Evaluation of Treatment Effect Modification by Post-randomization Biomarker-defined Principal Strata with Application to Vaccine Efficacy Trials

Yingying Zhuang

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

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Peter B. Gilbert, Chair

Ying Huang, Chair

M. Elizabeth Halloran

Ann Duerr (GSR)

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Evaluation of Treatment Effect Modification by Post-randomization Biomarker-defined Principal Strata with Application to Vaccine Efficacy Trials

Yingying Zhuang

Co-Chairs of the Supervisory Committee:
Professor Peter B. Gilbert
Department of Biostatistics

Professor Ying Huang
Department of Biostatistics

In vaccine studies, investigators are often interested in studying effect modifiers of clinical treatment efficacy by biomarker-based principal strata, which is useful for selecting biomarker study endpoints for evaluating treatments in new trials, exploring biological mechanisms of clinical treatment efficacy, and studying mediators of treatment efficacy. Such post-randomization effect modification research has been referred to as principal surrogate evaluation. Towards this goal, many methods have been developed for a two-phase sampling design, motivated by trials in HIV vaccine research. However, new challenges have arisen in Dengue vaccine efficacy trials, including pre-trial disease exposure and long time span between baseline and when post vaccination antibody neutralization titers are measured. Motivated by these challenges, my dissertation research aims at addressing some limitations of existing literature and developing new methods for principal stratification effect modification analysis to accommodate broader vaccine efficacy trial designs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are a number of people I would like to thank for supporting me throughout the time I was a student at University of Washington. First of all, I would like to thank my two advisors, Peter B Gilbert, and Ying Huang, for their direction, support, and encouraging remarks and, perhaps more so, for believing in me. I have found their insights, guidance and feedback to be extremely valuable. I would also like to thank my reading committee member M. Elizabeth Halloran for her comments and advice during the revision of this dissertation.

And I want to extend a big thank you to all members at the Fred Hutch SCHARP team, especially Zoe Moodie and Michal Juraska, with whom I have worked closely for my RA position on the Dengue trials that have become a major motivation for all my projects in this dissertation. It has been a treat working with them and brainstorming about statistics and science.

To my parents Jialiang and Yan, you helped nurture the seeds of science as I grew and continue to support me to this day. I would not be the person I am today without your influence in my life. To my puppy Drogo, who always sleeps by my feet while I work on this dissertation, you are a bundle of joy and a true blessing. To my dear classmates and colleagues, you have been immensely fun to do a PhD with and I am lucky to know all of you.

Last, and most, I want to thank my fiancé Leo, for all his love and support. You help me strive to always do my best and help pick me up when I fall short of that goal. I wouldn't be who I am today without you, both academically and as a person. You are everything.

DEDICATION

To my fiancé Leo

Chapter 1

INTRODUCTION

In vaccine research, identifying biomarkers that can be used as surrogate endpoints for clinical endpoints is an important question to address. A good surrogate can be used to guide the development of the vaccine and predict the vaccine's protective effect on the clinical endpoint in future settings before conducting efficacy trials. Oftentimes a candidate surrogate is measured at a fixed time after randomization. My dissertation research is motivated by the need to evaluate immune response biomarkers as effect modifiers of a vaccine's effect on the clinical endpoint of interest (i.e., clinical vaccine efficacy), which is one way to study biomarkers and develop their utility for predicting vaccine efficacy.

Within the principal stratification framework introduced by Frangakis and Rubin [5], Gilbert and Hudgens [11], henceforth GH, proposed the Causal Effect Predictiveness (CEP) surface as an estimand for evaluating candidate principal surrogates. The CEP surface is defined in terms of clinical risks conditional on the pair of potential biomarker values if assigned to vaccine or placebo, and it quantifies how well causal treatment effects on the biomarker predict causal treatment effects on the clinical endpoint. Based on the CEP surface, a useful biomarker is one that is a strong effect modifier, where the CEP surface varies over subgroups defined by potential outcome levels of the pair of biomarker responses[7]. More specifically, principal surrogate analysis is essentially principal stratification effect modification analysis, with an objective to characterize how clinical treatment efficacy varies over subgroups defined by principal strata and possibly also by baseline covariates. Candidate surrogates are compared and ranked by their strength of effect modification. In applications such as HIV vaccine efficacy trials where the biomarker of interest is an immune response to HIV but an enrollment criterion is being HIV negative at baseline, the biomarker values of placebo

recipients all take a constant value (zero), which has been named the "constant biomarkers" (CB) case. Under CB, the CEP surface simplifies to the marginal CEP (mCEP) curve conditional on the potential biomarker if assigned to vaccine, namely, the vaccine-induced biomarker. However, in other settings the CB condition does not hold. For example, in dengue or influenza vaccine efficacy trials, many participants enter the study having previously been infected with dengue or influenza, respectively. For the biomarker of interest defined as an immune response to dengue or influenza, this prior exposure causes variability in the biomarker across placebo recipients. Therefore, the CB condition does not hold, yet the mCEP curve still has a useful interpretation for comparing marginal conditional disease risks that average over the conditional distribution of the potential biomarker value under placebo [27]. Multiple methods have been developed to estimate the mCEP curve for randomized vaccine efficacy trials with two-phase sampling of biomarker data ([4]; [11]; [13]; [14]; [25]; [27]). Many of these methods were motivated by trials in HIV vaccine research.

In this dissertation research, we aim at developing new methods for principal stratification effect modification analysis motivated by challenges and problems that arose in two dengue phase III vaccine efficacy trials, CYD14 ([1]) and CYD15 ([26]). Unlike the HIV vaccine efficacy trials, placebo recipients of dengue vaccine commonly have variable levels of the immune response biomarkers, reflecting natural immunity arising from pre-trial exposure to the disease-causing pathogen. These baseline biomarkers may modify vaccine efficacy (they have been shown to do so for the dengue trials), and moreover vaccine efficacy may depend jointly on a biomarker measured at baseline and measured at the fixed time after vaccination used for candidate surrogate assessment. Therefore, it is of interest to estimate the mCEP curve as a function of baseline biomarker measurements. Previous methods allow this assessment if the biomarker is always measured at baseline when it is measured at the fixed time point post vaccination. However, in the dengue Phase III trials, the biomarker values at baseline were only measured in a fraction of those with the biomarker measured at the post-vaccination time point, thus necessitating development of new methods to utilize the subsampled baseline immune responses to predict unobserved potential post-randomization

immune responses and to estimate the mCEP curve as a function of baseline biomarker measurements. Another challenge of treatment effect modification analysis in the dengue trials is the plausibility of the Equal Early Clinical Risk assumption (A3) made by Gilbert and Hudgens ([11]) and other CEP based methods. This assumption - which assumes the vaccine has no individual-level causal effect on clinical endpoint occurrence in the period before the biomarker is measured post-vaccination - is needed for helping identify the CEP curve based on data from subjects observed to be at risk for disease when the biomarker is measured post vaccination. Inferences will be robust to this assumption if the biomarker is measured near baseline relative to the period of follow-up for clinical events such that almost all clinical endpoint events occur after the time of biomarker measurement. However, in the two Dengue phase III vaccine trials, post vaccination antibody neutralization titer measurements were collected 13 months after the first vaccination (as opposed to 8 weeks in some HIV trials, e.g., HVTN 502), and about half of the primary dengue disease endpoint events occurred before 13 months, therefore requiring new methods to be developed to relax the equal early clinical risk assumption.

Motivated by these challenges, my dissertation consists of the following three projects:

1. Evaluation of Treatment Effect Modification by Post-randomization Biomarker-defined Principal Strata Based on a Pseudo-Score Estimator Including Simultaneous Inference

   To estimate the mCEP curve in the presence of missing counter-factual vaccine-induced biomarker values in an efficacy trial, Huang and others [14], henceforth HGW, proposed a pseudo-score type estimator (inspired by Chatterjee and others [3]). This estimator utilizes the baseline predictor associated with the vaccine-induced immune response biomarker as an auxiliary variable to estimate the parameters in the assumed parametric risk model. HGW developed a procedure for inference about the mCEP curve under the condition that the risk model is independent of the baseline predictor after conditioning on the vaccine-induced biomarker. This condition may be reasonable in some settings, e.g., when the baseline predictor is the same immune response variable

as the vaccine-induced immune response (merely at different time points), but this is not expected to hold in general. In this chapter, we propose an approach to estimate the mCEP curve allowing the risk model to be dependent on the baseline predictor after conditioning on the vaccine-induced biomarker, built upon HGW's pseudo-score estimator. Additionally, we develop a perturbation resampling method to approximate the asymptotic distribution of our estimator and to calculate pointwise and simultaneous confidence bands. Making simultaneous inference about the mCEP curve is important for understanding biomarker-defined principal strata effect modification over a range of biomarker values but to our knowledge has not been addressed previously.

2. Evaluation of Bivariate Treatment Effect Modification by Post-randomization Biomarker-defined Principal Strata and Baseline Covariates in the Presence of Non-monotone Missingness

   In randomized vaccine trials where participants may enter the study with prior exposure therefore with variable baseline biomarker values (as in the dengue trials), clinical treatment efficacy may depend jointly on a biomarker measured at baseline and measured at a fixed time after vaccination. Therefore, it is of interest to conduct a bivariate effect modification analysis by biomarker-defined principal strata and baseline biomarker values. Previous methods allow this assessment if participants who have the biomarker measured at the the fixed time point post randomization would also have the biomarker measured at baseline. However, additional complications in study design could happen. For example, in the Dengue correlates study, baseline biomarker values were only available from a fraction of participants who have biomarkers measured post-randomization. How to conduct the bivariate effect modification analysis in these studies remains an open research question. In this project, we propose an estimated likelihood method to utilize the sub-sampled baseline biomarker in the effect modification analysis.

3. Evaluation of Treatment Effect Modification by Post-randomization Biomarker-defined Principal Strata Relaxing the Equal Early Clinical Risk Assumption

   One of the identifiability assumptions made by GH is equal early clinical risk up to a fixed time when the biomarker is measured, often referred to as assumption A3. A3 is helpful for identifying the causal estimand mCEP based on data from subjects observed to be at risk for disease at the time of biomarker measurement. Inferences will be robust to A3 if relatively few clinical events happen before candidate surrogates are measured. However, in trials where biomarkers are measured long time after baseline, as in the dengue trials, A3 is likely to be violated. In this project, we consider how to extend the mCEP curve estimation methods to the more plausible assumption: the monotonicity assumption (refereed to as A3-mon) which states that an individual who would stay at risk under placebo up to the time of biomarker measurement would also stay at risk under vaccine up to the time of biomarker measurement, based on an estimated maximum likelihood approach.

Motivated by challenges that arose in the dengue trials, all three projects seek to address some limitations of the existing literature on effect modification analysis by biomarker-defined principal strata and develop new methods to accommodate broader vaccine efficacy trial designs.

Chapter 2

THE SET-UP

In this chapter, I present notation and assumptions that will be adopted throughout the rest of the dissertation, unless specified otherwise. I also provide a brief background on the two Phase 3 dengue vaccine efficacy trials: CYD14[1] and CYD15[26] which will be used as examples for illustration in chapter 3, chapter 4, and chapter 5.

2.1 Notation

We consider data from a two-arm vaccine efficacy trial that randomizes a total of n participants to either the vaccine arm or the placebo arm, with Z being the binary indicator of assignment to the vaccine arm ($Z = 1$ for vaccine arm and $Z = 0$ for placebo arm). A vector of baseline covariates (e.g. gender and country) are recorded for everyone at baseline and are denoted by X. Trial participants are followed for the primary clinical endpoint for a predetermined period of time and let Y be the indicator of clinical endpoint event during the study follow-up period. At some fixed time $\tau > 0$ post randomization, an intermediate response endpoint, S, is measured. Because S must be measured prior to disease to evaluate its treatment effect modification, availability of S is conditional on remaining clinical endpoint free at time τ (denoted by $Y^\tau = 0$). If clinical endpoint occurs in the time interval $[0, \tau]$ ($Y^\tau = 1$), then S is undefined and we set $S = *$. Frequently a two-phase sampling design is used where S is only measured in a subcohort. Two-phase sampling is particularly useful when the cost for observing S is expensive and the event Y is rare [19]. In the first phase, baseline covariates X and the clinical outcome data Y and Y^τ are measured for everyone, and in the second phase, S is measured for all cases (defined by $Y = 1$ and $Y^\tau = 0$) as well as for a random subset of controls (defined by $Y = 0$), where the control sampling probabilities

may depend on X. In practice, the sampling of controls is done from participants who are observed to have $Y = 0$, i.e. those who complete follow-up free of the endpoint. Throughout this dissertation, we assume there is no dropouts. Therefore, the set of $Y = 0$ equals the set of participants who are observed to have $Y = 0$. To define the estimand of interest, we use potential outcomes, where all post-randomization measurements are considered under either $z = 0$ or $z = 1$ for each individual. Let $S(z)$, $Y^\tau(z)$, $Y(z)$ be the potential outcomes if the subject receives treatment z, for $z = 0$ or 1. If $Y^\tau(z) = 1$, $S(z)$ is undefined and we set $S(z) = *$. With the interest of evaluating effect modification in subgroups with S defined, the analysis is restricted to subjects with $Y^\tau = 0$ in whom S could be potentially measured.

GH defined the CEP surface as

$$CEP^{risk}(s_1, s_0) \equiv h(risk_1(s_1, s_0), risk_0(s_1, s_0)),$$

where

$$risk_z(s_1, s_0) \equiv P\left(Y(z) = 1 | S(1) = s_1, S(0) = s_0, Y^\tau(1) = Y^\tau(0) = 0\right),$$

for $z = 0, 1$ and the function $h(x, y)$ is a known contrast function satisfying $h(x, y) = 0$ if and only if $x = y$. It conditions on the counterfactual pair $(S(0), S(1))$ which forms a principal stratification and can be considered as an unobserved baseline characteristic of each subject. The latter condition $Y^\tau(1) = Y^\tau(0) = 0$ ensures that causal treatment effects on S are defined. Under the "constant biomarkers" (CB) case defined in chapter 1 and the Equal Early Clinical Risk Assumption (referred to as A3 in GH, which we define and discuss in detail in Section 2.2), the CEP surface equals the marginal CEP curve, which is defined as

$$mCEP^{risk}(s_1) \equiv h(risk_1(s_1), risk_0(s_1)),$$

where

$$risk_z(s_1) \equiv P\left(Y(z) = 1 | S(1) = s_1, Y^\tau(1) = Y^\tau(0) = 0\right).$$

In settings where the CB condition does not hold, such as our motivating Dengue example, the mCEP curve still has a useful interpretation for comparing marginal conditional disease risks that average over the conditional distribution of $S(0)$.

A common choice of the contrast function h is the vaccine efficacy function:

$$\text{VE}(s_1) \equiv 1 - \frac{risk_1(s_1)}{risk_0(s_1)},$$

which we refer to as the VE curve and it constitutes the estimand of interest throughout this dissertation research. The VE curve measures the percentage of reduction in the clinical endpoint rate for the subgroup of vaccine recipients at-risk for the clinical endpoint at τ under both treatment assignments and with immune response $S(1) = s_1$ compared to what the clinical endpoint rate would have been had they been assigned to the placebo arm. Large variability in the VE curve indicates strong effect modification. The VE curve can be a useful tool for the future development of vaccines by providing a ranking of immune response biomarkers by their strength of effect modification.

However, the marginal CEP curves are not identifiable from the standard assumptions made in randomized vaccine efficacy trials due to that $S(1)$ is missing for all placebo recipients. To address this problem, two augmented trial designs have been proposed by Follmann[4]. The first utilizes baseline immunogenicity predictors (BIPs) to develop an imputation model for the unobserved immune response biomarker values based on the estimable relationship between baseline covariates and biomarker values. However, as GH has pointed out, this approach only identifies the principal stratification estimands under an untestable assumption made on the risk model. The second augmented design vaccinates all or a fraction of placebo recipients who remain free of the clinical endpoint at the closeout of the trial, and the immune response biomarker S_c is measured at time τ after vaccination. These values S_c are then used to substitute for the missing biomarker values as if they originally had been assigned to the vaccine arm. Follmann named the second augmented design "closeout placebo vaccination"(CPV). A major advantage of including CPV is that it makes possible the evaluation of the risk model assumptions. We call the BIP design with no CPV component the BIP-only design and the BIP design with a nonzero CPV component the BIP+CPV design.

2.2 Key Assumptions

Throughout this dissertation, we adopt the following assumptions to help identify the estimators for the marginal CEP curves from the observed data.

(A1) Stable Unit Treatment Value Assumption (SUTVA) and Consistency;

(A2) Ignorable Treatment Assignment;

(A3) Equal early clinical risk: $P\left(Y^{\tau}(1) = Y^{\tau}(0)\right) = 1$.

For the BIP+CPV design only:

(A1-CPV) Time constancy of immune response: For event-free placebo recipients, $S(1) = S^{true} + U_1$, and $S_c = S^{true} + U_2$, for some underlying S^{true} and i.i.d. measurement errors U_1, U_2 that are independent of one another.

(A2-CPV) No placebo subjects event-free at closeout experience the endpoint over the next τ time-units.

A1 implies that the potential outcomes of each trial participant $(S_i(1), S_i(0), Y_i^{\tau}(1), Y_i^{\tau}(0), Y_i(1), Y_i(0))$ are not influenced by other subjects, and the potential outcomes under the assigned treatment arm equal the observed outcomes. In other words,

$$\begin{aligned} Y_i^{\tau}(z) &= Y_i^{\tau} \text{ if } Z_i = z, z = 0, 1 \\ Y_i(z) &= Y_i \text{ if } Z_i = z, z = 0, 1 \\ S_i(z) &= S_i \text{ if } Z_i = z \text{ and } Y_i^{\tau} = 0, z = 0, 1. \end{aligned} \tag{2.1}$$

A2 holds for blinded randomized trials, where randomization, possibly depending on the baseline covariates X, ensures that

$$(S(0), S(1), Y(0), Y(1), Y^{\tau}(0), Y^{\tau}(1)) \amalg Z|X \tag{2.2}$$

because $(S(0), S(1), Y(0), Y(1), Y^\tau(0), Y^\tau(1))$ can be considered an unobserved baseline characteristic of each subject, like genetic make-up.

A3 is needed so that the clinical risks conditional on Z, X, $S(1)$, and $Y^\tau(1) = Y^\tau(0) = 0$ can be identified based on subjects who are at risk at time τ and could potentially have S measured. A3 is plausible if the treatment cannot confer any effect on occurrence of Y (either beneficial or harmful) until time τ. However, A3 will fail in many trials, and Wolfson and Gilbert [27] showed that relaxing A3 makes it harder to identify and estimate $risk_z$ and for trials where A3 fails, its violations cause smaller bias for scenarios where relatively few clinical endpoints occur through time τ compared to after τ (and similarly in this situation violations of A2-CPV only minorly bias inferences). In our work presented in chapter 3 and chapter 4, we study our estimand under A3 (and A2-CPV for the BIP+CPV design) and we tacitly assume all probabilities condition on $Y^\tau(1) = Y^\tau(0) = 0$. In chapter 5, we develop a sensitivity analysis approach where A3 could be relaxed and the condition $Y^\tau(1) = Y^\tau(0) = 0$ is not tacitly assumed. We will explicitly write out if probabilities are conditioned on $Y^\tau(1) = Y^\tau(0) = 0$ in chapter 5.

Under A1-CPV and A2-CPV, S_c is substituted for $S(1)$ for subjects selected for CPV. For the rest of this dissertation we consolidate the notation of S and S_c , and let $S = S(1)$ be the vaccine-induced immune response measurement obtained either during the standard trial follow-up or during the CPV and let δ be the indicator for the availability of S.

We also assume that the study participants make up a random sample from a large population of interest, the observed data $O_i \equiv (Z_i, X_i, Y_i^\tau, Y_i, \delta_i, S_i\delta_i)$, $i = 1, ..., n$, are independently and identically distributed (i.i.d.) copies of a random vector $O \equiv (Z, X, Y^\tau, Y, \delta, S\delta)$.

2.3 the CYD14 and CYD15 trials

In chapter 3, 4, and 5, we demonstrate application of our proposed methods using data from two Phase 3 dengue vaccine efficacy trials: CYD14 [1] in 2–14 year olds in Asia and CYD15 [26] in 9–16 year olds in Latin America. Participants were randomly assigned in 2:1 allocation to receive three injections of a live attenuated tetravalent dengue vaccine

(containing one dengue strain each from the four serotypes of dengue) or placebo at months 0, 6, and 12 and were followed with active surveillance for occurrence of the primary study endpoint of symptomatic virologically confirmed dengue disease (VCD). Based on a Cox model, estimated vaccine efficacy against VCD due to any serotype diagnosed after Month 13 and by Month 25 was 56.5% (95% confidence interval, 43.8 to 66.4) for CYD14 and 60.8% (95% confidence interval, 52.0 to 68.0) for CYD15. Neutralizing antibody titers were measured to each of the four parental dengue serotype vaccine strains at Month 13 (time τ) in case-control samples. For all Dengue illustrations in this dissertation, we let S be the individual's sample average response to the four serotype-specific log10-transformed antibody titers (i.e., "average titer") and include the subjects' age and country in the baseline covariate vector X, and Y is the indicator of the VCD occurring more than 28 days after the third injection. All methods were conducted on the trial-pooled data set of 9–16 year olds to assess how VE varied with neutralizing antibody titers if assigned to receive the vaccine. Justification for pooling the data across the trials is provided in Moodie et al.[17], with main justification that the protocols were essentially the same.

Chapter 3

SIMULTANEOUS INFERENCE OF TREATMENT EFFECT MODIFICATION BY POST-RANDOMIZATION BIOMARKER-DEFINED PRINCIPAL STRATA BASED ON A PSEUDO-SCORE ESTIMATOR

3.1 Introduction

Multiple methods ([11]; [20]; [15];[13]) have been developed to identify and estimate the marginal CEP curve based on an estimated likelihood approach where the estimators are obtained by maximizing an estimated version of the observed likelihood. With the observed data being $O_i \equiv (Z_i, X_i, Y_i^\tau, Y_i, \delta_i, S_i\delta_i)$, the observed likelihood is defined as

$$L = \prod_{i=1}^{n} f(Y_i|Z_i, X_i, \delta_i, S_i)$$

where

$$f(Y|Z, X, \delta, S) \tag{3.1}$$

$$= \begin{cases} risk_1(S, X; \beta)^Y (1 - risk_1(S, X; \beta)^{1-Y}) \;\; if \;\; \delta = 1 \\ (\int risk_z(s, X; \beta) \times dF_1^{S|X}(s|X))^Y \times (1 - \int risk_z(s, X; \beta) \times dF_1^{S|X}(s|X))^{1-Y} \;\; o.w. \end{cases} \tag{3.2}$$

Define nuisance parameters $\nu \equiv F_1^{S|X}$. Consistent estimators of the nuisance parameters ν are obtained based on data from subset $\{i : Z_i = 1, \delta_i = 1\}$ and assumption (A2): $F(S|Z = 1, X) = F(S|Z = 0, X) = F(S|X)$. The CPV component is not used in this step even for a BIP+CPV design due to the fact that all infected placebo recipients have zero sampling probability for S. After the consistent estimator $\widehat{\nu}$ is obtained, the likelihood $L(\beta, \hat{\nu})$ is maximized in β, where the CPV component is included. In a sequential phase

2b HIV vaccine efficacy trial with a primary goal of evaluating vaccine efficacy and immune correlates for multiple HIV vaccine regimens, Gilbert and others [9] examined the power for detecting vaccine efficacy modification using a estimated likelihood approach. The results were seemingly counterintuitive. In some scenarios where X was strongly correlated with S, the BIP-only design was more powerful than the BIP+CPV design. HGW investigated this result and realized that the decreased efficiency caused by the CPV sampling was due to the fact that the CPV component is included in the step of maximizing the likelihood $L(\beta, \hat{\nu})$ but not in the estimation of ν.

To address the inconsistent use of the CPV component, HGW proposed a pseudo-score type estimator suitable for both the BIP-only design and the BIP+CPV design to identify the parameters β in the assumed parametric risk model

$$risk_Z\{S, X\} \equiv P\left(Y(Z) = 1 | S, X, Y^\tau(1) = Y^\tau(0) = 0\right) = g\{\beta; S, Z, X\}.$$

Furthermore, in order to estimate the marginal CEP curve, specifically the vaccine efficacy curve as a function of the vaccine-induced immune response S, HGW considered the scenario that the risk model is independent of X after conditioning on S: $risk_Z\{S, X\} = g\{\beta; S, Z\}$. Then it follows that the VE curve estimator can be represented as a function of the risk model directly:

$$\mathrm{VE}(s_1) = 1 - \frac{g\{\beta; S = s_1, Z = 1\}}{g\{\beta; S = s_1, Z = 0\}}.$$

In general, however, clinical risks under Z might not be independent of X after conditioning on S. Under these scenarios, the estimation of the VE curve requires information not only on the risk model, but also on the distribution of the baseline covariate X conditional on the vaccine-induced immune response biomarker S. Next, we propose an estimator for the VE curve applicable to both the BIP-only design and the BIP+CPV design that is built upon HGW's pseudo-score estimator for the risk model and allows the risk model to depend on X after conditioning on S.

3.2 Method

3.2.1 Identifiability Assumptions

Beside assumptions A1-A3, A1-CPV and A2-CPV we talked about in chapter 2, HGW also made the following assumptions to help identify the pseudo-score estimators from the observed data.

(A4) Risk functions have a generalized linear model form: $risk_{(z)}\{S(1), X\} = g\{\beta; S(1), Z, X\}$ for some known link function $g(\cdot)$, for $z = 0, 1$.

(A5) $\int\int P(\delta = 1|y, z, X)\mathrm{d}y\mathrm{d}z > \varepsilon$ for some constant $\varepsilon > 0$ for X almost everywhere.

(A6) $risk_Z\{S, X\} = risk_Z\{S\} = g\{\beta; S, Z\}$.

A5 is needed for the pseudo-score estimators that use inverse probability of sampling S; critically, non-zero sampling probabilities are not needed for each of the four individual subgroups defined by treatment arm (Z) crossed with case-control status (Y), thus applicable to both the BIP-only and the BIP+CPV designs.

HGW adopted assumption A6 that the risk model is independent of X after conditioning on vaccine-induced immune response S. Follmann [4] made exactly the same assumption that X has no effect on $Y(Z)$ once S and Z are in the model in his equation (1). Follmann also discussed that this assumption can be relaxed when using a maximum likelihood approach with generalizations to the risk model that include X as an additional main effect, or even allow for an interaction term between X and Z. This generalized risk model can be estimated using maximum likelihood when CPV is included. When CPV is not included, the generalized risk model is identifiable provided, for example, that there is no interaction between X and Z. Furthermore, A6 is similar to a standard assumption made in many "surrogate" measurement error statistical methods that the conditional distribution of Y given (S, Z, X) depends only on (S, Z), in which case X is said to be a *surrogate* [2].

With A6, it follows that

$$\mathrm{VE}(s_1) = 1 - \frac{g\{\beta; S = s_1, Z = 1\}}{g\{\beta; S = s_1, Z = 0\}}.$$

At each fixed s_1 value, $\mathrm{VE}(s_1)$ is a continuous function of the parameters β and can be estimated by plugging in the pseudo-score estimators of β. We call this the A6-based VE estimator $\widehat{\mathrm{VE}}^{(A6)}(s_1)$. Asymptotic normality of $\widehat{\mathrm{VE}}^{(A6)}(s_1)$ follows by applying the delta method. We investigate the effects of assumption A6 on estimation in further detail in Section 3.4.

