o,u) - r_{M,0}(u)\left(\varphi_{\S,0}(o,u+t) - \varphi_{\S,0}(o,u)\right) , \\
\mathcal{G}_{u,R} &\coloneqq \{g_{u,t} : |t| \leq R\} , \\
G_{u,R} &\coloneqq \text{envelope function of } \mathcal{G}_{u,R} .
\end{aligned}
$$

Given these definitions, it is assumed that

(A8) There exist constants $C > 0$ and $V \in [0, 2)$, for all $\epsilon \in (0, 1]$ and R small enough, such that either $\log N_{[]}(\epsilon ||G_{u,R}||_{P_0,2}, \mathcal{G}_{u,R}, L_2(P_0)) \leq C\epsilon^{-V}$ or $\log \sup_{\mathcal{Q}} N(\epsilon ||G_{u,R}||_{\mathcal{Q},2}, \mathcal{G}_{u,R}, L_2(\mathcal{Q})) \leq C\epsilon^{-V}$;

(A9) $P_0 G_{u,R}^2 = O(R)$ and $P_0 G_{u,R}^2 \{RG_{u,R} > \eta\} = o(R)$ as $R \to 0$ for all $\eta > 0$.

Conditions (A8) and (A9) are restatements of conditions WC.B1 and WC.B2, respectively, in our setting.

Satisfying condition WC.B3

Next, define

$$
\Sigma_0(t_1, t_2) \coloneqq E_0\left[\left\{\varphi_{\Gamma_0,0}(O, t_1) - r_{M,0}(u)\varphi_{\S_0,0}(O, t_1)\right\}\left\{\varphi_{\Gamma_0,0}(O, t_2) - r_{M,0}(u)\varphi_{\S_0,0}(O, t_2)\right\}\right] \quad (3.21)
$$

to be the covariance function of the Gaussian process to which $\{\mathbb{G}_n[\varphi_{\Gamma_0,0}(O,t) - r_{M,0}(u)\varphi_{\S_0,0}(O,t)] : t \in (0,1)\}$ weakly converges. The function $(t_1, t_2) \mapsto \Sigma_0(t_1, t_2)$ drives the asymptotic behavior of $r_{M,n}$. Also, define the following for convenience:

$$
\begin{aligned}
\varphi^{(1)}_{\Gamma,0}(o) &\coloneqq \frac{z}{\pi_0(o)} \left\{ \frac{\omega_0(o)}{g_0(x,s)} + r_{M,0}(s) \right\}, \\
\varphi^{(2)}_{\Gamma,0}(o,u) &\coloneqq \left\{ 1 - \frac{z}{\pi_0(o)} \right\} q_0(o,u) + \eta_0(x,u) - 2\Gamma_0(u), \\
\varphi^{(1)}_{\S,0}(o) &\coloneqq \frac{z}{\pi_0(o)}, \\
\varphi^{(2)}_{\S,0}(o,u) &\coloneqq -\frac{z F^S_0(u)}{\pi_0(o)}.
\end{aligned}
$$

This allows us to write the influence functions $\varphi_{\Gamma,0}$ and $\varphi_{\S,0}$ as

$$
\begin{aligned}
\varphi_{\Gamma,0}(o) &\coloneqq \varphi^{(1)}_{\Gamma,0}(o) I(s \le u) + \varphi^{(2)}_{\Gamma,0}(o,u), \\
\varphi_{\S,0}(o) &\coloneqq \varphi^{(1)}_{\S,0}(o) I(s \le u) + \varphi^{(2)}_{\S,0}(o,u).
\end{aligned}
$$

Next, define

$$
\begin{aligned}
\Sigma^*_0(t_1,t_2) \coloneqq\ & E_0 \left[I(S \le t_1) \varphi^{(1)}_{\Gamma,0}(O) \varphi^{(2)}_{\Gamma,0}(O,t_2) + \varphi^{(2)}_{\Gamma,0}(O,t_1) \left\{ I(S \le t_2) \varphi^{(1)}_{\Gamma,0}(O) + \varphi^{(2)}_{\Gamma,0}(O,t_2) \right\} \right] \\
& - r_{M,0}(u) E_0 \left[I(S \le t_1) \varphi^{(1)}_{\Gamma,0}(O) \varphi^{(2)}_{\S,0}(O,t_2) + \varphi^{(2)}_{\Gamma,0}(O,t_1) \left\{ I(S \le t_2) \varphi^{(1)}_{\S,0}(O) + \varphi^{(2)}_{\S,0}(O,t_2) \right\} \right] \\
& - r_{M,0}(u) E_0 \left[I(S \le t_2) \varphi^{(1)}_{\Gamma,0}(O) \varphi^{(2)}_{\S,0}(O,t_1) + \varphi^{(2)}_{\Gamma,0}(O,t_2) \left\{ I(S \le t_1) \varphi^{(1)}_{\S,0}(O) + \varphi^{(2)}_{\S,0}(O,t_1) \right\} \right] \\
& + \left\{ r_{M,0}(u) \right\}^2 E_0 \left[I(S \le t_1) \varphi^{(1)}_{\S,0}(O) \varphi^{(2)}_{\S,0}(O,t_2) + \varphi^{(2)}_{\S,0}(O,t_1) \left\{ I(S \le t_2) \varphi^{(1)}_{\S,0}(O) + \varphi^{(2)}_{\S,0}(O,t_2) \right\} \right].
\end{aligned}
$$

Then it is the case that

$$
\begin{aligned}
\Sigma_0(t_1,t_2) &= \Sigma^*_0(t_1,t_2) + E_0 \left[I(S \le t_1 \wedge t_2) \left\{ \varphi^{(1)}_{\Gamma,0}(O) - r_{M,0}(u) \varphi^{(1)}_{\S,0}(O) \right\}^2 \right] \\
&= \Sigma^*_0(t_1,t_2) + E_0 \left[I(S \le t_1 \wedge t_2) \frac{Z}{\{\pi_0(O)\}^2} \left\{ \frac{\omega_0(O)}{g_0(X,S)} + r_{M,0}(S) - r_{M,0}(u) \right\}^2 \right].
\end{aligned}
$$

Iterating the expectation twice, it follows that

$$\Sigma_0(t_1,t_2) = \Sigma_0^*(t_1,t_2)$$
$$+ E_0\left\{E_0\left(E_0\left[I(S \le t_1 \wedge t_2)\frac{Z}{\{\pi_0(O)\}^2}\left\{\frac{\omega_0(O)}{g_0(X,S)} + r_{M,0}(S) - r_{M,0}(u)\right\}^2 \,\middle|\, S, X\right]\middle| X\right)\right\}.$$

Written in terms of integrals, this can be expressed as

$$\Sigma_0(t_1,t_2) = \Sigma_0^*(t_1,t_2) + \int\int I(s \le t_1 \wedge t_2) E_0\left[\frac{Z g_0(x,s)}{\{\pi_0(x,Y,\Delta)\}^2}\right.$$
$$\left.\times\left\{\frac{\omega_0(x,Y,\Delta,s)}{g_0(x,s)} + r_{M,0}(s) - r_{M,0}(u)\right\}^2 \middle| S = s, X = x\right] dF_0^S(s) dF_0^X(x)\,.$$

This satisfies the form of equation (3) of Westling and Carone (2020), as it can be written as

$$\Sigma_0(t_1,t_2) = \Sigma_0^*(t_1,t_2) + \int\int_{-\infty}^{t_1\wedge t_2} A_0(t_1,t_2,s,x) H_0(ds,x) Q_0(dx)$$

where

$$A_0(t_1,t_2,s,x) \coloneqq E_0\left[\frac{Z g_0(x,s)}{\{\pi_0(x,Y,\Delta)\}^2}\left\{\frac{\omega_0(x,Y,\Delta,s)}{g_0(x,s)} + r_{M,0}(s) - r_{M,0}(u)\right\}^2 \middle| S = s, X = x\right],$$
$$H_0(s,x) \coloneqq F_0^S(s)\,,$$
$$Q_0(x) \coloneqq F_0^X(x)\,.$$

Condition WC.B3a is satisfied, since $\Sigma_0^*(t_1,t_2) = \Sigma_0^*(t_2,t_1)$ and is continuously differentiable within a neighborhood of u for each $u \in (0,1)$. Additionally, condition WC.B3b is easily satisfied; since $A_0(t_1,t_2,s,x)$ does not depend on t_1 or t_2, it trivially holds that A_0 is differentiable in t_1 and t_2, with a derivative of zero, and that $A_0(t_1,t_2,s,x) = A_0(t_2,t_1,s,x)$. Next, assume that

(A10) The following smoothness conditions hold:

(i) $s \mapsto F_0^{S|X}(s \,|\, x)$ is continuously differentiable over $(0,1)$ for F_0^X-almost each x;

(ii) $s \mapsto Q_0(t_0 \,|\, x, s)$ and $s \mapsto Q_0^C(t_0 \,|\, x, s)$ are both continuous in s for F_0^X-almost each x.

These conditions help us to verify WC.B3c and WC.B3d. By assumption, it holds that $s \mapsto r_{M,0}(s)$ is continuous over $(0,1)$. Also, since $s \mapsto F_0^{S|X}(s \,|\, x)$ is continuous over $(0,1)$ for F_0^X-almost each x, it follows that $s \mapsto g_0(x,s)$ is continuous. Finally, since $s \mapsto Q_0(t_0 \,|\, x,s)$ and $s \mapsto Q_0^C(t_0 \,|\, x,s)$ are both continuous in s for F_0^X-almost each x, it follows that $s \mapsto \omega_0(x,y,\delta,s)$ is continuous for F_0^X-almost each x. Since $s \mapsto A_0(u,u,s,x)$ is an elementary function of these component functions, it is continuous at $s = u$ for $u \in (0,1)$ uniformly in x over the support of F_0^X. Condition WC.B3d is satisfied since $s \mapsto H_0(s,x) = F_0^S(s)$ does not depend on x, is nondecreasing over its domain, and is continuously differentiable at each $s \in (0,1)$, since $s \mapsto F_0^{S|X}(s \,|\, x)$ is continuously differentiable over $(0,1)$ by assumption.

Satisfying conditions WC.B4 and WC.B5

Next, define:

$$K_n(\delta) \coloneqq n^{2/3} \sup_{|t| \leq \delta n^{-1/3}} |R_{\Gamma,n}(u+t) - R_{\Gamma,n}(u)|\,,$$

and assume that

(A11) $K_n(\delta) = o_P(1)$ for all $\delta > 0$ and for some $\alpha \in (1,2)$, $\delta \mapsto \delta^{-\alpha} E_0\{K_n(\delta)\}$ is decreasing for all δ small enough and n large enough.

This condition corresponds to conditions WC.B4 and WC.B5, and guarantees that the stochastic remainder term $R_{\Gamma,n}(u)$ does not affect the limit distribution. Finally, we must derive the form of the scale factor $\tau_0(u)$. This is defined as

$$\tau_0(u) \coloneqq \left[\frac{4 \dot{r}_{M,0}(u) \kappa_0(u)}{\{f_0^S(u)\}^2} \right]^{1/3}, \tag{3.22}$$

where $\dot{r}_{M,0}$ is the derivative of $r_{M,0}$, and where

$$\kappa_0(u) \coloneqq f_0^S(u) E_0\{A_0(u,u,u,X)\}$$
$$= E_0\left(E_0\left[\frac{Z\left\{ f_0^S(u)\omega_0(X,Y,\Delta,u)\right\}^2}{f_0^{S|X}(u\,|\,X)\left\{\pi_0(X,Y,\Delta)\right\}^2} \middle| S=u, X \right] \right).$$

Plugging in the form above for $\kappa_0(u)$ into (3.22), it follows that

$$\begin{aligned}
\tau_0(u) &\coloneqq \left\{ \frac{4\dot{r}_{M,0}(u)}{\{f_0^S(u)\}^2} E_0\left(E_0\left[\frac{Z\{f_0^S(u)\omega_0(X,Y,\Delta,u)\}^2}{f_0^{S|X}(u\,|\,X)\{\pi_0(X,Y,\Delta)\}^2} \,\middle|\, S=u, X \right] \right) \right\}^{1/3} \\
&= \left\{ \frac{4\dot{r}_{M,0}(u)}{\{f_0^S(u)\}^2} E_0\left(E_0\left[\frac{\{f_0^S(u)\omega_0(X,Y,\Delta,u)\}^2}{f_0^{S|X}(u\,|\,X)\{\pi_0(X,Y,\Delta)\}^2} \middle| Z=1, S=u, X \right] P_0(Z=1\,|\,X, S=u) \right) \right\}^{1/3} \\
&= \left\{ 4\dot{r}_{M,0}(u) E_0\left(\frac{1}{f_0^{S|X}(u\,|\,X)} E_0\left[\left\{ \frac{\omega_0(X,Y,\Delta,u)}{\pi_0(X,Y,\Delta)} \right\}^2 \middle| Z=1, S=u, X \right] g_{z,0}(x,s) \right) \right\}^{1/3} \\
&= \left\{ 4\dot{r}_{M,0}(u) \int \frac{\gamma_0(X,u) g_{z,0}(x,s)}{f_0^{S|X}(u\,|\,X)} dF_0^X(x) \right\}^{1/3}
\end{aligned}$$

This completes the proof. □

3.6.2 Proof of Theorem 4

If $r_{M,0}$ is constant on (0,1), then $\beta_0 = 0$

Suppose $s \mapsto r_{M,0}(s)$ is constant on $(0,1)$. Then $\Theta_0(x) \coloneqq \int_0^x r_{M,0}(t)dt = cx$ for some $c \in \mathbb{R}$. β_0 can then be evaluated as

$$\begin{aligned}
\beta_0 &= E_0^*\left[\{(\lambda_{1,*}\lambda_{2,*} - \lambda_{3,*})(U - \lambda_{1,*}) + (\lambda_{2,*} - \lambda_{1,*}^2)(U^2 - \lambda_{2,*})\}\Theta_0(U) \right] \\
&= cE_0^*\left[\{(\lambda_{1,*}\lambda_{2,*} - \lambda_{3,*})(U^2 - \lambda_{1,*}U) + (\lambda_{2,*} - \lambda_{1,*}^2)(U^3 - \lambda_{2,*}U)\} \right] \\
&= c\left(\lambda_{1,*}\lambda_{2,*}^2 - \lambda_{1,*}^3\lambda_{2,*} - \lambda_{2,*}\lambda_{3,*} + \lambda_{1,*}^2\lambda_{3,*} \right) + \left(\lambda_{2,*}\lambda_{3,*} - \lambda_{1,*}\lambda_{2,*}^2 - \lambda_{1,*}^2\lambda_{3,*} + \lambda_{1,*}^3\lambda_{2,*} \right) \\
&= 0
\end{aligned}$$

If $\beta_0 = 0$, then $r_{M,0}$ is constant on (0,1)

Suppose $\beta_0 = 0$ and assume that $s \mapsto r_{M,0}$ is nonincreasing. Since $r_{M,0}$ is nonincreasing, it must be the case that $\Theta_0(s) \coloneqq \int_0^s r_{M,0}(t)dt$ is concave. Also assume that the distribution P_0^* satisfies

$$(\lambda_{2,*}^2 - \lambda_{4,*})(\lambda_{1,*}^2 - \lambda_{2,*}) - (\lambda_{3,*} - \lambda_{1,*}\lambda_{2,*})^2 \neq 0\,, \tag{3.23}$$

where, as in section 3.2.1, $\lambda_{k,*} \coloneqq E_0^*[U^k]$. Our goal is to show that $r_{M,0}$ must be constant over the interval $(0,1)$. This is the case if $s \mapsto \Theta_0(s)$ is a linear function.

First, we prove that if $\beta_0 = 0$ and equality (3.23) is satisfied, then the P_0^*-least-squares projection of Θ_0 onto the space of quadratic functions has a quadratic coefficient of zero. We begin by finding the form of this coefficient. Note that this projection is given by the function

$$\Theta_0^Q : u \mapsto \alpha_0 + \alpha_1 u + \alpha_2 u^2\,,$$

where $(\alpha_0, \alpha_1, \alpha_2)$ is chosen to minimize

$$E_0^*\left[\left\{\Theta_0(U) - (\alpha_0 + \alpha_1 U + \alpha_2 U^2)\right\}^2\right] \tag{3.24}$$

The values $(\alpha_0, \alpha_1, \alpha_2)$ that minimize (3.24) can be found by solving the system of equations

$$\frac{\partial}{\partial \alpha_k} E_0^*\left[\left\{\Theta_0(U) - (\alpha_0 + \alpha_1 U + \alpha_2 U^2)\right\}^2\right] = 0 \text{ for } k \in \{0,1,2\}.$$

Solving this system, the quadratic coefficient is found to be

$$\frac{E_0^*\left[\{(\lambda_{1,*}\lambda_{2,*} - \lambda_{3,*})(U - \lambda_{1,*}) + (\lambda_{2,*} - \lambda_{1,*}^2)(U^2 - \lambda_{2,*})\}\Theta_0(U)\right]}{(\lambda_{2,*}^2 - \lambda_{4,*})(\lambda_{1,*}^2 - \lambda_{2,*}) - (\lambda_{3,*} - \lambda_{1,*}\lambda_{2,*})^2}\,. \tag{3.25}$$

By assumption, the numerator of (3.25) equals zero and the denominator of (3.25) is nonzero, and so Θ_0^Q is a linear function. Next, assume for contradiction that $r_{M,0}$ is not constant on $(0,1)$. Then Θ_0 is concave and nonlinear. We will show that Θ_0^Q cannot be a linear function, thus obtaining a contradiction. Let Θ_0^L represent the P_0^*-least-squares projection of Θ_0 onto the space of linear

functions. First, note that, since Θ_0 is concave and nonlinear, Θ_0^L necessarily intersects Θ_0 at two points; denote these points $p_1 \coloneqq (x_1, y_1)$ and $p_2 \coloneqq (x_2, y_2)$, with $x_1 < x_2$. These are represented by the two orange points in Figure 3.7.

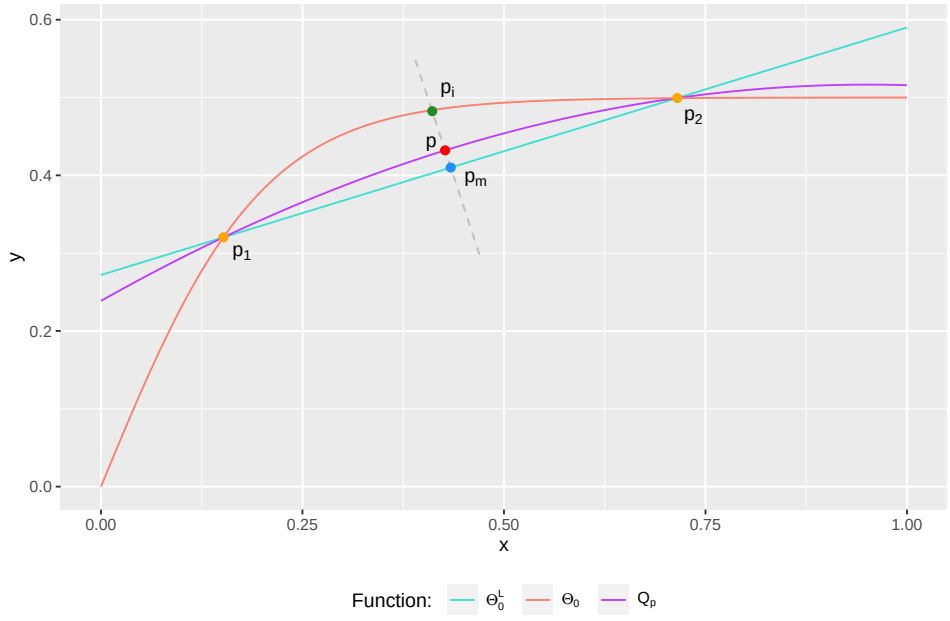


Figure 3.7: Quadratic pointwise improvement relative to a linear fit.

Our strategy will be to show that there exists a quadratic function Q that is a better approximation to Θ_0 pointwise, in the sense that

$$|Q(x) - \Theta_0(x)| < |\Theta_0^L(x) - \Theta_0(x)| \quad \forall x \in [0,1] \setminus \{x_1, x_2\}\,. \tag{3.26}$$

The midpoint of the points p_1 and p_2 along Θ_0^L (represented by the blue point in Figure 3.7) is the point $p_m \coloneqq (\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$. If a perpendicular line is extended from p_m towards Θ_0, this line will intersect Θ_0; denote this intersection point $p_i \coloneqq (x_i, y_i)$ (represented by the green point in Figure 3.7). Next, for any point p that lies on the line segment between p_i and p_m (such that $p \neq p_m$), a geometric fact is that there is a unique quadratic function that passes through these three points; denote this function by $x \mapsto Q_p(x)$. We will show that if a point p is chosen close enough to p_m, condition (3.26) are satisfied for $Q = Q_p$. First, note that as p moves towards p_m such that the distance $||p - p_m||_2$ decreases, $\dot{Q}_p(x_1)$ decreases and $\dot{Q}_p(x_2)$ increases, where $\dot{Q}_p$ is

the derivative of Q_p. Denote the slope of the line segment ℓ_1 between p_1 and p_i by s_1 and denote the slope of the line segment ℓ_2 between p_i and p_2 by s_2. Note that both of these line segments lie below Θ_0, since any line segment connecting two points of a concave function necessarily lies below that function. If p is selected such that $\dot{Q}_p(x_1) < s_1$, then it is the case that $Q_p(x) < \ell_1(x)$ over (x_1, x_i), and thus $Q_p(x) < \Theta_0(x) <$ over (x_1, x_i), since the second derivative of Q_p is negative. Similarly, if p is selected such that $\dot{Q}_p(x_2) > s_2$, it is the case that $Q_p(x) < \ell_2(x)$ over (x_i, x_2), and thus $Q_p(x) < \Theta_0(x) <$ over (x_i, x_2).

Next, consider the interval $(0, x_1)$. Select any line segment ℓ defined by the point p_1 and a second point $(0, y_\ell)$, with y_ℓ chosen such that the slope of ℓ is greater than the slope of Θ_0^L but less than the derivative of Θ_0 at x_1. By the concavity of Θ_0, it holds that Θ_0 lies entirely below ℓ over $[0, x_1)$. Next, note that if p is chosen such that $Q_p(0) > y_\ell$, by the concavity of Q_p, it must be true that ℓ lies entirely below Q_p over $[0, x_1)$. Thus, it holds that $\Theta_0(x) < Q_p(x)$ for $x \in [0, x_1)$. An analogous argument can be made to show that p can be chosen such that $\Theta_0(x) < Q_p(x)$ for $x \in (x_2, 1]$.

Therefore, if p is selected to be close enough to p_m, it holds that $\Theta_0(x) < Q_p(x)$ for $x \in [0, x_1) \cup (x_2, 1]$ and $Q_p(x) < \Theta_0(x)$ for $x \in (x_1, x_2)$. This implies that (3.26) holds with $Q = Q_p$, which in turn implies that Q_p is a better approximation to Θ_0 in a P_0^*-least-squares sense than Θ_0^L. Since Q_p lies in the space of quadratic functions and represents a better approximation to Θ_0 than Θ_0^L, the best P_0^*-least-squares approximation to Θ_0 in the space of linear functions, it must be the case that Θ_0^Q is nonlinear. Thus, we have obtained a contradiction, since we previously showed that Θ_0^Q must be linear. Therefore, Θ_0 must be linear, and so $r_{M,0}$ must be constant on $(0, 1)$. □

3.6.3 Proof of Theorem 5

Many arguments used in the proof of Theorem 5 are analogous to those used in section 3.6.1 to show that $\Gamma_n(u)$ is asymptotically linear, and so for brevity, we do not provide the same level of detail. To begin, define the remainder term

$$R_{\Theta,n}(u) \coloneqq \tilde{\Theta}_n(u) - \Theta_0(u) - E_0\left\{\varphi_{\Theta,n}(O, u)\right\},$$

where $\varphi_{\Theta,n}$ is an estimator of the influence function $\varphi_{\Theta,0}$ of $\Theta_n(u)$ given in (3.15), and where

$$\tilde{\Theta}_n(u) \coloneqq \int\int I(s \leq u)\{1 - Q_n(t_0 \,|\, x, s)\} dF_n^X(x) ds \,.$$

As with $\Gamma_n(u)$, the one-step estimator $\Theta_n(u)$ is asymptotically linear if the following three conditions hold:

1. $R_{\Theta,n}(u) = o_P(n^{-1/2})$,
2. $E_0 \left\{\varphi_{\Theta,n}(O,u) - \varphi_{\Theta,0}(O,u)\right\}^2 = o_P(1)$
3. $o \mapsto \varphi_{\Theta,n}(o,u) - \varphi_{\Theta,0}(o,u)$ falls in some Donsker class with probability tending to one.

First, conditions are provided under which $R_{\Theta,n}(u) = o_P(n^{-1/2})$. First, assume that

(B1) There exists an $\epsilon_f > 0$ such that $f_0^{S|X}(s \,|\, x) > \epsilon_f^{-1}$ for P_0-almost all (s,x).

(B2) $$\left[E_0 \left\{Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S)\right\}^2\right]^{1/2} \left[E_0 \left\{ \frac{1}{f_n^{S|X}(S \,|\, X)} - \frac{1}{f_0^{S|X}(S \,|\, X)} \right\}^2\right]^{1/2} = o_P(n^{-1/2}).$$

Next, the remainder term can be expressed as

$$\begin{aligned} R_{\Theta,n}(u) &= E_0 \left[\frac{Z I(S \leq u) \omega_n(O)}{\pi_0(O) f_n^{S|X}(S \,|\, X)} + \left\{1 - \frac{Z}{\pi_0(O)}\right\} \tilde{q}_n(O,u) + \eta_n^*(X,u) \right] - \Theta_0(u) \\ &= E_0 \left[\frac{I(S \leq u) \omega_n(O)}{f_n^{S|X}(S \,|\, X)} \right] - E_0 \left[\frac{I(S \leq u)\{Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S)\}}{f_0^{S|X}(S \,|\, X)} \right] \end{aligned}$$

Using arguments analogous to those used in section 3.6.1, it follows that

$$E_0 \left[\frac{I(S \leq u) \omega_n(O)}{f_n^{S|X}(S \,|\, X)} \right] = E_0 \left[\frac{I(S \leq u)\{Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S)\}}{f_n^{S|X}(S \,|\, X)} \right] + r_n^{(4)} \,,$$

where

$$r_n^{(4)} := E_0\Bigg[\frac{I(S \le u)}{f_n^{S|X}(S \,|\, X)} Q_n(t_0 \,|\, X, S) \\ \times \int_0^{t_0} \frac{Q_0(u \,|\, X, S) Q_0^C(u \,|\, X, S)}{Q_n(u \,|\, X, S)} \left\{ \frac{1}{Q_n^C(u \,|\, X, S)} - \frac{1}{Q_0^C(u \,|\, X, S)} \right\} (\Lambda_0 - \Lambda_n)(du \,|\, X, S)\Bigg].$$

This allows us to express the remainder as $R_{\Theta,n}(u) = r_n^{(4)} + r_n^{(5)}$, where

$$r_n^{(5)} := E_0\left[I(S \le u)\{Q_n(t_0 \,|\, X, S) - Q_0(t_0 \,|\, X, S)\} \left\{ \frac{1}{f_n^{S|X}(S \,|\, X)} - \frac{1}{f_0^{S|X}(S \,|\, X)} \right\} \right]$$

Conditions (B1) and (A2.i) imply that

$$\left| r_n^{(4)} \right| \le \epsilon_f E_0 \left| \int_0^{t_0} \frac{Q_0(u \,|\, X, S)}{Q_n(u \,|\, X, S)} \left\{ \frac{Q_0^C(u \,|\, X, S)}{Q_n^C(u \,|\, X, S)} - 1 \right\} (\Lambda_0 - \Lambda_n)(du \,|\, X, S) \right| = o_P(n^{-1/2})$$

The Cauchy-Schwartz inequality and (B2) imply that $r_n^{(5)} = o_P(n^{-1/2})$, and so $R_{\Theta,n}(u) = o_P(n^{-1/2})$ as well.

Next, conditions are provided to guarantee that $E_0\left\{\varphi_{\Theta,n}(O,u) - \varphi_{\Theta,0}(O,u)\right\}^2 = o_P(1)$. Assume that

(B3) The following terms all converge to zero in probability:

(i) $E_0 \left\{ \frac{1}{f_n^{S|X}(S \,|\, X)} - \frac{1}{f_0^{S|X}(S \,|\, X)} \right\}^2 = o_P(1).$

(ii) $E_0\left\{\tilde{q}_n(O,u) - \tilde{q}_0(O,u)\right\}^2 = o_P(1).$

(iii) $\int\int \{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\}^2 dF_0^X(x) ds = o_P(1).$

(iv) $\int\int I(s \le u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\} d(F_n^X - F_0^X)(x) ds = o_P(1).$

Note that

$$\varphi_{\Gamma,n}(o,u) - \varphi_{\Gamma,0}(o,u) = \sum_{j=1}^{4} E_{j,u,n}(o)\,,$$

where

$$
\begin{aligned}
E_{1,u,n}(o) &\coloneqq \frac{zI(s \leq u)}{\pi_0(o)} \left\{ \frac{\omega_n(o)}{f_n^{S|X}(s \,|\, x)} - \frac{\omega_0(o)}{f_0^{S|X}(s \,|\, x)} \right\}, \\
E_{2,u,n}(o) &\coloneqq \left\{ 1 - \frac{z}{\pi_0(o)} \right\} \left\{ \tilde{q}_n(o,u) - \tilde{q}_0(o,u) \right\}, \\
E_{3,u,n}(o) &\coloneqq \eta_n^*(x,u) - \eta_0^*(x,u), \\
E_{4,u,n}(o) &\coloneqq \Theta_0(u) - \tilde{\Theta}_n(u).
\end{aligned}
$$

By the triangle inequality, it holds that

$$
E_0 \left\{ \varphi_{\Theta,n}(O,u) - \varphi_{\Theta,0}(O,u) \right\}^2 \leq \left\{ \sum_{j=1}^{4} \left(P_0 E_{j,u,n}^2 \right)^{1/2} \right\}^2.
$$

We proceed by bounding each term $P_0 E_{j,u,n}^2$ individually. First, to bound $P_0 E_{1,u,n}^2$, (A4) implies that

$$
P_0 E_{1,u,n}^2 \leq \epsilon_\pi E_0 \left\{ \frac{\omega_n(O)}{f_n^{S|X}(S \,|\, X)} - \frac{\omega_0(O)}{f_0^{S|X}(S \,|\, X)} \right\}^2 \tag{3.27}
$$

Using arguments analogous to those used to bound $P_0 D_{1,u,n}^2$ in section 3.6.1, (B3.i) implies that $P_0 E_{1,u,n}^2 = o_P(1)$. To bound $P_0 E_{2,u,n}^2$, note that

$$
P_0 E_{2,u,n}^2 \leq (\epsilon_\pi - 1)^2 E_0 \left\{ \tilde{q}_n(O,u) - \tilde{q}_0(O,u) \right\}^2.
$$

Then by (B3.ii), it follows that $P_0 E_{2,u,n}^2 = o_P(1)$. To bound $P_0 E_{3,u,n}^2$, it follows from Jensen's inequality that

$$
\begin{aligned}
P_0 E_{3,u,n}^2 &= \int \left[\int I(s \leq u) \{ Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s) \} ds \right]^2 dF_0^X(x) \\
&\leq \int \int I(s \leq u) \{ Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s) \}^2 dF_0^X(x) ds \\
&\leq \int \int \{ Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s) \}^2 dF_0^X(x) ds.
\end{aligned}
$$

Then (B3.iii) implies that $P_0 E_{2,u,n}^3 = o_P(1)$. Finally, note that

$$\Theta_0(u) - \tilde{\Theta}_n(u) = \int\int I(s \leq u)\{1 - Q_n(t_0 \,|\, x, s)\} d(F_n^X - F_0^X)(x) ds + \int\int I(s \leq u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\} dF_0^X(x) ds \,.$$

The second term on the right-hand side of this equation is $o_P(1)$ by (B3.iii) since Jensen's inequality implies that

$$\left[\int\int I(s \leq u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\} dF_0^X(x) ds\right]^2 \leq \int\int I(s \leq u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\}^2 dF_0^X(x) ds$$

For the other term, it holds that

$$\begin{aligned}
&\int\int I(s \leq u)\{1 - Q_n(t_0 \,|\, x, s)\} d(F_n^X - F_0^X)(x) ds \\
&\quad = \int\int I(s \leq u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\} d(F_n^X - F_0^X)(x) ds \\
&\qquad + \int\int I(s \leq u)\{1 - Q_0(t_0 \,|\, x, s)\} d(F_n^X - F_0^X)(x) ds \\
&\quad = \int\int I(s \leq u)\{Q_0(t_0 \,|\, x, s) - Q_n(t_0 \,|\, x, s)\} d(F_n^X - F_0^X)(x) ds + o_P(1)
\end{aligned}$$

Then by (B3.iv), it holds that $P_0 E_{4,u,n}^2 = o_P(1)$, and so it is the case that $E_0\left\{\varphi_{\Theta,n}(O,u) - \varphi_{\Theta,0}(O,u)\right\}^2 = o_P(1)$. Finally, it is assumed that

(B4) $o \mapsto \varphi_{\Theta,n}(o,u) - \varphi_{\Theta,0}(o,u)$ falls in some Donsker class with probability tending to one.

Then conditions (A4), (A2.i), (A1.iii), and (A5.ii), as well as (B1) – (B4) jointly guarantee that $\Theta_n(u)$ is an asymptotically linear estimator of $\Theta_0(u)$ for each $u \in [0,1]$. As discussed in section 3.1.6, condition (B4) can be avoided through the use of cross-fitting.

This completes the proof. □

3.7 Additional simulation results

3.7.1 Estimation: point mass at the edge

This set of simulations is analogous to the set described in section 3.3.1, except the distribution of S included a point mass at zero, based on the function $\tilde{g}_{s,0}^{(3)}$ described in section 3.3.2, resulting in a point mass of roughly 40% at the edge. This gave us the opportunity to evaluate the operating characteristics of the edge-corrected estimator described in section 3.1.7. Figure 3.8 shows that the edge correction essentially eliminated the bias at the left edge (i.e., the lower limit of detection). This is expected, as the edge-specific estimator is theoretically unbiased. At values of s just above the edge, the bias was generally reduced as well. Towards the right edge of the interval, the edge correction did not have a noticeable effect, as expected.

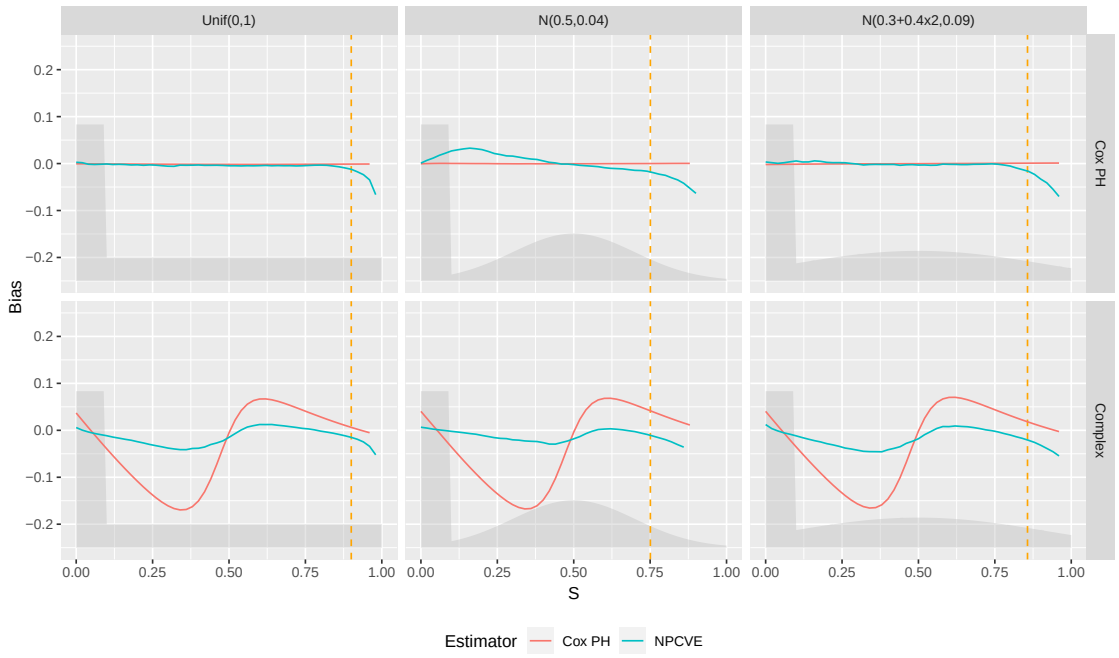


Figure 3.8: Bias of two estimators ("NPCVE" = nonparametric CVE, "Cox PH" = marginalized Cox proportional hazards), displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

Figure 3.9 highlights the effect of the edge estimator on confidence interval coverage. The effect on coverage was not as pronounced as the effect on bias, but it was generally closer to the

nominal 95% value at the left edge. The improvement was most pronounced when the marginal distribution of the marker was standard uniform, since for this simulation scenario, we observed poor undercoverage of the regular NPCVE estimator.

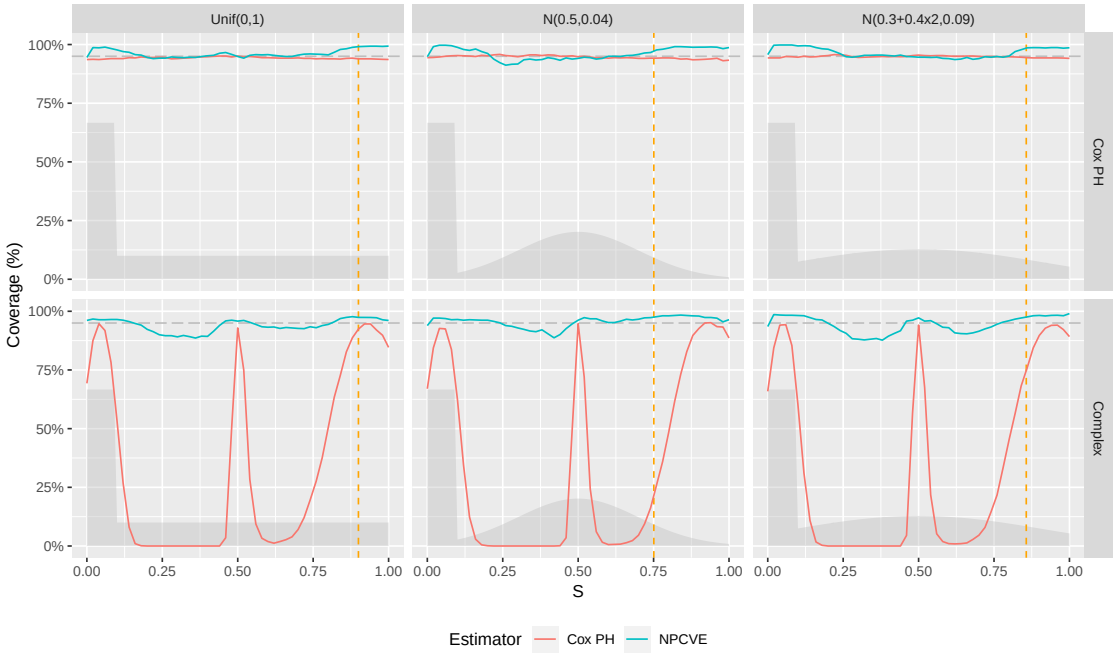


Figure 3.9: 95% confidence interval coverage of two estimators ("NPCVE" = nonparametric CVE, "Cox PH" = marginalized Cox proportional hazards), displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

Figure 3.10 shows that there were massive improvements in terms of the standard deviation of the estimator at the left edge. This is expected, since a substantial portion of the data were being used specifically to estimate $\text{CVE}_0(0)$ using an estimator that converges at root-n rate, and since the nonparametric estimator is highly unstable at the left edge.

3.7.2 Estimation: empirical standard deviation

Figure 3.11 displays the standard deviation of the nonparametric risk estimator across simulation replicates (labeled "empirical SD") against the true asymptotic standard deviation (labeled "true SD"). The two generally agreed in the interior of the unit interval, but the empirical SD was much greater than the true SD towards the endpoints of the interval.

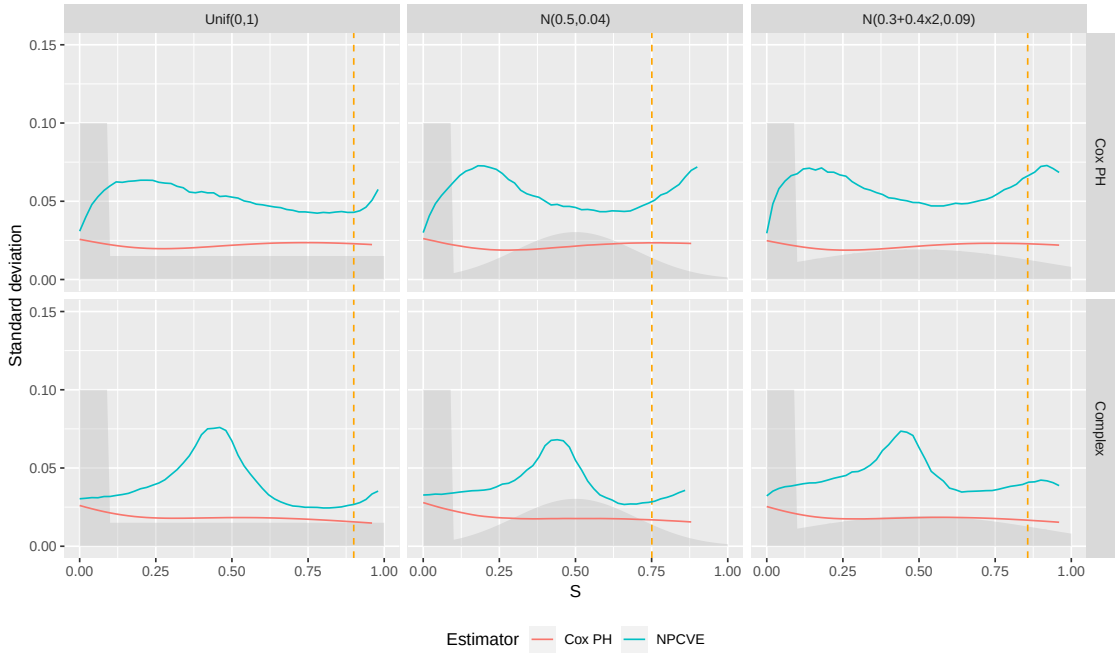


Figure 3.10: Standard deviation of two estimators ("NPCVE" = nonparametric CVE, "Cox PH" = marginalized Cox proportional hazards), displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

3.7.3 Hypothesis test with different marginal distributions of S

Figures 3.12 and 3.13 summarize results analogous to those given in section 3.3.2, but for different conditional distributions $f_0^{S|X}$. Figure 3.12 displays results corresponding to the setting in which this distribution was a mixture of a Normal distribution with mean 0.5 and variance 0.2 (truncated to lie within $[0, 1]$) and a point mass at zero, for three different propensity scores $\tilde{g}_{s,0}$ and two different conditional survival functions Q_0.

Figure 3.13 displays results corresponding to the setting in which this distribution was a mixture of a Normal distribution with mean $0.3 + 0.4X_2$ and variance 0.3 (truncated to lie within $[0, 1]$) and a point mass at zero, for three different propensity scores $\tilde{g}_{s,0}$ and two different conditional survival functions Q_0.