3.2.2 The Pseudo-Score Estimator

According to equation (7) in HGW, the pseudo-score type estimator is defined as the solution to the pseudo-score estimation equation

$$U(\beta, F_0, \pi_0) = \sum_{\delta_i=1} U_\beta(Y_i|S_i, Z_i, X_i) + \sum_{\delta_i=0} \frac{\int U_\beta(Y_i|s, Z_i, X_i) \frac{P(Y_i|s,Z_i,X_i)}{P(\delta=1)|s,X_i)} dF(s|X_i, \delta = 1)}{\int \frac{P(Y_i|s,Z_i,X_i)}{P(\delta=1)|s,X_i)} dF(s|X_i, \delta = 1)}. \quad (3.3)$$

For the BIP-only design, the cumulative distribution function $F_0 \equiv F(s|x, \delta = 1)$ is estimated using data from the second phase while for the BIP+CPV design it is estimated using data from both the second phase and the CPV component. That is, we estimate F_0 empirically by $F_N \equiv \frac{\sum_{\delta_i=1} I(S_i \le s, X_i = x)}{\sum_{\delta_i=1} I(X_i = x)}$. The probability $\pi_0 \equiv P(\delta = 1|S, X) = \int\int P(\delta = 1|y, z, S, X) P(y, z|S, X) dydz = \int\int P(\delta = 1|y, z, X) P(y|S, z, X) P(z) dydz$ is estimated by $\hat{\pi} \equiv \hat{P}(\delta = 1|S, X) = \sum_{z=0}^{1} \sum_{y=0}^{1} \hat{P}(\delta = 1|y, z, X) \hat{P}(y|S, z, X) P(z)$. By substituting F_0 and π_0 in expression (3.3) with their estimates F_N and $\hat{\pi}$, we obtain the estimating function $U(\beta, F_N, \hat{\pi})$. The pseudo-score estimator for parameters β in the risk model $risk_Z\{S, X\} = g\{\beta; S, Z, X\}$ is then obtained by solving the equation $U(\beta, F_N, \hat{\pi}) = 0$.

3.2.3 Estimating the Marginal CEP Curve for general settings

We consider one approach to estimating $\mathrm{VE}(s_1)$ with a two-phase sampling design and a possible CPV component, without assuming A6, and call this estimator $\widehat{\mathrm{VE}}^{new}(s_1)$.

Assuming the study participants make up a random sample from a large population of interest, the observed data $O_i \equiv (Z_i, X_i, Y_i^\tau, Y_i, \delta_i, S_i\delta_i)'$, $i = 1, ..., n$, are i.i.d. copies of a random vector $O \equiv (Z, X, Y^\tau, Y, \delta, S\delta)'$. We adopt all assumptions HGW made except for A6. We consider situations where X is categorical with D levels: $x_1, x_2, ..., x_D$. Then

$$\mathrm{VE}(s_1) = 1 - \frac{\sum_{j=1}^{D} g\left\{\beta; S = s_1, Z = 1, X = x_j\right\} \cdot P(X = x_j | S = s_1)}{\sum_{j=1}^{D} g\left\{\beta; S = s_1, Z = 0, X = x_j\right\} \cdot P(X = x_j | S = s_1)}.$$

The parameter β can be estimated by the pseudo-score type estimators following HGW's approach as summarized in section 3.2.2. We propose to model $P(X|S)$ with a multinomial logistic function. Suppose the multinomial logistic function can be represented as:

$$\begin{aligned}
ln(\frac{Pr(X = x_1)}{Pr(X = x_D)}) &= \gamma_{10} + \gamma_{11}h_1(s_1) + \gamma_{12}h_2(s_1) + ... + \gamma_{1T}h_T(s_1) = \boldsymbol{\gamma_1}^T \boldsymbol{h}(s_1) \\
ln(\frac{Pr(X = x_2)}{Pr(X = x_D)}) &= \boldsymbol{\gamma_2}^T \boldsymbol{h}(s_1) \\
&\vdots \\
ln(\frac{Pr(X = x_{D-1})}{Pr(X = x_D)}) &= \boldsymbol{\gamma_{D-1}}^T \boldsymbol{h}(s_1),
\end{aligned}$$

where h_1, h_2,..., h_T are pre-specified functions, such as polynomial functions or basis functions of the natural cubic spline. Then,

$$\begin{aligned}
P(X = x_1 | S = s_1) &= \frac{e^{\boldsymbol{\gamma_1^T h}(s_1)}}{1 + \sum_{d=1}^{D-1} e^{\boldsymbol{\gamma_d^T h}(s_1)}} \equiv p_1(s_1; \gamma) \\
P(X = x_2 | S = s_1) &= \frac{e^{\boldsymbol{\gamma_2^T h}(s_1)}}{1 + \sum_{d=1}^{D-1} e^{\boldsymbol{\gamma_d^T h}(s_1)}} \equiv p_2(s_1; \gamma) \\
&\vdots \\
P(X = x_D | S = s_1) &= \frac{1}{1 + \sum_{d=1}^{D-1} e^{\boldsymbol{\gamma_d^T h}(s_1)}} \equiv p_D(s_1; \gamma).
\end{aligned}$$

We propose to estimate γ by the Weighted Likelihood (WL) approach, with the weights being the inverse probabilities of sampling S within the cohort with $Y^\tau = Y^\tau(1) = Y^\tau(0) = 0$. To simplify the notation somewhat, we let $\pi_{0i} \equiv P(\delta_i = 1 | y_i, z_i, x_i)$ where $i = 1, 2, ..., n$. Given that X has D levels, we construct the likelihood function with D binary variables coded

as 0 or 1 to indicate the group membership of an observation regarding X: $G_1, G_2, ..., G_D$. If $X = x_d$, then $G_d = 1$ and all other $G's = 0$. Then the likelihood function for $P(X|S)$ can be derived as $L(\boldsymbol{\gamma}) = \prod_{i=1}^{n} p_1(s_{1i};\gamma)^{g_{1i}} \cdot p_2(s_{1i};\gamma)^{g_{2i}} \cdots p_D(s_{1i};\gamma)^{g_{Di}}$. In Appendix A, we show that the estimating function can be found by taking the first partial derivatives of the log-likelihood function $l(\boldsymbol{\gamma})$ with respect to each of the $D(T+1)$ unknown parameters: $\frac{\partial l(\boldsymbol{\gamma})}{\partial \gamma_{jt}} = \sum_{i=1}^{n} h_t(s_{1i})(g_{ji} - p_j(s_{1i};\gamma))$ where $j = 1, 2, ..., D$ and $t = 0, 1, 2, ..., T$, with $h_0(s_{1i}) = 1$ for each subject. Incorporating inverse probability weighting, the WL estimator $\hat{\gamma}$ is obtained by solving

$$\Psi(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi_{0i}} h_t(s_{1i})(g_{ji} - p_j(s_{1i};\gamma)) = 0. \tag{3.4}$$

When π_0 is unknown and needs to be estimated with a consistent estimator $\hat{\pi}_\alpha$, where α is the parameters estimated by ML from the Phase-I observations, π_0 will be substituted with $\hat{\pi}_\alpha$ in equation (3.4) to obtain $\hat{\gamma}$. Based on estimators $\hat{\beta}$ and $\hat{\gamma}$, we estimate $\text{VE}(s_1)$ with

$$\widehat{\text{VE}}^{(new)}(s_1) = 1 - \frac{\sum_{j=1}^{D} g\left\{\hat{\beta}; S = s_1, Z = 1, X = x_j\right\} \cdot p_j(s_1;\hat{\gamma})}{\sum_{j=1}^{D} g\left\{\hat{\beta}; S = s_1, Z = 0, X = x_j\right\} \cdot p_j(s_1;\hat{\gamma})}. \tag{3.5}$$

The asymptotic normality of $\widehat{\text{VE}}^{(new)}(s_1)$ is summarized in Theorem 1 in Appendix A followed by the proof.

3.3 Pointwise and Simultaneous Confidence Bands

Drawing simultaneous inference for the VE curve over a range of biomarker values is of interest in biomarker-defined principal strata effect modification analysis. For example, to evaluate whether conditional vaccine efficacy departs from a specific value, ve, for all S's in the range of $[s_l, s_u]$, the null hypothesis to be tested is H_0: $\text{VE}(s_1) = ve$ for all $s_1 \in [s_l, s_u]$. To evaluate whether the VE curve of one biomarker S equals the VE curve of another biomarker S' for values in the range of $[s_l, s_u]$, the null hypothesis to be tested is H_0: $\text{VE}(S = s_1) = \text{VE}(S' = s_1)$ for all $s_1 \in [s_l, s_u]$. Similarly, to evaluate whether the VE curve for trial 1 equals the VE curve for trial 2, the null hypothesis is H_0: $\text{VE}(s_1|X_1 = 1) = \text{VE}(s_1|X_1 = 0)$

for $s_1 \in [s_l, s_u]$, where X_1 is the indicator of trial 1 (i.e., $X_1 = 1$ for trial 1 and $X_1 = 0$ for trial 2). Such simultaneous inference typically involves estimation of the distribution of a process, which is often not tractable explicitly. Furthermore, explicit variance estimation might not be feasible and/or reliable. We propose a perturbation resampling method to approximate the distribution of our estimator for $\mathrm{VE}(s_1)$ and to draw simultaneous inference.

The construction of simultaneous confidence bands requires approximating the distribution of a Gaussian process $\widehat{W}_{s_1} \equiv \sqrt{n}\left\{\widehat{\mathrm{VE}}^{(new)}(s_1) - \mathrm{VE}_0(s_1)\right\}$. We propose a perturbation resampling procedure that provides a valid estimate for the distribution of $\widehat{W}_{s_1}$, based on which we construct the pointwise confidence intervals and simultaneous confidence bands for the VE curve. Because VE ranges from negative infinity to 1, we perform our estimation on the log scale of relative risk (RR), where $RR(s_1) = 1 - \mathrm{VE}(s_1)$. To be specific, the perturbation estimation can be carried out using the following resampling procedure:

1. Generate n random realizations of ϵ from a known distribution with mean of 1 and variance of 1 to create $\mathcal{E} \equiv \{\epsilon_i, i = 1, 2, ..., n\}$.

2. Use $\mathcal{E}$ to obtain the perturbed estimator $\hat{\beta}^{(\epsilon)}$ by solving $U^{(\epsilon)}(\beta, F_N^{(\epsilon)}, \hat{\pi}^{(\epsilon)}) = 0$, where

$$\begin{aligned}
U^{(\epsilon)}(\beta, F_N^{(\epsilon)}, \hat{\pi}^{(\epsilon)}) &= \sum_{\delta_i=1} U_\beta(Y_i|S_i, Z_i, X_i) \cdot \epsilon_i \\
&+ \sum_{\delta_i=0} \frac{\int \epsilon_i \cdot U_\beta(Y_i|s, Z_i, X_i) \frac{P(Y_i|s,Z_i,X_i)}{\hat{P}^{(\epsilon)}(\delta=1|s,X_i)} dF_N^{(\epsilon)}(s|X_i, \delta = 1)}{\int \frac{P(Y_i|s,Z_i,X_i)}{\hat{P}^{(\epsilon)}(\delta=1)|s,X_i)} dF_N^{(\epsilon)}(s|X_i, \delta = 1)}, \\
F_N^{(\epsilon)}(s|x, \delta = 1) &= \frac{\sum_{\delta_i=1} I(S_i \le s, X_i = x) \cdot \epsilon_i}{\sum_{\delta_i=1} I(X_i = x) \cdot \epsilon_i}, \\
\hat{\pi}^{(\epsilon)}(S_i, X_j) &= \hat{P}^{(\epsilon)}(\delta = 1|S_i, X_j) = \sum_{z=0}^{1} \sum_{y=0}^{1} \hat{P}^{(\epsilon)}(\delta = 1|y, z, X_j) P(y|S_i, z, X_j) P(z), \\
\hat{P}^{(\epsilon)}(\delta = 1|y, z, X_j) &= \frac{\sum_k I(\delta = 1, Y_k = y, Z_k = z, X_k = X_j) \cdot \epsilon_k}{\sum_k I(Y_k = y, Z_k = z, X_k = X_j) \cdot \epsilon_k}.
\end{aligned}$$

3. Use $\mathcal{E}$ to obtain the perturbed estimator $\hat{\gamma}^{(\epsilon)}$ by solving

$$\Psi^{(\epsilon)}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi_{0i}} h_t(s_{1i})(g_{ji} - p_j(s_{1i}; \gamma)) \cdot \epsilon_i = 0$$

when π_0 is known, or

$$\Psi^{(\epsilon)}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}^{(\epsilon)}(y_i, z_i, x_i)} h_t(s_{1i})(g_{ji} - p_j(s_{1i};\gamma)) \cdot \epsilon_i = 0$$

when π_0 is unknown, where $\hat{\pi}^{(\epsilon)}(y_i, z_i, x_i) = \hat{P}^{(\epsilon)}(\delta = 1 | y_i, z_i, x_i)$, which is defined in Step 2.

4. With $\hat{\beta}^{(\epsilon)}$ and $\hat{\gamma}^{(\epsilon)}$, we obtain the perturbed version of $\widehat{logRR}(s_1)$ as:

$$\widehat{logRR}^{(\epsilon)}(s_1) = log\left\{\frac{risk_1^{(\epsilon)}(s_1)}{risk_0^{(\epsilon)}(s_1)}\right\} = log\left\{\frac{\sum_{j=1}^{D} g\left\{\hat{\beta}^{(\epsilon)}; S = s_1, Z = 1, X = x_j\right\} \cdot p_j(s_1; \hat{\gamma}^{(\epsilon)})}{\sum_{j=1}^{D} g\left\{\hat{\beta}^{(\epsilon)}; S = s_1, Z = 0, X = x_j\right\} \cdot p_j(s_1; \hat{\gamma}^{(\epsilon)})}\right\}.$$

Repeat Steps 1–4 B_0 times to obtain B_0 realizations of $\widehat{logRR}^{(\epsilon)}(s_1)$, denoted by

$$\left\{\widehat{logRR}^{(b)}(s_1), b = 1, 2, ..., B_0\right\}.$$

The empirical distribution of $\widehat{W}_{logRR}(s_1)^{(b)} \equiv \sqrt{n}\left\{\widehat{logRR}^{(b)}(s_1) - \widehat{logRR}(s_1)\right\}$ conditional on the observed data can be used to approximate the distribution of

$$\widehat{W}_{logRR}(s_1) \equiv \sqrt{n}\left\{\widehat{logRR}(s_1) - logRR_0(s_1)\right\}.$$

In Appendix A, we provide theoretical justification for why the distribution of $\widehat{W}_{logRR}(s_1)^{(b)}$ given the observed data $O_i = (Z_i, X_i, Y_i^{\tau}, Y_i, \delta_i, S_i\delta_i)'$, $i = 1, ..., n$ can be used to approximate the unconditional distribution of $\widehat{W}_{logRR}(s_1)$. With the above resampling method, one may calculate the sample standard deviation, $\hat{\sigma}_{logRR}(s_1)$ of the B_0 realizations $\left\{\widehat{logRR}^{(b)}(s_1), b = 1, 2, ..., B_0\right\}$. A $100(1-\alpha)\%$ pointwise confidence interval and simultaneous confidence band for $\{logRR(s_1), s_1 \in \zeta\}$ may be constructed as $\widehat{logRR}(s_1) \pm \mathcal{Z}_{1-\alpha/2}\hat{\sigma}_{logRR}(s_1)$ and $\widehat{logRR}(s_1) \pm \mathcal{Q}_{1-\alpha}\hat{\sigma}_{logRR}(s_1)$, respectively, where $\mathcal{Z}_\eta$ is the 100ηth percentile of $N(0,1)$ and $\mathcal{Q}_\eta$ is the 100ηth percentile of $\left\{sup_{s_1\in\zeta}\hat{\sigma}_{logRR}(s_1)^{-1}\left|\widehat{W}_{logRR}(s_1)^{(b)}\right|, b = 1, 2, ..., B_0\right\}$. Finally, the Wald $100(1-\alpha)\%$ pointwise and simultaneous confidence bands for $\text{VE}^{(new)}(s_1)$ are obtained by transformation of the symmetric bounds from the $logRR$ scale back to the VE scale.

3.4 Simulation Studies

We conducted simulation studies to examine the finite-sample performance of our proposed estimator $\widehat{\mathrm{VE}}^{(new)}(s_1)$ and the perturbation resampling procedure for inference. For comparison, we also studied $\widehat{\mathrm{VE}}^{(A6)}(s_1)$, the estimator that makes the extra assumption A6 that after accounting for Z and S the conditional risk does not depend on X, as well as a traditional bootstrap procedure for making pointwise and simultaneous inference. Specifically, we assessed (1) the bias of $\widehat{\mathrm{VE}}^{(new)}(s_1)$ compared to $\widehat{\mathrm{VE}}^{(A6)}(s_1)$, (2) the distribution of the estimated standard errors of $log\widehat{RR}(s_1)$ obtained by the perturbation resampling procedure compared to the traditional bootstrap procedure, and (3) the empirical coverage probabilities of the resulting pointwise and simultaneous confidence bands based on perturbation compared to the bootstrap.

Simulated data followed a 2:1 vaccine:placebo randomized two-arm trial with 3000 subjects (2000 vaccine recipients and 1000 placebo recipients). The variable S was generated from a normal distribution with mean 3 and standard deviation 1. Simulated values of S less than 0 were truncated to 0. X was generated to have four categories, 1, 2, 3, 4, from the following multinomial model conditional on S:

$$ln\left(\frac{P(X=1)}{P(X=4)}\right) = -1.99{+}0.89S;\ ln\left(\frac{P(X=2)}{P(X=4)}\right) = -4.80{+}1.84S;\ ln\left(\frac{P(X=3)}{P(X=4)}\right) = -9.79{+}3.29S.$$

The parameters in the multinomial model were chosen such that there were about equal numbers of subjects in each of the four categories and the intraclass correlation [21] between X and S was 0.5. The risk model $P(Y=1|S,Z,X)$ was assumed to take the form $P(Y=1|S,Z,X) = \Phi(\beta_0+\beta_1 Z+\beta_2 S+\beta_3 SZ+\beta_4 X)$ with Φ denoting the cdf of the standard normal distribution. We chose the risk model parameters under three different scenarios: (1) the probability of $Y=1$ in the placebo arm (r_0) equals 0.090 and the probability of $Y=1$ in the vaccine arm (r_1) equals 0.042, (2) $r_0=0.055$ and $r_1=0.020$, (3) $r_0=0.0090$ and $r_1=0.0068$. We call these three scenarios the Non-Rare case, the Med-Rare case, and the Rare case, respectively, and they reflect the intention-to-treat cohort arm-specific probability of dengue disease in CYD14 (a phase 3 dengue vaccine efficacy trial conducted

in Asia [1]) and in CYD15 (a harmonized phase 3 dengue vaccine efficacy trial conducted in Latin America [26]) and of HIV infection in RV144 (a phase 3 HIV vaccine efficacy trial conducted in Thailand [22]), respectively. We also studied the performance of our estimators under different values of two ratios for two-phase sampling: r_v, the average ratio of sampled controls to cases in the vaccine arm; and r_p, the average ratio of sampled controls in the placebo arm to cases in the vaccine arm. Under the BIP-only design, $r_p = 0$ and S was treated as missing for all placebo recipients. We considered three values for r_v in our simulation: 5, 10, and *All*, where $r_v = All$ denotes the scenario where all vaccine recipients at-risk at τ had S measured. Under the BIP+CPV design, we considered three settings: $r_v = 5$ and $r_p = 5$; $r_v = 10$ and $r_p = 10$; and $r_v = All$ and $r_p = All$, the last of which denotes the scenario where all vaccine recipients had S measured and all event-free placebo recipients were included in the CPV component.

3.4.1 Different resampling strategies

We first study the coverage of simultaneous confidence band for three different resampling strategies: perturbation, bootstrap on case status and treatment arm (Y and Z), and bootstrap on treatment arm(Z) only. Because of the computational time it takes to perform resampling, we perform this comparison only in the BIP+CPV design Med-Rare case with $r_v = 5$ and $r_p = 5$; $r_v = 10$ and $r_p = 10$; or $r_v = All$ and $r_p = All$. The results are displayed in table 3.1, which shows that bootstrap on case status and treatment arm does not yield correct estimates. Next we studies in further detail the performance of perturbation resampling versus bootstrap resampling on treatment arm(Z) only. Henceforth, all bootstrap results refers to bootstrap resampling on Z only.

3.4.2 Bias of $\widehat{\mathrm{VE}}^{(A6)}(s_1)$ *and* $\widehat{\mathrm{VE}}^{(new)}(s_1)$

For each of 1000 simulated data sets, the estimates $\widehat{\mathrm{VE}}^{(A6)}(s_1)$ and $\widehat{\mathrm{VE}}^{(new)}(s_1)$ were computed. Figure 3.1 (3.2) displays the true VE curve, average $\widehat{\mathrm{VE}}^{(A6)}(s_1)$, and average $\widehat{\mathrm{VE}}^{(new)}(s_1)$ over the 1000 simulations for different sampling ratios in the Non-Rare, Med-Rare, and

Table 3.1: Coverage of 95% simultaneous confidence band for perturbation resampling, bootstrap resampling on case status and treatment arm (Y and Z), and bootstrap resampling on treatment arm(Z) only.

	$r_v = 5$ and $r_p = 5$	$r_v = 10$ and $r_p = 10$	$r_v = All$ and $r_p = All$
Perturbation	97.3%	96.9%	96.3%
Bootstrap on Y and Z	46.2%	50.0%	63.8%
Bootstrap resampling on Z only	98.3%	97.7%	97.3%

Rare case, respectively in a BIP+CPV (BIP-only) design. It also reports the average sampling proportions of S for each treatment arm $(\overline{\pi}_v, \overline{\pi}_p)$, with the sampling proportions in each of the 1000 simulated data sets calculated as: $\pi_v = \frac{\sum_{i=1}^{n} I(\delta_i=1, Y_i=0, Z_i=1)}{\sum_{i=1}^{n} I(Y_i=0, Z_i=1)}$ and $\pi_p = \frac{\sum_{i=1}^{n} I(\delta_i=1, Y_i=0, Z_i=0)}{\sum_{i=1}^{n} I(Y_i=0, Z_i=0)}$. It can been seen in Figure 3.1 that the $\widehat{\text{VE}}^{(A6)}(s_1)$ estimator could significantly underestimate the true VE for small values of S. For example, the bias in $\widehat{\text{VE}}^{(A6)}(s_1)$ at $s_1 = 0.5$ for $r_v = 5$ and $r_p = 5$ was -73% in the Rare case, -67% in the Med-Rare case, and -37% in the Non-Rare case. On the other hand, the bias for $\widehat{\text{VE}}^{(new)}(s_1)$ is generally negligible across all scenarios. Figure 3.2 shows a similar pattern. These results confirmed the superior performance of our proposed estimator over the estimator making the assumption of (A6), $risk_Z\{S, X\} = risk_Z\{S\} = g\{\beta; S, Z\}$, when the assumption is violated.

3.4.3 Standard error estimator

To examine the finite-sample performance of our proposed resampling procedure, we calculated the estimated standard errors of $\widehat{logRR}(s_1)$ and $\widehat{\sigma}_{logRR}(s_1)$, obtained by the perturbation resampling procedure. For each of the 1000 simulated datasets, $B_0 = 500$ perturbations were used and $\mathcal{E} = \{\epsilon_i, i = 1, 2, ..., n\}$ were generated from the exponential distribution with rate 1. The results were not sensitive to the choice of the distribution of $\mathcal{E}$ and $B_0 = 500$ was generally sufficient to approximate standard errors well. We also estimated standard errors of $\widehat{logRR}(s_1)$ based on 500 bootstrap samples. For comparison, we calculated the

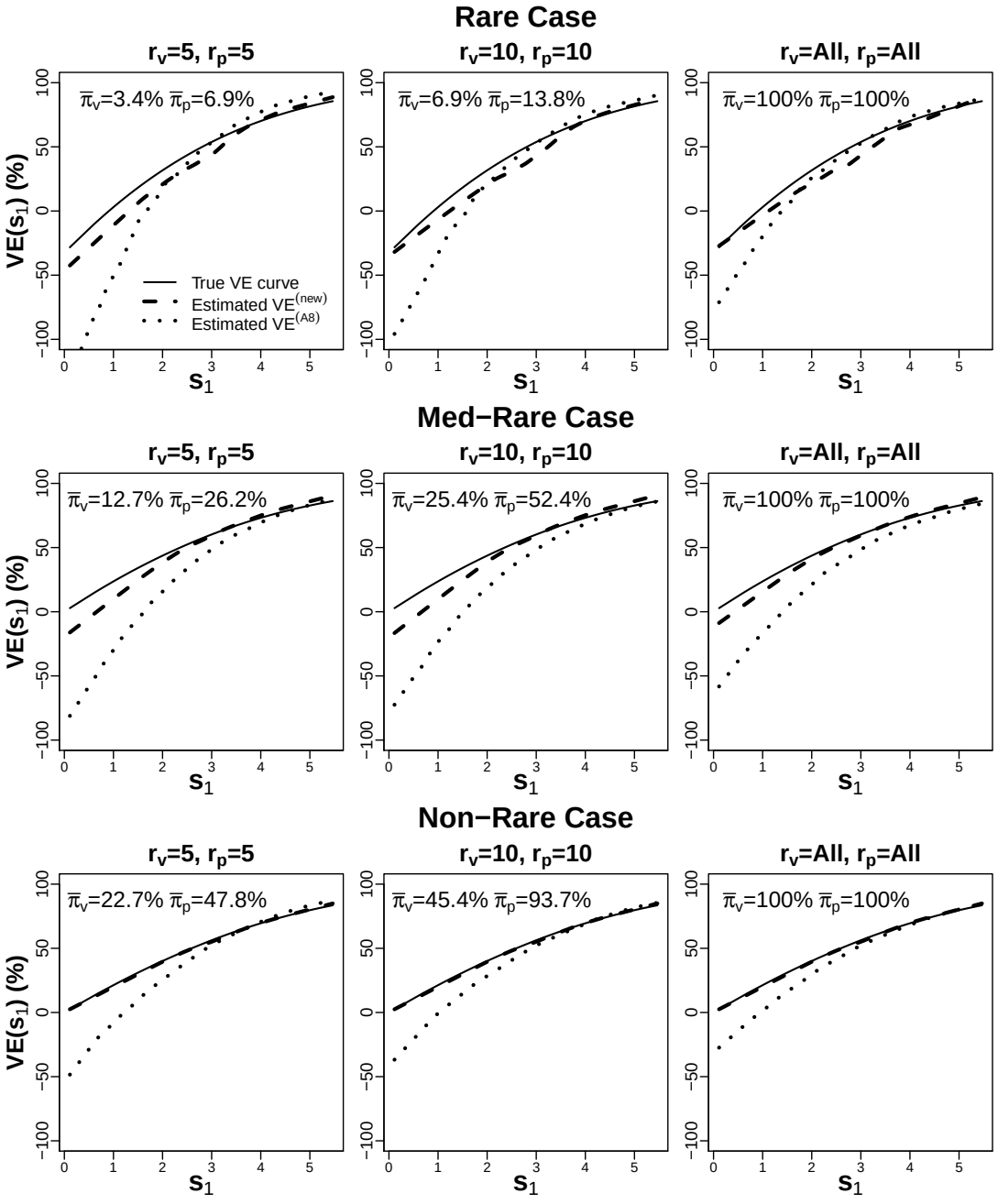

Figure 3.1: Estimated vaccine efficacy curve using our proposed method without assumption A6 and using the HGW method with assumption A6, compared to the true VE curve for checking the bias of these two estimators based on 500 simulated datasets for the Rare case where the probability of $Y = 1$ for $Z = 0$ (r_0) equals 0.090 and for $Z = 1$ (r_1) equals 0.042, the Med-Rare case where $r_0 = 0.055$ and $r_1 = 0.020$, and the Non-Rare case where $r_0 = 0.0090$ and $r_1 = 0.0068$ with a BIP+CPV design.

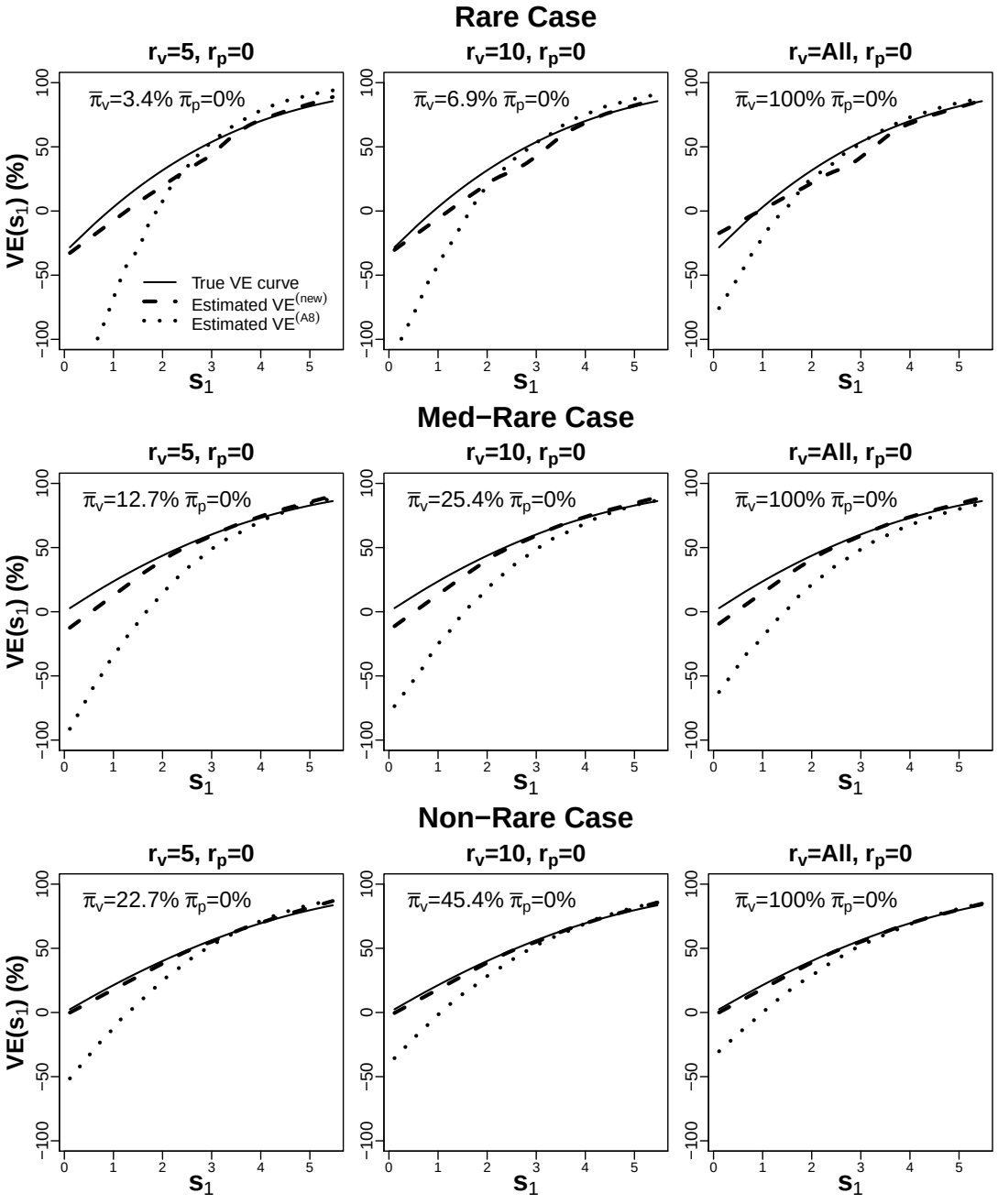

Figure 3.2: Estimated vaccine efficacy curve using our proposed method without assumption A6 and using the HGW method with assumption A6, compared to the true VE curve for checking the bias of these two estimators based on 500 simulated datasets for the Rare case where the probability of $Y = 1$ for $Z = 0$ (r_0) equals 0.090 and for $Z = 1$ (r_1) equals 0.042, the Med-Rare case where $r_0 = 0.055$ and $r_1 = 0.020$, and the Non-Rare case where $r_0 = 0.0090$ and $r_1 = 0.0068$ with a BIP-only design.

Monte Carlo standard errors as a benchmark for the correct standard errors. Averages of the estimated standard errors for various values of S are summarized in Figure 3.3 for a BIP+CPV design and Figure 3.4 for a BIP-only design. Generally the perturbation resampling procedure yielded standard error estimates closer to the Monte Carlo standard errors than the bootstrap procedure. When the disease endpoint is rare, perturbation-based SEs were remarkably smaller than the bootstrap-based SEs.

3.4.4 Coverage probabilities of pointwise and simultaneous confidence bands

We present coverage probabilities of the 95% pointwise confidence intervals (CIs) for $logRR(s_1)$ for different s_1 values in Figure 3.5 and Figure 3.6. In all scenarios, perturbation methods yielded coverage levels closer to the nominal 95% across most s_1 values compared to the bootstrap methods. The bootstrap CIs over-covered the truth, especially in the Rare case, due to the fact that the bootstrap-based SE estimates were remarkably large, yielding wide confidence intervals and high bootstrap CI coverage, while the perturbation-based SE estimator continued to perform well (coverage probability near the nominal level of 95%) as shown in Section 3.4.3. The empirical coverage levels of the 95% simultaneous confidence bands from the perturbation/bootstrap methods are also reported in Figure 3.5 and Figure 3.6 as "sim.cover.per"/"sim.cover.boot". Perturbation and bootstrap yielded similar simultaneous confidence band coverage, with coverage percentage close to the nominal 95% in the Med-Rare and the Non-Rare case. Based on these results, we conclude that the perturbation methods demonstrate a clear advantage over the the bootstrap methods, especially when the disease endpoint is rare, by producing smaller estimated SEs and therefore narrower pointwise confidence intervals and simultaneous confidence bands while maintaining nominal coverage.

In summary, our proposed procedure for estimating the VE curve had negligible bias and our proposed perturbation resampling method yielded confidence intervals with proper coverage probabilities. The perturbation-based SE estimators in general were smaller than the bootstrap-based SE estimators especially when the disease endpoint is rare.

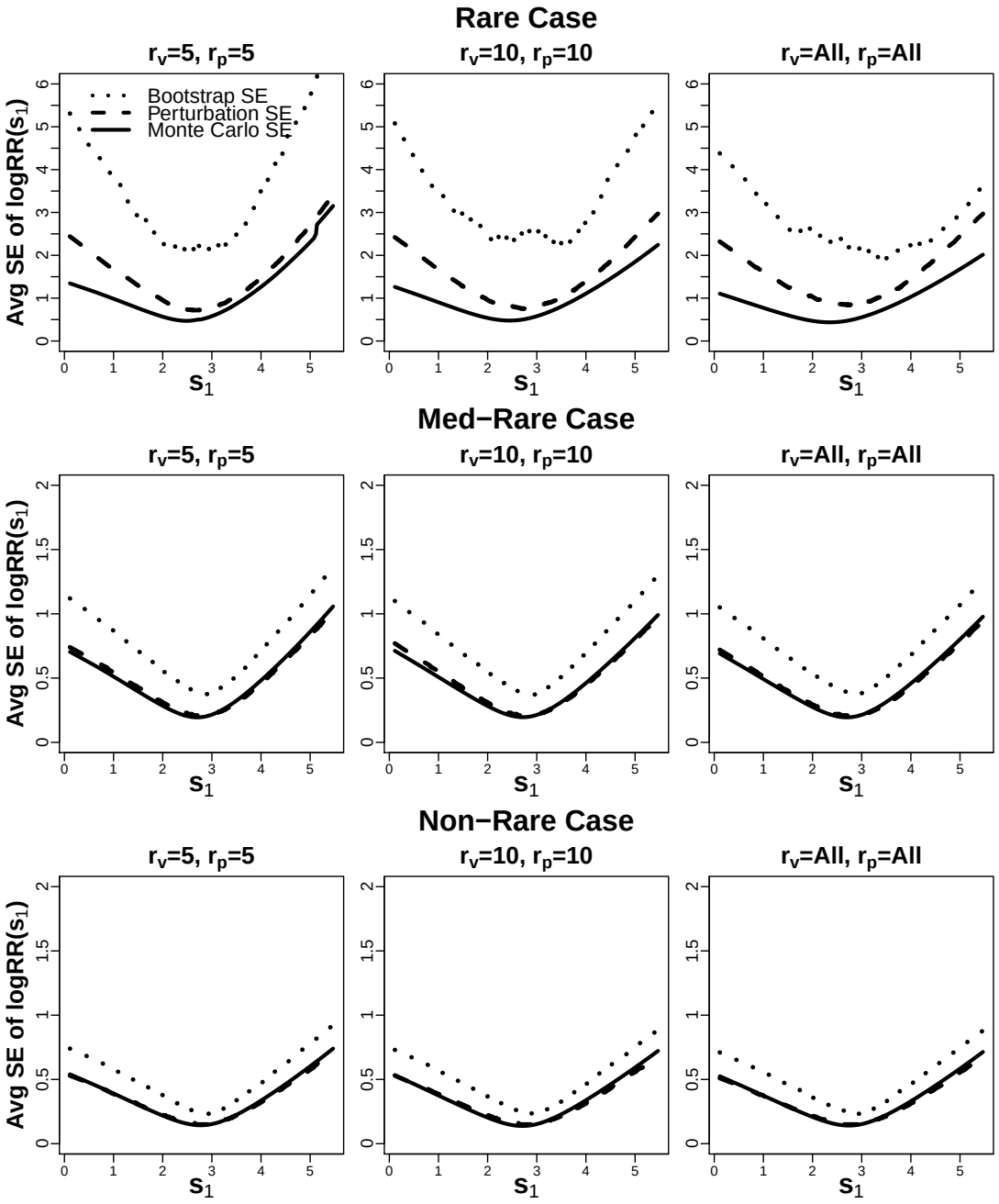

Figure 3.3: Estimated standard errors (SEs) of $log\widehat{RR}(s_1)$, solid for the Monte Carlo SEs, dashed for the perturbation resampling approach and dotted for the bootstrap approach, for the Rare case, the Med-Rare case, and the Non-Rare case with a BIP+CPV design.

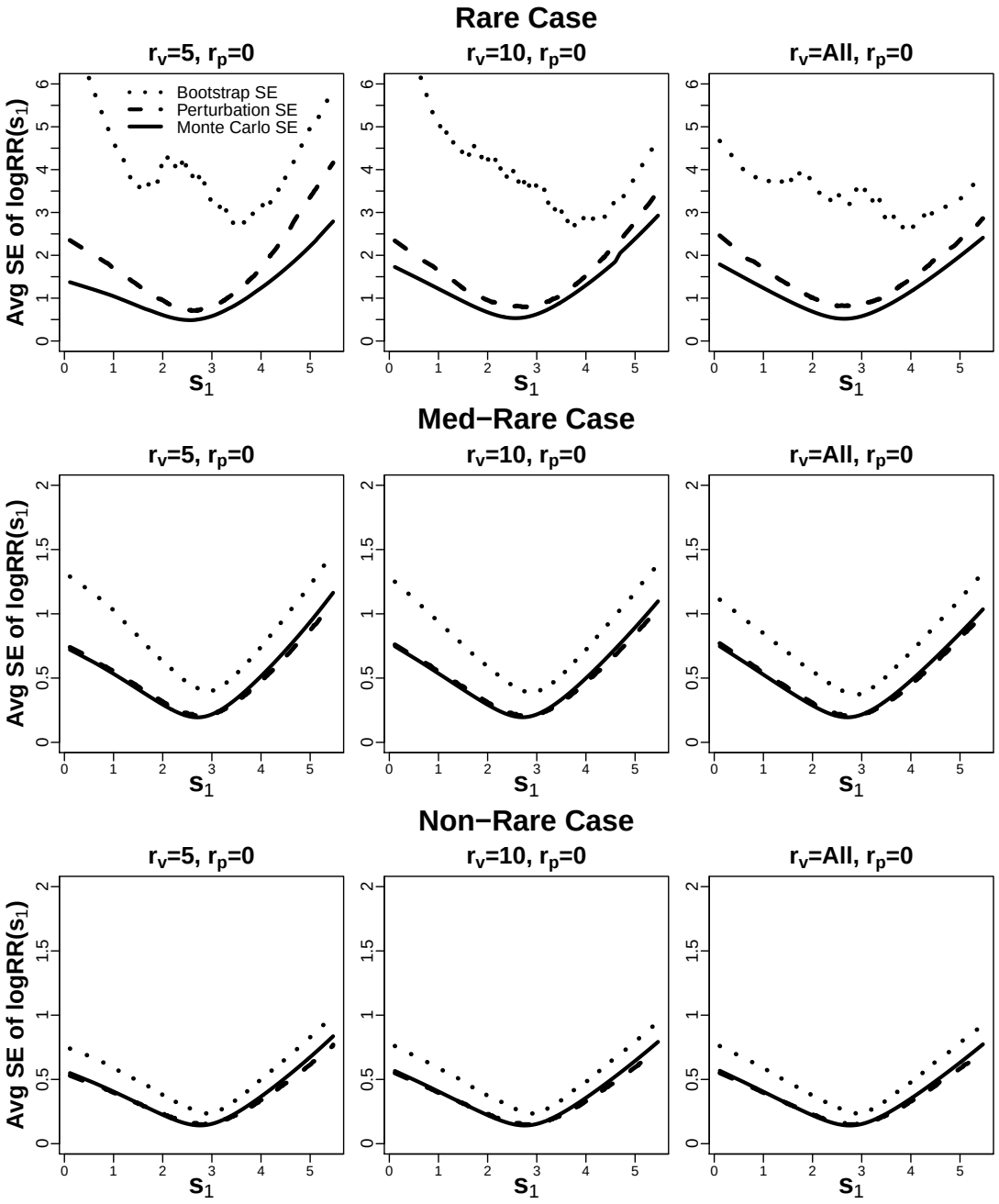

Figure 3.4: Estimated standard errors (SEs) of $log\widehat{RR}(s_1)$, solid for the Monte Carlo SEs, dashed for the perturbation resampling approach and dotted for the bootstrap approach, for the Rare case, the Med-Rare case and the Non-Rare case with a BIP-only design.

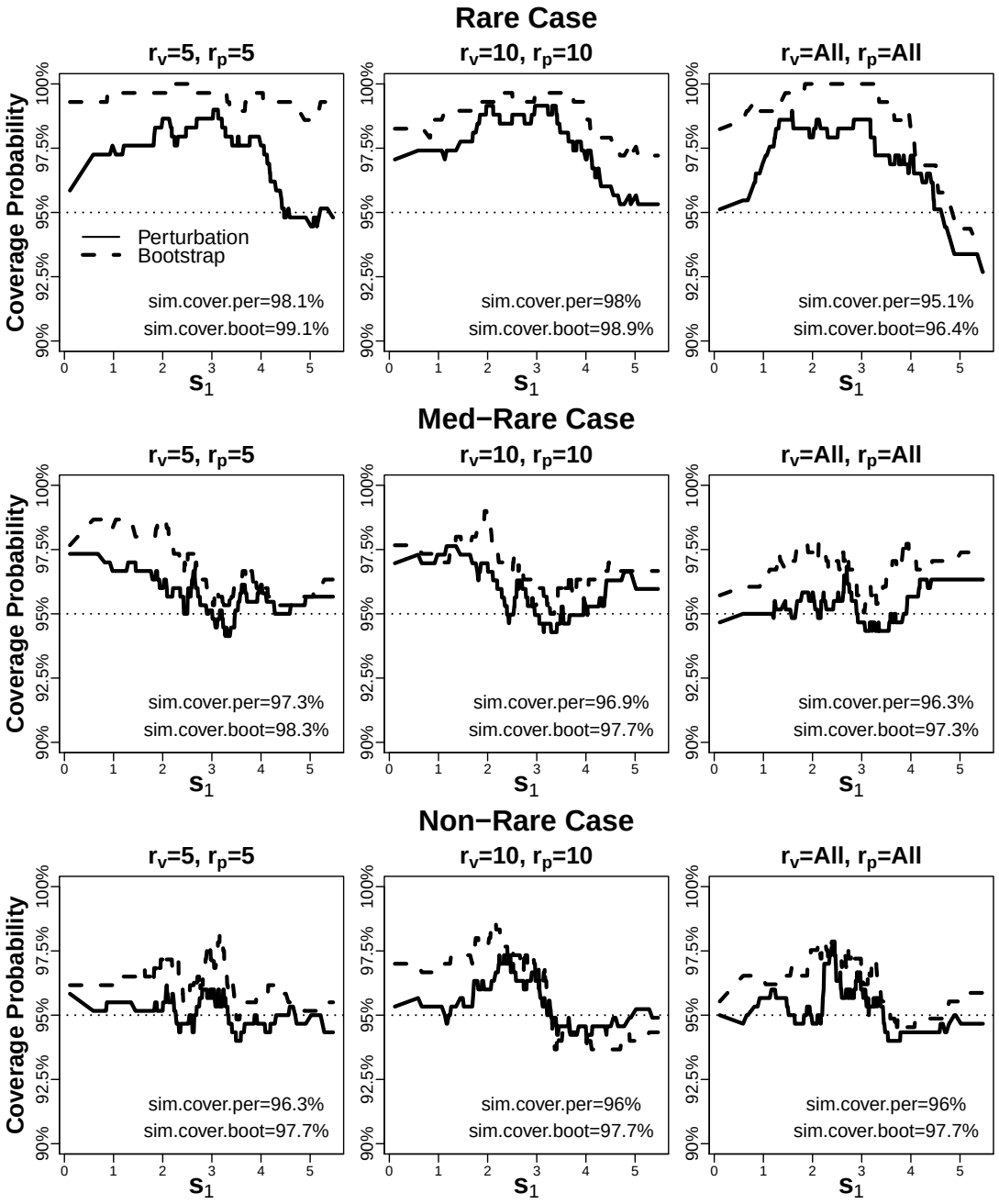

Figure 3.5: Empirical coverage probabilities of 95% pointwise confidence intervals and simultaneous confidence bands about VE(s_1), for the Rare case, the Med-Rare case, and the Non-Rare case with a BIP+CPV design.

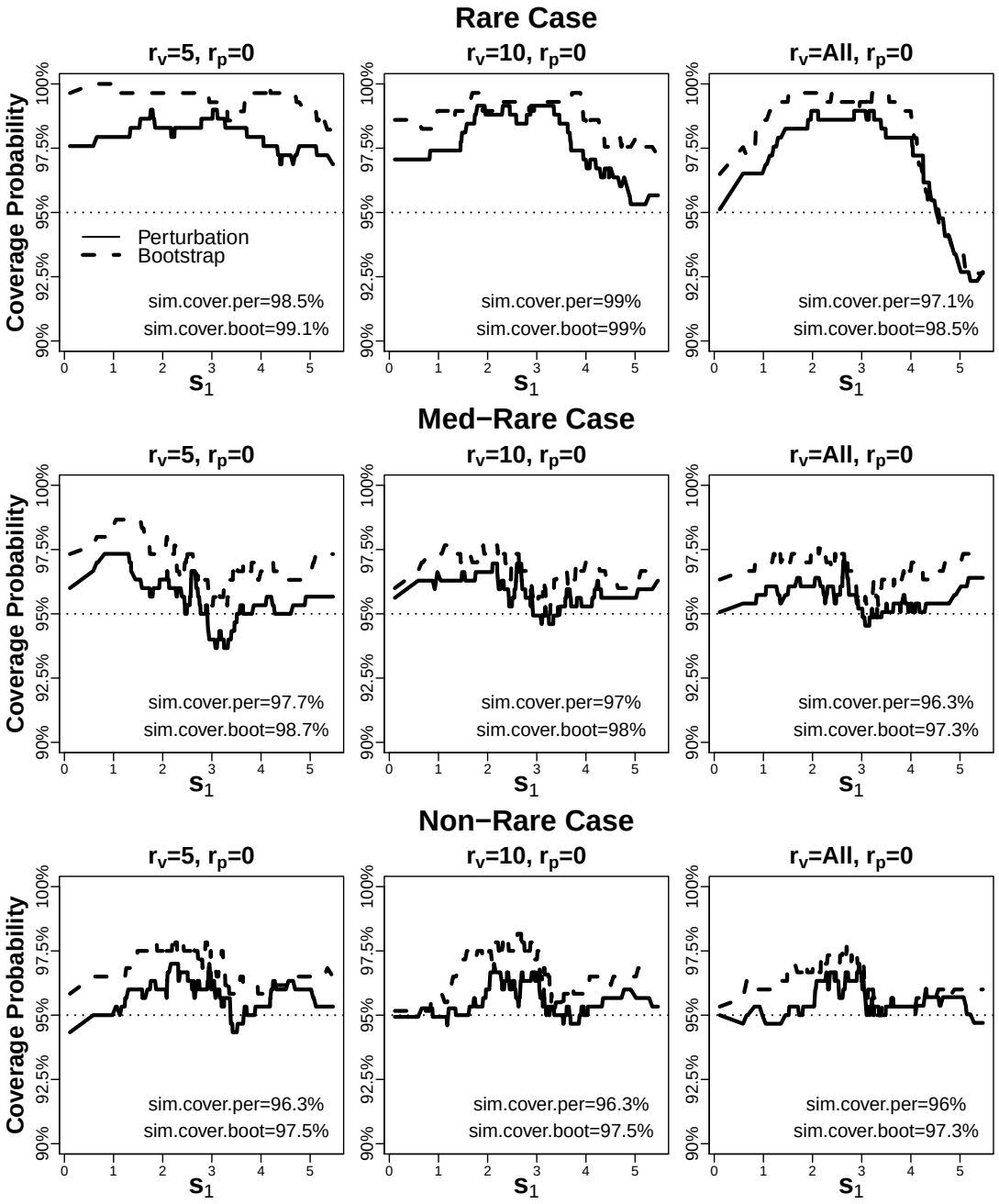

Figure 3.6: Empirical coverage probabilities of the 95% pointwise and simultaneous confidence bands about VE(s_1), for the Rare case, the Med-Rare case and the Non-Rare case with a BIP-only design.

3.5 Dengue Example

An brief background of trial CYD14 and CYD15 and meanings of notations Z, S, X, Y in the context of CYD14 and CYD15 9-16 year olds data are provided in chapter 2. We assume a probit risk model conditional on Z, S, and X: $risk_Z\{S, X\} = \Phi(\beta_0 + \beta_1 Z + \beta_2 S + \beta_3 SZ + \beta_4 X)$. Our proposed methods were conducted on the trial-pooled data set of 9–16 year olds to assess how VE varied with Month 13 neutralizing antibody titers if assigned to receive the vaccine. In Figure 3.7 we present the point estimates and pointwise and simultaneous confidence intervals for the VE curve, showing that VE against VCD increases with average titer. Estimated VE reached 42.1%, 85.6%, and 96.4%, respectively, at titer 100, 1000, and 10,000. The fact that a horizontal line cannot be drawn between the simultaneous confidence bands without intersecting both the lower and upper limits indicates that VE(s_1) significantly varies with s_1.

3.6 Discussion

In this chapter we have proposed a procedure for estimating the marginal CEP curve. Our proposed methods have the advantage of allowing clinical risks under Z being dependent on X after conditioning on S. We also developed procedures for obtaining pointwise and simultaneous confidence intervals about the marginal CEP curve via perturbation resampling. In addition, we have shown that pointwise and simultaneous interval estimates via perturbation resampling are more accurate and tighter than those obtained by the bootstrap, especially when the disease endpoint is rare.

While we have defined the baseline covariates X to be categorical and employed a multinomial model for $P(X|S)$, our proposed definition and estimation procedure can be extended to scenarios where X is continuous with a different parametric model for $P(X|S)$. Furthermore, this chapter assumed the baseline covariates X are phase-I variables that are available in all subjects. Extending our estimator to settings where only a subset of subjects are sampled for measurement of X is of interest, which is the topic of our resaerch in chapter 4.

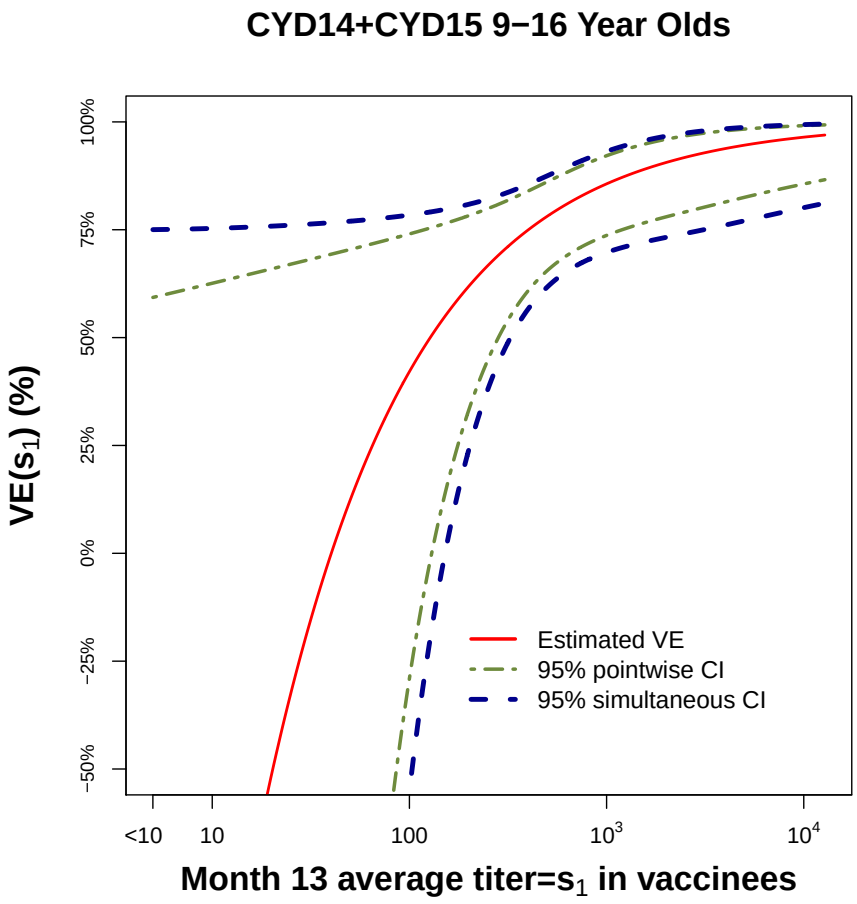

Figure 3.7: Estimated vaccine efficacy against dengue disease of any serotype through Month 25 by Month 13 average titers in vaccinees with 95% pointwise confidence intervals and simultaneous confidence bands in 9–16 year olds in the two Phase 3 dengue vaccine efficacy trials combined (CYD14 and CYD15).

Finally, like many other methods developed in the principal stratification/effect modification framework (e.g., [13], [6]), we have adopted assumption A3 in Gilbert and Hudgens (2008), which states equal early clinical risk: $P\left(Y^{\tau}(1) = Y^{\tau}(0)\right) = 1$. This assumption is important because the risk parameters of interest condition on $Y^{\tau}(1) = Y^{\tau}(0) = 0$. However, A3 is questionable if the treatment is efficacious early on or the biomarkers of interest are measured a long time after baseline. For the dengue application, assumption A3 is rejected based on the data, since the rate of dengue by Month 13 was significantly higher in the placebo group than the vaccine group ([1], [26]). Therefore the results should be interpreted with caution, and this example motivates the research in chapter 5 to relax A3.

Chapter 4

EVALUATION OF BIVARIATE TREATMENT EFFECT MODIFICATION BY BIOMARKER-BASED PRINCIPAL STRATA AND BASELINE COVARIATES IN THE PRESENCE OF NON-MONOTONE MISSINGNESS

4.1 Introduction

In chapter 2 and 3, we introduced the BIP design [4] that uses baseline predictor(s) to infer the unobserved potential biomarker values for evaluating the VE curve. Examples of good baseline predictors are baseline biomarker measurements in trials where participants may enter the study with prior exposure and biomarker measurements at baseline reflects natural immunity arising from pre-trial exposure to the disease-causing pathogen. Furthermore, some baseline predictors may modify the VE curve and contrasting clinical risks under each treatment assignment may depend jointly on those baseline predictors and the biomarker values measured at a fixed time after randomization. Therefore, it is of interest to estimate the VE curve in those baseline predictors defined subgroups. For example, a common question that emerges from dengue Phase 3 trials is whether it is the new biomarker response generated by the vaccine over the baseline value or the absolute biomarker value achieved following vaccination that predicts clinical treatment efficacy. Comparing the VE curve within the baseline seropositive subgroup (defined as baseline biomarker value equal or above the lower limit of quantification (LLOQ)) to that within the baseline seronegative subgroup (defined as baseline biomarker value below the LLOQ) could provide important insights to this question. Previous methods allow this assessment if baseline predictors are measured in everyone. Huang (2017) [12] studied a three-phase sampling design in which immune response is further measured among a subset of participants for whom the baseline predictors are available.

However, additional complications in study design could happen in practice. For example, in the CYD14 and CYD15 phase 3 dengue vaccine trials, the biomarker values at baseline were only measured in a fraction of those with the biomarker measured at the post-randomization time point. Our goal in this chapter is to propose methods for a bivariate treatment effect modification analysis by biomarker-based principal strata and baseline covariates in general settings without requirements of a nested sub-sampling relationship between the immune response biomarker and baseline predictors, in other words, in the presence of non-monotone missingness, applicable to both the BIP-only design and the BIP+CPV design.