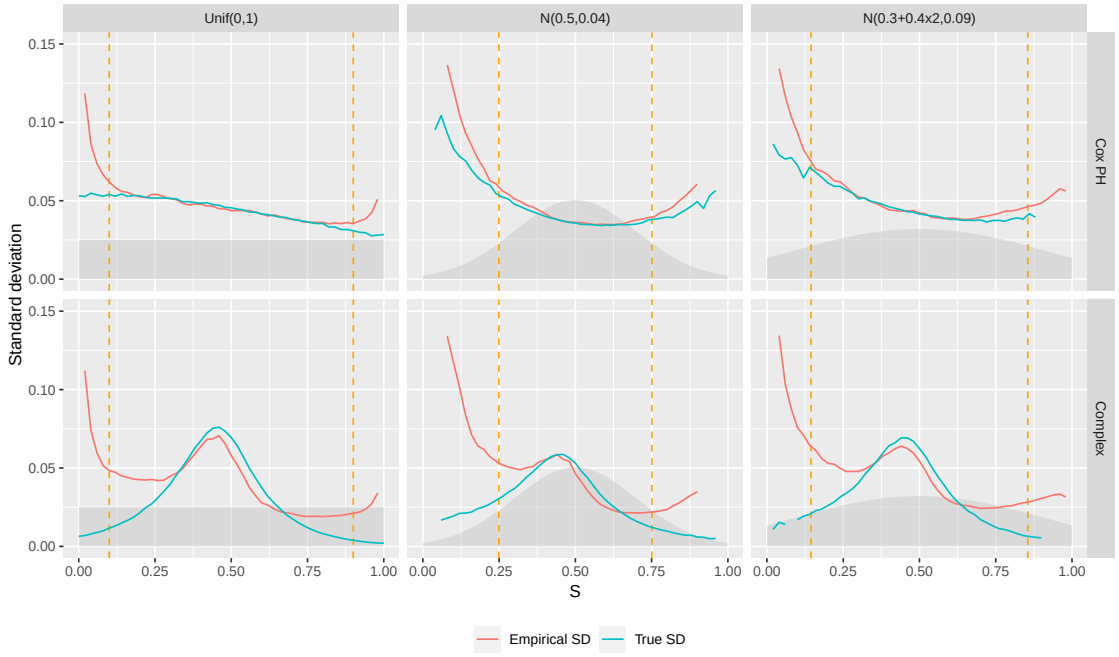


Figure 3.11: Empirical versus true standard deviation of nonparametric CVE estimator, displayed for three conditional distributions of $S|X$ and two conditional survival functions ("Cox PH" = $Q_0^{(1)}$, "Complex" = $Q_0^{(2)}$). Dashed vertical yellow lines represent the 10% and 90% quantiles of the marginal distribution of S.

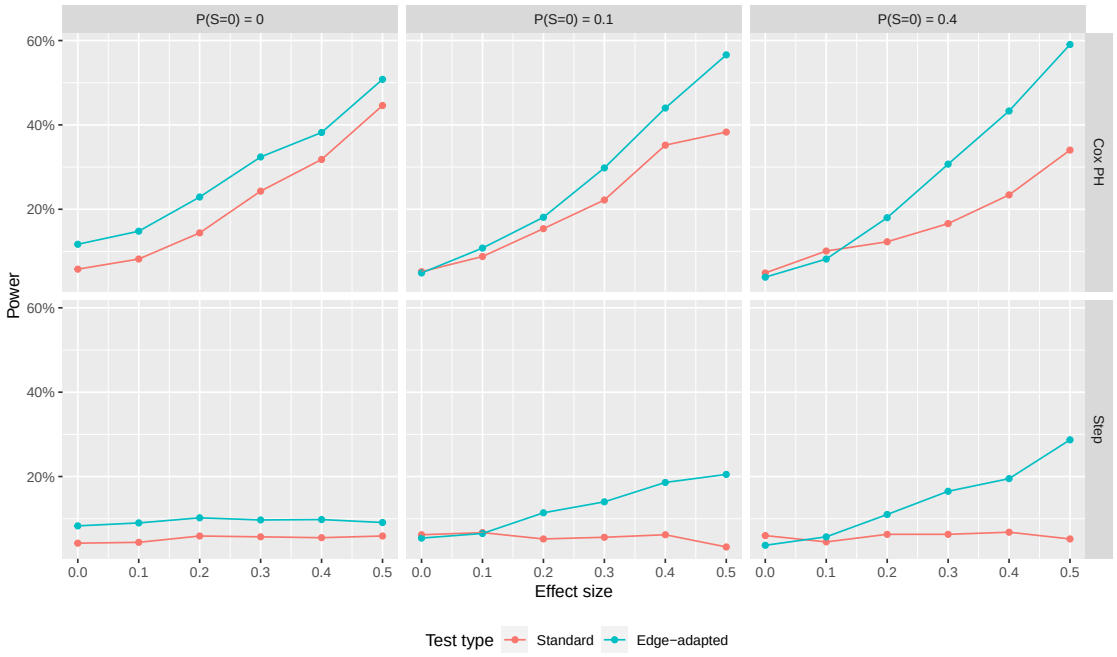


Figure 3.12: Power of two versions of hypothesis test, where the marker S follows a mixture of a $N(0.5, 0.04)$ distribution (truncated to lie within $[0, 1]$) and a point mass at zero, displayed for three propensity scores $\tilde{g}_{s,0}$ ("P(S=0)=0" $= \tilde{g}^{(1)}_{s,0}$, "P(S=0)=0.1" $= \tilde{g}^{(2)}_{s,0}$, "P(S=0)=0.4" $= \tilde{g}^{(3)}_{s,0}$) and two conditional survival functions ("Cox PH" $= Q^{(1)}_0$, "Step" $= Q^{(3)}_0$).

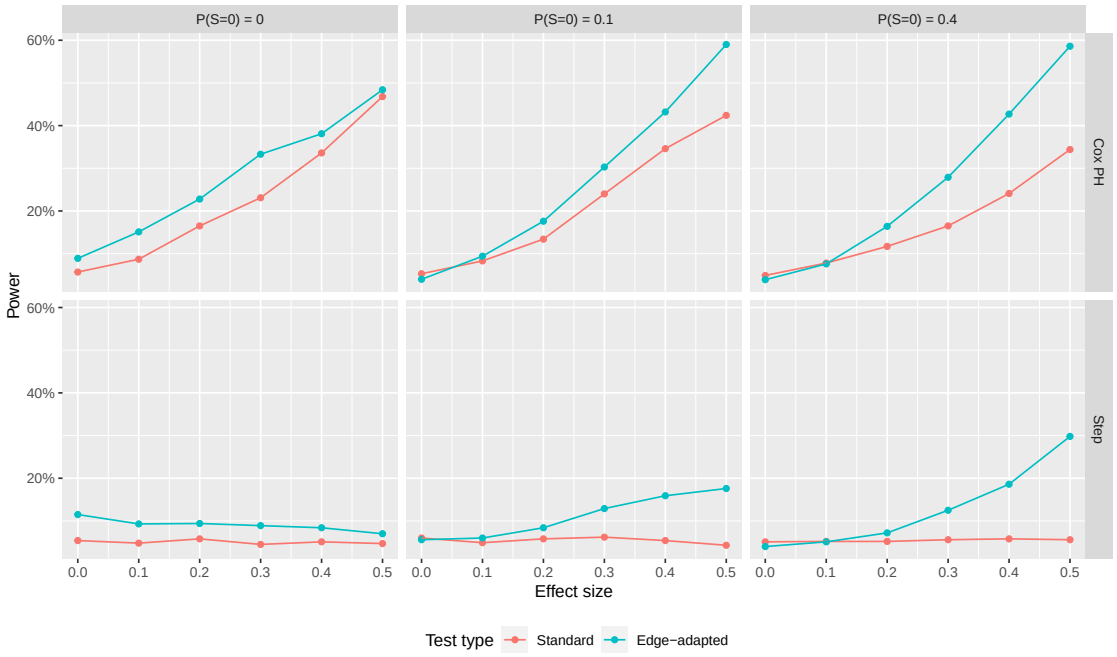


Figure 3.13: Power of two versions of hypothesis test, where the marker S follows a mixture of a $N(0.3+0.4X_2, 0.09)$ distribution (truncated to lie within $[0,1]$) and a point mass at zero, displayed for three propensity scores $\tilde{g}_{s,0}$ ("P(S=0)=0" $= \tilde{g}_{s,0}^{(1)}$, "P(S=0)=0.1" $= \tilde{g}_{s,0}^{(2)}$, "P(S=0)=0.4" $= \tilde{g}_{s,0}^{(3)}$) and two conditional survival functions ("Cox PH" $= Q_0^{(1)}$, "Step" $= Q_0^{(3)}$).

3.8 Additional data analysis results

In section 3.4, results were presented from data analyses of two of the four markers considered by Gilbert et al. (2022b) in their immune correlates analysis of the Coronavirus Efficacy (COVE) trial (NCT04470427) of the mRNA-1273 COVID-19 vaccine. Here, results are presented for the other two markers, IgG bAbs to the spike protein and 80% inhibitory dilution (ID80) nAb titer.

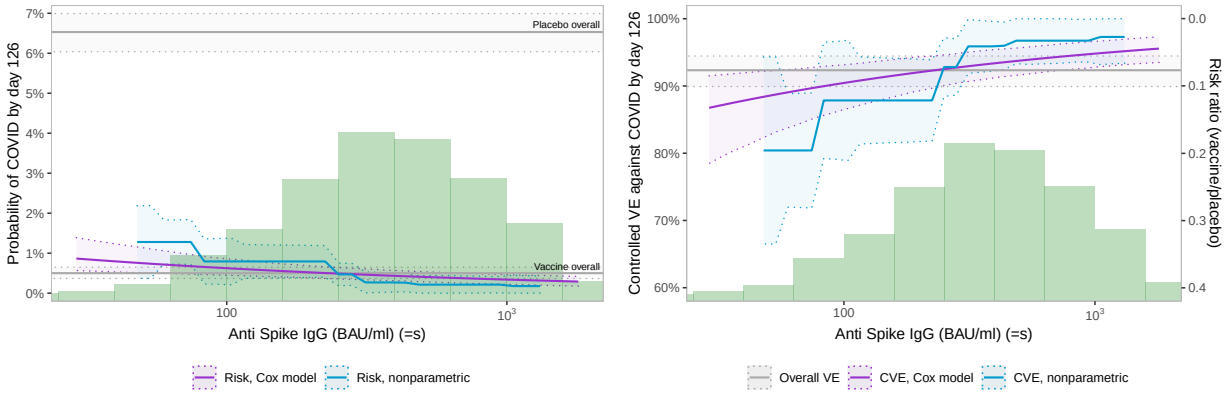
Figure 3.14 shows CR and CVE curves for the IgG bAbs to spike receptor binding domain (RBD) marker, estimated using both a Cox proportional hazards model and the nonparametric estimator, along with pointwise 95% confidence intervals. Subfigures (3.14a) and (3.14b) correspond to the day 29 marker measurement whereas subfigures (3.14c) and (3.14d) correspond to the day 57 measurement.

Figure 3.15 shows analogous results for the 80% inhibitory dilution (ID80) nAb titer marker. The edge correction described in section 3.1.7 was used for the day 29 ID80 measurements (subfigures 3.15a and 3.15b).

Table 3.2 displays hypothesis testing results for the two markers considered above. P-values for the ID-80 marker at both day 29 and day 57 are significant at a significance level of $\alpha = 0.05$, although this does not account for multiple testing. Note that we used the edge-adapted test for the ID-80 marker at day 29, as the estimated marginal distribution of the marker has greater than 10% mass at the edge.

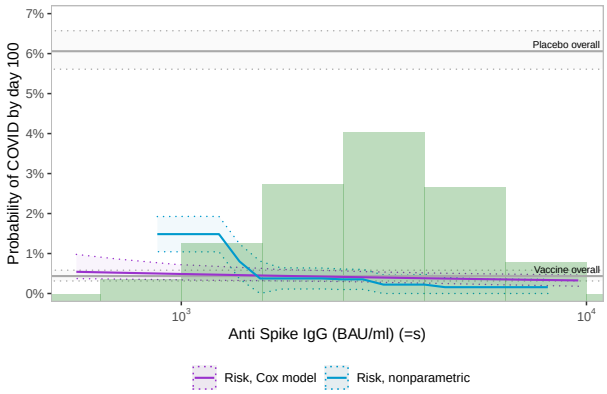
Marker	Day	Test version	P-value
IgG bAbs to the spike protein	29	Standard	0.6609
IgG bAbs to the spike protein	57	Standard	0.9283
80% inhibitory dilution (ID80) nAb titer	29	Edge-adapted	0.0190
80% inhibitory dilution (ID80) nAb titer	57	Standard	0.0303

Table 3.2: Hypothesis testing results for Spike and ID80 markers, measured at days 29 and 57.

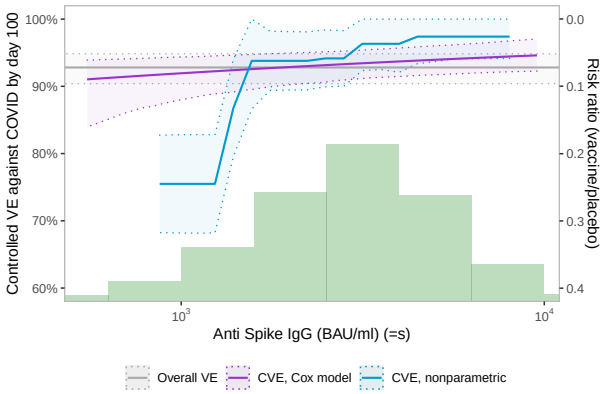


(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29

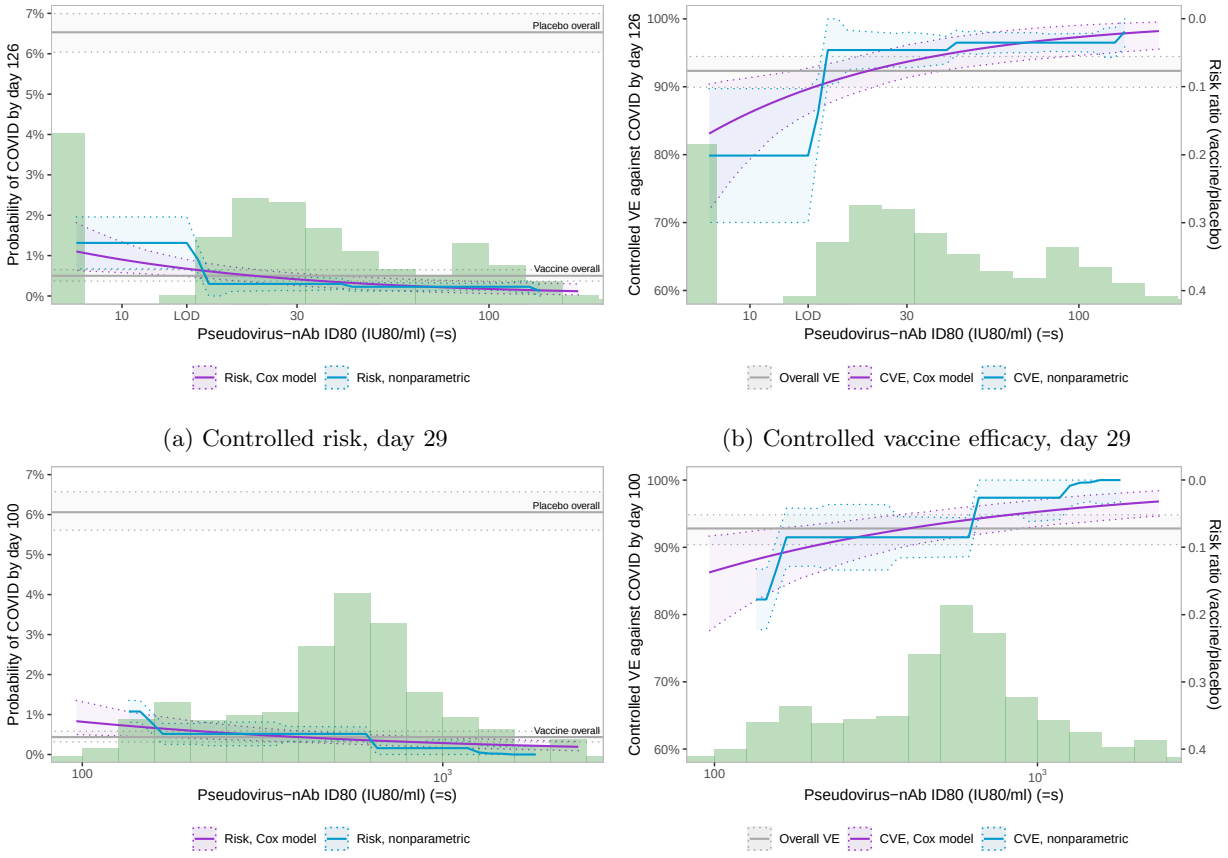


(c) Controlled risk, day 57



(d) Controlled vaccine efficacy, day 57

Figure 3.14: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the IgG bAbs to the spike protein marker, measured at days 29 and 57. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.



(a) Controlled risk, day 29

(b) Controlled vaccine efficacy, day 29

(c) Controlled risk, day 57

(d) Controlled vaccine efficacy, day 57

Figure 3.15: Controlled risk (CR) and controlled vaccine efficacy (CVE) curves for the 80% inhibitory dilution (ID80) nAb titer marker, measured at days 29 and 57. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). Grey lines in the CR plots represent the vaccine group risk and the placebo group risk, and the grey line in the CVE plot represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

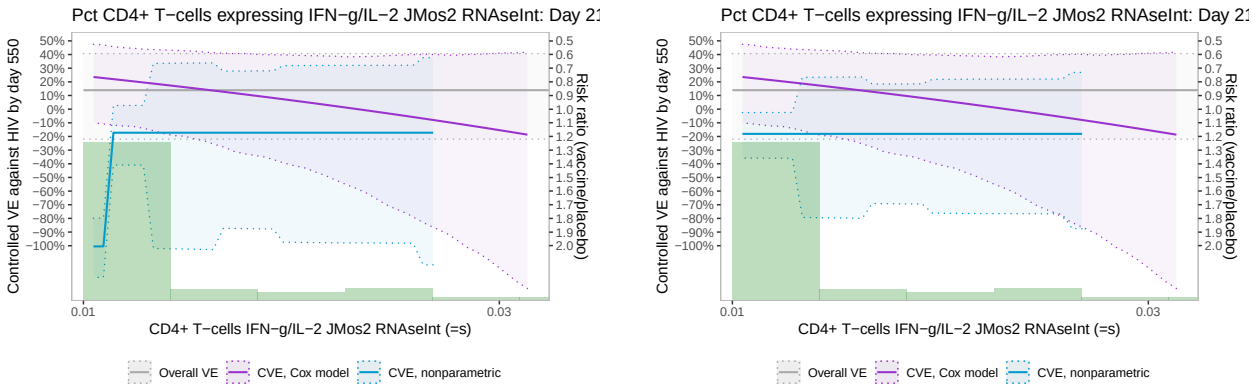
Chapter 4

Improving monotone function estimators using convex least squares

4.1 Problem statement

In general, most function estimators that assume monotonicity have a known asymptotic bias when evaluated at the endpoints of the target function domain, as discussed in section 3.1.7. Additionally, we have observed in simulation that, for points in the interior, there is a characteristic finite sample bias that increases in magnitude as the chosen point approaches the endpoint. This represents a serious issue with generalized Grenander-type estimators that can lead to an overestimation of the amount of change in the function over its domain. In data analyses, we have observed jumps in the CVE curve near the edge that we deem to be unrealistic; an example of this phenomenon is given in Figure 4.1a.

In section 3.1.7, we described two possible solutions to this issue, truncating the display of the estimated function and constructing an estimator of the function at the LLOD. The first method is useful, but somewhat ad-hoc and involves the subjective selection of cutoff points. The second only applies in settings in which there is a sufficient point mass at the LLOD. However, another solution may be possible to reduce bias in settings in which there is no point mass. The intuition for this idea is easiest to understand in the case of classical isotonic regression. In section 3.1.1,



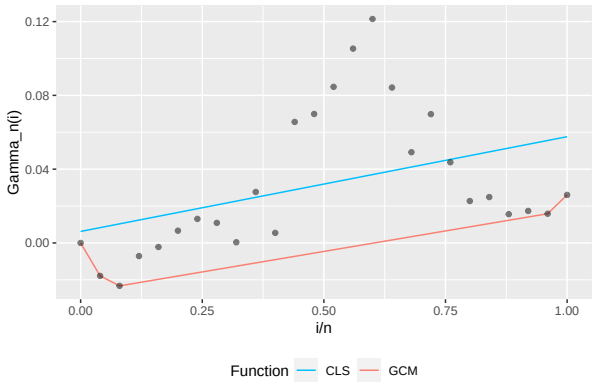
(a) Nonparametric estimator based on the greatest convex minorant (GCM) operator.

(b) Nonparametric estimator based on the convex least squares (CLS) function.

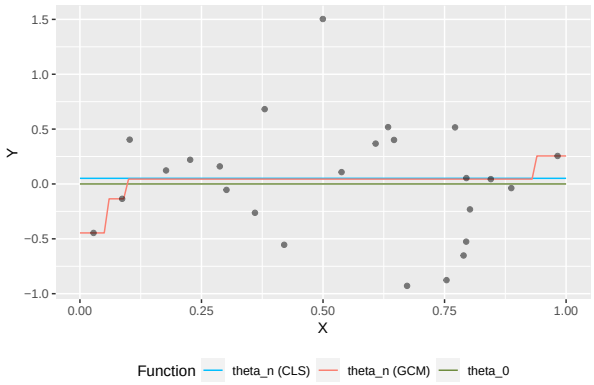
Figure 4.1: Controlled vaccine efficacy (CVE) curves for a CD4+ T-cell biomarker in the HVTN 705 HIV vaccine efficacy trial, estimated using a Cox proportional hazards model (purple) and a nonparametric estimator (blue). The grey line represents overall vaccine efficacy and the dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

we described how the isotonic regression estimator can be expressed in terms of the derivative of the GCM of the cusum diagram. The key observation is that the step of computing the GCM can instead be replaced by computing the convex least squares (CLS) function. The CLS function, relative to a set of points, is defined as the function within the class of all convex piecewise linear functions that minimizes the residual sum of squares with respect to the observed data points. It is necessary to restrict to a smaller function class than the set of all convex functions to guarantee the uniqueness of the function within that class that minimizes the residual sum of squares. This modified procedure also results in a nondecreasing final estimator, but simulation results suggest that it may also result in reduction of finite sample bias towards the edge, as well as elimination of the asymptotic bias at the edge itself. In particular, this may help avoid bias when the true function is constant. Intuitively, the reason for this can be seen in Figure 4.2, which represents a single sample of 25 iid pairs (X, Y), with X distributed at standard uniform and $Y|X$ normally distributed with mean zero and standard deviation 0.5. Panel (4.2a) shows the cusum diagram points with the GCM and CLS lines overlaid. Panel (4.2b) shows the resulting derivatives of the GCM and CLS lines, plotted against the original data points. In panel (4.2a), it can be seen that while these two lines are parallel for most of the interval $[0, 1]$, a single data point at the edge can

substantially influence the shape of the GCM, whereas the shape of the CLS line is more stable. This results in the estimators seen in panel (4.2b).