4.2 Method

Adopt notation $(Z, Y^\tau, Y, \delta, S)$ from Chapter 2. Furthermore, Let Q be a vector of baseline covariates used for modeling disease risk and Q can be partitioned into two part, X and B. X denotes baseline covariates recorded for everyone at baseline such as gender and country while B denotes baseline covariates that are only available in a subset of the n trial participants, such as baseline biomarker measurements. We also consider cases where S is continuous and subject to "limit of quantification" left censoring. The observable random variable $S \equiv max(S^*, c)$ where c is the limit of quantification and S^* has a continuous cdf with $Pr(S^* \leq c) > 0$. Similarly to S, if B denotes the baseline biomarker measurements, then observable random variable $B \equiv max(B^*, c)$ where B^* has a continuous cdf with $Pr(B^* \leq c) > 0$. We consider a general sampling framework where baseline covariates X and the clinical outcome data Y and Y^τ are measured for everyone, sampling probability of B depends on X, Z and Y, and sampling probability of S depends on X, Z and Y. We propose methods to estimate the causal estimand for bivariate treatment effect modification analysis $\text{VE}(\text{S}, \text{B}) \equiv 1 - \frac{\text{risk}_1(\text{S,B})}{\text{risk}_0(\text{S,B})}$, applicable to both the BIP-only design and the BIP+CPV design based on an estimated likelihood approach in the presence of non-monotone missingness. Furthermore, in the special case where B denotes the baseline biomarker values, we also derive the estimator for the baseline seropositive VE curve ($\text{VE}(\text{S}, \text{B} > \text{c}) \equiv 1 - \frac{\text{risk}_1(\text{S,B>c})}{\text{risk}_0(\text{S,B>c})}$) and the baseline seronegative VE curve ($\text{VE}(\text{S}, \text{B} = \text{c}) \equiv 1 - \frac{\text{risk}_1(\text{S,B=c})}{\text{risk}_0(\text{S,B=c})}$).

We adopt assumptions (A1), (A2), (A3), (A1-CPV) and (A2-CPV) from Chapter 2. These assumptions reduce the number of missing potential outcomes and help with identifiability of our estimands. Because S in this chapter is subject to LLOQ left-censoring and $S = max(S^*, c)$, a more accurate expression for assumption (A1-CPV) is:

(A1-CPV) Time constancy of immune response: For event-free placebo recipients, $S^*(1) = S^{*\text{true}} + U_1$, and $S^*_c = S^{*\text{true}} + U_2$, for some underlying $S^{*\text{true}}$ and i.i.d. measurement errors U_1, U_2 that are independent of one another.

We consolidate the notation and let S^* be the potential outcome of S^* under treatment arm $Z = 1$, either obtained during the standard trial follow-up for vaccine recipients or replaced by the CPV measurements for placebo recipients. We let $S \equiv max(S^*, c)$ and δ to be the indicator that S is measured. In addition, if B denotes the baseline biomarker values, we also replace a missing B with $S(0)$ if it is available for placebo recipients based on (A3) and the next assumption (A4):

(A4) $B^* = B^{*\text{true}} + U_3$, and $S^*(0) = B^{*\text{true}} + U_4$, for some underlying $B^{*\text{true}}$ and i.i.d. measurement errors U_3, U_4 that are independent of one another.

Henceforth, if B denotes the baseline biomarker values subject to lower quantification limit: $B = max(B^*, c)$, then we assume that B^* denotes the baseline biomarker values that could potentially being replaced by $S^*(0)$. We let $B \equiv max(B^*, c)$ and δ_B to be the indicator that B is available.

For the settings we consider in this chapter, $\{i : \delta_i = 1\}$ and $\{i : \delta_{Bi} = 1\}$ do not need to hold an inclusion relationship. In section 4.5, we discuss the special cases that $\delta = 1$ implies $\delta_B = 1$.

We assume the risk functions have a generalized linear model form:

(A5) $risk_z \{S, B, X\} = g \{\beta; S, B, Z, X\}$ for some known link function $g(\cdot)$, for $z = 0, 1$.

Lastly, we assume observed data $O_i \equiv (Z_i, X_i, \delta_i, \delta_i S_i, \delta_{Bi}, \delta_{Bi} B_i, Y_i^\tau, Y_i)', i = 1, ..., n$ are independent and identically distributed (i.i.d).

4.2.1 Risk Model Parameters Estimation

We propose an estimated likelihood estimator based on conditional likelihood for our risk model parameters β. Subjects with $\delta_{Bi} = \delta_i = 1$ contribute to likelihood $risk_{Z_i}(S_i, B_i, X_i; \beta)^{Y_i}(1 - risk_{Z_i}(S_i, B_i, X_i; \beta))^{1-Y_i}$. The likelihood contribution for subjects with $\delta_{Bi} = 1$ and $\delta_{Si} = 0$ is obtained by integrating $risk_{Z_i}(\cdot, B_i, X_i; \beta)$ over the conditional cdf $F^{S|B,X}$. The contribution for subjects with $\delta_i = 1$ and $\delta_{Bi} = 0$ is obtained by integrating $risk_{Z_i}(S_i, \cdot, X_i; \beta)$ over the conditional cdf $F^{B|S,X}$. The contribution for subjects with $\delta_i = \delta_{Bi} = 0$ is obtained by integrating $risk_{Z_i}(\cdot, \cdot, X_i; \beta)$ over the conditional cdf $F^{(B,S)|X}$. Define nuisance parameter $\nu \equiv \left(F^{B|X}, F^{S|B,X}, F^{B|S,X}\right)$. Then the condition likelihood is

$$L(\beta, \nu) \equiv \prod_{i=1}^{n} f(Y_i | Z_i, X_i, \delta_{Bi}, \delta_i, \delta_{Bi} B_i, \delta_i S_i)$$

where

$$\begin{aligned}
&f(Y|Z, X, \delta_B, \delta, \delta_B B, \delta S) \\
&= \left\{risk_Z(S, B, X; \beta)^Y (1 - risk_Z(S, B, X; \beta))^{1-Y}\right\}^{\delta_B \delta} \\
&\times \left\{ \left(\int risk_Z(s_1, B, X; \beta) dF^{S|B,X}(s_1|B, X) \right)^Y \right. \\
&\qquad \left. \times \left(1 - \int risk_Z(s_1, B, X; \beta) dF^{S|B,X}(s_1|B, X) \right)^{1-Y} \right\}^{\delta_B(1-\delta)} \\
&\times \left\{ \left(\int risk_Z(S, b, X; \beta) dF^{B|S,X}(b|S, X) \right)^Y \right. \\
&\qquad \left. \times \left(1 - \int risk_Z(S, b, X; \beta) dF^{B|S,X}(b|S, X) \right)^{1-Y} \right\}^{(1-\delta_B)\delta} \\
&\times \left\{ \left(\int \int risk_Z(s_1, b, X; \beta) dF^{S|B,X}(s_1|b, X) dF^{B|X}(b|X) \right)^Y \right. \\
&\qquad \left. \times \left(1 - \int \int risk_Z(s_1, b, X; \beta) dF^{S|B,X}(s_1|b, X) dF^{B|X}(b|X) \right)^{1-Y} \right\}^{(1-\delta_B)(1-\delta)}
\end{aligned}$$

We consider the estimated likelihood approach by Pepe and Fleming (1991) [18] where consistent estimates of ν are obtained first and then $L(\beta, \hat{\nu})$ is maximized in β. Here we assume $F^{B|X}$ and $F^{S|B,X}$ have particular parametric distribution. For example, we might assume $F^{B|X}$ is censored normal and $F^{S|B,X}$ is also censored normal, with left-censoring of values below c. Then according to Bayes' theorem, we have $f(B|S,X) = \frac{f(S|B,X)f(B|X)}{\int f(S|b,X)\cdot f(b|X)db}$. We obtain the maximum likelihood estimator (MLE) for $F^{B|X}$ using data from all individuals with B measured, $\{i : \delta_{Bi} = 1\}$. For estimation of $F^{S|B,X} = F^{S|B,X,Z=1}$, we use data from vaccine recipients who have both S and B measured, with inverse probability weighting (IPW) used to account for biased sampling of S. Even if there is a CPV component in the study design, we can not use S for placebo recipients obtained during CPV because placebo recipients who are infected at study closeout have zero probability of obtaining S thus IPW is not applicable. The estimator of β is then derived as the maximizer of the estimated likelihood $L(\beta, \hat{\nu})$ and $\widehat{\text{VE}}(S,B,X) = 1 - \frac{g\{\widehat{\beta};S,B,Z=1,X\}}{g\{\widehat{\beta};S,B,Z=0,X\}}$ provides an estimate of bivariate treatment effect modification by S and B, adjusting for X. Standard errors for $\widehat{\beta}$ can be estimated using a perturbation resampling technique, similar to chapter 3.3. In essence, one can generate n random realizations of ϵ from a known distribution with mean of 1 and variance of 1 to create $\mathcal{E} \equiv \{\epsilon_i, i = 1, 2, ..., n\}$. Let $L^{(\epsilon)}(\beta, \hat{\nu}^{(\epsilon)})$ be a perturbed version of $L(\beta, \hat{\nu})$, where $L^{(\epsilon)}(\beta, \nu) \equiv \prod_{i=1}^{n} f(Y_i|Z_i, X_i, \delta_{Bi}, \delta_i, \delta_{Bi}B_i, \delta_i S_i) \cdot \epsilon_i$ and $\nu^{(\epsilon)}$ is the perturbed estimator of ν with $\mathcal{E}$ being the weights. Then the perturbed estimator $\beta^{(\epsilon)}$ is derived as the maximizer of $L^{(\epsilon)}(\beta, \hat{\nu}^{(\epsilon)})$. In practice, one may obtain a variance estimator of $\hat{\beta}$ based on the empirical variance of M_0 realizations of $\beta^{(\epsilon)}$. In our simulation and example, we use $M_0 = 500$.

4.2.2 Baseline Seropositive/Seronegative VE Curves

In this section, we study the special case where B denotes the baseline biomarker values subject to the lower limit of quantification, c, and derive the estimator for the VE curve ($\text{VE(S)} \equiv 1 - \frac{\text{risk}_1(\text{S})}{\text{risk}_0(\text{S})}$), baseline seropositive VE curve ($\text{VE(S, B} > \text{c)} \equiv 1 - \frac{\text{risk}_1(\text{S,B>c})}{\text{risk}_0(\text{S,B>c})}$) and the baseline seronegative VE curve ($\text{VE(S, B} = \text{c)} \equiv 1 - \frac{\text{risk}_1(\text{S,B=c})}{\text{risk}_0(\text{S,B=c})}$). With some cal-

culations, the risk functions in our estimands of interest can be expressed as: marginal risk function $risk_Z(S,X) = \int_{c-}^{\infty} risk_Z(S,b,X)dF^{B|S,X}(b|S,X)$; seropositive risk function $risk_Z(S,B>c,X) = \frac{P(Y(Z)=1,B>c|S,X)}{P(B>c|S,X)} = \frac{\int_{c+}^{\infty} risk_Z(S,b,X)dF^{B|S,X}(b|S,X)}{P(B>c|S,X)}$; and seronegative risk function $risk_Z(S,B=c,X)$. All three risk functions can be estimated based on $\hat{\beta}$ and $\hat{\nu}$.

We consider situations where X is categorical with D levels: $x_1, x_2, ..., x_D$. Then

$$risk_Z(S) = \sum_{j=1}^{D} risk_Z(S,x_j) \cdot P(X=x_j|S) \tag{4.1}$$

$$risk_Z(S,B>c) = \sum_{j=1}^{D} risk_Z(S,B>c,x_j) \cdot P(X=x_j|S,B>c) \tag{4.2}$$

$$risk_Z(S,B=c) = \sum_{j=1}^{D} risk_Z(S,B=c,x_j) \cdot P(X=x_j|S,B=c). \tag{4.3}$$

In Appendix B.2, we show that based on the Bayes' theorem and with some calculations, we have

$$P(X|S) = \frac{\left(\int f(S|b,X)f(b|X)db\right) \cdot P(X)}{\sum_{i=1}^{D}\left(\int f(S|b,x_i)f(b|x_i)db\right) \cdot P(x_i)} \tag{4.4}$$

$$P(X|S,B>c) = \frac{\frac{\int_c^{\infty} f(S|b,X)f(b|X)db}{\int_c^{\infty} f(b|X)db}P(X|B>c)}{\sum_{i=1}^{D}\frac{\int_c^{\infty} f(S|b,x_i)f(b|x_i)db}{\int_c^{\infty} f(b|x_i)db}P(x_i|B>c)} \tag{4.5}$$

$$P(X|S,B=c) = \frac{f(S|B=c,X)P(X|B=c)}{\sum_{i=1}^{D} f(S|B=c,x_i)P(x_i|B=c)}. \tag{4.6}$$

Suppose probabilities $P(X)$, $P(X|B>c)$ and $P(X|B=c)$ can be modeled parametrically with parameters γ_1, γ_2, and γ_3, respectively. We substitute γ_1, γ_2, and γ_3 with their maximum likelihood estimators (MLEs) $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$ to obtain $\hat{P}(X)$, $\hat{P}(X|B>c)$, and $\hat{P}(X|B=c)$. For example, in the simulation studies described next, we apply a saturated model for the sampling probability of X conditional on B, such that $\hat{P}(X)$ equals the observed sampling fractions of X, $\hat{P}(X|B>c)$ equals the observed sampling fractions of X on the cohort with $\delta_B = 1$ and $B > c$, and $\hat{P}(X|B=c)$ equals the observed sampling fractions of X on the cohort with $\delta_B = 1$ and $B = c$. Because sampling of B may depend on other phase-I variables such as Y, inverse probability weighting (IPW) (Horvitz and Thompson, 1952) can

be implemented. Probabilities $P(X|S)$, $P(X|S, B > c)$, and $P(X|S, B = c)$ can then be estimated by $\hat{\nu}$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$ based on expression 4.4, 4.5, and 4.6.

Appendix B.3 provides a detailed estimation procedure of VE(S), VE(S, B > c), and VE(S, B = c) for the case where $F^{B|X}$ is assumed censored normal, $F^{S|B,X}$ is assume censored normal, and the risk functions take the form $risk_z\{S, B, X\} = g\{\beta; S, B, Z, X\} = \Phi(\beta_0 + \beta_1 Z + \beta_2 S + \beta_3 Z \cdot S + \beta_4 B + \beta_5 Z \cdot B + \beta_6 X)$.

A perturbation resampling method can be used to make simultaneous inference of the VE curve, baseline seropositive VE curve and baseline seronegative VE curve. Resampling procedures are similar to section 3.3. Because VE ranges from negative infinity to 1, we perform our estimation on the log scale of relative risk (RR), where $RR(s_1) = 1 - \text{VE}(s_1)$. To be specific, perturbed estimators $\beta^{(\epsilon)}$ and $\nu^{(\epsilon)}$ are obtained based on $\mathcal{E}$. Then the corresponding perturbed estimators $\widehat{logRR}^{(\epsilon)}(s_1)$, $\widehat{logRR}^{(\epsilon)}(s_1, b > c)$, and $\widehat{logRR}^{(\epsilon)}(s_1, b = c)$ are obtained by plugging in $\beta^{(\epsilon)}$ and $\nu^{(\epsilon)}$ in equation 4.1, 4.2, and 4.3. Repeat this process M_0 times to obtain M_0 realizations of $\widehat{logRR}^{(\epsilon)}(s_1)$, $\widehat{logRR}^{(\epsilon)}(s_1, b > c)$, and $\widehat{logRR}^{(\epsilon)}(s_1, b = c)$, denoted by $\left\{\widehat{logRR}^{(m)}(s_1), m = 1, 2, ..., M_0\right\}$, $\left\{\widehat{logRR}^{(m)}(s_1, b > c), m = 1, 2, ..., M_0\right\}$, and $\left\{\widehat{logRR}^{(m)}(s_1, b = c), m = 1, 2, ..., M_0\right\}$. Then calculate the sample standard deviations $\hat{\sigma}_{logRR}(s_1)$, $\hat{\sigma}_{logRR}(s_1, b > c)$, $\hat{\sigma}_{logRR}(s_1, b = C)$. $100(1-\alpha)\%$ pointwise confidence intervals can be constructed as

$$\widehat{logRR}(s_1) \pm \mathcal{Z}_{1-\alpha/2}\hat{\sigma}_{logRR}(s_1)$$
$$\widehat{logRR}(s_1, b > c) \pm \mathcal{Z}_{1-\alpha/2}\hat{\sigma}_{logRR}(s_1, b > c)$$
$$\widehat{logRR}(s_1, b = c) \pm \mathcal{Z}_{1-\alpha/2}\hat{\sigma}_{logRR}(s_1, b = c).$$

And $100(1-\alpha)\%$ simultaneous confidence bands for $s_1 \in \zeta$ can be constructed as

$$\widehat{logRR}(s_1) \pm \mathcal{Q}_{1-\alpha}\hat{\sigma}_{logRR}(s_1)$$
$$\widehat{logRR}(s_1, b > c) \pm \mathcal{Q}'_{1-\alpha}\hat{\sigma}_{logRR}(s_1, b > c)$$
$$\widehat{logRR}(s_1, b = c) \pm \mathcal{Q}''_{1-\alpha}\hat{\sigma}_{logRR}(s_1, b = c),$$

where $\mathcal{Z}_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of $N(0,1)$, $\mathcal{Q}_{1-\alpha}$ is the $100(1-\alpha)$th percentile of $\left\{ sup_{s_1\in\zeta} \left| \frac{\sqrt{n}\left\{\widehat{mCEP}^{(m)}(S) - \widehat{mCEP}(S)\right\}}{\hat{\sigma}_{mCEP}(S)} \right|, m = 1, 2, ..., M_0 \right\}$, and $\mathcal{Q}'_{1-\alpha}$ and $\mathcal{Q}''_{1-\alpha}$ defined similar to $\mathcal{Q}_{1-\alpha}$ with the VE estimator, VE perturbed estimator and standard error estimator replaced by its own version. Finally, the Wald $100(1-\alpha)\%$ pointwise and simultaneous confidence bands for $\mathrm{VE}(s_1)$, $\mathrm{VE}(s_1, b > c)$, and $\mathrm{VE}(s_1, b = c)$ are obtained by transformation of the symmetric bounds from the $logRR$ scale back to the VE scale.

Furthermore, simultaneous inference enables evaluation of the hypothesis testing of H_0: $\mathrm{VE}(\mathrm{s}_1, \mathrm{b} > \mathrm{c}) = \mathrm{VE}(\mathrm{s}_1, \mathrm{b} = \mathrm{c})$ for $s_1 \in \zeta$. We first construct the simultaneous confidence band for $\widehat{logRR}(s_1, b > c) - \widehat{logRR}(s_1, b = c)$. Let $\widehat{\sigma}(\widehat{logRR}(s_1, b > c) - \widehat{logRR}(s_1, b = c))$ denote the sample standard deviation of the perturbed estimates $\widehat{logRR}^{(\epsilon)}(s_1, b > c) - \widehat{logRR}^{(\epsilon)}(s_1, b = c)$. Let $\mathcal{Q}'''_{1-\alpha}$ be the $100(1-\alpha)$th percentile of $\left\{ sup_{S\in\zeta} \left| \frac{\sqrt{n}\left\{\widehat{logRR}^{(m)}(s_1,b>c) - \widehat{logRR}^{(m)}(s_1,b=c) - \left(\widehat{logRR}(s_1,b>c) - \widehat{logRR}(s_1,b=c)\right)\right\}}{\widehat{\sigma}(\widehat{logRR}(s_1,b>c) - \widehat{logRR}(s_1,b=c))} \right|, m = 1, 2, ..., M_0 \right\}$. Subsequently, the $100(1-\alpha)\%$ simultaneous confidence bands for $\widehat{logRR}(s_1, b > c) - \widehat{logRR}(s_1, b = c)$, $s_1 \in \zeta$ is

$$\widehat{logRR}(s_1, b > c) - \widehat{logRR}(s_1, b = c) \pm \mathcal{Q}'''_{1-\alpha}\widehat{\sigma}(\widehat{logRR}(s_1, b > c) - \widehat{logRR}(s_1, b = c)).$$

Lastly, the Wald $100(1-\alpha)\%$ pointwise and simultaneous confidence bands for $\mathrm{VE}(\mathrm{s}_1, \mathrm{b} > \mathrm{c}) - \mathrm{VE}(\mathrm{s}_1, \mathrm{b} = \mathrm{c})$ are obtained by transformation of the symmetric bounds from the $logRR$ scale back to the VE scale, denoted by $(l_\alpha(S), u_\alpha(S))$. Then, the two-sided p-value for the testing H_0: $\mathrm{VE}(\mathrm{s}_1, \mathrm{b} > \mathrm{c}) = \mathrm{VE}(\mathrm{s}_1, \mathrm{b} = \mathrm{c})$ is defined as the minimum of α_1 and α_2 that satisfy

$$\begin{array}{ll} \inf_{\mathrm{S}\in\zeta}\mathrm{u}_{\alpha_1}(\mathrm{S}) = 0, & \sup_{\mathrm{S}\in\zeta}\mathrm{l}_{\alpha_1}(\mathrm{S}) \leq 0 \\ \sup_{\mathrm{S}\in\zeta}\mathrm{l}_{\alpha_2}(\mathrm{S}) = 0, & \inf_{\mathrm{S}\in\zeta}\mathrm{u}_{\alpha_2}(\mathrm{S}) \geq 0 \end{array}.$$

Note that at least one of α_1 and α_2 always exists.

4.3 Simulation Studies

Through simulation studies, we evaluate the finite-sample performance of our proposed estimators. Simulation data are generated with 10,000 subjects randomized to vaccine and placebo by a ratio of

2:1. Baseline covariate X was generated with a multinomial distribution to have four categories, 1, 2, 3, and 4 with corresponding probabilities of 0.25, 0.25, 0.25 and 0.25. X_2, X_3, and X_4 are dummy variables indicating category 2, 3, or 4, respectively. Baseline biomarker values B were generated from a normal distribution with mean of $1.38+0.93X_2+1.25X_3-0.25X_4$ and standard deviation of 0.86. S were generated from a normal distribution with mean of $1.5+0.5B+0.2X_2-0.1X_3+0.4X_4$ and standard deviation of 0.4, which indicates a correlation of 0.7 between S and B. Let the the lower limit of quantification be 1. Simulated values of S and B less than 1 were set equal to 1. We assume a probit risk model of the clinical outcome Y conditional on S, B, Z, and X: $P(Y = 1|S, B, Z, X) = \Phi(\beta_0+\beta_1 Z+\beta_2 S+\beta_3 Z\cdot S+\beta_4 B+\beta_5 Z\cdot B+\beta_6 X)$. We set $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ as $(-0.50, 0.16, -0.34, -0.21, -0.25, 0, (0.24, 0.11, 0.20))$ so that the probability of infection equals 0.04 in the placebo arm and 0.02 in the vaccine arm. These simulation parameters were chosen to reflect the characteristics of the two Phase 3 Dengue trials. To achieve a non-monotone sampling design, 35% of study participants have B retained. For the BIP-only design, S is set missing for all placebo recipients and retained in all cases and all subjects with B measured in the vaccine arm, that is $\{i : Z_i = 1, Y_i = 1\} \cup \{i : Z_i = 1, \delta_{Bi} = 1\}$. For the BIP+CPV design, 70% of event-free placebo recipients are included in the CPV component and have S retained. Simulation results are based on 500 Monte-Carlo simulations and for each simulation 250 perturbation iterations are generated to construct point-wise confidence intervals and simultaneous confidence bands.

We then evaluate the finite-sample performance of our proposed estimators for the marginal VE curve ($\mathrm{VE}(S = s_1)$), baseline seropositive VE curve ($\mathrm{VE}(S = s_1, B > c)$), and baseline seronegative VE curve ($\mathrm{VE}(S = s_1, B = c)$). Results are presented in Figure 4.1 for BIP-only design and Figure 4.2 for BIP+CPV design. The empirical coverage levels of the 95% simultaneous confidence bands from the perturbation methods are also reported as "simultaneous.cover" in Figure 4.1 and 4.2. They demonstrate satisfactory performance of our proposed estimators, including nominal coverage probabilities of the confidence intervals given fixed s_1 values and simultaneous confidence band across all s_1 values.

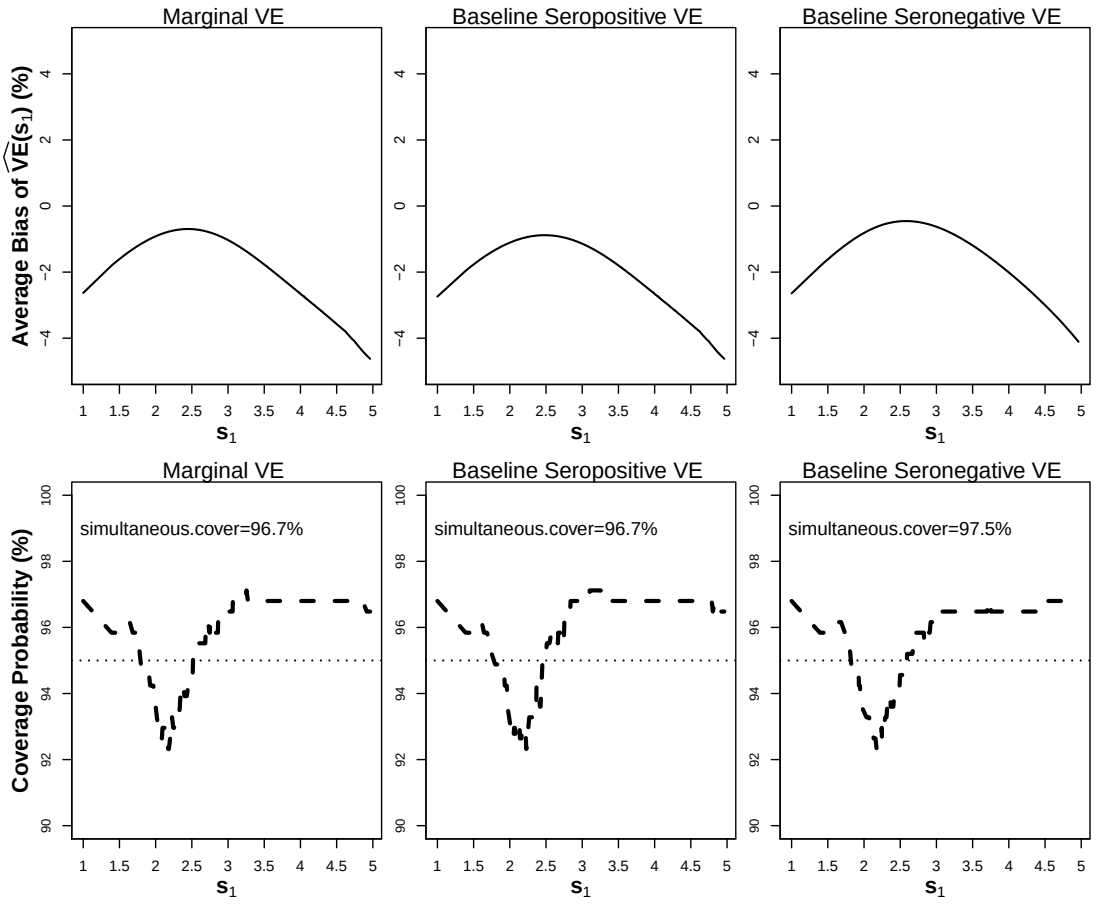

Figure 4.1: Average bias for our proposed estimators $\widehat{\mathrm{VE}}\,(S = s_1)$, $\widehat{\mathrm{VE}}\,(S = s_1, B > c)$ and $\widehat{\mathrm{VE}}\,(S = s_1, B = c)$ and coverage probabilities of 95% perturbation Wald confidence intervals in a BIP-only design.

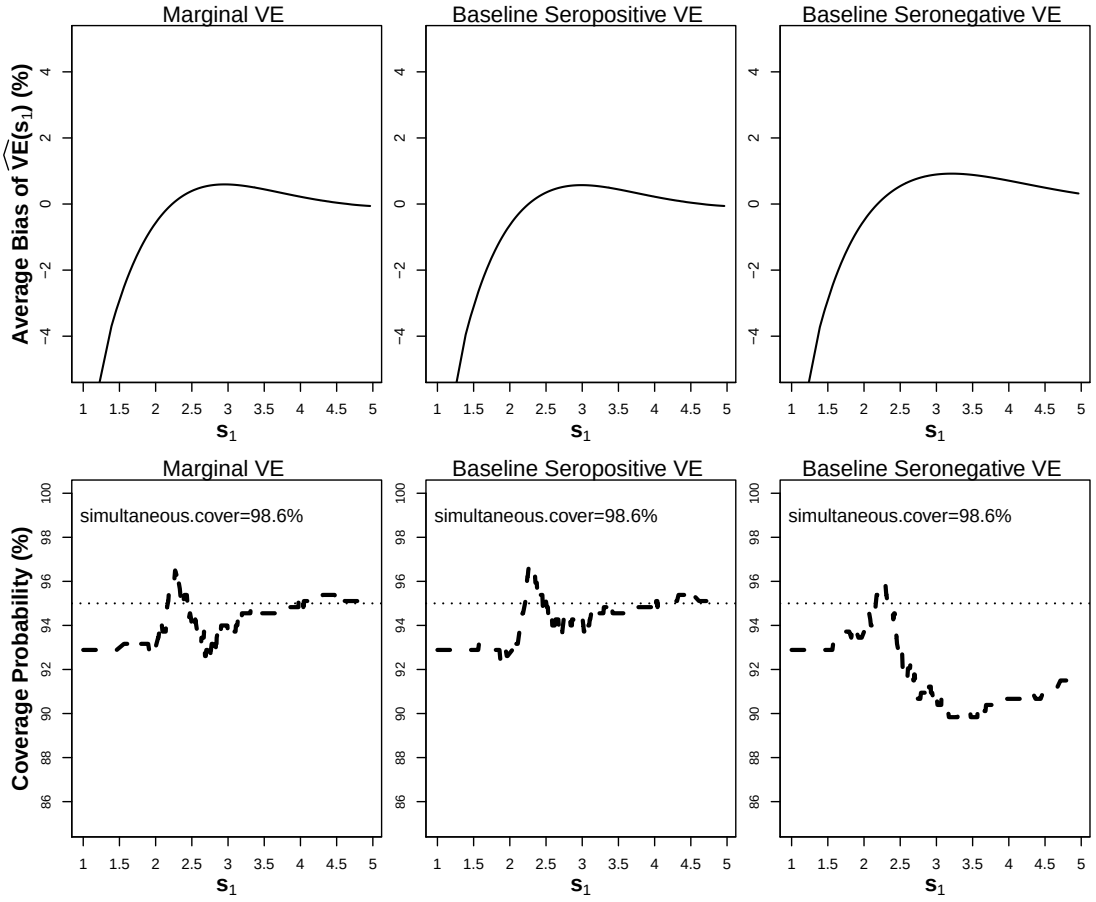

Figure 4.2: Average bias for our proposed estimators $\widehat{\mathrm{VE}}\,(S = s_1)$, $\widehat{\mathrm{VE}}\,(S = s_1, B > c)$ and $\widehat{\mathrm{VE}}\,(S = s_1, B = c)$ and coverage probabilities of 95% perturbation Wald confidence intervals in a BIP+CPV design.

4.4 Application to the CYD14 and 15 Trials

We demonstrate our proposed method to data pooling across CYD14 and CYD15 9-16 year olds to assess how VE varied by Month 13 titers, and how VE varied by Month 13 titers within baseline seropositive and seronegative subgroups. An brief background of trial CYD14 and CYD15 and meanings of notations Z, S, X, Y in the context of CYD14 and CYD15 9-16 year olds data are provided in chapter 2. In addition, we let B be the average of the log10-transformed titers at baseline. Baseline seropositive and seronegative subgroups are defined as $B > 1$ and $B = 1$. CYD14 and CYD15 hold a non-monotone sampling design. B is available for everyone in the immunogenicity subset ($\{i : \delta_B = 1\}$) and S is available for all vaccine recipients who are either in the immunogenicity subset or are cases ($\{i : Z_i = 1, \delta_{Bi} = 1\} \cup \{i : Z_i = 1, Y_i = 1\}$). These two sets do not have an inclusion relationship.

Figure 4.3 shows the estimated VE curve and 95% CIs and CBs based on 500 perturbation iterations. VE curves were similar for baseline seropositive and baseline seronegative subgroups, with estimated VE approximately 25% for vaccine recipients with no seroresponse at Month 13. For vaccine recipients with Month 13 average titers of 500 and 10,000, estimated VE was 79.3% and 97.3% for the baseline seropositive subgroup compared to 70.4% and 91.8% for the baseline seronegative subgroup, respectively. Furthermore, we tested the null hypothesis H_0: BL seropositive VE(S) =BL seronegative VE(S) for $S \in$ range of month 13 average titer in vaccinees in the data using procedure provided in section 4.2.2, which gave a p-value of 0.35. This suggests that the seropositive VE curve was not significantly different from the seronegative VE curve, implying that it is not the new neutralization response generated by the vaccine over baseline value that predicts VE, but rather the absolute titer achieved following vaccination. See Moodie and others[17] for the reporting of the full analysis to a clinical audience.

4.5 Discussion

In this chapter, we developed an estimated likelihood approach to evaluate the bivariate treatment effect modification analysis by biomarker-based principal strata and baseline covariates in general settings without requirements of a nested sub-sampling relationship between the immune response biomarker and baseline predictors suitable to both the BIP-only design and the BIP+CPV design.

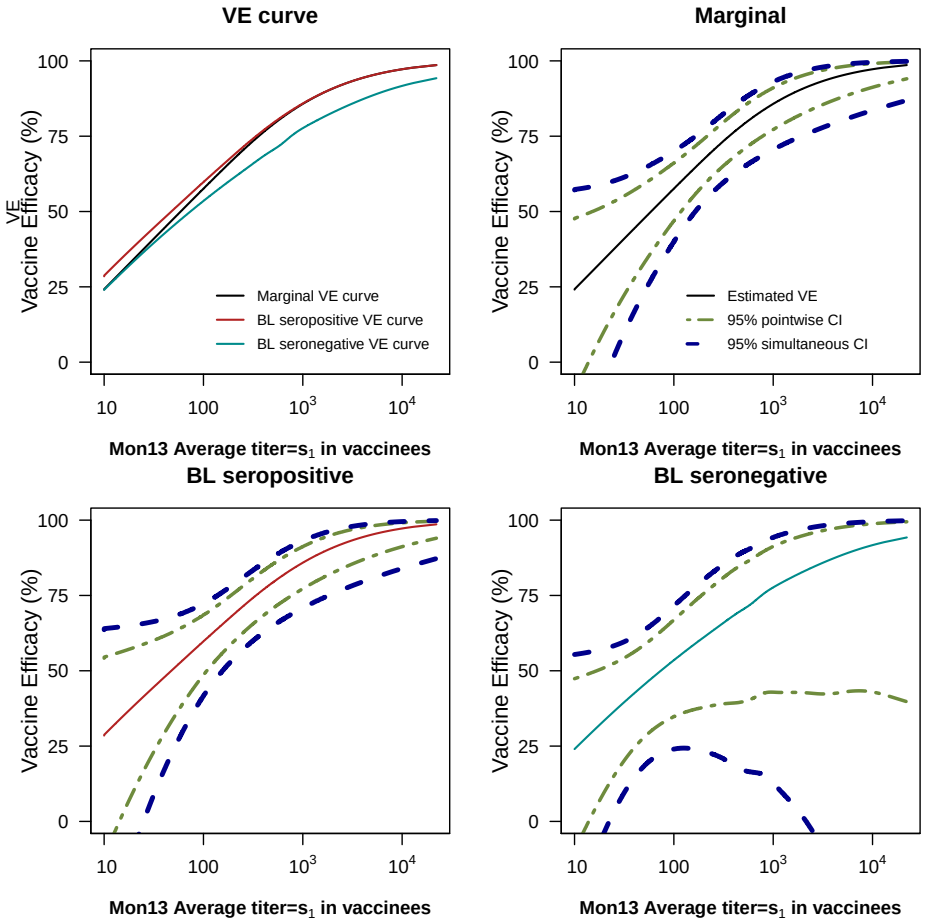

Figure 4.3: Estimated vaccine efficacy by average $\log_{10}$ titer at Month 13 with 95% pointwise confidence intervals and simultaneous confidence bands in CYD14 and CYD15 9-16-year-olds.

In our settings, the biomarker sampled set $\{i : \delta_i = 1\}$ and the baseline covariates sampled set $\{i : \delta_{Bi} = 1\}$ do not need to be a subset of the other. Dengue vaccine trials CYD14 and CYD15 are examples of this non-inclusion relationship.

Our proposed method can apply to the special case where $\{i : \delta_i = 1\}$ is a subset of $\{i : \delta_{Bi} = 1\}$. An example would be a three-phase sampling design when lab-assay-based baseline covariates are only measured from a subset of the trial participants due to high costs of acquiring lab assay and the vaccine-induced immune response is further measured among a subset of participants for whom the lab-assay-based baseline covariates are available. The phase 3 Zostavax Efficacy and Safety Trial (ZEST) adopted such a three-phase sampling design to study the effect of the Zostavax vaccine against varicella zoster virus (VZV) [23]. Under this sampling framework, Huang (2017) [12] proposed a semiparametric pseudo-score estimator based on conditional likelihood and also develop several alternative semiparametric estimated likelihood estimators when B is discrete. One can think of our work in this chapter as an extension of Huang (2017) [12] that our proposed method can incorporate different sampling settings including, but not limited to, the one in Huang (2017) and our baseline covariate B can be either discrete or continuous, possibly subject to lower limit of quantification left censoring. In general, our methods are applicable to intervention studies where a bivariate effect modification analysis is of interest where the bivariate is a post-randomization measurement and a baseline covariate, and measurements of one do not necessarily imply measurements of the other.

Chapter 5

EVALUATION OF TREATMENT EFFECT MODIFICATION BY POST-RANDOMIZATION BIOMARKER-DEFINED PRINCIPAL STRATA RELAXING THE EQUAL EARLY CLINICAL RISK ASSUMPTION

The methods proposed in chapter 3 and 4 and many other methods proposed in recent literature developed in the principal stratification/effect modification framework (e.g., [13], [14], [8], [6], [12]) adopted assumption A3 in GH, which states equal early clinical risk: $P\left(Y^{\tau}(1) = Y^{\tau}(0)\right) = 1$. This assumption is important because the risk parameters of interest condition on $Y^{\tau}(1) = Y^{\tau}(0) = 0$. However, A3 is questionable if the treatment is efficacious early on or the biomarkers of interest are measured a long time after baseline. For the dengue application, post vaccination antibody neutralization titer measurements were collected 13 months after the first vaccination and about half of the primary dengue disease endpoint events occurred before 13 months. Table 5.1 summarizes the number and rate of cases observed before Month 13 (early cases) for CYD14 and CYD15 ([1], [26]). Rate of dengue by Month 13 was significantly higher in the placebo group than the vaccine group in both trials. Therefore, assumption A3 is rejected based on the data. In this chapter, we consider how to extend the mCEP curve estimation methods to the more plausible assumption: A3-mon (standing for monotonicity): $P(Y^{\tau}(1) \geq Y^{\tau}(0)) = 0$, by adapting the approach of Shepherd et al. ([24]) to our application.

5.1 Assumptions

We adopt notation $(Z, X, Y^{\tau}, Y, \delta, S)$ and assumption (A1) and (A2) from Chapter 2 and add the following two assumptions:

(A3-mon) Monotonicity: $P(Y^{\tau}(1) \leq Y^{\tau}(0)) = 1$

(A4) A model for the mixing probabilities of the early always-at-risk (EAAR) subgroup and the

Table 5.1: Number and Rate of Cases Observed by Month 13 for CYD14 and CYD15.

	Vaccine Arm: N=6848		Placebo Arm: N=3424	
CYD14	Dengue Cases by Month 13	Dengue Rate by Month 13	Dengue Cases by Month 13	Dengue Rate by Month 13
	169	0.025	176	0.051
	Vaccine Arm: N=13920		Placebo Arm: N=6949	
CYD15	Dengue Cases by Month 13	Dengue Rate by Month 13	Dengue Cases by Month 13	Dengue Rate by Month 13
	101	0.007	164	0.024

early protect (EP) subgroup among all individuals at risk at time τ in the vaccine group. Specifically,

$$P(Y^\tau(0) = 0 | Y^\tau(1) = 0, Y(1), S, X) = w(S, X, Y(1); \beta),$$

where

$$w(s_1, x, y; \beta) = \Omega\left\{m(s_1, x) + g(s_1, x, y; \beta)\right\}.$$

β is fixed and known, $\Omega(\cdot)$is a known cdf, $m(\cdot, \cdot)$ is an unspecified function of (S, X), and for each β, $g(\cdot, \cdot, \cdot; \beta)$ is a known function of S, X, and $Y(1)$.

Each subject can be classified into one of four principal strata defined in terms of the counterfactual pair $(Y^\tau(0), Y^\tau(1))$:

- Early Always-at-risk (EAAR): $Y^\tau(0) = Y^\tau(1) = 0$
- Early Protected (EP): $Y^\tau(0) = 1, Y^\tau(1) = 0$
- Early Harmed (EH): $Y^\tau(0) = 0, Y^\tau(1) = 1$
- Early Never-at-risk (ENAR): $Y^\tau(0) = Y^\tau(1) = 1$

A3-mon states that an individual who would stay at risk under placebo up to time τ would also stay at risk under vaccine up to time τ. It implies that the vaccine effect on the risk of clinical endpoint occurrence up to time τ is either beneficial or harmless, and that all individuals at risk at time τ in the placebo group belong to the early always-at-risk subgroup and the early harmed

subgroup is empty. Under A1 and A2, A3-mon will holds if vaccination does not increase infection probability for any subject up to time τ and can be checked by testing if the infection rate from time 0 to time τ is higher in vaccine than placebo for any participant subgroup.

Given the randomization assignment (Z_i) and the observed early-at-risk status (Y_i^τ) of a trial participant, Table 5.2 summarized the possible strata to which a participant could belong. Because biomarker S is only measured among those who remain event-free untill τ, S is undefined if $Y^\tau = 1$. Table 5.2 also lists the information available on potential biomarker and makes clear that the causal estimand CEP surface is defined only in the Early always-at-risk (EAAR) stratum because it conditions on $S(1)$ and $S(0)$.

A3-mon ensures our previous methods applicable on risk under placebo because $\{i : Y_i^\tau(0) = 0\} = \{i : Y_i^\tau(1) = Y_i^\tau(0) = 0\}$, such that $\text{risk}_0(\text{s}_1) = \text{P}(\text{Y}(0) = 1|\text{S} = \text{s}_1, \text{Y}^\tau(1) = \text{Y}^\tau(0) = 0) = \text{P}(\text{Y}(0) = 1|\text{S} = \text{s}_1, \text{Y}^\tau(0) = 0)$. However, new work is needed to estimate risk under vaccine because, as shown in table 5.2, the subgroup $\{i : Y_i^\tau(1) = Y_i^\tau(0) = 0\}$ is not identifiable from subjects in the vaccine group due to the fact that subjects at-risk under vaccine at τ is the union of the EAAR subgroup and the EP subgroup. A4 is introduced to address this problem, which models the mixing probabilities of the EAAR subgroup and the EP subgroup among all individuals at risk at time τ in the vaccine group.

If X and S are categorical variables with a small number of values, then we can apply previous methods within each level of X and S. However, if the distribution of X and/or S are continuous or discrete with a large support, performing analysis within each cell is unfeasible because the data are too sparse to conduct cell-specific estimation. We address this problem by imposing the following additional modeling assumptions:

(M1) $P(Y^\tau(1) = 0|X) = \theta_v(X; \mu)$, where μ is unknown and for each μ, $\theta_v(\cdot; \mu)$ is a known function.

(M2) The function $m(s_1, x)$ in (A4) follows a parametric model $m(S, X) = m(S, X; \alpha)$, where α is an unknown parameter vector and for each α, $m(\cdot, \cdot; \alpha)$ is a known function. Thus function $w(\cdot)$ in (A4) is a function of unknown parameters (α, β).

(M3) The distribution of immune response under vaccine, S, given X is known up to a finite dimensional parameter γ; that is, $f_{S|Y^\tau,X}(s_1|Y^\tau(1) = 0, X = x) = f(s_1|x; \gamma)$, where γ is unknown; and for each γ, $f(s_1|x; \gamma)$ is a known function.

Table 5.2: For each trial participant i, the table lists the possible strata to which the participant could belong for four combinations for Z_i and Y_i^τ. $S_i(z)$ is only defined if $Y_i^\tau(z) = 0$, for $z = 0, 1$.

Treatment Assignment Z_i	Observed Early-at-risk Status Y_i^τ	Principal Stratum $\{Y_i^\tau(0), Y_i^\tau(1)\}$ and information on potential biomarker $S_i(1)$, $S_i(0)$		
$Z_i = 1$	$Y_i^\tau = 0$	**Early protected**	or	**Early always-at-risk**
		$\{Y_i^\tau(0) = 1, Y_i^\tau(1) = 0\}$		$\{Y_i^\tau(0) = 0, Y_i^\tau(1) = 0\}$
		$S_i(0)$ undefined		$S_i(0)$ possibly observed
		$S_i(1)$ possibly observed		$S_i(1)$ possibly observed
$Z_i = 0$	$Y_i^\tau = 0$		or	**Early always-at-risk**
		$\{Y_i^\tau(0) = 0, Y_i^\tau(1) = 1\}$		$\{Y_i^\tau(0) = 0, Y_i^\tau(1) = 0\}$
		(empty set by A3-mon)		$S_i(0)$ possibly observed
				$S_i(1)$ possibly observed
$Z_i = 1$	$Y_i^\tau = 1$		or	**Early never-at-risk**
		$\{Y_i^\tau(0) = 0, Y_i^\tau(1) = 1\}$		$\{Y_i^\tau(0) = 1, Y_i^\tau(1) = 1\}$
		(empty set by A3-mon)		$S_i(0)$ undefined
				$S_i(1)$ undefined
$Z_i = 0$	$Y_i^\tau = 1$	**Early protected**	or	**Early never-at-risk**
		$\{Y_i^\tau(0) = 1, Y_i^\tau(1) = 0\}$		$\{Y_i^\tau(0) = 1, Y_i^\tau(1) = 1\}$
		$S_i(0)$ undefined		$S_i(0)$ undefined
		$S_i(1)$ possibly observed		$S_i(1)$ undefined

(M4) The risk function under placebo given S and X in the early always-at-risk stratum is known up to a finite dimensional parameter η_p; that is, $P(Y(0) = 1|Y^\tau(0) = 0, S, X) = P(Y(0) = 1|Y^\tau(0) = Y^\tau(1) = 0, S, X) = g_p(S, X; \eta_p)$, where η_p is unknown; and for each η_p, $g_p(S, X; \eta_p)$ is a known risk function.

We also make one of the following two assumptions:

(M5a) The risk function under vaccine given S and X in the population comprised of both early always-at-risk individuals and the early protect individuals, is known up to a finite dimensional parameter η_v^a; that is, $P(Y(1) = 1|Y^\tau(1) = 0, S, X) = g_v(S, X; \eta_v^a)$ where η_v^a is unknown and for each η_v^a, $g_v(S, X; \eta_v^a)$ is a known risk function.

(M5b) The risk function under vaccine given S and X in the early always-at-risk (EAAR) stratum is known up to a finite dimensional parameter η_v^b; that is, $P(Y(1) = 1|Y^\tau(1) = Y^\tau(0) = 0, S, X) = g_v^{EAAR}(S, X; \eta_v^b)$ where η_v^b is unknown and for each η_v^b, $g_v^{EAAR}(S, X; \eta_v^b)$ is a known risk function.

$\mathcal{M}_a$ is a natural approach to modeling the potential outcomes. One assigns parametric distributions to the observed clinical outcome Y and then through $w(s_1, x, y; \beta, \alpha)$, models the portion of the early at-risk vaccinees who would have been early at-risk regardless of treatment. Therefore, the two modeling assumptions, (A4) and (M5a), together induce the distribution of $Y(1)$ in the EAAR principal stratum.

Rather than modeling the clinical outcomes for early at risk vaccinees, a different approach assumes a distribution for clinical outcomes in early at risk vaccinees in the EAAR principal stratum where we assume $P(Y(1) = 1|Y^\tau(1) = Y^\tau(0) = 0, S, X) = g_v^{EAAR}(S, X; \eta_v^b)$. If $w(s_1, x, y; \alpha, \beta)$ is greater than 0 for all s_1, x, and y, then one can write

$$\begin{aligned} &P(Y(1) = 1|Y^\tau(1) = 0, S = s_1, X = x) = g_v^*(s_1, x; \alpha, \beta, \eta_v^b) \\ =&\frac{w^{-1}(s_1, x, y = 1; \beta, \alpha) g_v^{EAAR}(s_1, x; \eta_v^b)}{w^{-1}(s_1, x, y = 1; \beta, \alpha) g_v^{EAAR}(s_1, x; \eta_v^b) + w^{-1}(s_1, x, y = 0; \beta, \alpha)(1 - g_v^{EAAR}(s_1, x; \eta_v^b))} \end{aligned} \tag{5.1}$$

For ease of reference, we call the model defined by assumptions (A1)-(A4), (M1)-(M4) and (M5a) model M_a and model defined by assumptions (A1)-(A4), (M1)-(M4) and (M5b) model M_b.

5.2 Sensitivity Analysis

The function $w(s_1, x, y; \beta, \alpha)$ in assumption A4 is the probability that a subject stayed at risk untill τ with clinical outcome y, biomarker response s_1, baseline covariate x under vaccine would also stay at risk untill τ if randomized to placebo. However, $w(\cdot)$ is unknown and it is not possible to test whether any particular function assumption of $w(\cdot)$ is correctly specified from the data plus A1-A3. Our approach to this problem assumes $\Omega(\cdot)$, $m(\cdot, \cdot)$, and $g(\cdot, \cdot, \cdot)$ are fixed and known. To carry out a sensitivity analysis, β is assumed fixed and known, and inference about the marginal CEP curve in the EAAR stratum is repeated for a variety of fixed choices of β in a plausible range, Γ, which reveals how estimates for the marginal CEP curve vary over different values for the sensitivity parameter.

Given fixed β, α is determined by the following procedures if X and S are discrete and low-dimensional: For a specific x and s, with some calculations, one can show that

$$P(Y(1) = y|Y^\tau(0) = Y^\tau(1) = 0) = \frac{w(s, x, y; \beta, \alpha)P(Y(1) = y|Y^\tau(1) = 0, X = x, S = s)}{P(Y^\tau(0) = 0|Y^\tau(1) = 0, X = x, S = s)}.$$

$P(Y(1) = y|Y^\tau(1) = 0, X = x, S = s)$ can be estimated nonparametrically on set $\{i : Z_i = 1, Y_i^\tau = 1, X_i = x, S_i = s\}$;

$P(Y^\tau(0) = 0|Y^\tau(1) = 0, X = x, S = s) = \frac{P(Y^\tau(0)=0, X=x, S=s)}{P(Y^\tau(1)=0, X=x, S=s)}$ by A3-mon, which can also be estimated nonparametrically from the observed data.

For a fixed β, α is determined as the solution to the equation $P(Y(1) = 1|Y^\tau(1) = 0, X = x, S = s) + P(Y(1) = 0|Y^\tau(1) = 0, X = x, S = s) = 1$.

However, in applications where X and S are either continuous or discrete with a large support, α can not be determined for a fixed β, we propose estimating α together with risk parameters η through a maximized likelihood approach. We detail the estimation procedure of (α, η) in section 5.3.

We recommend reporting an estimated ignorance interval, that is a set of point estimates for the marginal CEP curve with β varied over region Γ and it expresses ambiguity about the parameter of interest due to partial identifiability. In section 5.5, we describe procedures for constructing the estimated uncertainty intervals (EUIs) that incorporate imprecision due to sampling variability as well as lack of identifiability.

If results hold in one direction for a plausible range of the sensitivity parameter β, then a causal conclusion in that direction may be drawn. Otherwise, the analysis remains inconclusive.

5.3 Maximum Likelihood Estimation

5.3.1 Constructing the Likelihood

To derive the ML estimator we express the joint density of the observables $O = (Z, X, Y^\tau, Y, \delta, S\delta)'$,

$$\begin{aligned} f_O(O) =& P_{\delta|Y,Y^\tau,S,Z,X}(\delta|Y^\tau, Y, S, Z, X) P_{Y|Y^\tau,S,Z,X}(Y|Y^\tau, S, Z, X) \\ & f_{S|Y^\tau,Z,X}(S|Y^\tau, Z, X) P_{Y^\tau|Z,X}(Y^\tau|Z, X) P_{Z|X}(Z|X) f_X(X) \end{aligned}$$

in terms of the model parameters. Specifically, in Appendix C.1 we show that

$$\begin{aligned} & P(Y = 1|Y^\tau = 0, S = s_1, Z = 1, X = x) \\ =& \begin{cases} g_v(s_1, x; \eta_v^a) & \text{under } \mathcal{M}_\text{a} \\ \frac{w^{-1}(s_1,x,y=1;\beta,\alpha) g_v^{EAAR}(s_1,x;\eta_v^b)}{w^{-1}(s_1,x,y=1;\beta,\alpha) g_v^{EAAR}(s_1,x;\eta_v^b) + w^{-1}(s_1,x,y=0;\beta,\alpha)(1-g_v^{EAAR}(s_1,x;\eta_v^b))} \equiv g_v^*(s_1, x; \beta, \alpha, \eta_v^b) & \text{under } \mathcal{M}_\text{b} \end{cases} \end{aligned}$$

$$\begin{aligned} & P(Y(1) = 1|Y^\tau(0) = Y^\tau(1) = 0, S = s_1, X = x) \\ =& \begin{cases} \frac{w(s_1,x,y=1;\beta,\alpha) g_v(s_1,x;\eta_v^a)}{w(s_1,x,y=1;\beta,\alpha) g_v(s_1,x;\eta_v^a) + w(s_1,x,y=0;\beta,\alpha)(1-g_v(s_1,x;\eta_v^a))} \equiv g_v^{*EAAR}(s_1, x; \beta, \alpha, \eta_v^a) & \text{under } \mathcal{M}_\text{a} \\ g_v^{EAAR}(s_1, x; \eta_v^b) & \text{under } \mathcal{M}_\text{b} \end{cases} \end{aligned}$$

$$P(Y = 1|Y^\tau = 0, S = s_1, Z = 0, X = x) = g_p(s_1, x; \eta_p) \quad \text{under } \mathcal{M}_\text{a} \text{ or } \mathcal{M}_\text{b}$$

$$\begin{aligned} & f_{S|Y^\tau,Z,X}(s_1|Y^\tau = 0, Z = 1, X = x) \\ =& f_{S|Y^\tau(1),X}(s_1|Y^\tau(1) = 0, X = x) \equiv f_v(s_1|x; \gamma) \qquad \text{under } \mathcal{M}_\text{a} \text{ or } \mathcal{M}_\text{b} \end{aligned}$$

$$
\begin{aligned}
&f_{S|Y^\tau,Z,X}(s_1|Y^\tau=0,Z=0,X=x)\\
&=\begin{cases}
\frac{\frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x;\eta_v^a)}{g_v^{*EAAR}(s_1,x;\beta,\alpha,\eta_v^a)}f(s_1|x;\gamma)}{\int\frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x;\eta_v^a)}{g_v^{*EAAR}(s_1,x;\beta,\alpha,\eta_v^a)}f(s_1|x;\gamma)ds_1}\equiv f_p^a(s_1|x;\beta,\alpha,\eta_v^a,\gamma) & \text{under } \mathcal{M}_{\rm a}\\
\frac{\frac{w(s_1,x,y=1;\beta,\alpha)g_v^*(s_1,x;\beta,\alpha,\eta_v^b)}{g_v^{EAAR}(s_1,x;\eta_v^b)}f(s_1|x;\gamma)}{\int\frac{w(s_1,x,y=1;\beta,\alpha)g_v^*(s_1,x;\beta,\alpha,\eta_v^b)}{g_v^{EAAR}(s_1,x;\eta_v^a)}ds_1}\equiv f_p^b(s_1|x;\beta,\alpha,\eta_v^b,\gamma) & \text{under } \mathcal{M}_{\rm b}
\end{cases}
\end{aligned}
$$

$$
P(Y^\tau=0|Z=1,X=x)=\theta_v(x;\mu)\qquad\qquad\text{under } \mathcal{M}_{\rm a} \text{ or } \mathcal{M}_{\rm b}
$$

$$
\begin{aligned}
&P(Y^\tau=0|Z=0,X=x)\\
&=\begin{cases}
\theta_v(x;\mu)\int\frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x;\eta_v^a)}{g_v^{*EAAR}(s_1,x;\beta,\alpha,\eta_v^a)}f(s_1|x;\gamma)ds_1\equiv\theta_p^a(x;\mu,\beta,\alpha,\eta_v^a,\gamma) & \text{under } \mathcal{M}_{\rm a}\\
\theta_v(x;\mu)\int\frac{w(s_1,x,y=1;\beta,\alpha)g_v^*(s_1,x;\beta,\alpha,\eta_v^b)}{g_v^{EAAR}(s_1,x;\eta_v^b)}f(s_1|x;\gamma)ds_1\equiv\theta_p^b(x;\mu,\beta,\alpha,\eta_v^b,\gamma) & \text{under } \mathcal{M}_{\rm b}
\end{cases}
\end{aligned}
$$

Appendix C.2 shows that the likelihoods $\mathcal{L}_a(\mu,\beta,\alpha,\eta_v^a,\eta_p,\gamma)$ and $\mathcal{L}_b(\mu,\beta,\alpha,\eta_v^b,\eta_p,\gamma)$ under models $\mathcal{M}_a$ and $\mathcal{M}_b$ respectively are

$$
\begin{aligned}
&\mathcal{L}_a(\mu,\beta,\alpha,\eta_v^a,\eta_p,\gamma)\\
&\propto\prod_{i=1}^n\left\{\pi_0(Y_i,Z_i,X_i)\left[g_v(S_i,X_i;\eta_v^a)\right]^{Y_i}\left[1-g_v(S_i,X_i;\eta_v^a)\right]^{1-Y_i}f_v(S_i|X_i;\gamma)\theta_v(X_i;\mu)\right\}^{(1-Y_i^\tau)Z_i\delta_i}\\
&\times\left\{\pi_0(Y_i,Z_i,X_i)\int\left[1-g_v(s_1,X_i;\eta_v^a)\right]f_v(s_1|X_i;\gamma)ds_1\theta_v(X_i;\mu)\right\}^{(1-Y_i^\tau)Z_i(1-\delta_i)}\\
&\times\left\{\int\left[g_p(s_1,X_i;\eta_p)\right]^{Y_i}\left[1-g_p(s_1,X_i;\eta_p)\right]^{1-Y_i}f_p^a(s_1|X_i;\beta,\alpha,\eta_v^a,\gamma)ds_1\theta_p^a(X_i;\mu,\beta,\alpha,\eta_v^a,\gamma)\right\}^{(1-Y_i^\tau)(1-Z_i)}\\
&\times\left\{\left[1-\theta_v(X_i;\mu)\right]^{Z_i}\left[1-\theta_p^a(X_i;\mu,\beta,\alpha,\eta_v^a,\gamma)\right]^{1-Z_i}\right\}^{Y_i^\tau},
\end{aligned}
$$

where $\pi_0(y,z,x)\equiv P(\delta=1|Y^\tau=0,Y=y,S(1)=s_1,Z=z,X=x)=P(\delta=1|Y^\tau=0,Y=y,Z=z,X=x)$. When π_0 is unknown, we can substitute π_0 with a consistent estimator $\hat{\pi}_\lambda$ where λ is a parameter to be estimated by ML from observations $\{i:Y_i^\tau=0\}$.