(a) The cumulative sum (cusum) diagram, with the greatest convex minorant (GCM) and convex least squares (CLS) lines overlaid.



(b) A sample of 25 iid pairs (X, Y), with X $U(0,1)$ and $Y|X$ $N(0, 0.25)$, along with the isotonic regression estimators resulting from the use of the GCM and the CLS.

Figure 4.2: Illustration of the use of the greatest convex minorant (GCM) and the convex least squares (CLS) lines in isotonic regression.

Although we described this idea in the context of classical isotonic regression, the same underlying idea applies to the more complex setting of generalized Grenander-type estimators, which includes the estimator studied in detail in chapter 3. This method is illustrated in Figure 4.1b; it can be seen that the use of the CLS instead of the GCM eliminates the unrealistic large jump.

4.2 Methods

In this section, we consider the setting of an arbitrary generalized Grenander-type estimator and consider the goal of making inference about a monotone nondecreasing function $\theta_0 : [0,1] \to \mathbb{R}$. Consider a generalized Grenander-type estimator $\theta_n(x)$ of $\theta_0(x)$ for fixed $x \in (0,1)$ constructed using an estimator $\Gamma_n(t)$ of the primitive function $\Gamma_0(t)$ such that $\Gamma_n(t)$ is asymptotically linear for each $t \in [0,1]$, and suppose that sufficient conditions hold such that the asymptotic results of Westling and Carone (2020) apply. Here, we consider an alternative estimator $\theta_n^*(x)$ of $\theta_0(x)$, obtained as a generalized Grenander-type estimator based instead on the primitive estimator Γ_n^*, which equals the least-squares projection of Γ_n onto the space of convex piecewise linear functions.

Of note, the GCM step involved in the construction of this generalized Grenander-type estimator is still done, but has no effect, since Γ_n^* is already convex.

We conjecture that $\theta_n(x)$ and $\theta_n^*(x)$ are asymptotically equivalent such that $n^{1/3}\{\theta_n(x) - \theta_0(x)\}$ and $n^{1/3}\{\theta_n^*(x) - \theta_0(x)\}$ both converge in distribution to the same scaled Chernoff distribution for each $x \in (0,1)$. One way of proving this would be to show that $\Gamma_n(t)$ and $\Gamma_n^*(t)$ have the same influence function for each $t \in [0,1]$, and then to verify conditions (B4) and (B5) of Westling and Carone (2020). Condition (B4) states that

(B4) $K_n(\delta) = o_P(1)$ for each fixed $\delta > 0$,

where

$$K_n(\delta) \coloneqq n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} \left| (H^*_{x+u,n} - H^*_{x,n}) - \theta_0(x)(R_{x+u,n} - R_{x,n}) \right| ,$$

and where $H^*_{x,n}$ and $R_{x,n}$ are the remainder terms corresponding to the asymptotically linear representations of the primitive function estimator Γ_n^* and the transformation function estimator, respectively. Condition (B5) states that

(B5) For some $\alpha \in (1,2)$, $\delta \mapsto \delta^{-\alpha} E_0\{K_n(\delta)\}$ is decreasing for all δ small enough and n large enough.

To proceed with studying (B4), we assume that

$$n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} \left| R_{x+u,n} - R_{x,n} \right| = o_P(1)$$

for each fixed $\delta > 0$, such that by the triangle inequality, it suffices to show that

$$n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} \left| H^*_{x+u,n} - H^*_{x,n} \right| = o_P(1) \tag{4.1}$$

for each fixed $\delta > 0$. Let $H_{x,n}$ be the remainder corresponding to Γ_n such that

$$\Gamma_n(x) - \Gamma_0(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\Gamma,x,0}(O_i) + H_{x,n}\,.$$

If it can be shown that $\Gamma_n(x) - \Gamma_n^*(x) = o_P(n^{-1/2})$, then adding and subtracting $\Gamma_n^*(x)$, it holds that

$$\Gamma_n^*(x) - \Gamma_0(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\Gamma,x,0}(O_i) + H_{x,n}^*\,,$$

with $H_{x,n}^* = H_{x,n} + \Gamma_n^*(x) - \Gamma_n(x)$. Since we assumed that conditions (B4) and (B5) hold for the estimator $\theta_n(x)$ derived from Γ_n, it suffices to show that

$$n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} \left| \{\Gamma_n^*(x+u) - \Gamma_n(x+u)\} - \{\Gamma_n^*(x) - \Gamma_n(x)\} \right| = o_P(1) \tag{4.2}$$

for each fixed $\delta > 0$. A term of this form is known as an *oscillation modulus*; historically, these have been studied using strong approximations (see, for example, Csörgö and Révész, 1981). Unfortunately, simulations appear to suggest that $\Gamma_n(x) - \Gamma_n^*(x) \neq o_P(n^{-1/2})$; See Figures 4.8 and 4.8 in the next section. Therefore, an alternative strategy that we will pursue is to derive the influence function of $\Gamma_n^*(x)$ and show that all conditions of Westling and Carone (2020) hold for this estimator.

4.3 Simulation study

To assess the potential benefits and feasibility of the CLS approach, we conducted a simulation study. We considered both the setting of monotone density function estimation as well as the setting of classical isotonic regression. We first focus on the regression scenario. Data was generated by first sampling X from either a standard uniform distribution or a Normal distribution with mean 0.5 and standard deviation 0.3 (truncated to lie within $[0, 1]$). The true regression function was the identity function, and Y values were generated as $X + \epsilon$, where ϵ is sampled from a mean-zero Normal distribution with standard deviation 0.2. Both the GCM-based estimator (i.e. classical isotonic regression) and the CLS-based estimator were computed for each point over a grid of

data points on the unit interval. The relative performance of the two estimators was assessed by computing their bias and standard deviation. We ran 10,000 simulation replicates for every level combination.

To begin, it is useful to plot realizations of both estimators against one another to get a basic sense of their distributions and how well correlated they are; this is displayed in Figure 4.3.

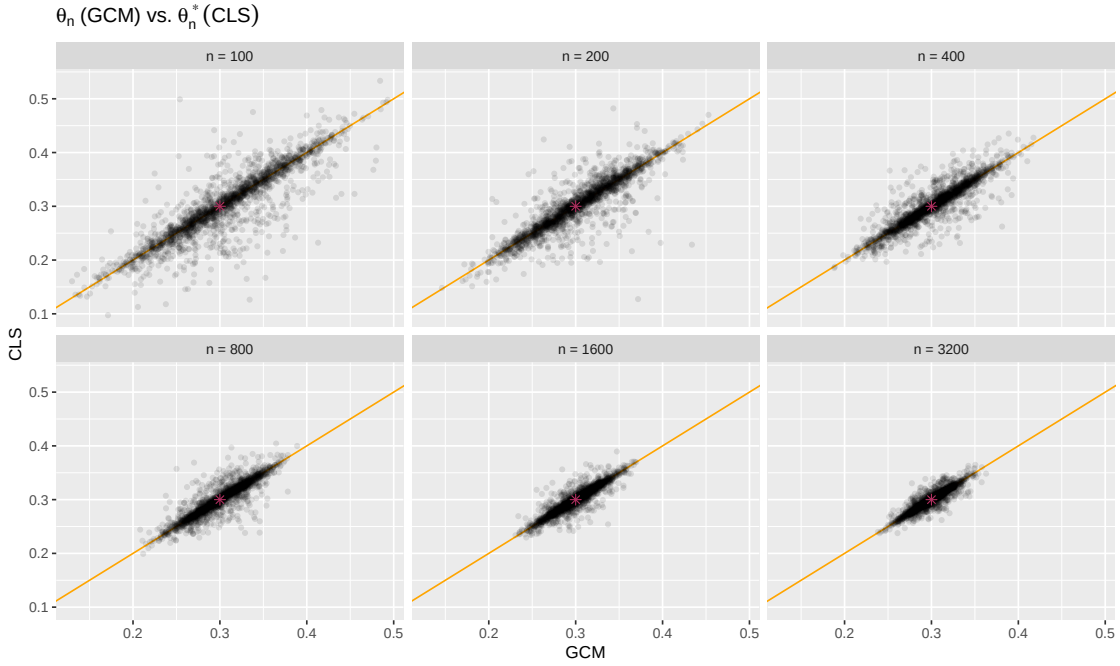


Figure 4.3: Realizations of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (isotonic regression) evaluated at $x = 0.3$, displayed for 1,000 simulation replicates and six different sample sizes; X follows a standard uniform distribution. The orange line is the identity function, and the red star at $(0.3, 0.3)$ represents the true regression value.

In Figure 4.4, the bias of both estimators is displayed across different values of the unit interval. The two estimators perform similarly in the interior, but the CLS estimator performs much better towards the endpoints.

In Figure 4.5, we show the same results as in Figure 4.4, but scaled by $n^{1/3}$. In particular, this illustrates the asymptotic bias associated with GCM-based estimator and shows that this bias appears to be far smaller with the CLS-based estimator.

In Figure 4.6, the standard deviation of both estimators is displayed across different values of the unit interval. Again, the two estimators perform similarly in the interior, with the GCM-based

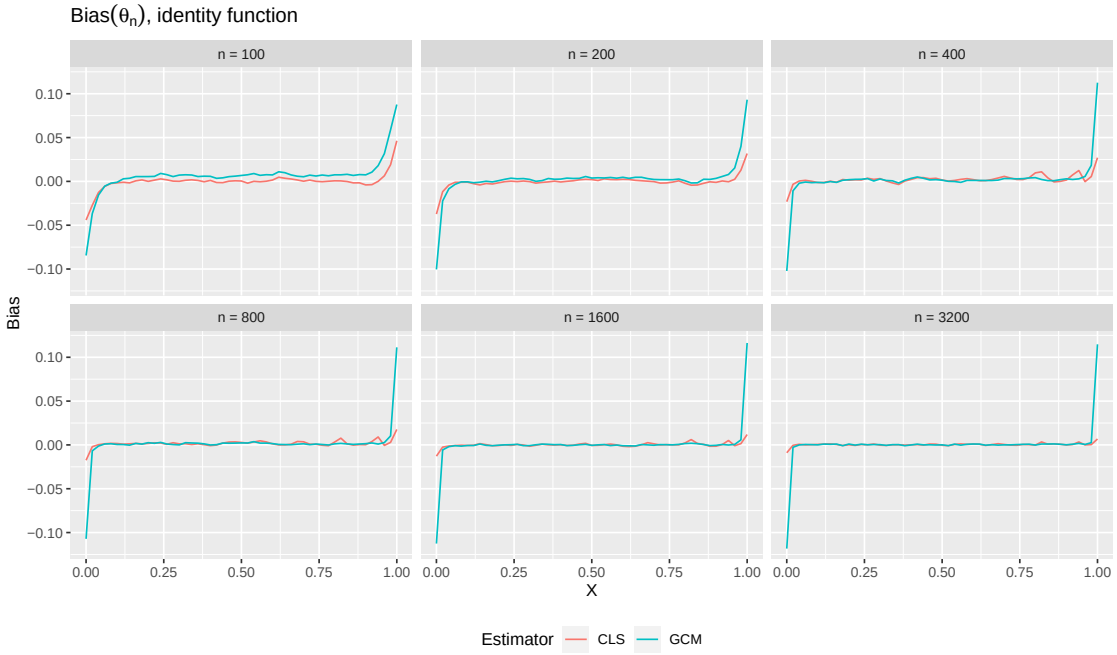


Figure 4.4: Bias of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (isotonic regression), displayed for six different sample sizes; X follows a standard uniform distribution.

estimator appearing to have a slightly lower standard deviation for lower sample sizes. However, towards the endpoints, the CLS-based estimator has far lower variance for nearly all sample sizes. Furthermore, the standard deviation of the CLS-based estimator evaluated at the endpoints appears to tending to zero at roughly a rate of $n^{1/3}$, whereas the standard deviation of the GCM-based estimator at the endpoints does not appear to be decreasing with sample size.

These simulations suggest that, at least in the case of classical isotonic regression, the CLS approach is highly advantageous in terms of reducing bias and variance at the endpoints, with minimal cost in terms of variance in the interior.

Next, we attempt to determine whether the estimators $\theta_n(x)$ and $\theta_n^*(x)$ have the same limiting distribution. As a first step, we can plot the estimated distributions of both $n^{1/3}\{\theta_n(x) - \theta_0(x)\}$ and $n^{1/3}\{\theta_n^*(x) - \theta_0(x)\}$, estimated using kernel density estimators, for fixed x. We display results for $x = 0.3$ in Figure 4.7, but plots for other interior points are qualitatively similar.

Figure 4.7 suggests that the two estimators approach a similar (but perhaps not identical)

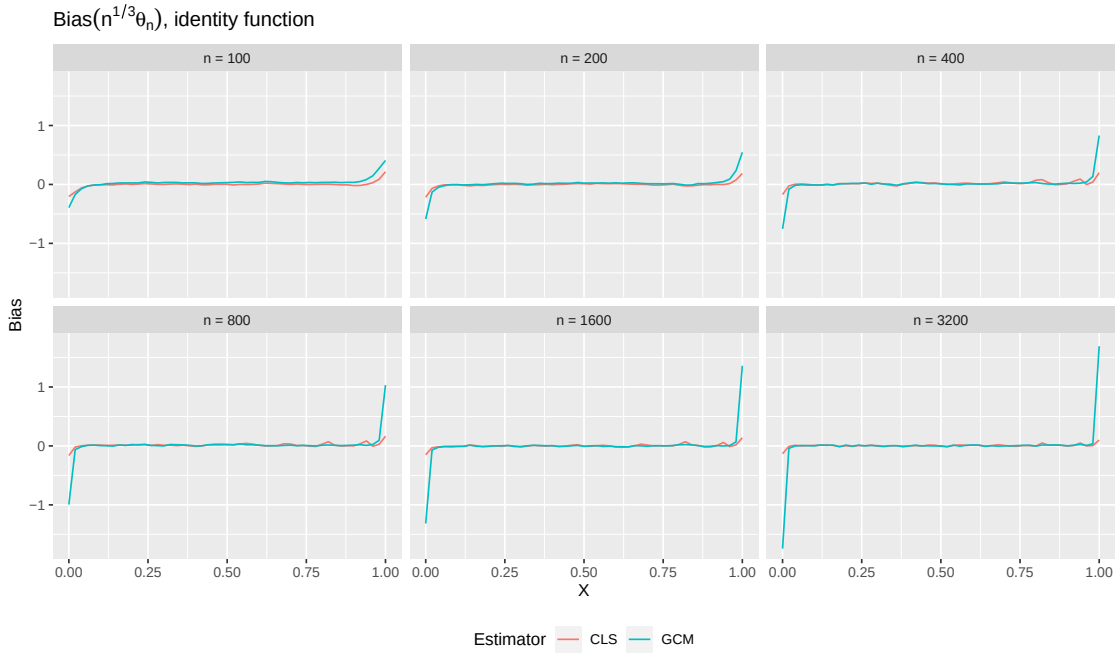


Figure 4.5: Bias of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (isotonic regression), displayed for six different sample sizes, scaled by $n^{1/3}$; X follows a standard uniform distribution.

limiting distribution. Also, for smaller sample sizes, the CLS-based estimator appears to have slightly greater variance, which is consistent with Figure 4.6.

Next, given the analysis in section 4.2, it is useful to compare the limiting behavior of the difference of the two primitives, $\Gamma_n(x) - \Gamma_n^*(x)$, to see whether this quantity tends to zero at root-n rate. Results are shown in Figure 4.8

Unfortunately, this figure suggests that the difference $\Gamma_n(x) - \Gamma_n^*(x)$ does not tend to zero at root-n rate, which implies that $\Gamma_n(x)$ and $\Gamma_n^*(x)$ have different influence functions. Therefore, the asymptotic analysis approach described in section 4.2 will not be possible.

At this stage, it is also useful to determine whether we expect the difference between the two primitive influence functions to result in a difference in the limiting distributions of the resulting estimator, as this is not necessarily the case (as described in section 4 of Westling and Carone, 2020). To assess this, we can plot the estimated density of the scaled difference $n^{1/3}\{\theta_n(x) - \theta_n^*(x)\}$ to see if it appears to tend to zero; results are given in Figure 4.9.

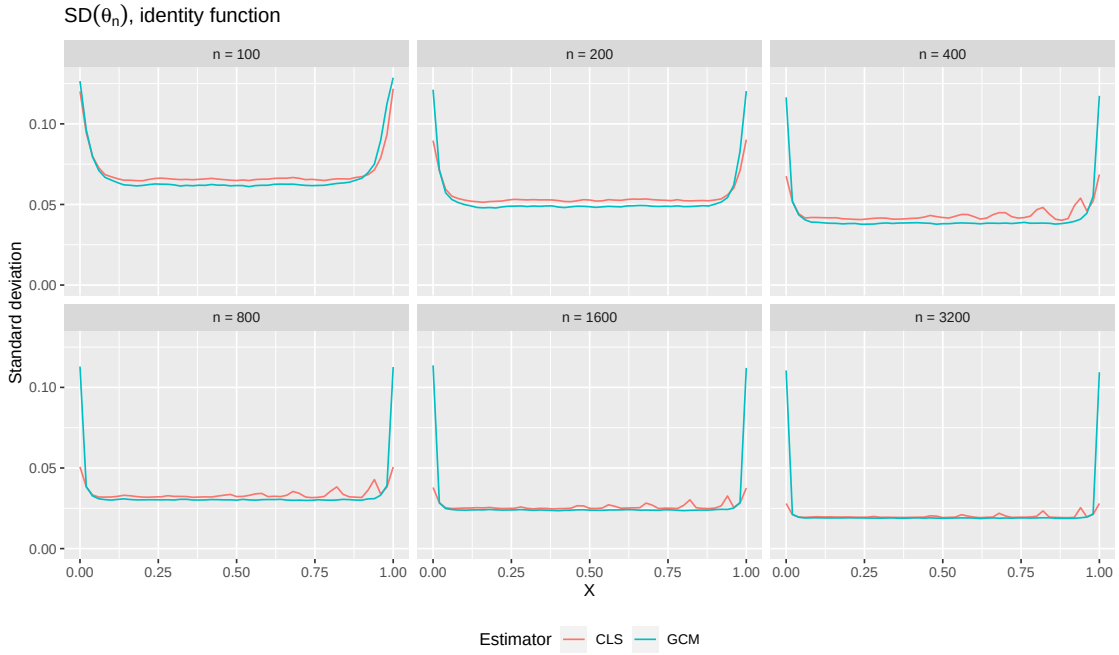


Figure 4.6: Bias of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (isotonic regression), displayed for six different sample sizes; X follows a standard uniform distribution.

Intriguingly, even though Figure 4.7 suggested that the two estimators $\theta_n(x)$ and $\theta_n^*(x)$ have similar limiting distributions, Figure 4.9 suggests that the difference between the two estimators does not disappear at a $n^{1/3}$ rate.

In conclusion, the CLS-based estimator appears to have considerable advantages relative to the GCM-based estimator in terms of mitigating the bias issues associated with the latter, but even though the two estimators seem to have similar limiting distributions, the difference between the estimators does not appear to disappear at a $n^{1/3}$ rate, and so further theoretical work is needed in order to establish precise asymptotic results.

We also studied the setting of monotone density estimation; here, the GCM-based estimator is equivalent to the classical Grenander estimator. Our simulation setup was simple; we sampled data values X from a Beta$(1, 5)$ distribution and compared the GCM-based estimator described in Westling and Carone (2020) to the corresponding CLS-based estimator. As in section 4.3, 10,000 simulation replicates were run for every level combination.

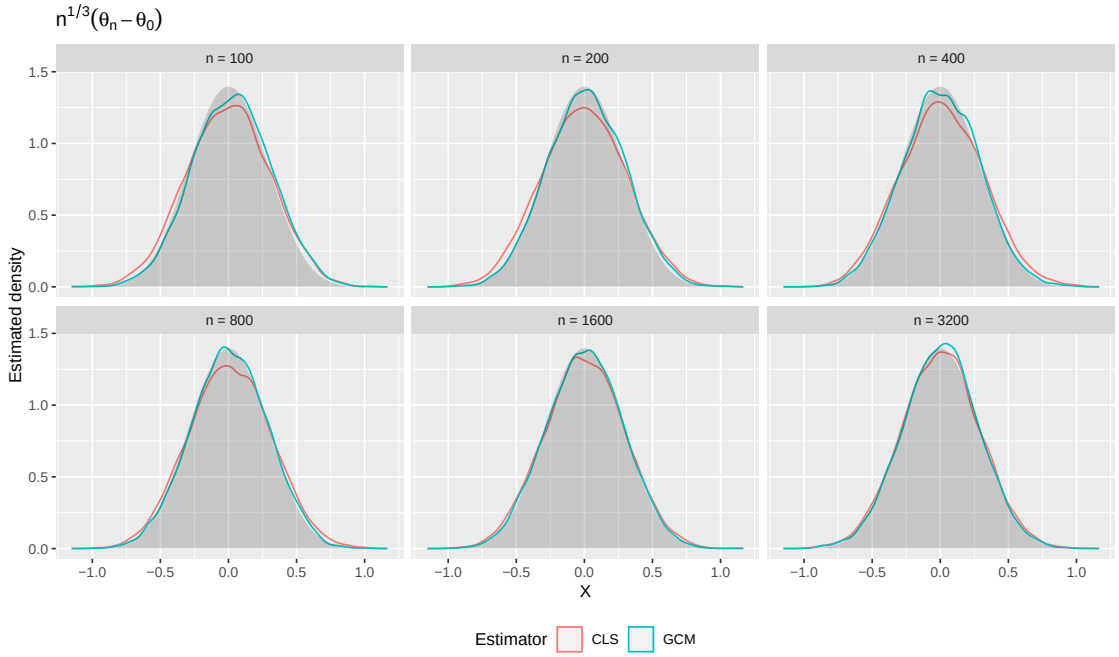


Figure 4.7: Kernel density estimates of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (isotonic regression), scaled by $n^{1/3}$ and centered, displayed for six different sample sizes; X follows a standard uniform distribution. The grey plot in the background represents the true (scaled Chernoff) asymptotic distribution.

Figure 4.10 shows realizations of both estimators plotted against one another to get a sense of their distributions and correlation. As in the regression case, the two estimators were highly correlated; compare to Figure 4.3.