And $\mathcal{L}_b(\mu, \beta, \alpha, \eta_v^b, \eta_p, \gamma)$ is defined like $\mathcal{L}_a$ with $g_v^*(s_1, x; \beta, \alpha, \eta_v^b)$ replacing $g_v(s_1, x; \eta_v^a)$, $f_p^b(s_1|x; \beta, \alpha, \eta_v^b, \gamma)$ replacing $f_p^a(s_1|x; \beta, \alpha, \eta_v^a, \gamma)$, and $\theta_p^b(x; \mu, \beta, \alpha, \eta_v^b, \gamma)$ replacing $\theta_p^a(x; \mu, \beta, \alpha, \eta_v^a, \gamma)$.

5.3.2 Risk Parameters Estimation

We consider the estimated likelihood approach by Pepe and Fleming (1991) [18] where consistent estimates of (μ, γ) are obtained first. Then for a fixed and known β, $\mathcal{L}(\hat{\mu}, \beta, \alpha, \eta_v, \eta_p, \hat{\gamma})$ is maximized in (α, η_v, η_p). Here $\mathcal{L}$ denotes $\mathcal{L}_a$ under models $\mathcal{M}_a$ and $\mathcal{L}_b$ under models $\mathcal{M}_b$; η_v denotes η_v^a under models $\mathcal{M}_a$ and η_v^b under models $\mathcal{M}_b$. μ is estimated by obtaining the MLE of the likelihood $L_1 = \prod_{i:z_i=1}^n [\theta_v(x_i; \mu)]^{1-Y_i^\tau} [1 - \theta_v(x_i; \mu)]^{Y_i^\tau}$. γ is estimated using vaccine recipients with S measured. Because sampling of S depends on other phase-I variables such as Y, inverse probability weighting (IPW) can be implemented.

5.3.3 VE Curve Estimation

By plugging in the MLEs $(\widehat{\eta_v}, \widehat{\eta_p})$, we have estimates for

$$\mathrm{risk_0(s_1, x)} \equiv \mathrm{P(Y(0) = 1|S = s_1, X = x, Y^\tau(0) = Y^\tau(1) = 0) = g_p(s_1, x)}$$

and

$$\mathrm{risk_1(s_1, x)} \equiv \mathrm{P(Y(1) = 1|S = s_1, X = x, Y^\tau(0) = Y^\tau(1) = 0) = g_v^{EAAR}(s_1, x).}$$

To integrate out X in the risk functions, we first need to estimate the distribution $X|S, Y^\tau(0) = Y^\tau(1) = 0$. We consider situations where X is categorical. In Appendix C.3, we show that

$$\begin{aligned} &P(X = x|S = s_1, Y^\tau(0) = Y^\tau(1) = 0) \\ =&\frac{P(Y^\tau(0) = 0|X = x, S = s_1, Y^\tau(1) = 0) \cdot P(X = x|S = s_1, Y^\tau(1) = 0)}{\sum_{all\ x} P(Y^\tau(0) = 0|X = x, S = s_1, Y^\tau(1) = 0) \cdot P(X = x|S = s_1, Y^\tau(1) = 0)}. \end{aligned}$$

$$\begin{aligned} &P(Y^\tau(0) = 0|X = x, S = s_1, Y^\tau(1) = 0) \\ =&\begin{cases} w(s_1, x, y = 1) g_v(s_1, x; \eta_v^a) + w(s_1, x, y = 0)(1 - g_v(s_1, x; \eta_v^a)) \text{ under } \mathcal{M}_\mathrm{a} \\ w(s_1, x, y = 1) g_v^*(s_1, x; \beta, \alpha, \eta_v^b) + w(s_1, x, y = 0)(1 - g_v^*(s_1, x; \beta, \alpha, \eta_v^b)) \text{ under } \mathcal{M}_\mathrm{b} \end{cases} \end{aligned}$$

which can be estimated with $(\widehat{\alpha}, \widehat{\eta_v}^a)$ under $\mathcal{M}_a$ or $(\widehat{\alpha}, \widehat{\eta_v}^b)$ under $\mathcal{M}_b$.

$$P(X|S, Y^\tau(1) = 0) = \frac{P(S|X, Y^\tau(1) = 0) \cdot P(X|Y^\tau(1) = 0)}{\sum_{all\ x} P(S|X, Y^\tau(1) = 0) \cdot P(X|Y^\tau(1) = 0)}$$

where $P(S|X, Y^\tau(1) = 0)$ can be estimated by $\widehat{\gamma}$ and $P(X|Y^\tau(1) = 0)$ can be estimated non-parametrically on subjects with $\{i : Z_i = 1, Y_i^\tau = 0\}$. Alternatively, $P(X|S, Y^\tau(1) = 0)$ can be estimated using a multinomial distribution by vaccine recipients with S measured with inverse probability weighting.

Finally, we calculate $\text{VE}(s_1)$ through:

$$\begin{aligned}&\text{VE}(s_1)\\ =&1 - \frac{\sum_{all\ x} P(Y(1) = 1|S = s_1, X = x, Y^\tau(0) = Y^\tau(1) = 0) \cdot P(X = x|S = s_1, Y^\tau(0) = Y^\tau(1) = 0)}{\sum_{all\ x} P(Y(0) = 1|S = s_1, X = x, Y^\tau(0) = Y^\tau(1) = 0) \cdot P(X = x|S = s_1, Y^\tau(0) = Y^\tau(1) = 0)}.\end{aligned}$$

5.4 Specific Parameterizations

In our simulation and example, we consider the following specific parameterizations:

$$f(s_1|x; \gamma) = \phi(s_1; \text{mean} = \text{x}^{\text{T}}\gamma_{\text{x}}, \text{sd} = \gamma_\sigma) \tag{5.2}$$

$$\theta_v(x; \mu) = \frac{exp(x^T\mu)}{1 + exp(x^T\mu)} \tag{5.3}$$

$$w(s_1, x, y; \beta, \alpha) = \frac{exp(\alpha_0 + \alpha_s s_1 + \alpha_x x + \beta y)}{1 + exp(\alpha_0 + \alpha_s s_1 + \alpha_x x + \beta y)} \tag{5.4}$$

$$\begin{aligned}\text{under } \mathcal{M}_\text{a} \quad & g_v(s_1, x; \eta_v^a) = \Phi(\eta_{v0} + \eta_{v1} s_1 + \eta_{v2} x)\\ & g_p(s_1, x; \eta_p) = \Phi(\eta_{p0} + \eta_{p1} s_1 + \eta_{p2} x)\end{aligned}$$

which can be consolidated as:

$$P(Y = 1|Y^\tau = 0, Z = z, S = s_1, X = x) = \Phi(\eta_0 + \eta_1 z + \eta_2 s_1 + \eta_3 s_1 z + \eta_4 x + \eta_5 xz). \tag{5.5}$$

$$\text{under } \mathcal{M}_{\text{b}} \quad g_v^{EAAR}(s_1, x; \eta_v^b) = \Phi(\eta_{v0} + \eta_{v1} s_1 + \eta_{v2} x) \tag{5.6}$$

$$g_p(s_1, x; \eta_p) = \Phi(\eta_{p0} + \eta_{p1} s_1 + \eta_{p2} x). \tag{5.7}$$

where ϕ and Φ are pdf and cdf of a normal distribution $N(\text{mean}, \text{sd})$, $x = (1, x_1, \cdots, x_q)^T$ and γ_x and μ are parameter vectors of length $q + 1$.

Here we assume a logistic selection bias model for $w(\cdot)$: the probability that a subject who remains early at risk under vaccine is an always at risk individual depends on his/her $Y(1), S$, and X through the expit function $w(s_1, x, y; \beta, \alpha) = \frac{exp(\alpha_0 + \alpha_s s_1 + \alpha_x x + \beta y)}{1 + exp(\alpha_0 + \alpha_s s_1 + \alpha_x x + \beta y)}$ and regards β as fixed and known. This expit function allows $w(\cdot)$ to be constant or smoothly monotone increasing or decreasing. The selection bias parameter β is interpretable through that e^β is the odds ratio of remaining early at risk under randomization to placebo given early at risk under randomization to vaccine with $Y(1) = 1$ versus with $Y(1) = 0$, adjusting for S and X. Interpretability of e^β allows the choice of plausible range Γ to be guided by beliefs about plausible degrees of selection bias. Γ should be chosen independently from the data. Choosing β such that $g(S, X, Y(1); \beta) = 0$ is the same as assuming that $Y(1) \amalg Y^\tau(0) | Y^\tau(1) = 0, S, X$.

5.5 Simulation Study

5.5.1 Simulations under $\mathcal{M}_a$

Simulation studies were performed to investigate finite sample properties and robustness of our estimators. Data were generated using a simulation experiment under $\mathcal{M}_a$ or $\mathcal{M}_b$ where $\beta = 0$, -1, or -3.

Simulation data generation procedures:

Step 1 Generate Z: The first 5,000 observations were set to $Z = 0$, the second 10,000 were set to $Z = 1$.

Step 2 Generate X from a normal distribution with mean of 11 and standard deviation of 1. X was binned into four categories using its quartiles. Dummy variables X_2, X_3, and X_4 were created indicating the 2^{nd}, 3^{rd}, and 4^{th} quartile, respectively.

Step 3 Generate $Y^\tau(1)$: Given X, $Y^\tau(1)$ was drawn from a Bernoulli($\theta_v(X;\mu)$) distribution where $\theta_v(X;\mu) = \frac{exp((1,X_2,X_3,X_4)^T\mu)}{1+exp((1,X_2,X_3,X_4)^T\mu)}$ with $\mu = (4.4, -0.23, -0.46, -0.69)$ so that probability of early infection under vaccine equals 0.02.

Step 4 Generate $S(1)$: Given X and $Y^\tau(1) = 0$, $S(1)$ was drawn from a $Normal((1, X_2, X_3, X_4)^T\gamma_x, \gamma_\sigma^2)$ distribution where γ_x, γ_σ were in (M3) with $\gamma_x = (2.4, -0.4, 0.6, 0.9)$ and $\gamma_\sigma = 0.68$ so that correlation coefficient between X and $S(1)$ $\rho = 0.5$. Simulation values of S less than 0 were set equal to 0.

Step 5 Generate $Y(1)$:

- For realizations with $Y^\tau(1) = 1$, $Y(1)$ was set to 1;
- For realizations with $Y^\tau(1) = 0$, $Y(1)$ was drawn from a Bernoulli($g_v(s_1, x; \eta_v^a)$) distribution where $g_v(s_1, x; \eta_v^a) = \Phi(\eta_{v0} + \eta_{v1}s_1 + \eta_{v2}x)$. The parameter $\eta_v^a = (\eta_{v0}, \eta_{v1}, \eta_{v2}) = (-0.755, -0.4650, (-0.13, -0.19, -0.31))$, which resembles the probability of infection being 0.026 in vaccine arm on the cohort of $Y^\tau(1) = 0$.

Step 6 Generate $Y^\tau(0)$:

- For each realization with $Y^\tau(1) = 1$, $Y^\tau(0)$ was set to 1;
- For each realization with $Y^\tau(1) = 0$ given $S(1)$, X and $Y(1)$, $Y^\tau(0)$ was drawn from a Bernoulli($w(S(1), X, Y(1); \beta, \alpha)$) where $w(s_1, x, y; \beta, \alpha) = \frac{exp(\alpha_0+\alpha_s s_1+\alpha_x x+\beta y)}{1+exp(\alpha_0+\alpha_s s_1+\alpha_x x+\beta y)}$ and $\alpha = (\alpha_0, \alpha_s, \alpha_x)$. In all simulations,$\alpha_s = -log(2)$ so that for 1 unit decrease in $S(1)$, the odds of being in the EAAR stratum doubled, and $\alpha_x = (log(1.1), log(1.3), log(1.7))$. Parameter value, α_0, was chosen so that $P(Y^\tau(0) = 0|Y^\tau(1) = 0) \approx 0.96$, yielding α_0 =5.4 when β =-3; α_0 =5.15 when β =-1; α_0 =4.95 when β =0.

Step 7 Generate $Y(0)$:

- For realizations with $Y^\tau(0) = 1$, $Y(0)$ was set to 1;
- For realizations with $Y^\tau(0) = 0$, $Y(0)$ was drawn from a Bernoulli($g_p(s_1, x; \eta_p)$) distribution where $g_p(s_1, x; \eta_p) = \Phi(\eta_{p0} + \eta_{p1}s_1 + \eta_{p2}x)$. The parameter $\eta_p = (\eta_{p0}, \eta_{p1}, \eta_{p2}) =$

$(-0.755, -0.22, (-0.13, -0.19, -0.31))$, which resembles the probability of infection being 0.07 in placebo arm on the $Y^\tau(0) = 0$ cohort.

Step 8 Only record observables:

- If $Z = 1$, then $S = S(1), Y = Y(1), Y^\tau = Y^\tau(1)$
- If $Z = 0$, then $S = NA, Y = Y(0), Y^\tau = Y^\tau(0)$

Step 9 Set $S = *$ if $Y^\tau = 1$

Step 10 For all vaccine recipient cases with $Z = 1, Y^\tau = 0$, and $Y = 1$, the value of S was retained (100% case sampling). For vaccine recipient controls with $Z = 1, Y = 0$, a random sample of 80% had their value of S set to NA, to create a 20% random sample of controls into the second-phase immunogenicity subset.

MLEs were obtained using quasi-newton methods implemented in R using the function $optim(\cdot)$. Obtaining MLEs requires maximizing over a likelihood containing an integral that is not in closed form. In our simulations, we noticed that there were several local maxima, therefore, it is important to specify good initial parameter values. Our analyses were run using multiple initial values, and the one that gave the largest estimated likelihood was chosen to be our MLE.

We examined the performance of the ML estimator of VE curve with its variance estimated from 250 bootstrap repetitions. Wald-based 95% point-wise confidence intervals (CIs) and simultaneous confidence bands (CBs) were calculated on the logRR scale. We also compared bootstrap estimated standard errors (SEs) with perturbation SEs and the results were very similar.

Figure 5.1 displays the performance of the maximum estimated likelihood estimators of VE(s_1) when sensitivity parameters are correctly specified to match the values used to generate the data. In each of 250 simulations, 95% Wald point-wise CIs and simultaneous CBs are computed from 250 bootstrap replicates on the logRR scale . The proposed method yields up to 10% bias when s_1 is close to 0. Bias is around 0 when s_1 is larger than 2.5. The empirical coverage levels of the 95% pointwise CIs and simultaneous CBs are close to the nominal level of 95%.

Figure 5.2 displays the sensitivity analyses on one simulated data, which reveal how estimates for the VE curve vary over different values of the sensitivity parameter β. This plot is called

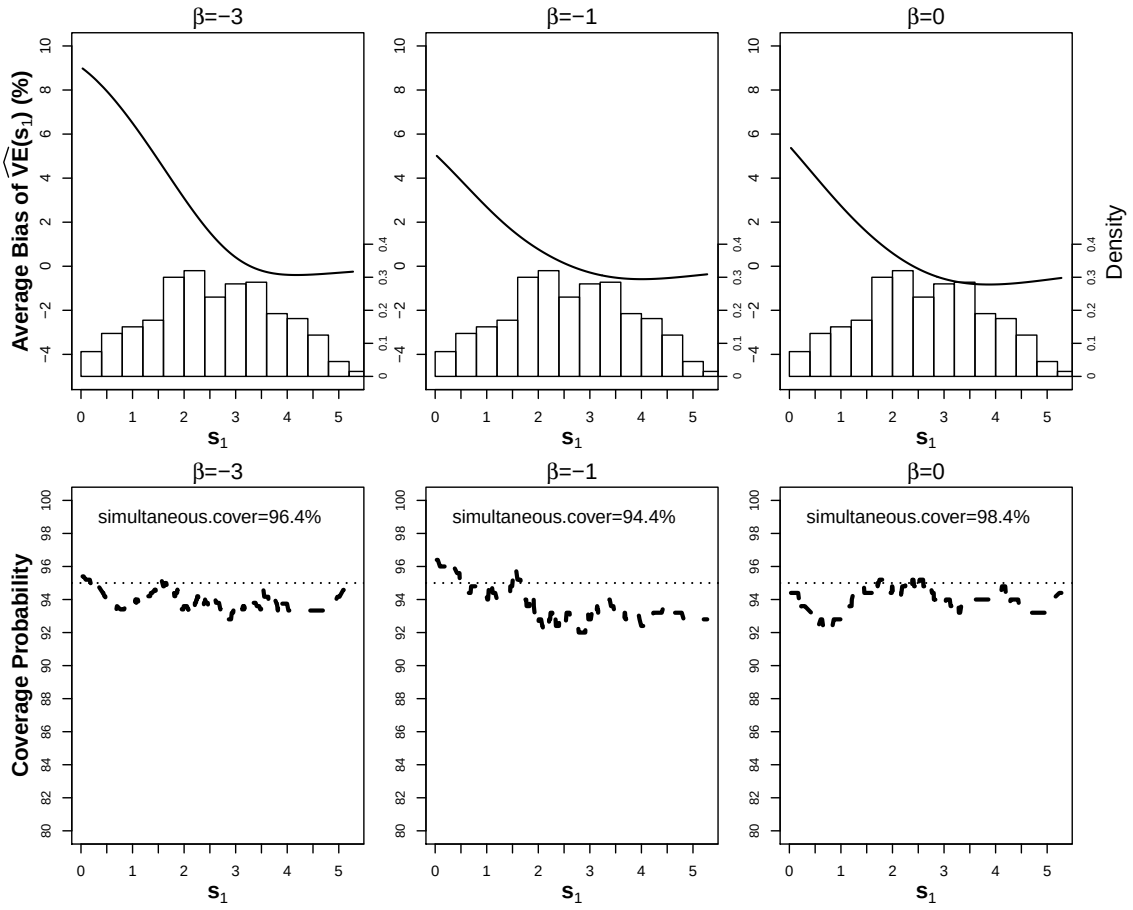

Figure 5.1: Histogram of S in the vaccine group, and bias and coverage for $\widehat{VE}(s_1)$ averaging over the 250 simulations, estimated via maximum estimated likelihood when β is correctly specified to match the three values(-3, -1, 0) used to generate the data under $\mathcal{M}_a$.

the Estimated Region of Ignorance by Molenberghs et. al. ([16]). It is the range of estimates corresponding to a plausible range of sensitivity parameter values and it expresses ignorance due to the missing data and untestable assumptions.

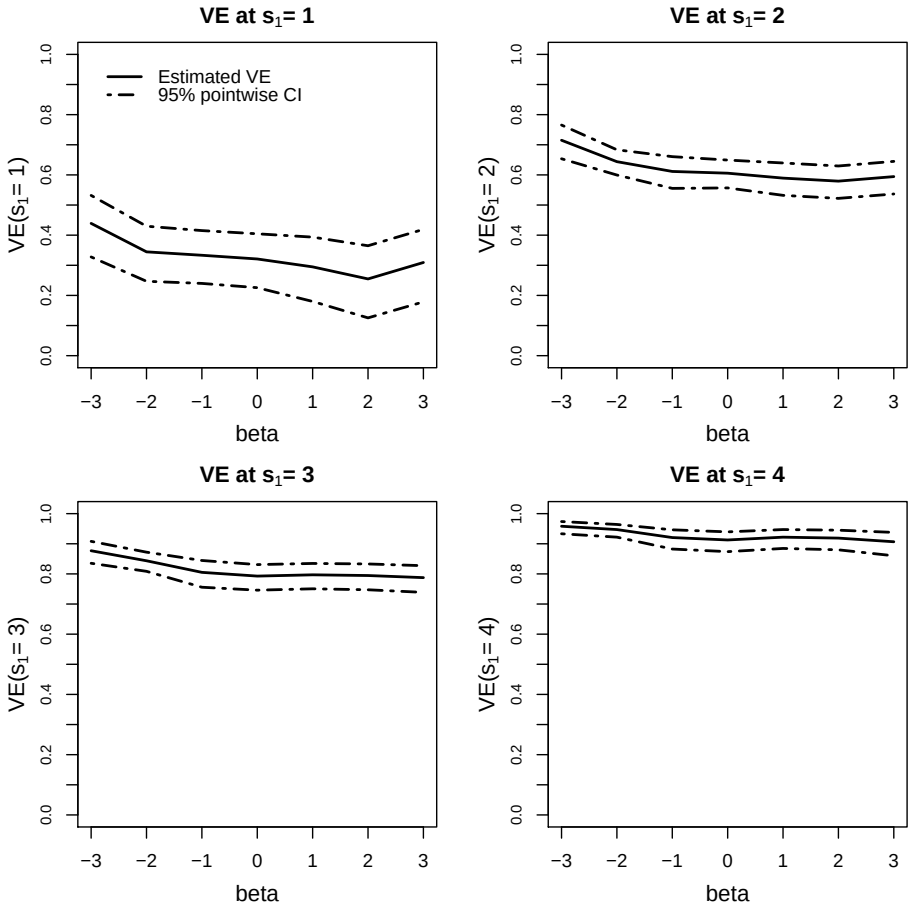

Figure 5.2: Estimated Region of Ignorance with β in the range of (-3,3) for VE($s_1 = 1$), VE($s_1 = 2$), VE($s_1 = 3$), and VE($s_1 = 4$) under $\mathcal{M}_a$.

Table 5.3 summarizes the performance over 250 simulations under three different true β values (-3, -1, 0). For each true β value, we consider three possible presumed sensitivity parameter settings: presumed β=-3, -1, and 0. Therefore, each simulation yields three estimates of VE(s_1) for each true β value. Wald C.I. for these three estimates of VE(s_1) are also calculated (on the logRR

scale) based on 250 bootstrap replicates. The Median 95% Estimated Uncertainty Interval Width (MEUIW) reported in the table acknowledges sampling imprecision in addition to the structural lack of information.

Table 5.3: Sensitivity analysis in 250 simulated trials under three different true β values (-3, -1, 0) under $\mathcal{M}_a$. For each simulation, we consider presumed $\beta = -3, -1, 0$, yielding three estimated of VE(s_1) from which the minimum, maximum, and range are computed. (Min, Max)[Rng] give the medians of these three statistics over the 250 simulations. MEUIW= median 95% estimated uncertainty interval (EUI) width, where the EUI is the union of the 95% Wald C.I for all presumed sensitivity parameter settings. And the median is taken over the 250 simulations.

	Est. VE($s_1 = 1$) (%)		Est. VE($s_1 = 3$) (%)		Est. VE($s_1 = 5$) (%)	
True β	(Min, Max)[Rng]	MEUIW	(Min, Max)[Rng]	MEUIW	(Min, Max)[Rng]	MEUIW
0	(34.6, 44.9)[9.7]	28.0	(81.0, 89.0)[7.8]	14.5	(97.0, 99.1)[2.0]	4.6
-1	(34.9, 46.0)[10.9]	29.2	(81.1, 88.8)[7.7]	14.5	(97.1, 99.1)[2.0]	5.5
-3	(35.1, 47.8)[13.0]	33.5	(81.0,88.6)[7.4]	14.6	(97.0, 98.9)[1.9]	5.1

5.5.1.1 Effects of infection probabilities on bias

It is worthwhile noting that in Figure 5.1, the bias is worse near zero. We investigate the effects of infection probabilities on bias in this section. The parameters in our simulation settings were chosen to reflect the characteristics of the two Phase 3 Dengue trials. The specific infection probabilties are sumarized in Table 5.4.

Table 5.4: Original probabilities of infection used to choose simulation parameters under $\mathcal{M}_a$

	By time τ	from τ to end of follow-up
Under Vaccine	0.02	0.026
Under Placebo	0.04	0.070

In a new set of simulations, we choose simulation parameters to reflect much higher infection probabilities. To be specific, infection probabilities in this new setting are summarized in Table 5.5.

Table 5.5: New probabilities of infection used to choose simulation parameters under $\mathcal{M}_a$

	By time τ	from τ to end of follow-up
Under Vaccine	0.10	0.13
Under Placebo	0.20	0.35

The results on estimation bias under this new simulation setting are displayed in Figure 5.3, which showed negligible bias across all s_1 values. In summary, our estimation procedure for estimating the VE curve had worse bias near zero when the disease is rare while it yielded negligible bias as the rate of disease gets larger.

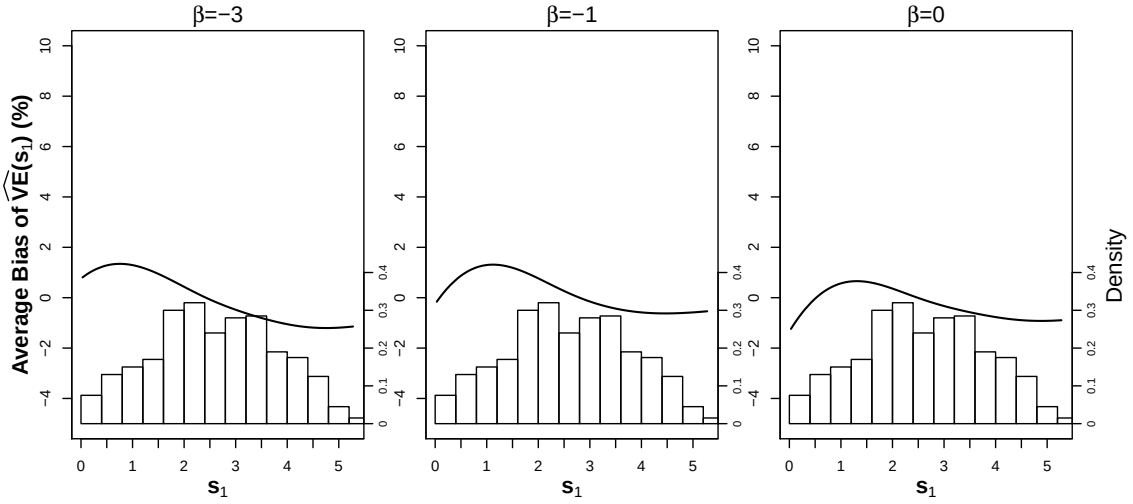

Figure 5.3: Histogram of S in the vaccine group, and bias for $\widehat{VE}(s_1)$ averaging over the 250 simulations, estimated via maximum estimated likelihood when β is correctly specified to match the three values(-3, -1, 0) used to generate the data under the new simulation setting under $\mathcal{M}_a$.