Figure 4.11 displays the bias of both estimators across different values of the unit interval. Density was sometimes massively overestimated at zero, and so we excluded bias estimates at zero from this plot due to the instability of the estimates; the next largest point considered was 0.02. The pattern of bias observed is different than in the regression setting, with overestimates towards both endpoints. As expected, bias in the interior tended towards zero as sample size increased. The CLS-based estimator appeared to perform better for values close to zero, but had slightly higher bias towards one. Compare to Figure 4.4.

In Figure 4.12, we show the same results as in Figure 4.11, but scaled by $n^{1/3}$. Compare with Figure 4.5.

In Figure 4.13, the standard deviation of both estimators is displayed across different values

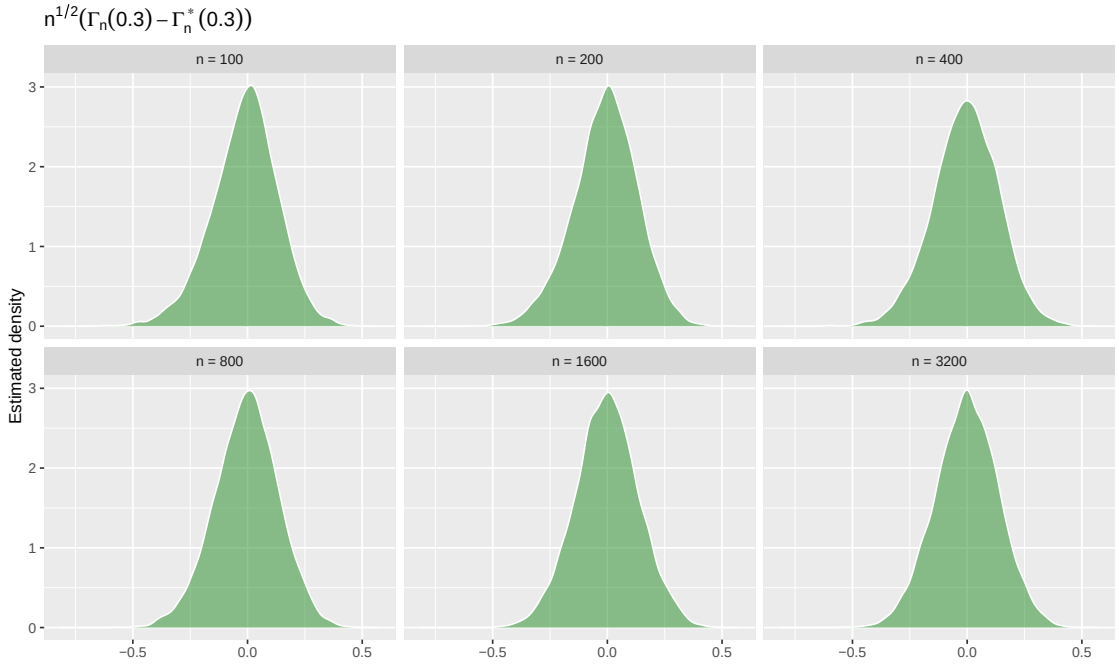


Figure 4.8: Kernel density estimates of the difference $\Gamma_n(x) - \Gamma_n^*(x)$, scaled by $n^{1/2}$, displayed for six different sample sizes; X follows a standard uniform distribution.

of the unit interval. Towards the left-hand side of the unit interval, the CLS-based estimator had slightly higher standard deviation across all sample sizes considered. Compare to Figure 4.6.

In Figure 4.14, we plot the estimated distributions of both $n^{1/3}\{\theta_n(x) - \theta_0(x)\}$ and $n^{1/3}\{\theta_n^*(x) - \theta_0(x)\}$, with results displayed for $x = 0.3$. Compare to Figure 4.7.

Figure 4.15 depicts the limiting behavior of the scaled difference of the two primitives $\Gamma_n(x) - \Gamma_n^*(x)$. As in the regression case, the difference does not seem to tend to zero when scaled by $n^{1/2}$. Compare to Figure 4.8.

In Figure 4.16, we plot the estimated density of the scaled difference $n^{1/3}\{\theta_n(x) - \theta_n^*(x)\}$. As in the regression case, the difference does not seem to tend to zero when scaled by $n^{1/3}$. Compare to Figure 4.9.

All in all, the results for the density case are qualitatively similar to those of the regression case, other than the different patterns of bias and variance.

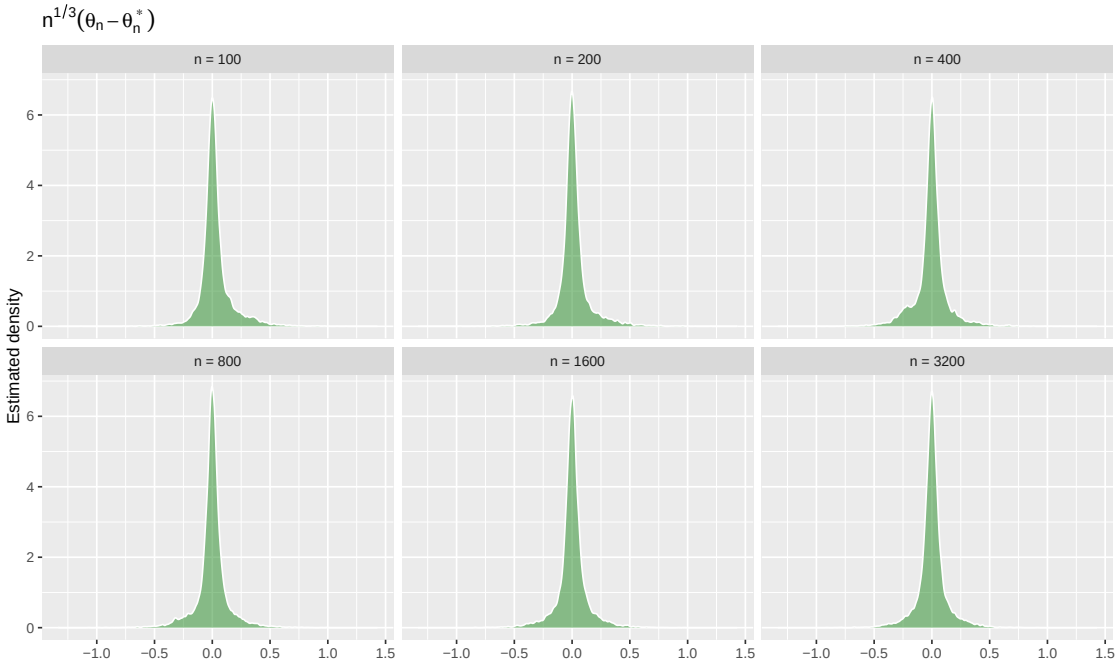


Figure 4.9: Kernel density estimates of the difference $\theta_n(x) - \theta_n^*(x)$, scaled by $n^{1/3}$, displayed for six different sample sizes; X follows a standard uniform distribution.

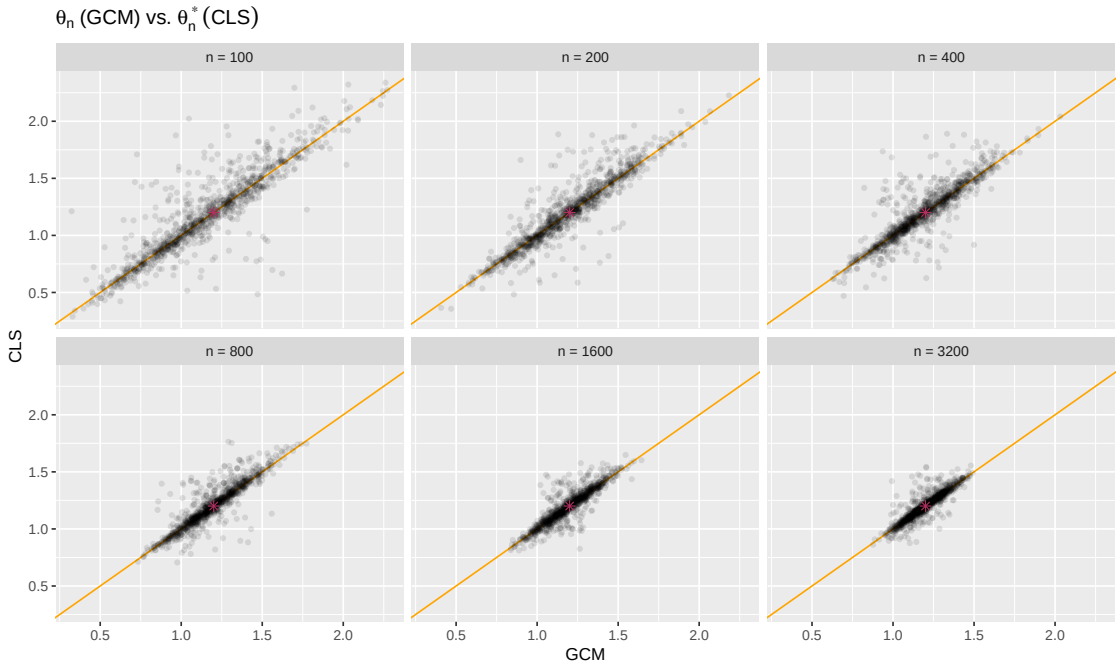


Figure 4.10: Realizations of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (Grenander estimator) evaluated at $x = 0.3$, displayed for 1,000 simulation replicates and six different sample sizes. The orange line is the identity function, and the red star at $(1.2, 1.2)$ represents the true density value.

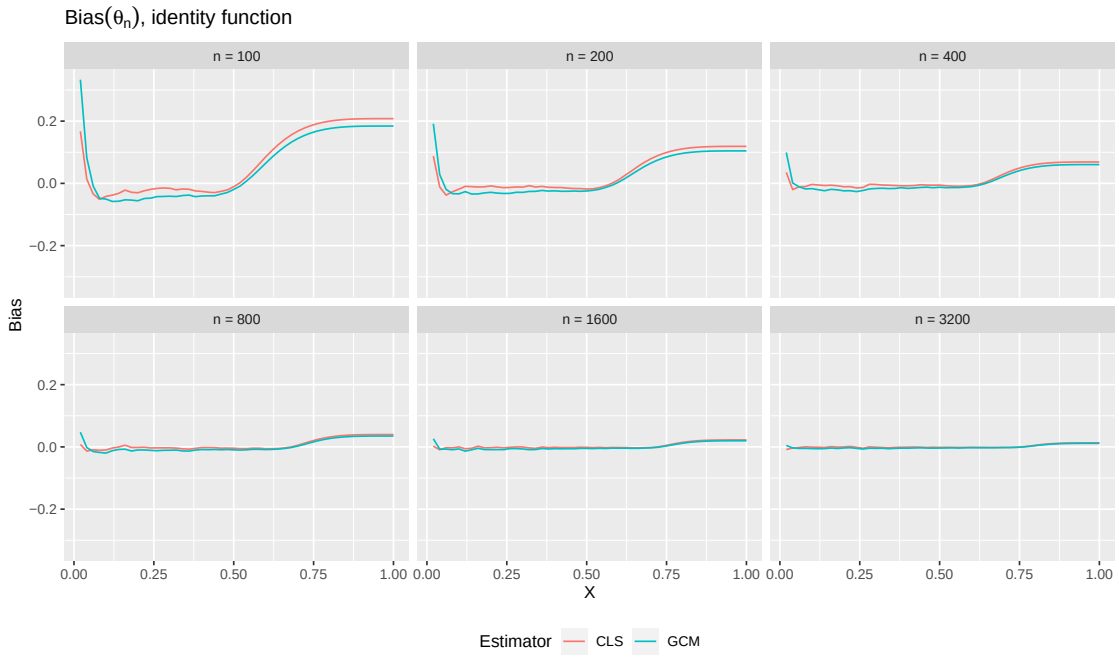


Figure 4.11: Bias of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (Grenander estimator), displayed for six different sample sizes.

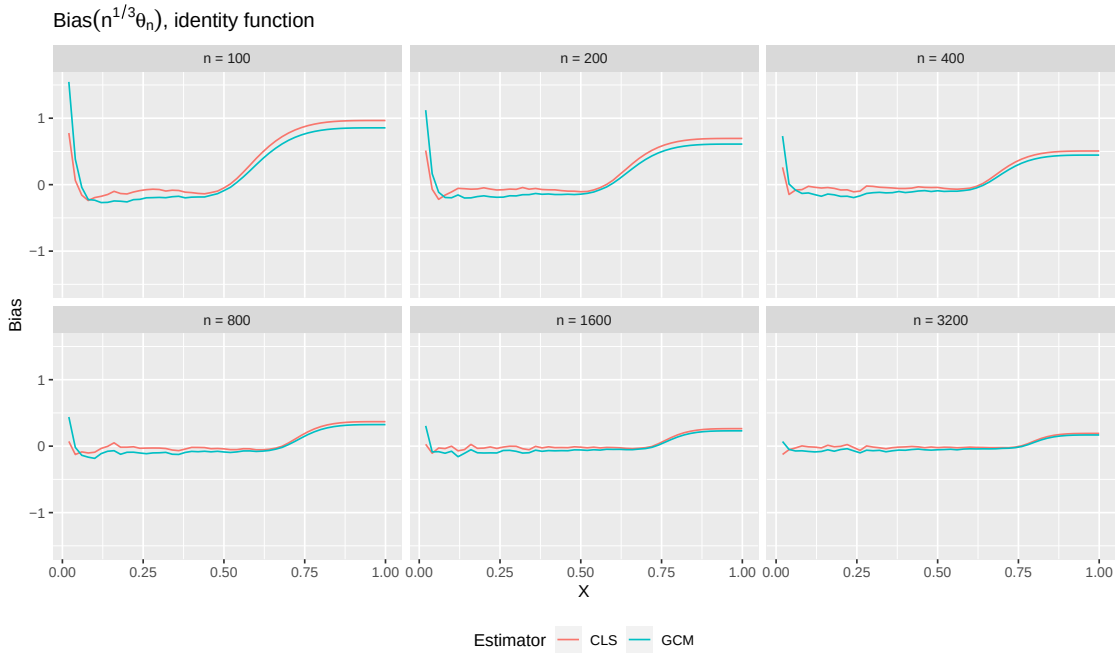


Figure 4.12: Bias of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (Grenander estimator), displayed for six different sample sizes, scaled by $n^{1/3}$.

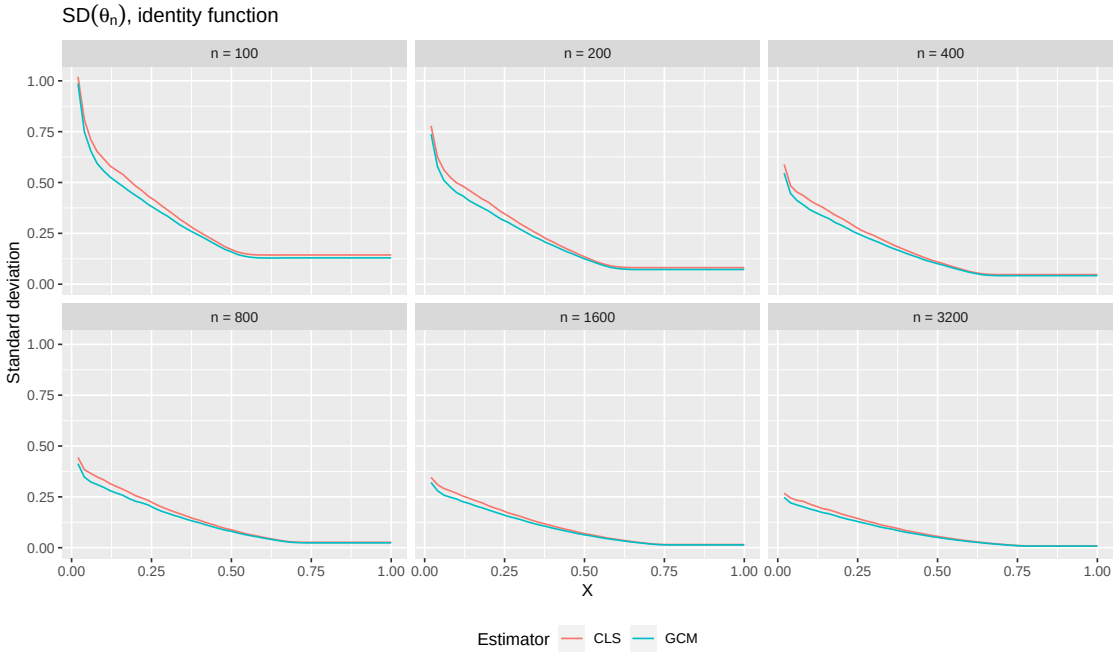


Figure 4.13: Bias of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (Grenander estimator), displayed for six different sample sizes.

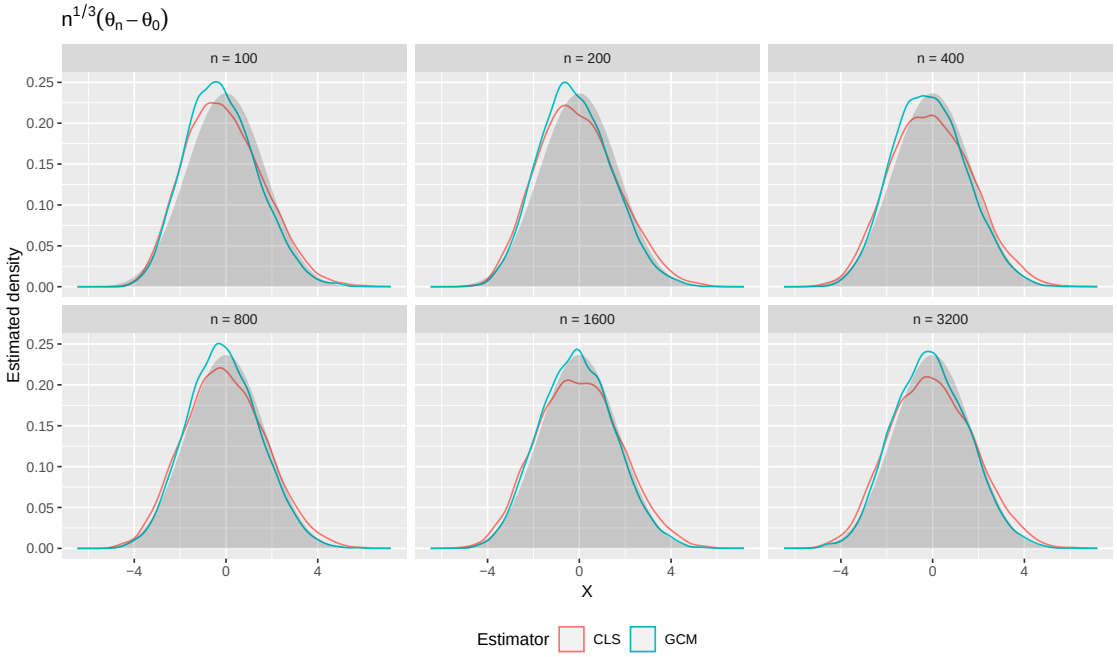


Figure 4.14: Kernel density estimates of two estimators ("CLS" = convex least squares, "GCM" = greatest convex minorant (Greanander estimator), scaled by $n^{1/3}$ and centered, displayed for six different sample sizes. The grey plot in the background represents the true (scaled Chernoff) asymptotic distribution.

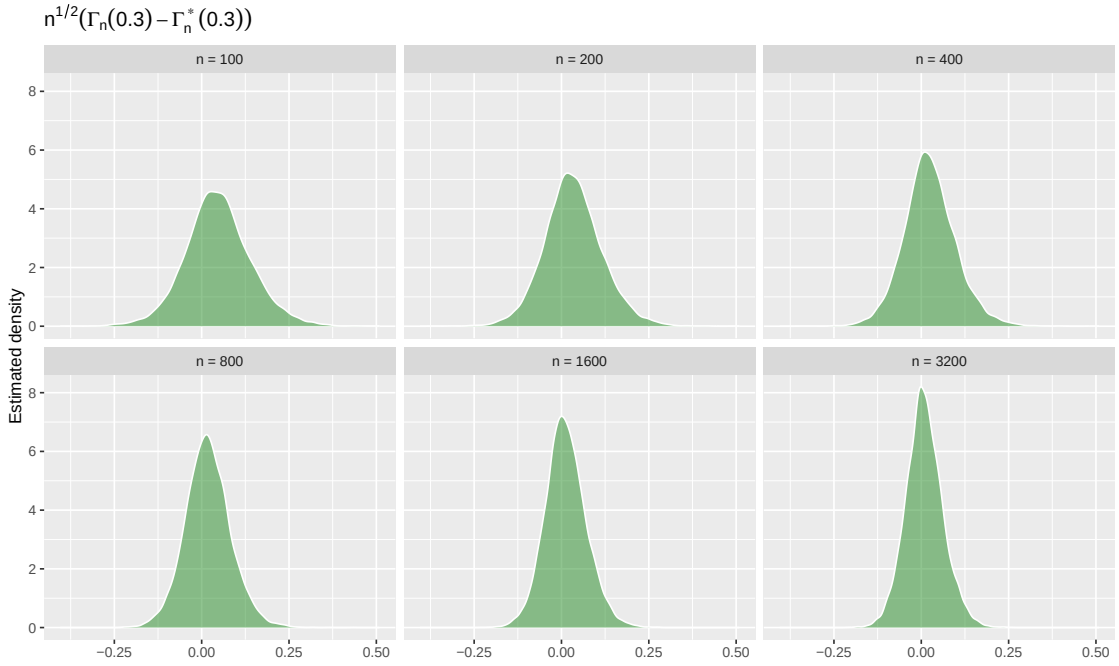


Figure 4.15: Kernel density estimates of the difference $\Gamma_n(x) - \Gamma_n^*(x)$, scaled by $n^{1/2}$, displayed for six different sample sizes.