5.5.2 Simulations under $\mathcal{M}_b$

Simulation data generation procedures: Step 1-4 are identical to procedures under $\mathcal{M}_a$.

Step 5 Generate $Y(1)$:

- For realizations with $Y^\tau(1) = 1$, $Y(1)$ was set to 1;
- For realizations with $Y^\tau(1) = 0$, $Y(1)$ was drawn from a Bernoulli($g_v^*(s_1, x; \alpha, \beta, \eta_v^b)$) distribution where $g_v^*(s_1, x; \alpha, \beta, \eta_v^b) = \frac{w^{-1}(s_1,x,y=1;\beta,\alpha) g_v^{EAAR}(s_1,x;\eta_v^b)}{w^{-1}(s_1,x,y=1;\beta,\alpha) g_v^{EAAR}(s_1,x;\eta_v^b) + w^{-1}(s_1,x,y=0;\beta,\alpha)(1-g_v^{EAAR}(s_1,x;\eta_v^b))}$, with $w(s_1, x, y; \beta, \alpha) = \frac{exp(\alpha_0+\alpha_s s_1+\alpha_x x+\beta y)}{1+exp(\alpha_0+\alpha_s s_1+\alpha_x x+\beta y)}$ and $g_v^{EAAR}(s_1, x; \eta_v^b) = \Phi(\eta_{v0} + \eta_{v1} s_1 + \eta_{v2} x)$. We detail out choices of $\alpha = (\alpha_0, \alpha_s, \alpha_x)$ and η_v^b in Step 6.

Step 6 Generate $Y^\tau(0)$:

- For each realization with $Y^\tau(1) = 1$, $Y^\tau(0)$ was set to 1;
- For each realization with $Y^\tau(1) = 0$ given $S(1)$, X and $Y(1)$, $Y^\tau(0)$ was drawn from a Bernoulli($w(S(1), X, Y(1); \beta, \alpha)$). In all simulations,$\alpha_s = -log(2)$ so that for 1 unit decrease in $S(1)$, the odds of being in the EAAR stratum doubled, and $\alpha_x = (log(1.1), log(1.3), log(1.7))$. Parameter values, α_0 and η_v^b, were chosen together so that $P(Y^\tau(0) = 0 | Y^\tau(1) = 0) \approx 0.96$ and probability of infection in vaccine arm $\approx$ 0.026, yielding $\eta_{v2} = (-0.13, -0.19, -0.31)$ and $\eta_{v1} = -0.465$ for all β values. α_0 =-1.18 and η_{v0} =-1.84 when β =-3; α_0 =-1.48 and η_{v0} =-1.15 when β =-1; α_0 =-1.8 and η_{v0} =-0.76 when β =0.

Step 7 Generate $Y(0)$:

- For realizations with $Y^\tau(0) = 1$, $Y(0)$ was set to 1;
- For realizations with $Y^\tau(0) = 0$, $Y(0)$ was drawn from a Bernoulli($g_p(s_1, x; \eta_p)$) distribution where $g_p(s_1, x; \eta_p) = \Phi(\eta_{p0} + \eta_{p1} s_1 + \eta_{p2} x)$. The parameter $\eta_p = (\eta_{p0}, \eta_{p1}, \eta_{p2}) = (-0.77, -0.22, (-0.13, -0.19, -0.31))$, which resembles the probability of infection being 0.07 in placebo arm on the $Y^\tau(0) = 0$ cohort.

Step 8-10 are identical to procedures under $\mathcal{M}_a$.

Figure 5.5, Figure 5.6, and Table 5.6 are an analog of Figure 5.1, Figure5.2, and Table 5.3 but for $\mathcal{M}_b$. The simulation results for $\mathcal{M}_b$ are similar to $\mathcal{M}_a$.

5.5.2.1 Effects of infection probabilities on bias

We observed a similar pattern on bias as $\mathcal{M}_a$ where the bias is worse near zero. And similarly, we investigate the effects of infection probabilities on bias under $\mathcal{M}_b$ in this section. In a new set of simulations, we choose simulation parameters to reflect much higher infection probabilities summarized in Table 5.5.

The results on estimation bias under this new simulation setting are displayed in Figure 5.4, which showed negligible bias across all s_1 values. In summary, our estimation procedure for estimating the VE curve had worse bias near zero when the disease is rare while it yielded negligible bias as the rate of disease gets larger.

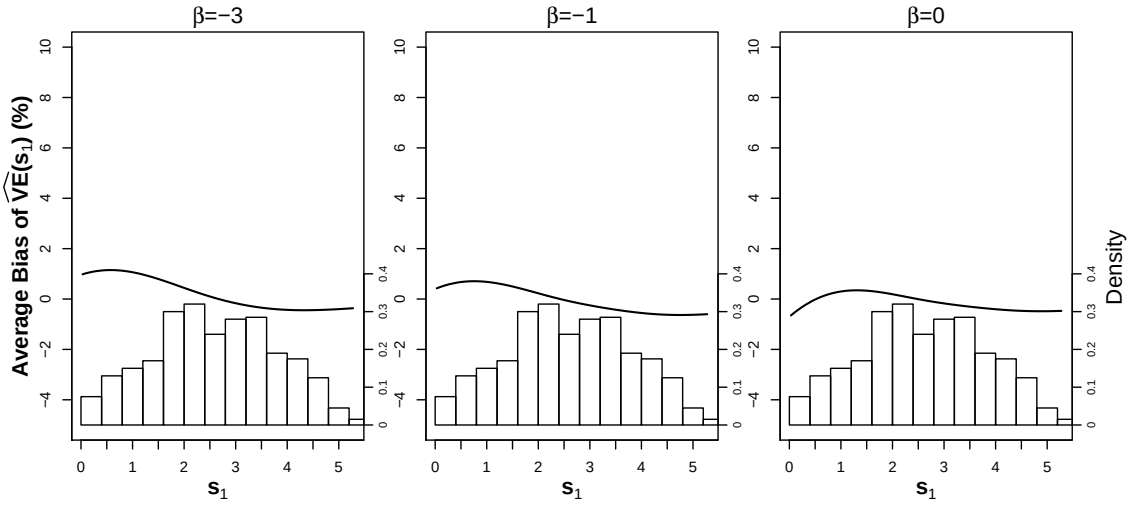

Figure 5.4: Histogram of S in the vaccine group, and bias for $\widehat{VE}(s_1)$ averaging over the 250 simulations, estimated via maximum estimated likelihood when β is correctly specified to match the three values(-3, -1, 0) used to generate the data under the new simulation setting under $\mathcal{M}_b$.

To conclude, our proposed procedure had negligible bias when s_1 was larger than 2.5 and less

than 10% bias when s_1 was close to 0. The empirical coverage levels of the 95% pointwise CIs and simultaneous CBs were close to the nominal level of 95%. The 95% estimated uncertainty interval (EUI) width increased as the magnitude of β increased, indicating that trials with large absolute value of true β (meaning there is a large difference between the odds of remaining early at risk under placebo given early at risk under vaccine for $Y(1) = 1$ and the odds for $Y(1) = 0$) may have low power to detect vaccine efficacy moderation. EUI widths were consistently narrower under $\mathcal{M}_a$ compared to $\mathcal{M}_b$.

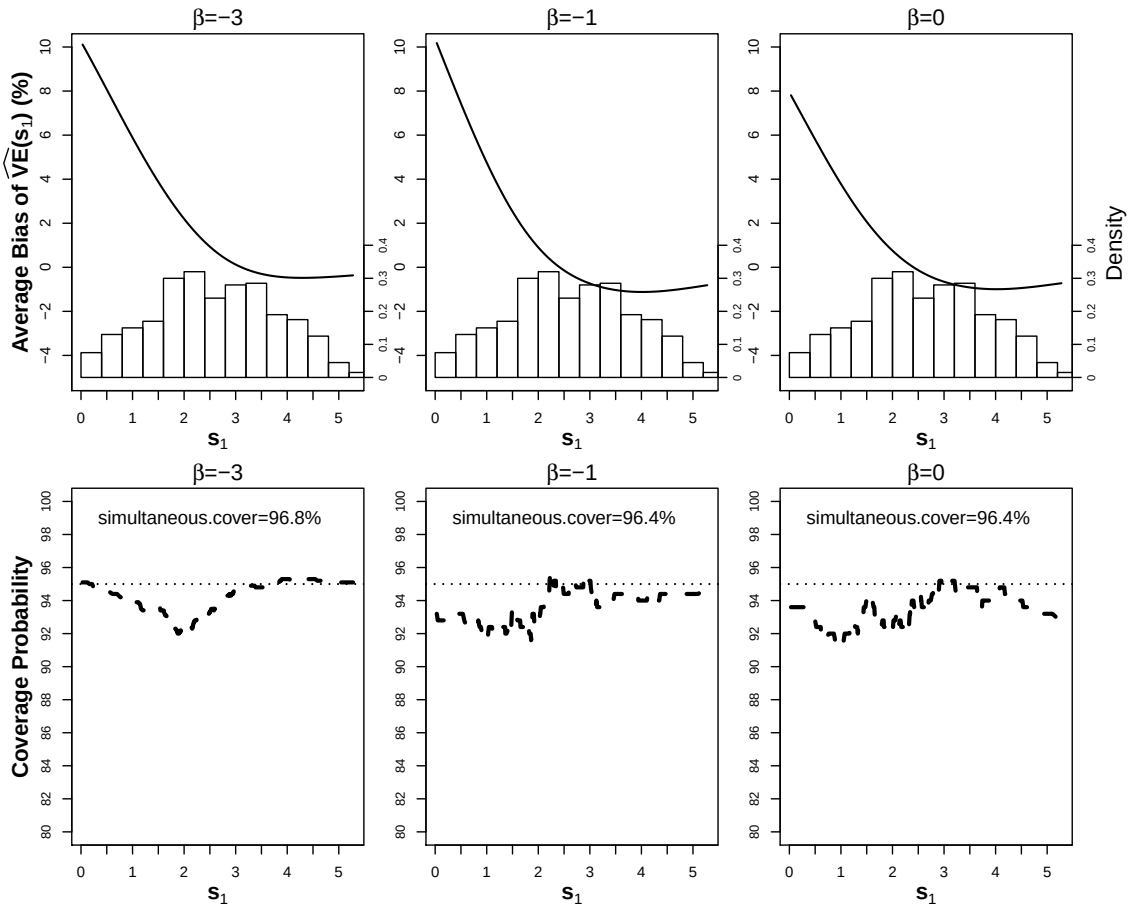

Figure 5.5: Histogram of S in the vaccine group, and bias and coverage for $\widehat{VE}(s_1)$ averaging over the 250 simulations, estimated via maximum estimated likelihood when β is correctly specified to match the three values(-3, -1, 0) used to generate the data under $\mathcal{M}_b$.

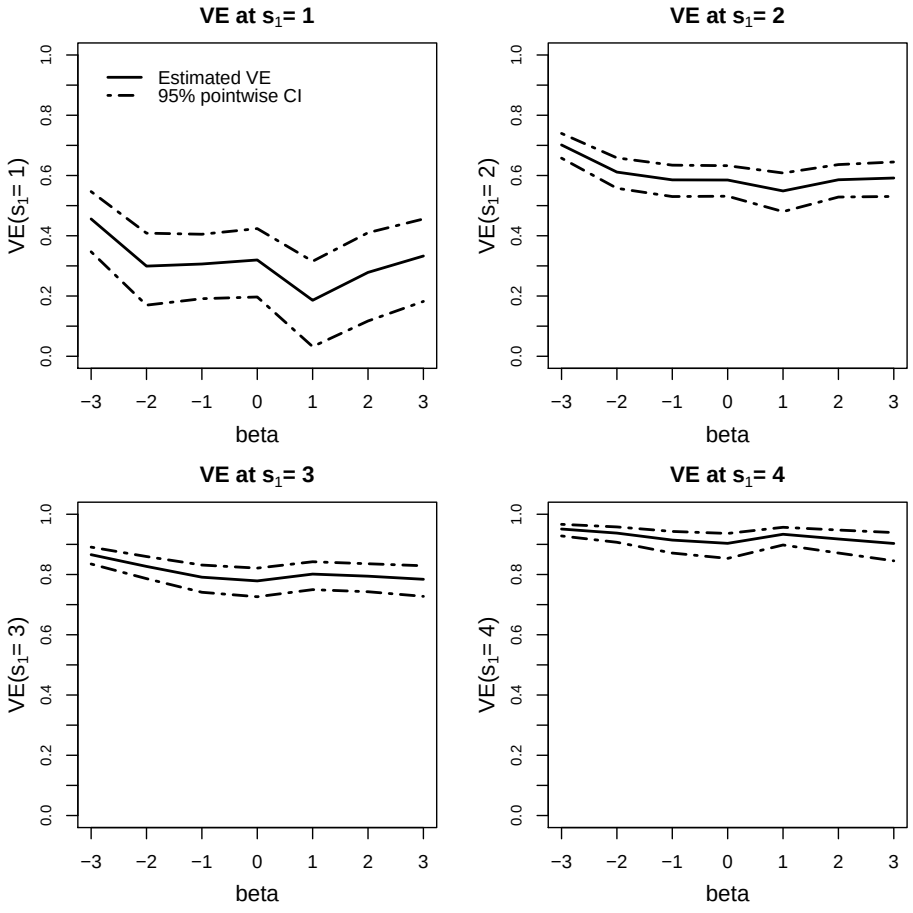

Figure 5.6: Estimated Region of Ignorance with β in the range of (-3,3) for VE($s_1 = 1$), VE($s_1 = 2$), VE($s_1 = 3$), and VE($s_1 = 4$) under $\mathcal{M}_b$.

Table 5.6: Sensitivity analysis in 250 simulated trials under three different true β values (-3, -1, 0) under $\mathcal{M}_b$. For each simulation, we consider presumed $\beta = -3, -1, 0$, yielding three estimated of VE(s_1) from which the minimum, maximum, and range are computed. (Min, Max)[Rng] give the medians of these three statistics over the 250 simulations. MEUIW= median 95% estimated uncertainty interval (EUI) width, where the EUI is the union of the 95% Wald C.I for all presumed sensitivity parameter settings. And the median is taken over the 250 simulations.

	Est. VE($s_1 = 1$) (%)		Est. VE($s_1 = 3$) (%)		Est. VE($s_1 = 5$) (%)	
True β	(Min, Max)[Rng]	MEUIW	(Min, Max)[Rng]	MEUIW	(Min, Max)[Rng]	MEUIW
0	(33.8, 46.2)[12.0]	34.6	(80.1, 88.0)[7.8]	15.0	(96.7, 98.8)[2.0]	5.8
-1	(34.4, 47.4)[13.1]	38.0	(79.4, 87.3)[7.9]	15.6	(96.4, 98.6)[2.1]	7.1
-3	(39.0, 52.3)[12.8]	37.2	(77.6,86.6)[9.0]	17.7	(94.7, 98.2)[3.5]	11.8

5.6 Dengue Example

We demonstrate our proposed method to data pooling across CYD14 and CYD15 9-16 year olds to assess how VE varied with neutralizing antibody titers if assigned to receive the vaccine. An brief background of trial CYD14 and CYD15 are provided in chapter 2. Neutralizing antibody titers were measured to each of the four dengue serotype strains at Month 13 (time τ) in case-control samples. We let S be the individual's sample average response to the four serotype-specific log10-transformed antibody titers and include the subjects' age and country in the baseline covariate vector X. Y^τ is the indicator of virologically confirmed dengue (VCD) disease occurring prior to month 13 and Y is the indicator of (VCD) occurring anytime from baseline to end of the active phase of follow-up (Month 25).

Figure 5.7, Figure 5.8, Figure 5.9, and Figure 5.9 show the estimated VE for β in [-3,3] using both model $\mathcal{M}_a$ and $\mathcal{M}_b$. The range of [-3,3] for β was chosen to reflect various possibilities about the relationship between the distribution of $Y(1)$ in the early always at risk stratum and early protected stratum. Absolute values of β as large as 3 correspond to up to an odds ratio of $e^3 \approx 20$. Thus I chose the plausible range for sensitivity analysis of β to be [-3,3]. However, intuitively I believe β is non-positive because e^β is the odds ratio of remaining early at risk under placebo given early at risk under vaccine and infected under vaccine ($Y^\tau(1) = 0, Y(1) = 1$) versus given early at

risk under vaccine and non-infected under vaccine ($Y^{\tau}(1) = 0, Y(1) = 0$), adjusting for $S(1)$ and X. If someone stayed at risk up to time τ but got infected during follow-up when assigned to vaccine, his chance of staying at risk up to time τ when assigned to placebo is likely to be smaller than those who stayed at risk up to time τ and did not get infected during follow-up when assigned to vaccine, implying β is more likely to be in the negative range. This belief is yet to be confirmed by experts on this topic. For now, I vary β between -3 and 3. The fact that a horizontal line cannot be drawn between the simultaneous confidence bands without intersecting both the lower and upper limits in any β value between -3 and 3 indicates that VE(s_1) significantly varies with s_1.

5.7 Discussion

In this chapter we have proposed sensitivity analysis methods for evaluating the causal treatment effect modification by post-randomization biomarker-defined principal strata with a more plausible assumption: A3-mon: $P(Y^{\tau}(1) \geq Y^{\tau}(0)) = 0$, based on a maximum likelihood approach. Under this plausible assumption of early vaccine efficacy, we first constructed a likelihood. This likelihood is general, the outcome of interest (Y) could be continuous or discrete. In our example we studied Y being discrete, indicating infection status but the likelihood is applicable to Y being continuous as well. One may also incorporate different models for $w(\cdot)$, the probability of a subject who remains early at risk under vaccine being an early always at risk individual. Lastly, it should be emphasized that these sensitivity analyses are meant to supplement other analyses as a tool to address whether and how treatment efficacy varies by the post-randomization biomarker and obtaining consensus among experts about the plausible range of sensitivity parameters could be useful to make the analyses more meaningful and interpretable.

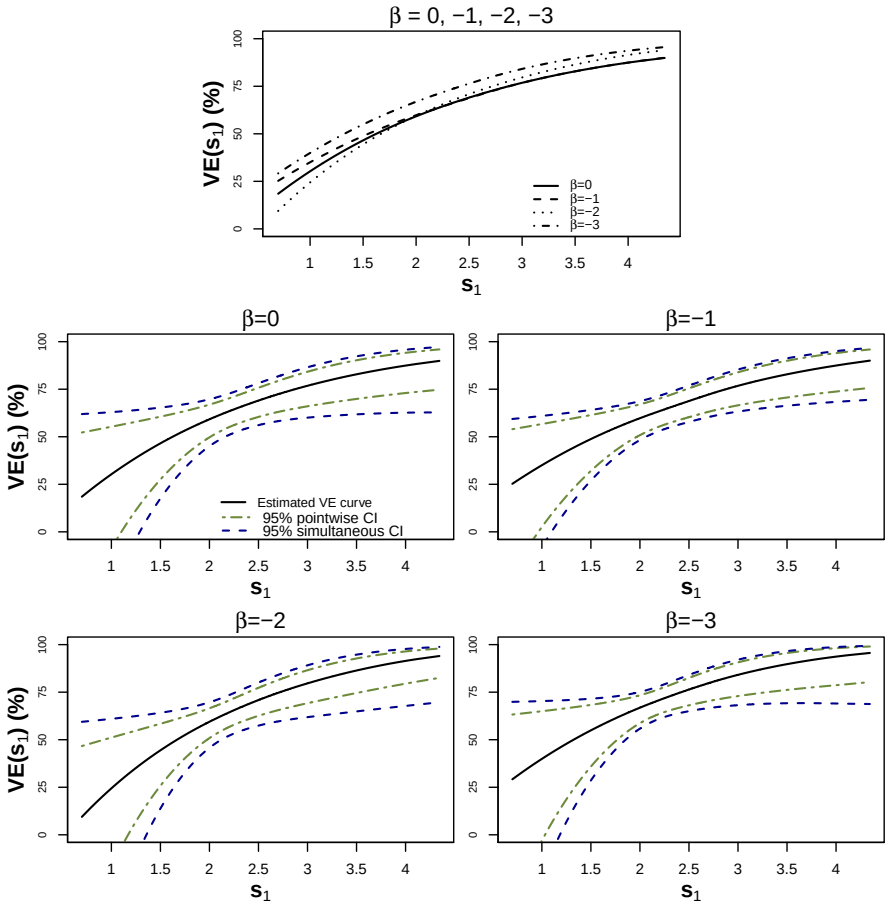

Figure 5.7: Sensitivity analysis of vaccine efficacy against dengue disease of any serotype by Month 13 average titer under vaccine in 9-16 year olds in the two Phase 3 dengue vaccine efficacy trials combined (CYD14 and CYD15) under $\mathcal{M}_a$. Estimated VE for $\beta = 0, -1, -2, -3$ values are shown together in the first panel. Estimated VE with 95% pointwise CI and 95% simultaneous CI for each of the four β values are shown in the following four panels.

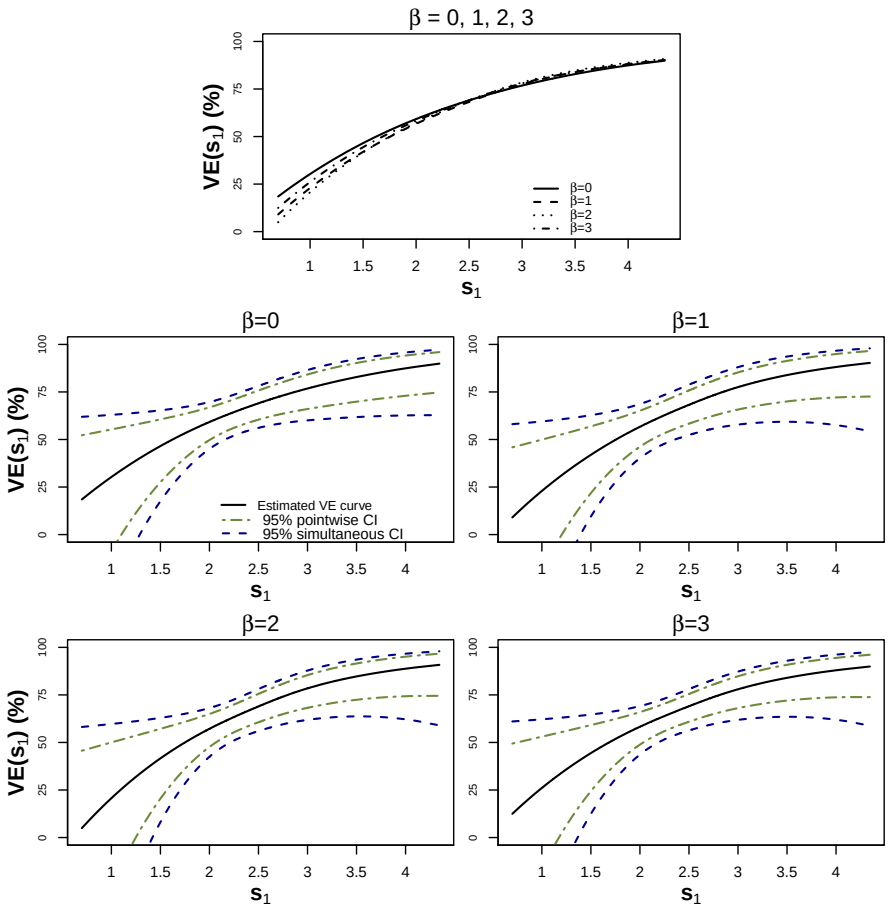

Figure 5.8: Continued Sensitivity analysis of vaccine efficacy against dengue disease of any serotype by Month 13 average titer under vaccine in 9-16 year olds in the two Phase 3 dengue vaccine efficacy trials combined (CYD14 and CYD15) under $\mathcal{M}_a$. Estimated VE for $\beta = 0, -1, -2, -3$ values are shown together in the first panel. Estimated VE with 95% pointwise CI and 95% simultaneous CI for each of the four β values are shown in the following four panels.

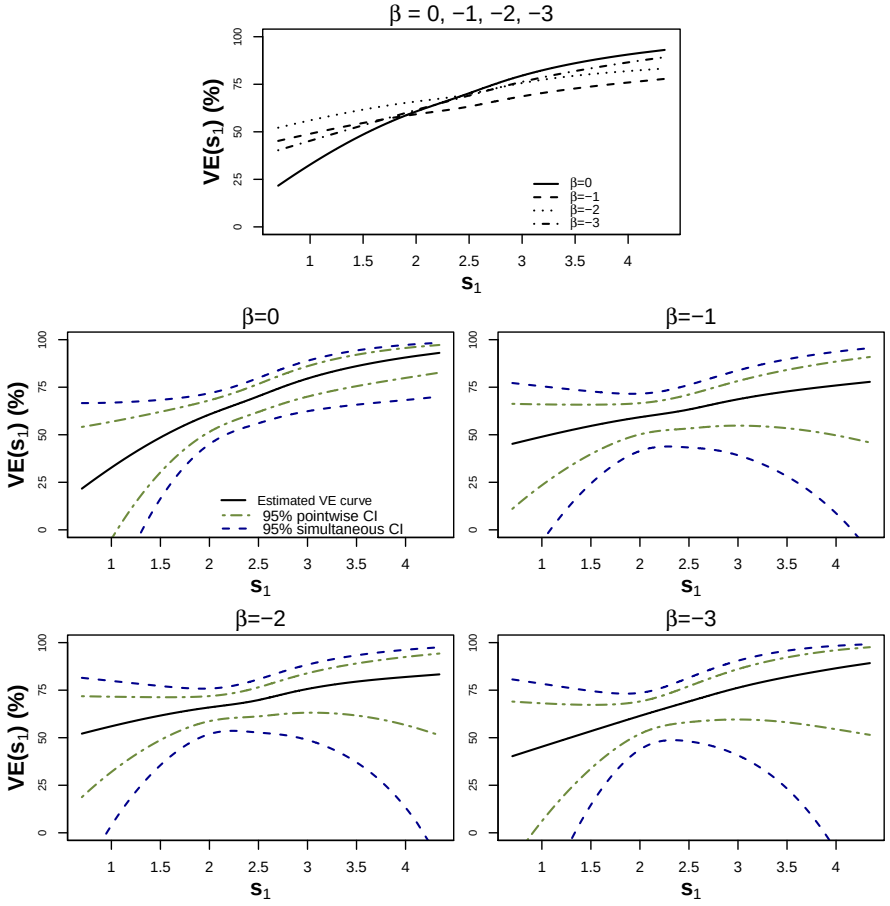

Figure 5.9: Sensitivity analysis of vaccine efficacy against dengue disease of any serotype by Month 13 average titer under vaccine in 9-16 year olds in the two Phase 3 dengue vaccine efficacy trials combined (CYD14 and CYD15) under $\mathcal{M}_b$. Estimated VE for $\beta = 0, -1, -2, -3$ values are shown together in the first panel. Estimated VE with 95% pointwise CI and 95% simultaneous CI for each of the four β values are shown in the following four panels.