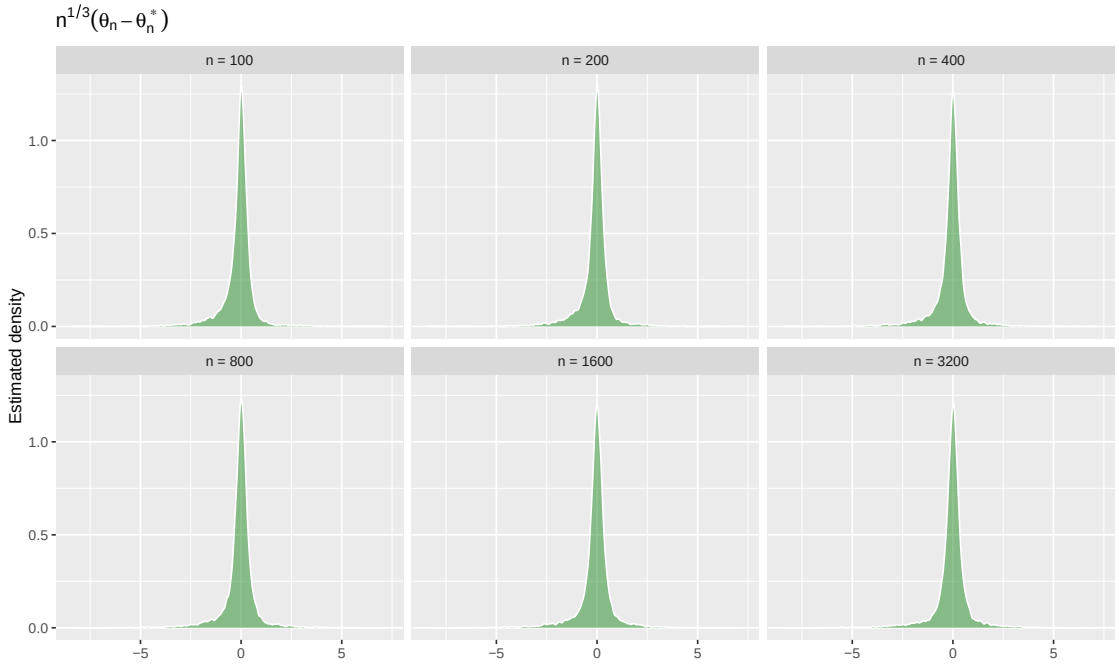


Figure 4.16: Kernel density estimates of the difference $\theta_n(x) - \theta_n^*(x)$, scaled by $n^{1/3}$, displayed for six different sample sizes.

Chapter 5

Implementation of methods in the `vaccine` R package

Developing new statistical methods is of limited value if researchers do not have the ability to use them in practice. As such, we developed an R package called `vaccine` to implement the methods described in Chapters 2 and 3. The package is available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/vaccine/index.html, the standard repository for R packages. The source code of the R package can be found on GitHub at https://github.com/Avi-Kenny/vaccine. At the time of writing, version 0.1.0 is the latest stable CRAN version. In this chapter, we briefly review the interface and functionality of the package.

The latest stable version of `vaccine` can be installed from CRAN using `install.packages()`. The current development version can be installed using `devtools::install_github()`.

```
install.packages("vaccine")
devtools::install_github(repo="Avi-Kenny/vaccine")
```

To illustrate the basic workflow of the package, we will leverage publicly-available data from the HVTN 505 vaccine efficacy trial, which assessed the safety and effectiveness of a DNA prime-recombinant adenovirus type 5 boost vaccine regimen in the United States Hammer et al. (2013). The data is available at https://atlas.scharp.org/cpas/project/\HVTN%20Public%20Data/HVTN%20505/begin.view. The data is bundled into the package and can be accessed using the

command `data(hvtn505)`.

Data is loaded into the package using the `load_data` function, which takes in raw data and creates a data object that can be accepted by various estimation functions. This function will perform input validation, store metadata needed by other package functions, and allow for unified documentation about expected variables and data types. This function can be used as follows.

```
data(hvtn505)

dat <- load_data(
  time = "HIVwk28preunblfu",
  event = "HIVwk28preunbl",
  vacc = "trt",
  marker = "IgG_env",
  covariates = c("age","BMI","bhvrisk"),
  weights = "wt",
  ph2 = "casecontrol",
  data = hvtn505
)
```

The `summary_stats` function, as the name suggests, calculates select summary statistics related to the dataset. We may choose to expand this function in the future to include other statistics.

```
summary_stats(dat)
#> Number of subjects (vaccine group, phase-1): 1161
#> Number of subjects (placebo group, phase-1): 1141
#> Number of subjects (vaccine group, phase-2): 150
#> Number of subjects (placebo group, phase-2): 39
#> Number of events (vaccine group, phase-1): 27
#> Number of events (placebo group, phase-1): 21
#> Number of events (vaccine group, phase-2): 25
#> Number of events (placebo group, phase-2): 19
#> Proportion of subjects with an event (vaccine group, phase- 1): 0.02326
#> Proportion of subjects with an event (placebo group, phase- 1): 0.0184
#> Proportion of subjects with an event (vaccine group, phase- 2): 0.16667
#> Proportion of subjects with an event (placebo group, phase- 2): 0.48718
```

The `est_overall` function allows us to estimate overall risk in the placebo and vaccine groups, as well as estimate vaccine efficacy, using either a nonparametric Kaplan-Meier estimator or a marginalized Cox model.

```
est_overall(dat=dat, t_0=578, method="KM")
#>   stat   group          est          se    ci_lower   ci_upper
#> 1 risk vaccine  0.04067009 0.008230842  0.02506853 0.05602199
#> 2 risk placebo  0.02879861 0.006563785  0.01622360 0.04121288
#> 3   ve    both -0.41222411 0.430451788 -1.56659984 0.22294979

est_overall(dat=dat, t_0=578, method="Cox")
#>   stat   group          est          se    ci_lower   ci_upper
#> 1 risk vaccine  0.04177642 0.008111679  0.02847302 0.06090588
#> 2 risk placebo  0.02938706 0.006486545  0.01901930 0.04514638
#> 3   ve    both -0.42159246 0.417915188 -1.52937491 0.20101796
```

The `est_ce` function is the main workhorse of the package. This function allows for calculation of controlled vaccine efficacy (CVE) curves and controlled risk (CR) curves using the Cox-model based methods described in chapter 2 and the nonparametric methods described in chapter 3. Three arguments are required; `dat` is a data object returned by the `load_data` function, `type` controls which method (Cox or nonparametric) is used, and `t_0` is used to specify t_0, the time of interest. Basic usage is as follows:

```
ests_cox <- est_ce(dat=dat, type="Cox", t_0=578)
ests_np <- est_ce(dat=dat, type="NP", t_0=578)
```

Beyond this, a number of options are available to customize the behavior of each estimator. For the Cox model estimator, spline terms and/or an indicator function corresponding to the LLOD can be included in the linear predictor (as described in section 2.2.2) through the `params_cox` argument. For the nonparametric estimator, many options are available through the `params_np` argument, including the direction of monotonicity (nondecreasing vs. nonincreasing), whether or not the edge correction (described in section 3.1.7) is to be performed, a grid size parameter that controls the accuracy of several numerical approximations, the type of conditional survival function (and conditional censoring function) estimator for the nuisance Q_n (and Q_n^C), the type of

conditional density function estimator for the nuisance $f_n^{S|X}$, and the type of derivative estimator for the function $\dot{r}_{M,0}$. Additionally, if the CVE curve is being estimated (in addition to the CR curve), both estimators allow for either a Kaplan-Meier estimator or a marginalized Cox model estimator to be used to estimate overall risk in the placebo group. For example, a Cox model with a spline term involving four degrees of freedom can be specified as:

```
ests_cox <- est_ce(
  dat = dat,
  type = "Cox",
  t_0 = 578,
  params_cox = params_ce_cox(spline_df=4)
)
```

As an additional example, a nonparametric estimator using the edge correction described in section 3.1.7 and the global survival stacking method of Wolock et al. (2022) to estimate the conditional survival and censoring functions can be specified as:

```
ests_np <- est_ce(
  dat = dat,
  type = "NP",
  t_0 = 578,
  params_np = params_ce_np(
    edge_corr = TRUE,
    surv_type = "survML-G"
  )
)
```

Some basic graphing functionality is provided in the package to plot CR and CVE curves via the `plot_ce` function as follows, which creates the plot given in Figure 5.1:

```
plot_ce(ests_cox, ests_np)
```

If CR or CVE curves are estimated using the nonparametric method, it is sometimes useful to examine some of the underlying nuisance estimators; this can be done with the `diagnostics` function as follows, which creates the composite plot given in Figure 5.2:

```
diagnostics(ests_np)
```

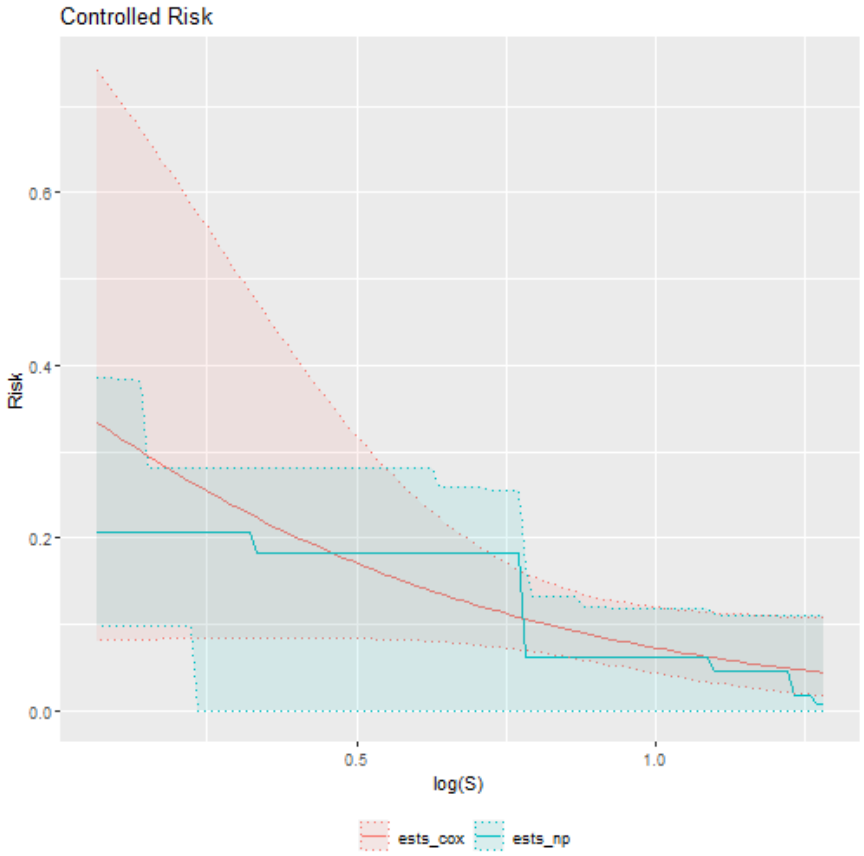


Figure 5.1: Controlled risk curves illustrating the use of the `plot_ce` function of the `vaccine` package.

The current version of the package implements many of the methods described in this dissertation, but there is more work to be done. Major changes and additions include implementing the hypothesis testing methods described in section 3.2, improving the graphing functionality, increasing the number of estimation parameters controlled by the user, and building out a unit testing and continuous integration framework to ensure continued code functionality.

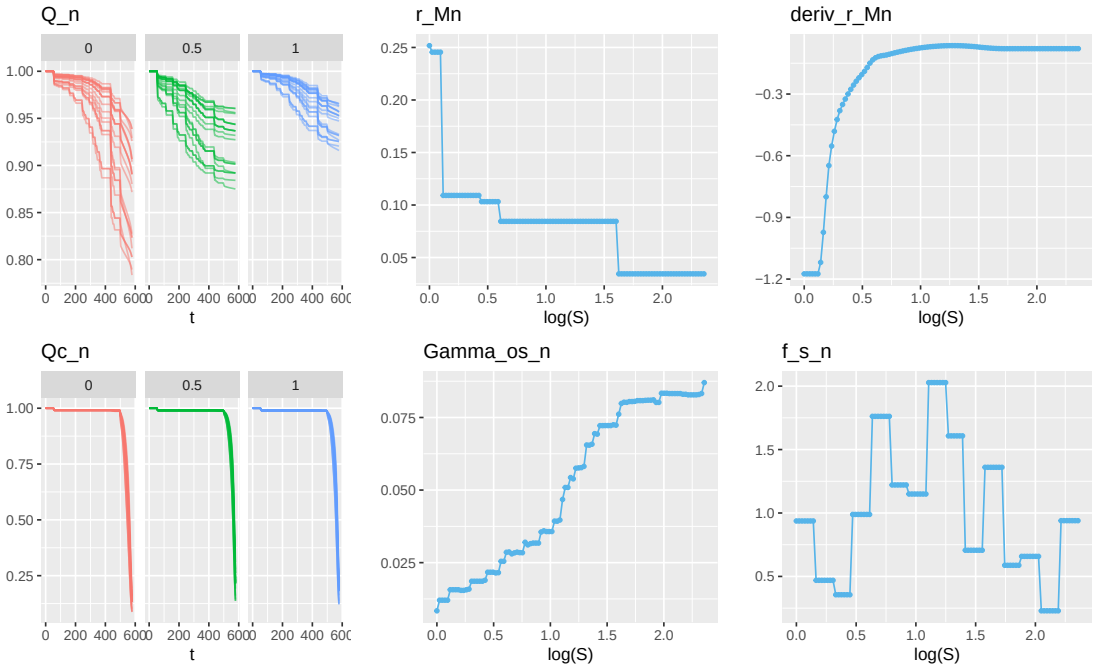


Figure 5.2: Diagnostics plot illustrating the use of the `diagnostics` function of the `vaccine` package.

Chapter 6

Applications of the methods to several vaccine efficacy trials

In sections 2.4 and 3.4, we displayed data analysis results for the Immune correlates analysis of the Coronavirus Efficacy (COVE) trial of the mRNA-1273 COVID-19 vaccine. Over the past two years, we have applied the methods described in Chapters 2 and 3 to data from six phase-3 vaccine efficacy trials. In this chapter, we present select results from several of these trials to illustrate the versatility of these methods, as well as to point out several interesting challenges that arose.

To begin, we give a few key details of each analysis in Table 6.1. Beyond this, we do not focus on the technical details related to the trials; for further information, see Benkeser et al. (2023) (COVE), Fong et al. (2022) (ENSEMBLE), Kenny et al. (2022) (HVTN 705), Seaton et al. (2023) (AMP), Palacios et al. (2020) (PROFISCOV), and Haynes et al. (2012) (RV 144).

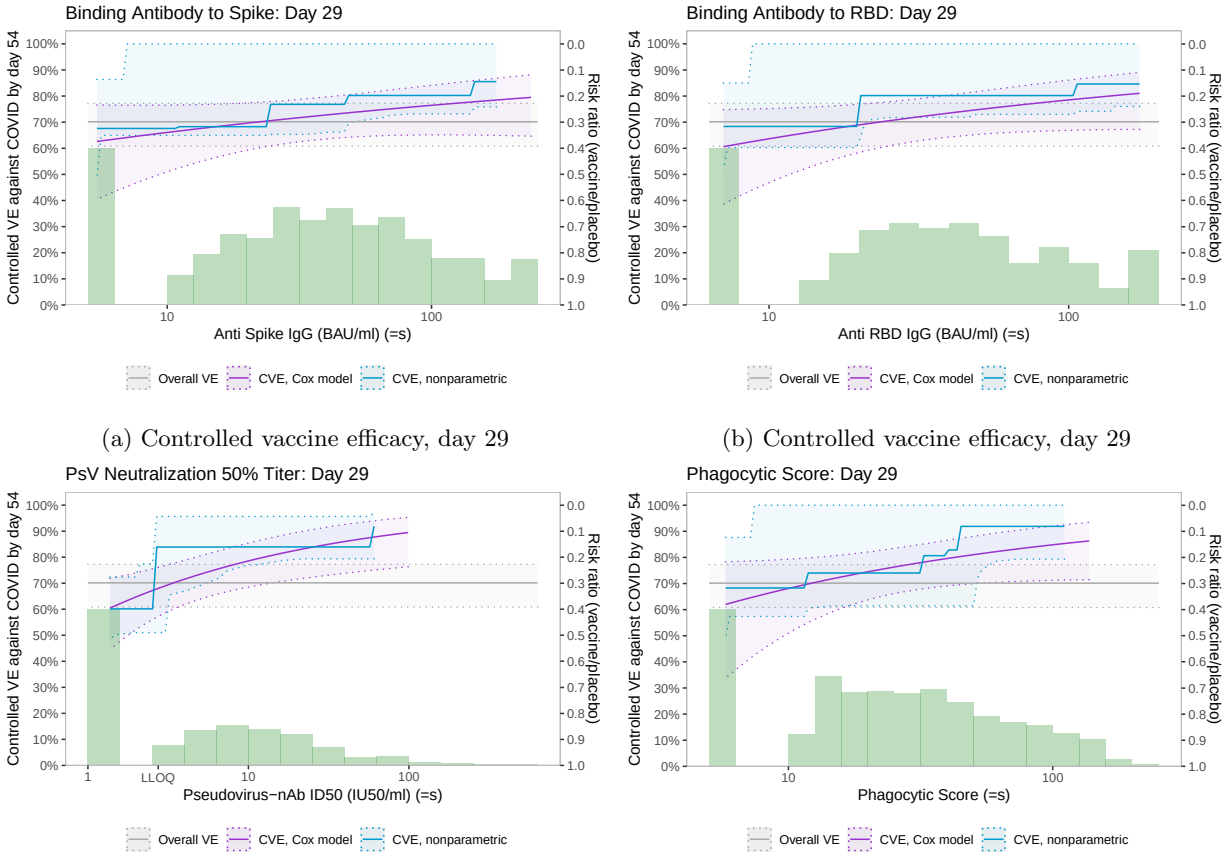
Trial name	Endpoint	Primary markers	Number of markers	Marker time points
COVE	COVID disease	Binding Antibody to Spike, Binding Antibody to RBD, PsV Neutralization 50% Titer, PsV Neutralization 80% Titer, Live Virus Micro Neut 50%	5	Day 29, Day 57
ENSEMBLE	COVID disease	Binding Antibody to Spike, Binding Antibody to RBD, PsV Neutralization 50% Titer, PsV Neutralization 80% Titer	4	Day 29
HVTN 705	HIV-1 infection	ELISA VT-C, ELISpot Env, ADCP Vx-C97ZA, BAMA IgG3 V1V2 breadth score, BAMA IgG3 Env breadth, Expanded Multi-epitope functions	41	Day 210
AMP	HIV-1 infection	Body Weight (kg)	1	Day 0
PROFISCOV	COVID disease	Binding Antibody to Spike, Binding Antibody to RBD, Binding Antibody to Nucleocapsid, Live Virus Micro Neut 50%	9	Day 43, Day 91
RV 144	HIV-1 infection	IgG3 Net MFI to AE.A244 V1V2 Tags 293F, IgG3 Net MFI to C.1086C V1V2 Tags, IgG3 Net MFI to gp70-Ce1086 B2 V1V2, IgG3 Net MFI to gp70-B.CaseAV1V2, IgA A1.con.env03 140 CF	5	Day 182

Table 6.1: Summary of phase-three vaccine efficacy trials for which correlates analysis was conducted.

6.1 ENSEMBLE

The ENSEMBLE clinical trial assessed the safety and efficacy of a single dose Ad26.COV2.S vaccine to prevent SARS-CoV-2-Mediated COVID-19 disease. In the primary immune correlates analysis of the ENSEMBLE trial, four primary immune markers were measured; in Figure 6.1, we display controlled vaccine efficacy curves for these four markers, all measured at day 29.

Figure 6.1 illustrates several features of the nonparametric estimation process. First, it was an ideal application for the edge-corrected estimator described in section 3.1.7, as all four markers had



(a) Controlled vaccine efficacy, day 29

(b) Controlled vaccine efficacy, day 29

(c) Controlled vaccine efficacy, day 57

(d) Controlled vaccine efficacy, day 57

Figure 6.1: Controlled vaccine efficacy (CVE) curves for four markers, including Binding Antibody to Spike, Binding Antibody to RBD, PsV Neutralization 50% Titer, and Phagocytic Score, measured at day 29. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey lines represent overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

significant mass at the lower limit of detection. These plots illustrate a key difference between the Cox model approach and the nonparametric approach; the former is constrained to always give a smooth log-linear curve, whereas the latter can take on a wider variety of shapes and can detect so-called threshold effects. In Subfigure 6.1c, we see a large jump in estimated CVE immediately after the limit of detection, whereas in the other three subfigures, we see estimated curves that follow the smooth Cox model curve more closely.

Next, we show select plots from the ESNEMBLE analysis that illustrate key challenges with the nonparametric approach. These plots come from an analysis in which estimates were computed for

the same set of markers, but for subgroups of the trial population defined by age and geography. Because of the low levels of disaggregation, the sample sizes for many of these plots were much smaller.

In Figure 6.2, we show data for the PsV Neutralization 50% Titer marker in the subset of the trial population defined as senior citizens based on their age at baseline. This plot illustrates one challenge with the nonparametric approach, which is that the CI bands sometimes appear too narrow for certain biomarker values (in the sense that they are sometimes more narrow than the Cox-based CI bands). In particular, this tends to happen when the estimated CVE curve has sections that are relatively flat; this is because the Chernoff scale factor (which determines the asymptotic variance) depends on the estimated derivative of the target function.

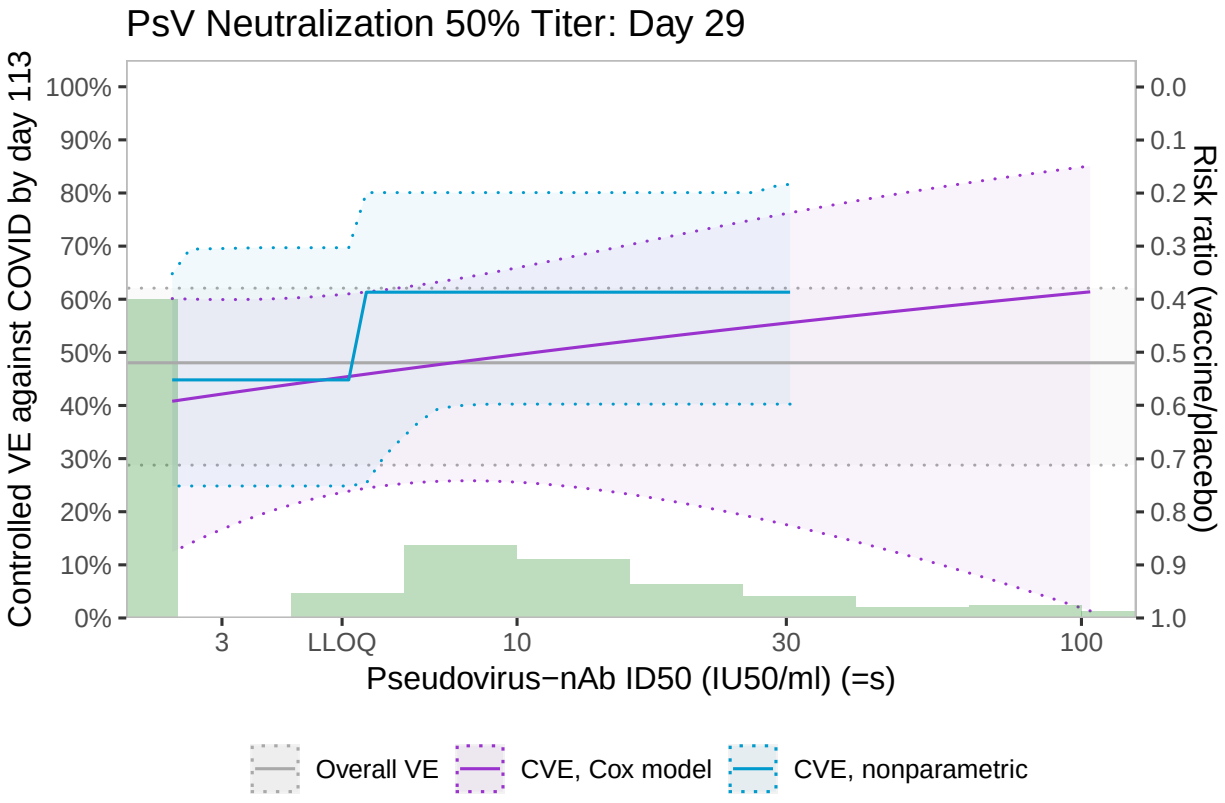


Figure 6.2: Controlled vaccine efficacy (CVE) curves for the 50% inhibitory dilution (ID50) nAb titer marker, measured at day 29, restricted to seniors. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

In Figure 6.3, we show data for the PsV Neutralization 50% Titer marker in the North America subset of the trial population. This illustrates on issue with the edge estimator, which is that it sometimes leads to estimated curves that are entirely flat. The reason this sometimes happens is that the edge estimator may have high variance if there is a relatively low proportion of volunteers who had a marker level below the LLOD. Then, the edge estimator value may lie entirely above the CVE curve estimated from the Grenander estimator. This is a limitation of the edge-corrected estimator.

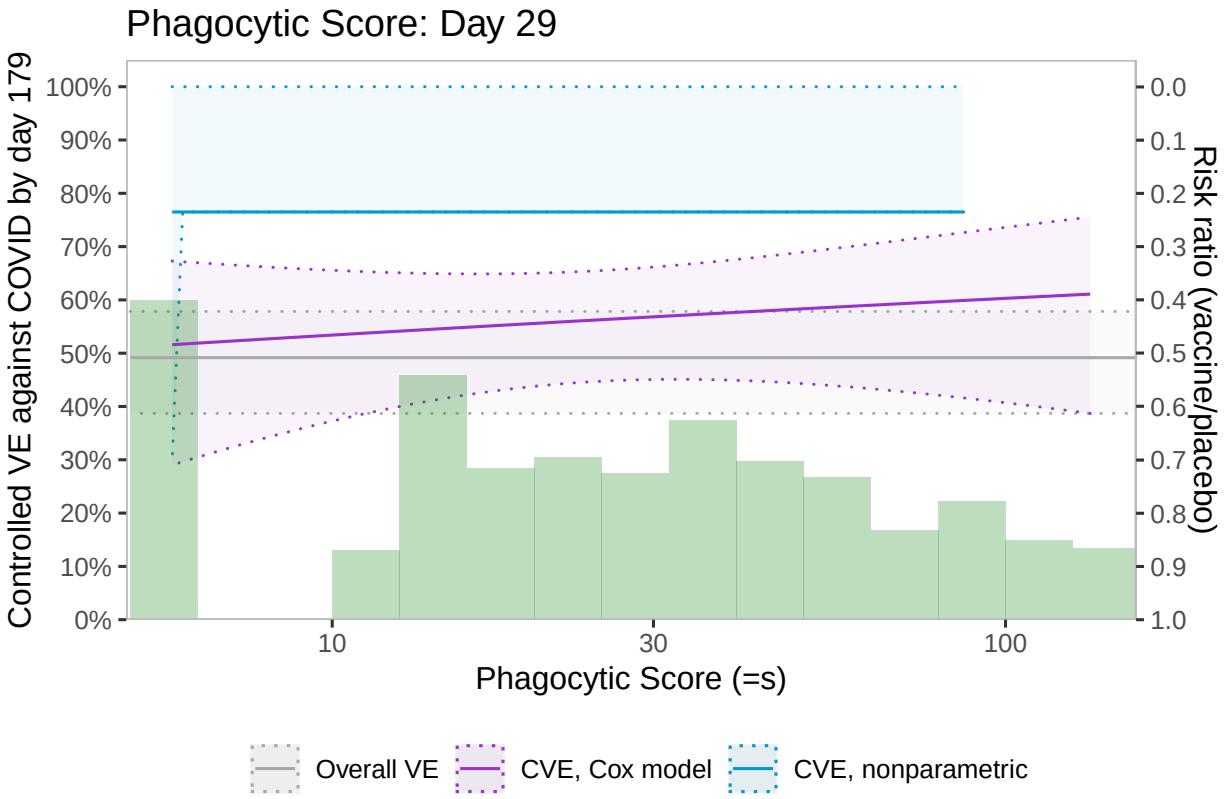


Figure 6.3: Controlled vaccine efficacy (CVE) curves for the 50% inhibitory dilution (ID50) nAb titer marker, measured at day 29, restricted to North Americans. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

In Figure 6.4, we show data for the Phagocytosis Score marker in the North American seniors subset of the trial population. The issue here is that the overall change in the CVE curve appears

to be overestimated, particularly towards the right-hand side of the curve. While this result is not impossible, it is implausible, and likely happens due to the bias issue that is discussed further in Chapter 4. This issue can also theoretically happen if highly flexible estimators (e.g., tree-based methods) are used to estimate the conditional survival function, and in practice, we have used a fairly restricted library for estimating the conditional survival function due to the increased stability of the resulting estimates.

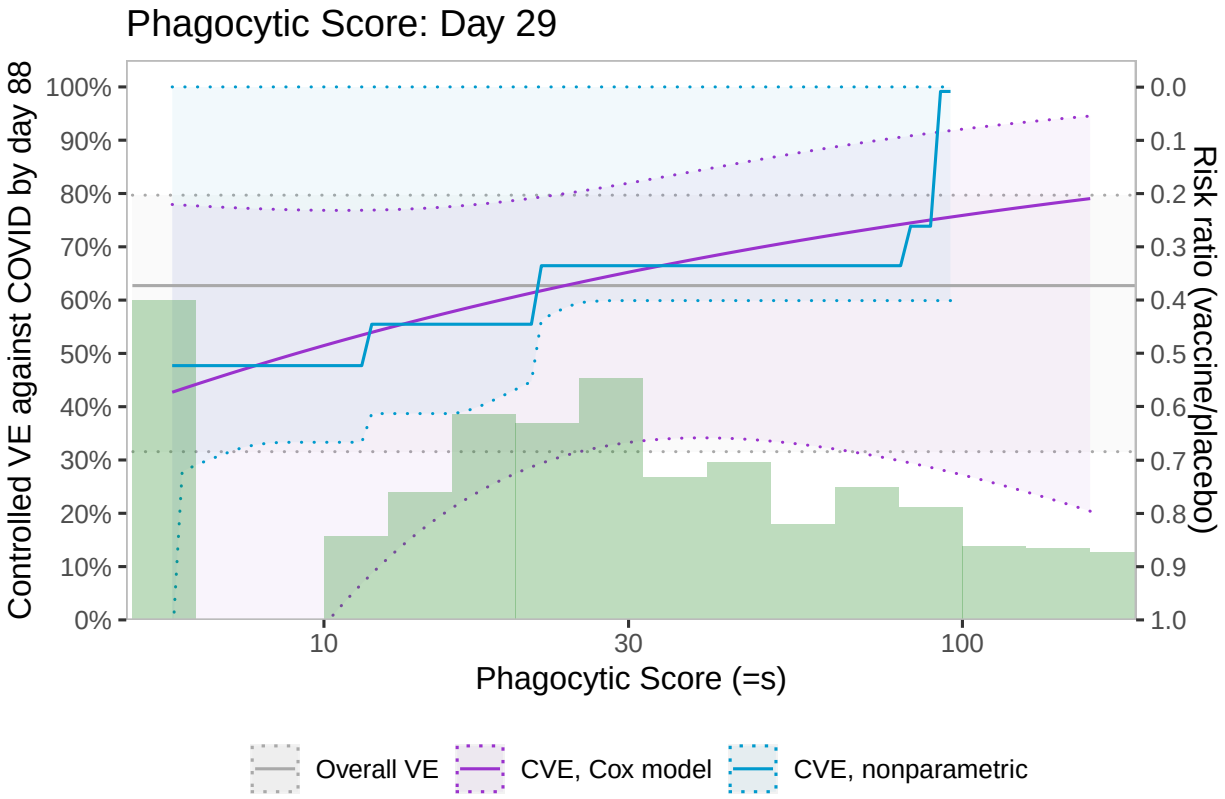


Figure 6.4: Controlled vaccine efficacy (CVE) curves for the phagocytosis score marker, measured at day 29, restricted to North American seniors. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

In Figure 6.5, we show data for the PsV Neutralization 50% Titer marker in the North American nonseniors subset of the trial population. In this plot, we see that confidence bands are not displayed when CVE is estimated to be equal to 100% for certain marker values. The reason this happens is

simply that log-transformed confidence limits are undefined when the relative risk estimate equals to zero. Because of this, it may be worth considering a finite-sample fix, such as setting CVE estimates to 99% if they are estimated above this value.

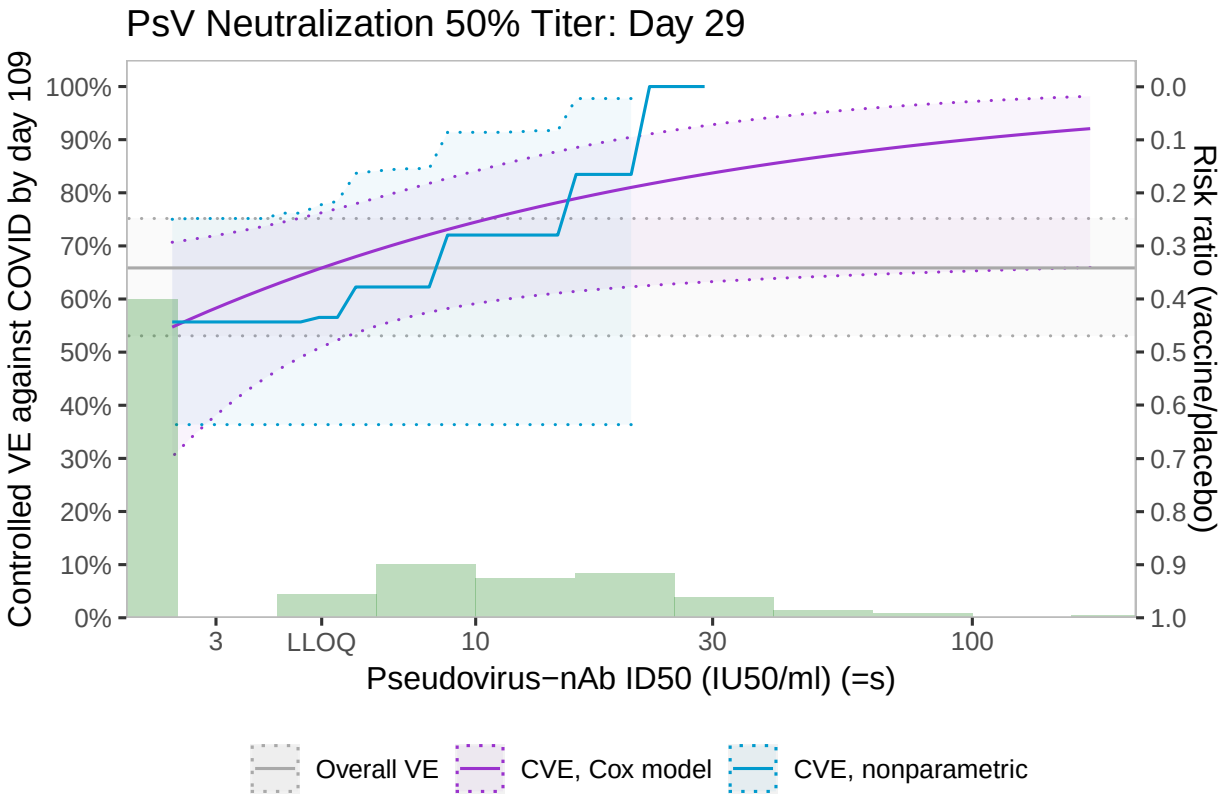


Figure 6.5: Controlled vaccine efficacy (CVE) curves for the 50% inhibitory dilution (ID50) nAb titer marker, measured at day 29, restricted to North American non-seniors. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

6.2 HVTN 705

In this section, we display a few representative graphs from the immune correlates of the HVTN 705 clinical trial. This trial assessed the safety and efficacy of a heterologous prime/boost regimen utilizing Ad26.Mos4.HIV and aluminum-phosphate adjuvanted Clade C gp140 against an endpoint of HIV-1 infection. In Figure 6.6, we display results for the IgG3 Net MFI to gp70-001428.2.42

V1V2 marker, measured at day 210. In this plot, both the Cox model curve and the nonparametric curve lie roughly on top of one another. The fact that both estimators result in similar curves gives us greater confidence that the trend is real, as opposed to an artefact of one particular method.

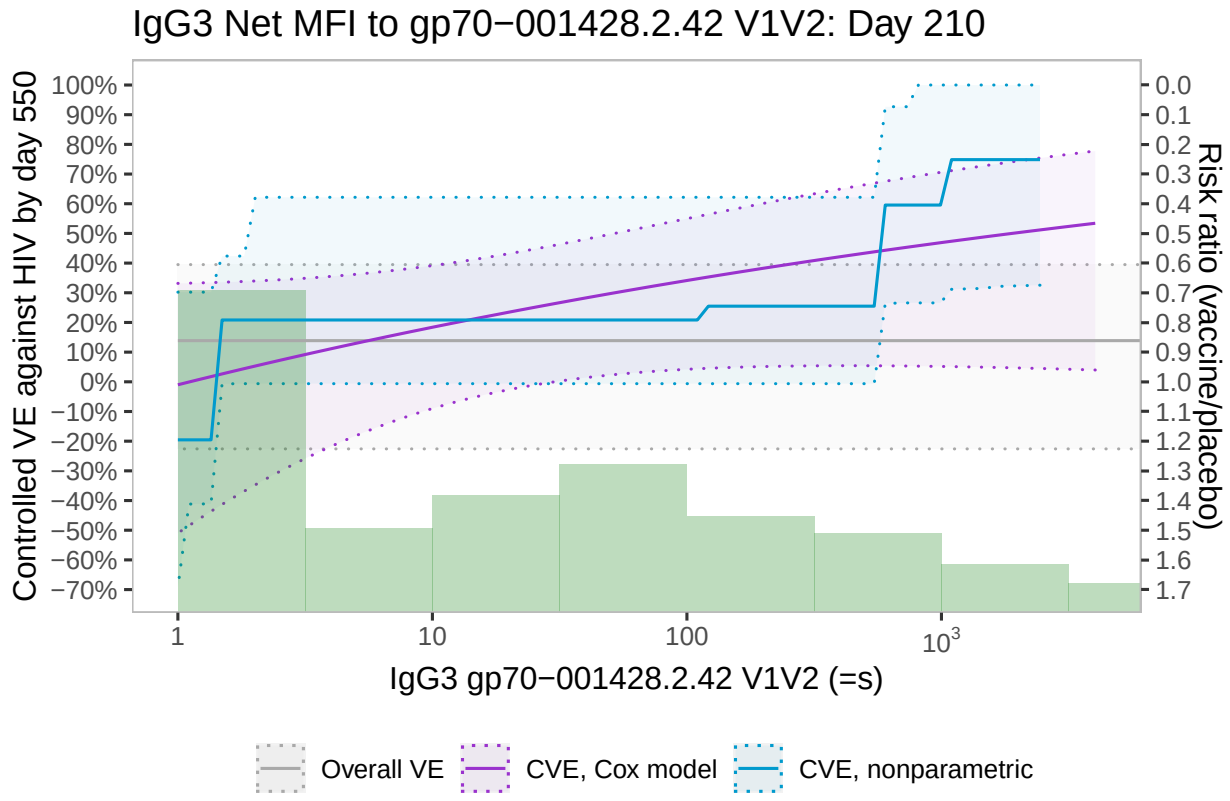


Figure 6.6: Controlled vaccine efficacy (CVE) curves for the IgG3 Net MFI to gp70-001428.2.42 V1V2 marker, measured at day 210. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

In Figure 6.7, we display results for the IgG3 V1V2 breadth (Weighted avg log10 Net MFI) marker, measured at day 210. This was one of the primary markers studied in the HVTN 705 immune correlates analysis, in part because it was found to be a correlate of protection in previous HIV vaccine studies, including the RV 144 trial. As previously discussed, the nonparametric method can pick up threshold effects, and this is well-illustrated in this plot, as the nonparametrically-estimated curve is flat up until a V1V2 breadth score of roughly 700, after which there is a large

jump in CVE. This leads one to hypothesize that if a vaccine could induce large V1V2 breadth scores in participants, it could potentially be very efficacious.

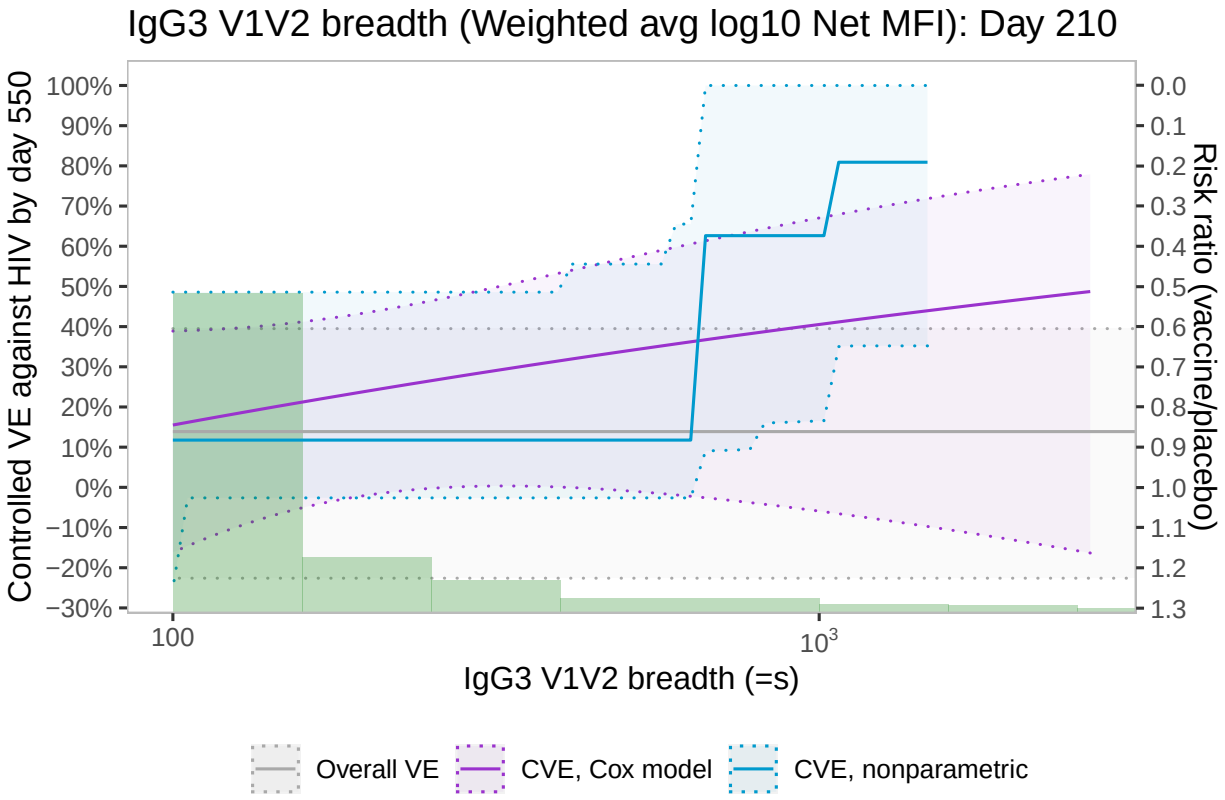


Figure 6.7: Controlled vaccine efficacy (CVE) curves for the IgG3 V1V2 breadth (Weighted avg log10 Net MFI) marker, measured at day 210. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

In Figure 6.8, we display results for the Peak baseline-subtracted pct loss luc activity to WITO marker, measured at day 210. If the Cox model results are to be believed, this marker is inversely correlated with protection against HIV-1 infection, and thus the monotonicity assumption of the nonparametric method is violated; this is why the nonparametric curve is flat. In theory, one could reverse the monotonicity assumption, but care has to be taken to avoid data dredging; ideally, a reversal of the usual monotonicity assumption should be prespecified.

In Figure 6.9, we display results for the IgG3 gp140 breadth (Weighted avg log10 Net MFI)

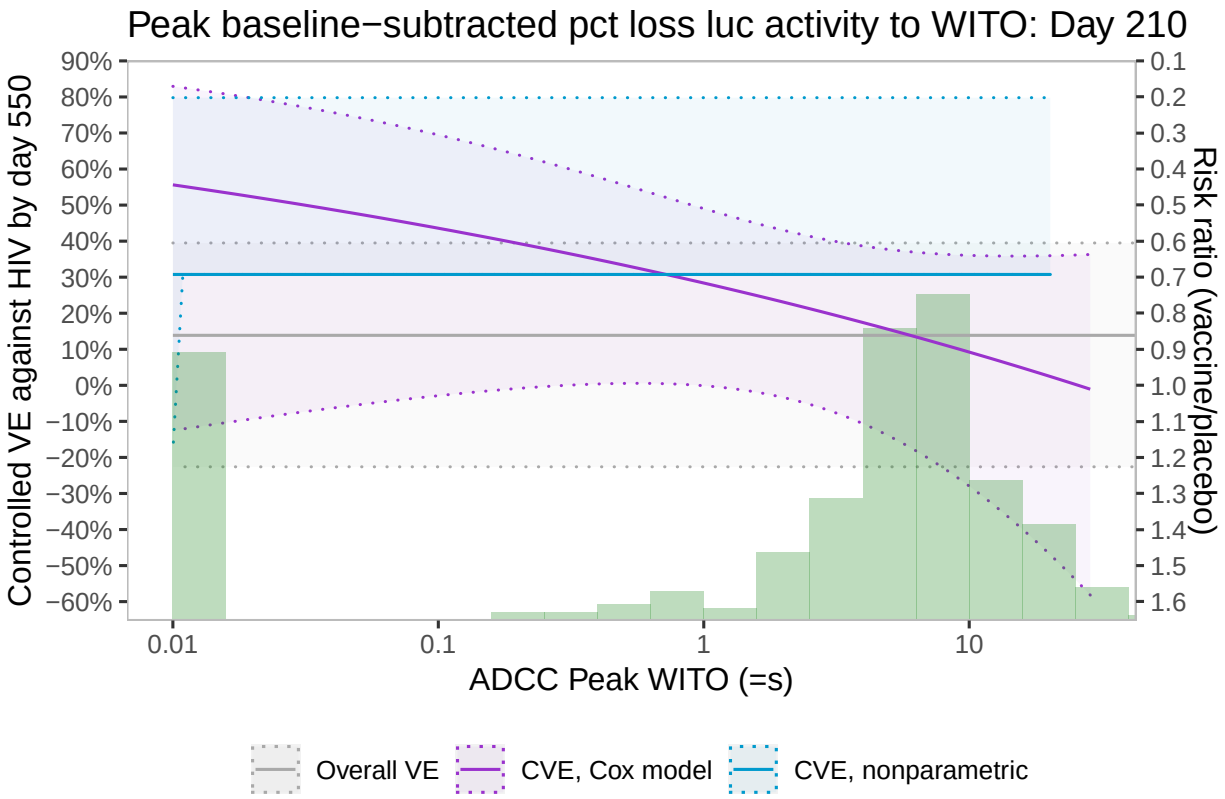


Figure 6.8: Controlled vaccine efficacy (CVE) curves for the Peak baseline-subtracted pct loss luc activity to WITO marker, measured at day 210. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

marker, measured at day 210. In practice, we sometimes obtain estimated curves that tell different stories; in this case, the nonparametric curve is nondecreasing whereas the Cox model curve is decreasing. However, the confidence bands are fairly wide, and so it is difficult to make any conclusions based on this plot. When we obtain contradictory results like this, it is sometimes beneficial to run additional models, such as the Cox model with spline terms, to provide additional evidence that may help guide interpretation.

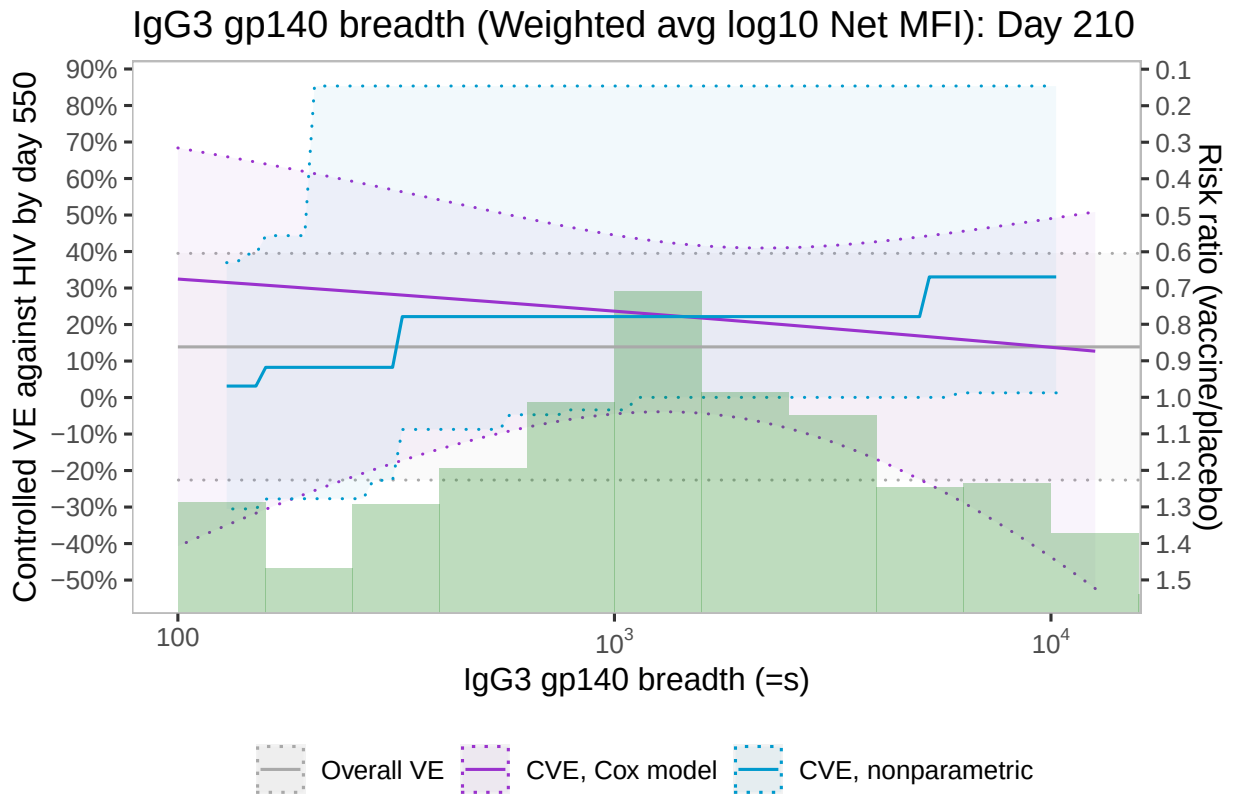


Figure 6.9: Controlled vaccine efficacy (CVE) curves for the IgG3 gp140 breadth (Weighted avg log10 Net MFI) marker, measured at day 210. Curves are estimated using a Cox proportional hazards model (purple) and the nonparametric estimator (blue). The grey line represents overall vaccine efficacy. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

6.3 AMP

Finally, we present select controlled risk plots from the analysis of the AMP trial, which tested a regimen of broadly neutralizing monoclonal antibody injections on HIV-1 infection. The AMP trial was divided into two protocols, the HVTN 703 study among women in sub-Saharan Africa and the HVTN 704 study among men and transgender persons who have sex with men in North America, South America, and Switzerland. In this trial, the primary marker of interest was body weight, highlighting the fact that these methods can be applied to settings involving any continuous marker, even though we typically are interested in complex immunological markers. Figure 6.10 depicts controlled risk curves within the HVTN 703 study. Across all trial arms, these curves suggest

that the probability of HIV-1 infection decreases as a function of body weight. Furthermore, this relationship appears to be modified by the VRC01 monoclonal antibody regimen, as these curves are quite similar to one another but different from the control group curve.

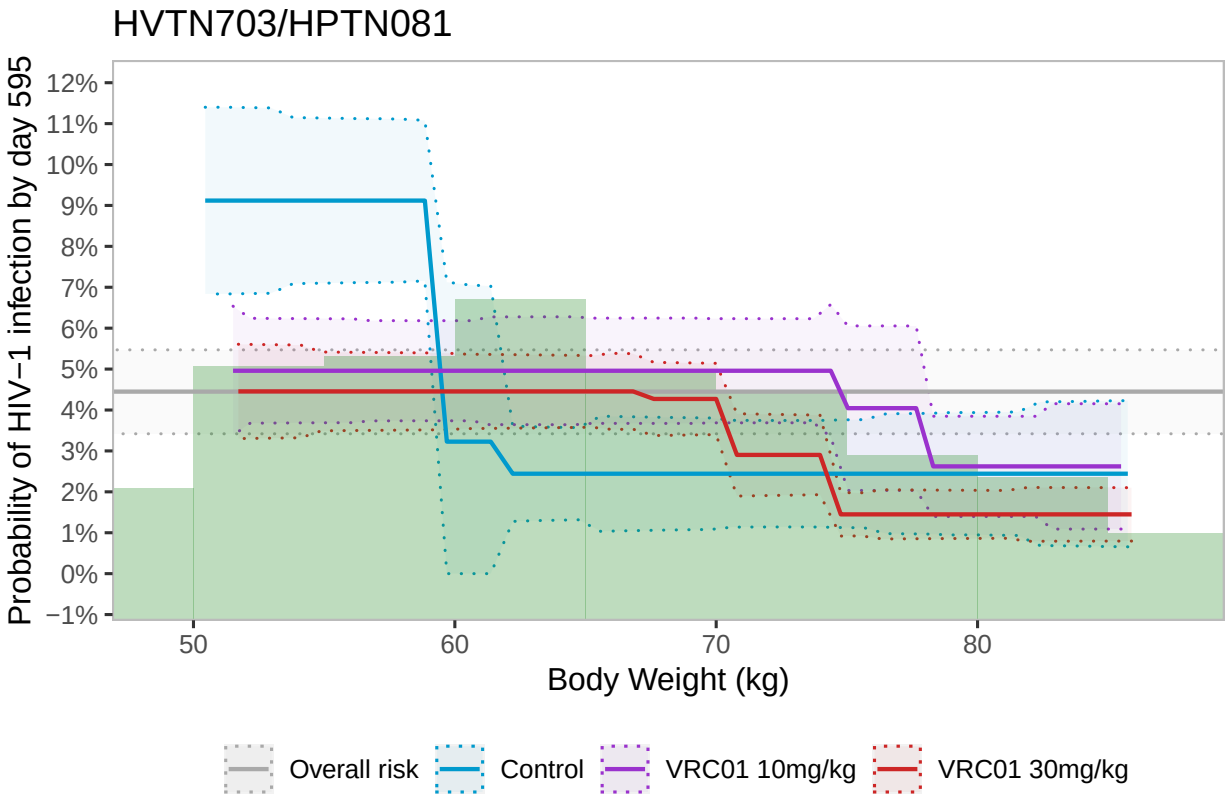


Figure 6.10: Controlled risk (CR) curves for the body weight (kg) marker, measured at baseline. Curves are estimated using the nonparametric method, and are displayed separately for the 10 mg/day dosing schedule (purple), the 30 mg/day dosing schedule (red), and the control group (blue). The grey line represents overall risk across all groups. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

Figure 6.11 displays analogous curves for the HVTN 704 study. Here, we similarly see that the probability of HIV-1 infection decreases as a function of body weight, but the differences between the arm-specific curves appears less pronounced. One thing that these graphs highlight is the need for tools that allow us to quantitatively compare two controlled risk curves to determine whether they are different from one another in terms of overall change across the domain, whether they differ in shape, and whether threshold effects differ; this was discussed in section 3.5.

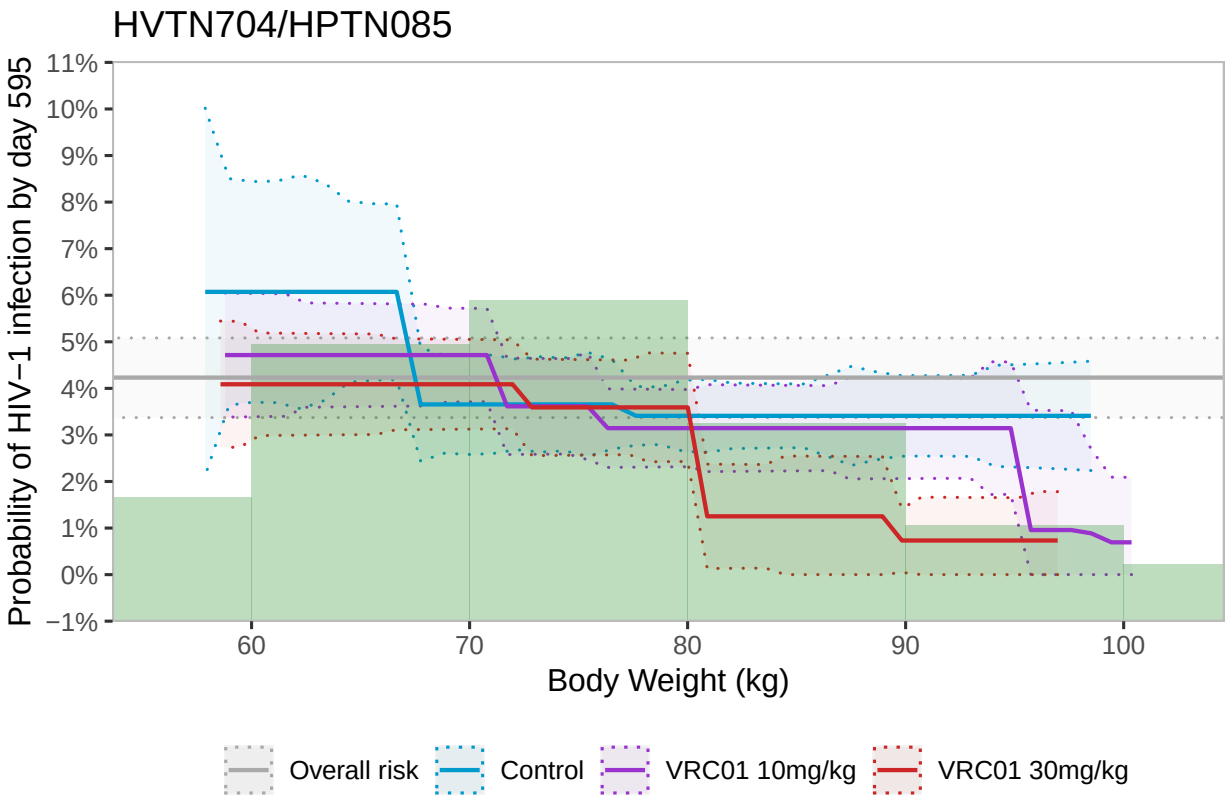


Figure 6.11: Controlled risk (CR) curves for the body weight (kg) marker, measured at baseline. Curves are estimated using the nonparametric method, and are displayed separately for the 10 mg/day dosing schedule (purple), the 30 mg/day dosing schedule (red), and the control group (blue). The grey line represents overall risk across all groups. Dotted lines represent pointwise 95% confidence intervals. Green histograms represent the estimated marginal distribution of the marker.

Bibliography

Lindsey R Baden, Hana M El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, Stephen A Spector, Nadine Rouphael, C Buddy Creech, et al. Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England journal of medicine*, 2020.

Richard E Barlow and Hugh D Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

David Benkeser, Iván Díaz, and Jialu Ran. Inference for natural mediation effects under case-cohort sampling with applications in identifying covid-19 vaccine correlates of protection. *arXiv preprint arXiv:2103.02643*, 2021.

David Benkeser, David C Montefiori, Adrian B McDermott, Youyi Fong, Holly E Janes, Weiping Deng, Honghong Zhou, Christopher R Houchens, Karen Martins, Lakshmi Jayashankar, et al. Comparing antibody assays as correlates of protection against covid-19 in the cove mrna-1273 vaccine efficacy trial. *Science translational medicine*, 15(692):eade9078, 2023.

Peter J Bickel. On adaptive estimation. *The Annals of Statistics*, pages 647–671, 1982.

NE Breslow, T Lumley, CM Ballantyne, LE Chambless, and M Kulich. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistical Biosciences*, 1:32–49, 2009.

Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

Norman E Breslow and Thomas Lumley. Semiparametric models and two-phase samples: Applications to cox regression. *IMS collections*, 9:65–77, 2013.

Norman E Breslow and Jon A Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102, 2007.

Norman E Breslow, Jie Hu, and Jon A Wellner. Z-estimation and stratified samples: application to survival models. *Lifetime data analysis*, 21:493–516, 2015.

B.J. Cowling, W.W. Lim, R.A. Perera, V.J. Fang, G.M. Leung, J.M. Peiris, and E.J. Tchetgen Tchetgen. Influenza hemagglutination-inhibition antibody titer as a mediator of vaccine-induced protection for influenza B. *Clinical Infectious Diseases*, 68(10):1713–7, 2019.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

M Csörgö and Pál Révész. *Two approaches to constructing simultaneous confidence bounds for quantiles*. Carleton University, Department of Mathematics and Statistics, 1981.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

D. Follmann. Augmented designs to assess immune response in vaccine trials. *Biometrics*, 62: 1161–1169, 2006.

Youyi Fong, Adrian B McDermott, David Benkeser, Sanne Roels, Daniel J Stieh, An Vandebosch, Mathieu Le Gars, Griet A Van Roey, Christopher R Houchens, Karen Martins, et al. Immune correlates analysis of the ensemble single ad26. cov2. s dose vaccine efficacy clinical trial. *Nature Microbiology*, pages 1–15, 2022.

Frederick N Fritsch and Ralph E Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.

Peter B Gilbert, Youyi Fong, Avi Kenny, and Marco Carone. A controlled effects approach to assessing immune correlates of protection. *Biostatistics*, 2022a.

Peter B Gilbert, David C Montefiori, Adrian B McDermott, Youyi Fong, David Benkeser, Weiping Deng, Honghong Zhou, Christopher R Houchens, Karen Martins, Lakshmi Jayashankar, et al. Immune correlates analysis of the mrna-1273 covid-19 vaccine efficacy clinical trial. *Science*, 375 (6576):43–50, 2022b.

Richard D Gill and Soren Johansen. A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, 18(4):1501–1555, 1990.

Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press, 2014.

Scott M Hammer, Magdalena E Sobieszczyk, Holly Janes, Shelly T Karuna, Mark J Mulligan, Doug Grove, Beryl A Koblin, Susan P Buchbinder, Michael C Keefer, Georgia D Tomaras, et al. Efficacy trial of a dna/rad5 hiv-1 preventive vaccine. *New England Journal of Medicine*, 369(22): 2083–2092, 2013.

Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, 2001.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Barton F Haynes, Peter B Gilbert, M Juliana McElrath, Susan Zolla-Pazner, Georgia D Tomaras, S Munir Alam, David T Evans, David C Montefiori, Chitraporn Karnasuta, Ruengpueng Sutthent, et al. Immune-correlates analysis of an hiv-1 vaccine efficacy trial. *New England Journal of Medicine*, 366(14):1275–1286, 2012.

Marshall M Joffe and Tom Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

A Kenny, A Luedtke, O Hyrien, Y Fong, R Burnham, J Heptinstall, S Sawant, S Stanfield-Oakley, FL Omar, S Khuzwayo, et al. Immune correlates analysis of the imbokodo hiv-1 vaccine efficacy trial. In *Journal of the International AIDS Society*, volume 25, pages 214–214, 2022.

Avi Kenny and Charles Wolock. Simengine: an r package for structuring statistical simulations, 2021. URL https://avi-kenny.github.io/SimEngine.

Phillip S Kott. Model-based finite population correction for the horvitz-thompson estimator. *Biometrika*, pages 797–799, 1988.

Haiqun Lin and Daniel Zelterman. Modeling survival data: extending the cox model, 2002.

G. Molenberghs, T. Burzykowski, A. Alonso, P. Assam, A. Tilahum, and M. Buyse. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *Journal of Statistical Planning and Inference*, 138:432–449, 2008.

Z. Moodie, M. Juraska, Y. Huang, Y. Zhuang, Y. Fong, L.N. Carpp, S.G. Self, L. Chambonneau, R. Small, N. Jackson, et al. Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *Journal of Infectious Diseases*, 217(5):742–753, 2018.

Ricardo Palacios, Elizabeth González Patiño, Roberta de Oliveira Piorelli, Monica Tilli Reis Pessoa Conde, Ana Paula Batista, Gang Zeng, Qianqian Xin, Esper G Kallas, Jorge Flores, Christian F Ockenhouse, et al. Double-blind, randomized, placebo-controlled phase iii clinical trial to evaluate the efficacy and safety of treating healthcare professionals with the adsorbed covid-19 (inactivated) vaccine manufactured by sinovac–profiscov: A structured summary of a study protocol for a randomised controlled trial. *Trials*, 21:1–3, 2020.

Stanley A Plotkin. Correlates of protection induced by vaccination. *Clinical and vaccine immunology*, 17(7):1055–1065, 2010.

Stanley A Plotkin and Peter B Gilbert. Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Diseases*, 54(11):1615–1617, 2012.

R.L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8:431–440, 1989.

Ross L Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11, 1986.

Lihong Qi, CY Wang, and Ross L Prentice. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100(472):1250–1263, 2005.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and statistics: essays in honor of David A. Freedman*, 2:335–421, 2008.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The international journal of biostatistics*, 7(1), 2011.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Alastair J Scott and Chris J Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997.

Kelly E Seaton, Yunda Huang, Shelly Karuna, Jack R Heptinstall, Caroline Brackett, Kelvin Chiong, Lily Zhang, Nicole L Yates, Mark Sampson, Erika Rudnicki, et al. Pharmacokinetic serum concentrations of vrc01 correlate with prevention of hiv-1 acquisition. *EBioMedicine*, 93, 2023.

Steven G Self and Ross L Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, pages 64–81, 1988.

Lars van der Laan, Wenbo Zhang, and Peter B Gilbert. Nonparametric estimation of the causal effect of a stochastic threshold-based intervention. *Biometrics*, 2022.

Aad van der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, 1996.

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

Wei Wang and Dylan S Small. Monotone b-spline smoothing for a generalized linear model response. *The American Statistician*, 69(1):28–33, 2015.

Ted Westling and Marco Carone. A unified study of nonparametric inference for monotone functions. *Annals of statistics*, 48(2):1001, 2020.

Ted Westling, Peter Gilbert, and Marco Carone. Causal isotonic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):719–747, 2020.

Ted Westling, Alex Luedtke, Peter Gilbert, and Marco Carone. Inference for treatment-specific survival curves using machine learning. *arXiv preprint arXiv:2106.06602*, 2021.

Charles J Wolock, Peter B Gilbert, Noah Simon, and Marco Carone. A framework for leveraging machine learning tools to estimate personalized survival curves. *arXiv preprint arXiv:2211.03031*, 2022.

X Xin, J Horrocks, and GA Darlington. Ties between event times and jump times in the cox model. *Statistics in Medicine*, 32(14):2374–2389, 2013.