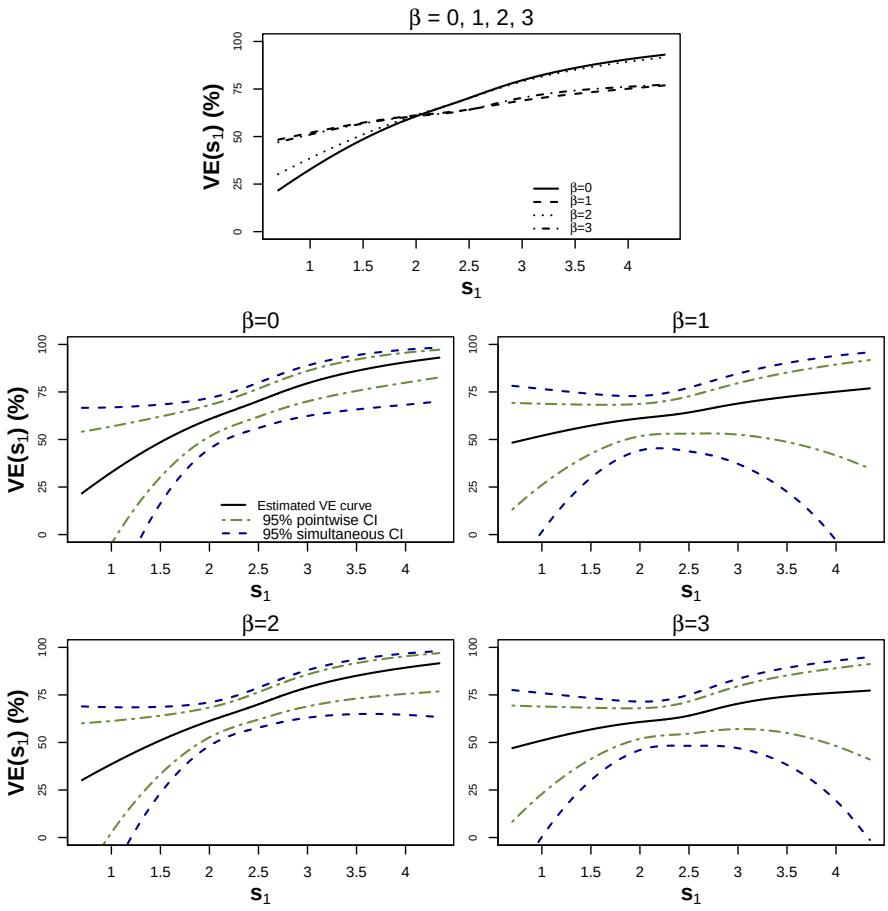

Figure 5.10: Continued Sensitivity analysis of vaccine efficacy against dengue disease of any serotype by Month 13 average titer under vaccine in 9-16 year olds in the two Phase 3 dengue vaccine efficacy trials combined (CYD14 and CYD15) under $\mathcal{M}_b$. Estimated VE for $\beta = 0, -1, -2, -3$ values are shown together in the first panel. Estimated VE with 95% pointwise CI and 95% simultaneous CI for each of the four β values are shown in the following four panels.

Chapter 6
FUTURE RESEARCH

The identification and evaluation of response biomarkers as surrogate endpoints and as effect modifiers are important questions to address in many biomedical research areas. A response type biomarker shown to be a strong modifier of clinical treatment efficacy can be used to help with trial design, trial implementation, and exploration of biological mechanisms of clinical treatment efficacy. Furthermore, clinical treatment effect modification analysis by a post-randomization biomarker can serve as an important tool in bridging or predicting the vaccines effect in further trial settings [10].

In this dissertation I first proposed a procedure for evaluating treatment effect modification by post-randomization biomarker-defined principal strata based on a pseudo-score estimator, allowing clinical risks under Z being dependent on X after conditioning on S. I also developed procedures for obtaining pointwise and simultaneous confidence intervals about the marginal CEP curve via perturbation resampling. I then developed an estimated likelihood approach to evaluate the bivariate treatment effect modification by the post-randomization biomarker-based principal strata and baseline covariates in general settings without requirements of a nested sub-sampling relationship between the immune response biomarker and baseline predictors. Lastly, I turned my attention to relaxing the equal early clinical risk assumption to a more plausible monotone early clinical risk assumption.

I believe that there are many other interesting questions either stemming from the work in this dissertation or closely related to it. In this section I briefly mention a few ideas for future study.

In chapter 5 we presented our work for a BIP-only design but it could be easily extended to allow for a CPV component. I would like to perform simulation studies to evaluate how precision compares between the BIP-only design and BIP+CPV design. Regarding sensitivity analyses, in its fullest sense, we would consider not just varying β, but also varying α and the form of $w(s_1, x, y; \beta, \alpha)$. Furthermore, different risk modeling approaches could be explored. For example, in addition to models $\mathcal{M}_a$ and $\mathcal{M}_b$, one could imagine a third parametric model, say $\mathcal{M}_c$ based on assumptions A1,

A2, and A3-mon, that models $P(Y(0) = 1|Y^{\tau}(0) = 0, S, X)$, $P(Y(1) = 1|Y^{\tau}(1) = Y^{\tau}(0) = 0, S, X)$, and instead of $w(\cdot)$, it models $P(Y(1) = 1|Y^{\tau}(1) = 1, Y^{\tau}(0) = 0, S, X)$, which is the risk function for the early protected. One advantage to $\mathcal{M}_c$ over $\mathcal{M}_b$ is that we no longer need to require $w(s_1, x, y; \beta, \alpha) > 0$ for all s_1, x, and y. To be able to yield fruitful results, any modeling approach should use a sensitivity parameter with a meaningful interpretation and plausible range of the sensitivity parameter should be chosen intuitively, and independently from the data.

Throughout this dissertation I have used the outcome variable Y to indicate the clinical endpoint event during the study follow-up period and one important research direction is extending these methods to time-to-event outcomes to handle independently right censored data. This will allow one to assess how VE varies with a post-randomization biomarker for a given time t. In chapter 3, we proposed a perturbation resampling method to make simultaneous inference about the VE curve over a range of biomarker values. When Y is extended to a time-to-event variable, it would be interesting (and fun) to develop a procedure to draw simultaneous inference over a range of biomarker values AND over a range of t.

There are other issues in practice that may warrant further study as well. For example, the observable infection diagnosis time is different from the unobservable infection time. People who have the same infection time may have very different diagnosis time and sensitivity analysis may be of interest to account for this unequal diagnosis time.

BIBLIOGRAPHY

[1] Maria Rosario Capeding, Ngoc Huu Tran, Sri Rezeki S Hadinegoro, Hussain Imam HJ Muhammad Ismail, Tawee Chotpitayasunondh, Mary Noreen Chua, Chan Quang Luong, Kusnandi Rusmil, Dewa Nyoman Wirawan, Revathy Nallusamy, et al. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *The Lancet*, 384(9951):1358–1365, 2014.

[2] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

[3] Nilanjan Chatterjee, Yi-Hau Chen, and Norman E Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168, 2003.

[4] Dean Follmann. Augmented designs to assess immune response in vaccine trials. *Biometrics*, 62(4):1161–1169, 2006.

[5] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[6] Erin E Gabriel and Peter B Gilbert. Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics*, 15(2):251–65, 2014.

[7] Peter B Gilbert, Erin E Gabriel, Ying Huang, and Ivan SF Chan. Surrogate endpoint evaluation: Principal stratification criteria and the prentice definition. *Journal of causal inference*, 3(2):157–175, 2015.

[8] Peter B Gilbert, Douglas Grove, Erin Gabriel, Ying Huang, Glenda Gray, Scott M Hammer, Susan P Buchbinder, James Kublin, Lawrence Corey, and Steven G Self. A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple hiv vaccine regimens. *Statistical Communications in Infectious Diseases*, 2(1):1, 2010.

[9] Peter B Gilbert, Douglas Grove, Erin Gabriel, Ying Huang, Glenda Gray, Scott M Hammer, Susan P Buchbinder, James Kublin, Lawrence Corey, and Steven G Self. A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple hiv vaccine regimens. *Statistical communications in infectious diseases*, 3(1), 2011.

[10] Peter B Gilbert and Ying Huang. Predicting overall vaccine efficacy in a new setting by re-calibrating baseline covariate and intermediate response endpoint effect modifiers of type-specific vaccine efficacy. *Epidemiologic Methods*, 5(1):93–112, 2016.

[11] Peter B Gilbert and Michael G Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4):1146–1154, 2008.

[12] Ying Huang. Evaluating principal surrogate markers in vaccine trials in the presence of multiphase sampling. *Biometrics*, 2017.

[13] Ying Huang and Peter B Gilbert. Comparing biomarkers as principal surrogate endpoints. *Biometrics*, 67(4):1442–1451, 2011.

[14] Ying Huang, Peter B Gilbert, and Julian Wolfson. Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics*, 69(2):301–309, 2013.

[15] Michael G Hudgens and Peter B Gilbert. Assessing vaccine effects in repeated low-dose challenge experiments. *Biometrics*, 65(4):1223–1232, 2009.

[16] Geert Molenberghs, Michael G Kenward, and Els Goetghebeur. Sensitivity analysis for incomplete contingency tables: the slovenian plebiscite case. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(1):15–29, 2001.

[17] Z Moodie. Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in asia and latin america. 2017.

[18] Margaret Sullivan Pepe and Thomas R Fleming. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413):108–113, 1991.

[19] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

[20] Li Qin, Peter B Gilbert, Dean Follmann, and Dongfeng Li. Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the cox model. *The annals of applied statistics*, 2(1):386, 2008.

[21] Fisher RA. *Statistical methods for research workers*. OLiver and Boyd, 1970.

[22] Supachai Rerks-Ngarm, Punnee Pitisuttithum, Sorachai Nitayaphan, Jaranit Kaewkungwal, Joseph Chiu, Robert Paris, Nakorn Premsri, Chawetsan Namwat, Mark de Souza, Elizabeth Adams, et al. Vaccination with alvac and aidsvax to prevent hiv-1 infection in thailand. *New England Journal of Medicine*, 361(23):2209–2220, 2009.

[23] Kenneth E Schmader, Myron J Levin, John W Gnann Jr, Shelly A McNeil, Timo Vesikari, Robert F Betts, Susan Keay, Jon E Stek, Nickoya D Bundick, Shu-Chih Su, et al. Efficacy, safety, and tolerability of herpes zoster vaccine in persons aged 50–59 years. *Clinical infectious diseases*, 54(7):922–928, 2012.

[24] Bryan E Shepherd, Peter B Gilbert, Yannis Jemiai, and Andrea Rotnitzky. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to hiv vaccine trials. *Biometrics*, 62(2):332–342, 2006.

[25] Jeremy MG Taylor, Yue Wang, and Rodolphe Thiébaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4):1102–1111, 2005.

[26] Luis Villar, Gustavo Horacio Dayan, José Luis Arredondo-García, Doris Maribel Rivera, Rivaldo Cunha, Carmen Deseda, Humberto Reynales, Maria Selma Costa, Javier Osvaldo Morales-Ramírez, Gabriel Carrasquilla, et al. Efficacy of a tetravalent dengue vaccine in children in latin america. *New England Journal of Medicine*, 372(2):113–123, 2015.

[27] Julian Wolfson and Peter Gilbert. Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics*, 66(4):1153–1161, 2010.

Appendix A

APPENDIX FOR CHAPTER 3

A.1 The estimating function for γ

The likelihood function for $P(X|S)$ can be derived as

$$L(\gamma) = \prod_{i=1}^{N} p_{1i}^{g_{1i}} \times p_{2i}^{g_{2i}} \times \cdots \times p_{Di}^{g_{Di}}.$$

Taking the log and using the fact that $\sum_j g_{ji} = 1$ for each i, the log-likelihood function is

$$l(\gamma) = \sum_{i=1}^{n} \left[y_{1i}\ ln\ p_{1i} + \cdots + y_{D-1i}\ ln\ p_{D-1i} - ln(1 + p_{1i} + \cdots + p_{D-1i}) \right].$$

The likelihood equations are found by taking the first partial derivatives of $l(\gamma)$ with respect to each of the $(D-1) \times (T+1)$ unknown parameters. The general form of these equations is: $\frac{\partial l(\gamma)}{\partial \gamma_{jt}} = \sum_{i=1}^{N} h_t(s_{1i})(g_{ji} - p_{ji})$ for $j = 1, 2, .., D-1$ and $t = 0, 1, 2, .., T$.

A.2 Asymptotic Distribution for the proposed VE estimator $\widehat{\mathrm{VE}}^{(new)}(s_1)$

A.2.1 Regularity conditions to be satisfied:

1. $risk_{(z)}(S, X) > 0$ and $P(\delta = 1|S, X) > 0$ almost surely.

2. $risk_{(z)}(S, X; \beta)$ is thrice differentiable with respect to β. For β in a neighborhood of the true value β_0, the third derivatives are bounded by an integrable function of (Y, Z, S, W).

3. $P(\delta = 1|Y, Z, X)$ as a function of α is thrice differentiable with respect to α. For α in a neighborhood of the true value α_0, the third derivatives are bounded by an integrable function of (Y, Z, S, X).

4. Ψ_β is nonsingular.

5. For all y, x, z, under β_0 and α_0,

$$0 < \int \frac{P(y|s,z,x)}{P(\delta=1|s,x)} dF(s|x,\delta=1) < \infty$$
$$0 < \int |U_\beta(y|s,z,x)| \frac{P(y|s,z,x)}{P(\delta=1|s,x)} dF(s|x,\delta=1) < \infty.$$

6. For all y, s, x, z, $P(y|s,z,x)/P(\delta=1|s,x)$ and $U_\beta(y|s,z,x)P(y|s,z,x)/P(\delta=1|s,x)$ are twice differentiable with respect to β, α, with the second derivatives uniformly integrable with respect to $F(s|x,\delta=1)$ for (β,α) within a neighborhood of (β_0,α_0).

7. Ψ_γ is nonsingular.

8. $P\|\Psi_{\gamma_0}\|^2 < \infty$ and that the map $\gamma \mapsto P\Psi_\gamma$ is differentiable at a zero γ_0, with a nonsingular derivative matrix.

   For convenient notation, we let $\pi_0 = \pi_{\alpha_0}(Y,Z,X) = P(\delta=0|Y,Z,X;\alpha_0)$, and $\hat{\pi} = \pi_{\hat{\alpha}}(Y,Z,X) = P(\delta=0|Y,Z,X;\hat{\alpha})$.

9. $P\|\Psi_{\alpha_0}\|^2 < \infty$ and that the map $\alpha \mapsto P\Psi_\alpha$ is differentiable at a zero α_0, with a nonsingular derivative matrix.

10. For α in a neighborhood of α_0, where $\zeta > 0$ and ψ satisfies $E\psi^2 < \infty$:

$$\left| \frac{1}{\hat{\pi}} - \frac{1}{\pi_0} - \frac{-\dot{\pi_0}^T}{\pi_0^2}(\alpha - \alpha_0) \right| \leq \psi \left|\alpha - \alpha_0\right|^{1+\zeta}.$$

Conditions 1–6 strictly followed HGW for the asymptotic distribution of the pseudo-score estimator $\hat{\beta}$. Conditions 7–8 are needed to establish the asymptotic distribution of the WL estimator $\hat{\gamma}$. Conditions 9–10 are needed for estimation with estimated sampling weights, where we substitute π_0 with a consistent estimator $\hat{\pi}_\alpha$, where α is a parameter to be estimated by ML from the Phase-I observations. Condition 10 typically follows from Condition 9 provided that π_α has a continuous second derivative.

A.2.2 Asymptotic normality:

Theorem 1: Under specified regularity conditions, as the sample size $N \to \infty$, we have

1.
$$\sqrt{N}(\hat{\beta}-\beta)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_1(\delta_i,Y_i,S_i,Z_i,X_i)+o_p(1)\rightarrow_d N(0,V), \tag{A.1}$$

2.
$$\sqrt{N}(\hat{\gamma}(\hat{\alpha})-\gamma)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_2(\delta_i,Y_i,S_i,Z_i,X_i)+o_p(1)\rightarrow_d N(0,K), \tag{A.2}$$

3.
$$\sqrt{N}\left(\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix}-\begin{pmatrix}\beta\\ \gamma\end{pmatrix}\right)\underset{d}{\rightarrow}N\left(\mathbf{0},\begin{bmatrix}\mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}}\\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}}\end{bmatrix}\right), \tag{A.3}$$

where $\Sigma_{11}=V,\Sigma_{22}=K,\Sigma_{12}=\Sigma_{21}=cov(\phi_1(\delta,Y,S,Z,X),\phi_2(\delta,Y,S,Z,X))$.

Sketch of Proof:

A.2.2.1 Asymptotic normality of $\hat{\beta}$:

The pseudo-score estimator $\hat{\beta}$ is then obtained by solving the equation $U(\beta,F_N,\hat{\pi})=0$, and HGW proved that

$$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta)&=-\dot{\Psi}_{\beta}^{-1}(\beta_0,F_0^*,\pi_0)\sqrt{N}\left\{\Psi_N(\beta_0;F_0^*,\pi_0)+\sum_{i=1}^{K}\Psi_{F_k^*}[F_{Nk}-F_{k0}^*]+\Psi[\hat{\alpha}-\alpha_0]\right\}+o_p(1)\\&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_1(\delta_i,Y_i,S_i,Z_i,X_i)+o_p(1),\end{aligned}$$

where

$$\phi_1(\delta,Y,S,Z,X)=a_0(\delta,Y,S,Z,X)+a_1(\delta,S,X)+a_2(\delta,Y,Z,X),$$

with

$$a_0(\delta,Y,S,Z,X)=\delta U_\beta(Y|S,Z,X)+(1-\delta)E\left\{U_\beta(Y|S,Z,X)|Y,Z,X,\delta=1\right\},$$

$$a_1(\delta,S,X)=\sqrt{N}\dot{\Psi}_{F_k^*}(\beta_0,F_0^*)(F_{Nk}^*-F_{k0}^*),$$

$$a_2(\delta,Y,Z,X)=\dot{\Psi}_\alpha(\beta_0,F_0^*,\pi(\alpha_0))I_\alpha^{-1}U_\alpha(\delta|Y,Z,X)+o_p(1),$$

and

$$V=\dot{\Psi}_\beta^{-1}Var(\phi_0)\dot{\Psi}_\beta^{-t}.$$

A.2.2.2 Asymptotic normality of $\hat{\gamma}$:

The weighted likelihood estimator $\hat{\gamma}$ is obtained by solving:

$$\Psi_N(\gamma) \equiv \frac{\partial l(\gamma)}{\partial \gamma_{jt}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\delta_i}{\pi_0(y_i, z_i, x_i)} h_t(s_{1i})(g_{ji} - p_{ji}) = 0.$$

Furthermore,

$$\frac{\partial^2 l(\gamma)}{\partial \gamma_{jt}\partial \gamma_{jt'}} = -\sum_{i=1}^{N} \frac{\delta_i}{\pi_0(y_i, z_i, x_i)} h_{t'}(s_{1i}) \cdot h_t(s_{1i}) \cdot p_{ji} \cdot (1 - p_{ji}) \tag{A.4}$$

$$\frac{\partial^2 l(\gamma)}{\partial \gamma_{jt}\partial \gamma_{j't'}} = \sum_{i=1}^{N} \frac{\delta_i}{\pi_0(y_i, z_i, x_i)} h_{t'}(s_{1i}) \cdot h_t(s_{1i}) \cdot p_{ji} \cdot p_{j'i}. \tag{A.5}$$

The information matrix $I(\gamma_0)$ can be estimated by the observed information matrix $\hat{I}(\hat{\gamma})$, whose elements are the negatives of the values in equations (A.4) and (A.5) evaluated at $\hat{\gamma}$.

Now following similar steps in (Breslow & Wellner 2006), we apply Theorem 19.26 of van der Vaart (1998) to conclude that

$$\sqrt{N}(\hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{\delta_i}{\pi_0(y_i, z_i, x_i)} \widetilde{l}_0(s_{1i}) + o_p(1)$$

where $\widetilde{l}_0$ denotes the efficient influence function $\widetilde{l}_0(s_{1i}) = I^{-1}(\gamma_0)\dot{l}_0(\gamma_0|s_{1i})$.
The asymptotic variance is therefore:

$$\begin{aligned} Var\sqrt{N}(\hat{\gamma} - \gamma_0) &= Var(\frac{\delta}{\pi_o}\widetilde{l}_0) \\ &= VarE(\frac{\delta}{\pi_o}\widetilde{l}_0|S(1)) + EVar(\frac{\delta}{\pi_o}\widetilde{l}_0|S(1)) \\ &= Var(\widetilde{l}_0) + E\left[\frac{\widetilde{l}_0^{\otimes 2}}{\pi_o^2} Var(\delta|S(1))\right] \\ &= I(\gamma_0)^{-1} + P_0(\frac{1-\pi_0}{\pi_0}\widetilde{l}_0^{\otimes 2}). \end{aligned}$$

A.2.2.2.1 Estimation with estimated sampling weights: In this section, we show that under mild assumptions,

$$\begin{aligned} \sqrt{N}\left[\hat{\gamma}(\hat{\alpha}) - \gamma_0\right] &= \sqrt{N}\left[\hat{\gamma}(\hat{\alpha}) - \hat{\gamma}(\alpha_0)\right] + \sqrt{N}\left[\hat{\gamma}(\alpha_0) - \gamma_0\right] \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \phi_2(\delta_i, Y_i, S_i(1), Z_i, X_i) + o_p(1) \to_d N(0, K). \end{aligned} \tag{A.6}$$

When we substitute π_0 with a consistent estimator $\hat{\pi}_\alpha$, where α is a parameter to be estimated by ML from the Phase-I observations, under regularity Condition 9, the ML estimator $\hat{\alpha}$ is consistent and asymptotically normal with influence function $\widetilde{l}_0^{\alpha}$ so that

$$\sqrt{N}\begin{pmatrix}\hat{\gamma}(\alpha_0)-\gamma_0\\ \hat{\alpha}-\alpha_0\end{pmatrix}=\sqrt{N}\begin{pmatrix}\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\pi_{0i}}\widetilde{l}_0(s_{1i})\\ \frac{1}{N}\sum_{i=1}^{N}\widetilde{l}_0^{\alpha}\end{pmatrix}. \tag{A.7}$$

Moreover,

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}(\frac{\delta_i}{\hat{\pi}_i}-\frac{\delta_i}{\pi_0})\widetilde{l}_0(s_{1i})\\ =\;&\frac{1}{N}\sum_{i=1}^{N}\delta_i\widetilde{l}_0(S_i)\left[\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_0}-\frac{-\dot{\pi_0^T}}{\pi^2}(\hat{\alpha}-\alpha_0)\right]+\frac{1}{N}\sum_{i=1}^{N}\delta_i\widetilde{l}_0(s_{1i})\left[\frac{-\dot{\pi_0^T}}{\pi^2}(\hat{\alpha}-\alpha_0)\right]\\ =\;&R_N-\frac{1}{N}\sum_{i=1}^{N}\delta_i\widetilde{l}_0(s_{1i})\left[\frac{\dot{\pi_0^T}}{\pi^2}(\hat{\alpha}-\alpha_0)\right].\end{aligned} \tag{A.8}$$

By regularity Condition 10,

$$\begin{aligned}|R_N| &= \frac{1}{N}\sum_{i=1}^{N}\delta_i\widetilde{l}_0(S_i)\left[\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_0}-\frac{-\dot{\pi_0^T}}{\pi^2}(\hat{\alpha}-\alpha_0)\right]\\ &\le \frac{1}{N}\sum_{i=1}^{N}\psi\left|\widetilde{l}_0(S_i)\right||\hat{\alpha}-\alpha_0|^{1+\zeta}\\ &= O_p(1)\,|\hat{\alpha}-\alpha_0|\,|\hat{\alpha}-\alpha_0|^{\zeta}=O_p(1)O_p(N^{-1/2})o_p(1).\end{aligned}$$

Multiplying through (A.8) by $\sqrt{N}$, we conclude that equation (A.6) holds by virtue of $\sqrt{N}R_N = o_p(1)$ and the strong law of large numbers.

Furthermore,

$$K=Var\sqrt{N}\left[\hat{\gamma}(\hat{\alpha})-\gamma_0\right]=Var(\frac{\delta}{\pi_0}\widetilde{l}_0)-\frac{\delta}{\pi_0}\cdot\frac{\widetilde{l}_0\dot{\pi_0}^T}{\pi_0}(\frac{\delta}{\pi_0}\cdot\frac{\dot{\pi_0}^{\otimes 2}}{\pi_0(1-\pi_0)})^{-1}\frac{\delta}{\pi_0}\cdot\frac{\dot{\pi_0}\widetilde{l_0}^T}{\pi_0}.$$

A.2.2.3 Asymptotic normality of $(\hat{\beta},\hat{\gamma})$:

$$\sqrt{n}\left(\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix}-\begin{pmatrix}\beta\\ \gamma\end{pmatrix}\right)\underset{d}{\to}N\left(\mathbf{0},\begin{bmatrix}\mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}}\\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}}\end{bmatrix}\right) \tag{A.9}$$

where $\Sigma_{11}=V,\Sigma_{22}=K,\Sigma_{12}=\Sigma_{21}=cov(\phi_1(\delta,Y,S,Z,X),\phi_2(\delta,Y,S,Z,X))$.

The proposed VE estimator $\widehat{\mathrm{VE}}^{(new)}(s_1)$ is a continuous function of $\hat{\beta}$ and $\hat{\gamma}$; therefore, the regular delta method applies.

A.3 Theoretical justification for perturbation resampling methods

A.3.1

In this section, we show that

$$\begin{aligned}\sqrt{N}(\hat{\beta}^{(\epsilon)}-\beta) &\equiv -[\dot{\Psi}_{\beta}^{(\epsilon)}]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_1(\delta_i,Y_i,S_i,Z_i,X_i)^{(\epsilon)}+o_p(1)\\ &= -[\dot{\Psi}_{\beta}]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_1(\delta_i,Y_i,S_i,Z_i,X_i)\epsilon_i+o_p(1). \qquad (A.10)\end{aligned}$$

We shall use $\mathcal{F}$ to denote the σ-field generated by the original data $(\delta_i, Y_i, S_i, Z_i, X_i)$.
First, consider the unconditional version of $\hat{\beta}^{(\epsilon)}$ with respect to the joint probability space of $\mathcal{F}$ and $\epsilon_i (i = 1, .., N)$:
Under suitable equicontinuity conditions and smoothness conditions (van der Vaart and Wellner, 1996), we have:

$$\begin{aligned}\hat{W}_s^{(\epsilon)} &= \sqrt{N}(\hat{\beta}^{(\epsilon)}-\beta_0)\\ &= \left[-\dot{\Psi}_{\beta}^{(\epsilon)}(\beta_0,F_0^*,\pi_0)\right]^{-1}\sqrt{N}\left\{\Psi_N^{(\epsilon)}(\beta_0;F_0^*,\pi_0)+\sum_{i=1}^{K}\Psi_{F_k^*}^{(\epsilon)}[F_{Nk}^{(\epsilon)}-F_{k0}^*]+\dot{\Psi}_{\alpha}^{(\epsilon)}[\hat{\alpha}-\alpha_0]\right\}+o_p(1)\\ &= \left[-\dot{\Psi}_{\beta}^{(\epsilon)}\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_1^{(\epsilon)}(\delta_i,Y_i,S_i,Z_i,X_i)+o_p(1),\end{aligned}$$

where $\phi_1^{(\epsilon)} = a_0^{(\epsilon)} + a_1^{(\epsilon)} + a_2^{(\epsilon)}$.
To show that $\hat{W}_s^{(\epsilon)} = \sqrt{N}(\hat{\beta}^{(\epsilon)} - \beta_0) = -[\dot{\Psi}_{\beta}]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi_1(\delta_i, Y_i, S_i, Z_i, X_i)\epsilon_i + o_p(1)$, it suffices to show the following:

1. $\dot{\Psi}_{\beta}^{(\epsilon)} = \dot{\Psi}_{\beta}$
2. $a_0^{(\epsilon)} = \epsilon a_0$
3. $a_1^{(\epsilon)} = \epsilon a_1 + o_p(1)$
4. $a_2^{(\epsilon)} = \epsilon a_2 + o_p(1)$.

A.3.1.0.1 1. To show that $\dot{\Psi}_\beta^{(\epsilon)} = \dot{\Psi}_\beta$

$$\begin{aligned}\dot{\Psi}_\beta^{(\epsilon)}(\beta_0, F_0^*, \pi_0) &= \frac{\partial}{\partial\beta}\Psi^{(\epsilon)}(\beta_0, F_0^*, \pi_0) \\ &= -E_0[\epsilon\delta I_\beta(Y|S,Z,X) - \epsilon(1-\delta)\frac{\partial\int U_\beta(Y|s,Z,X)h(s|Y,Z,X)ds}{\partial\beta}].\end{aligned}$$

Since ϵ is independent of (δ, Y, S, Z, X) and $E(\epsilon) = 1$, $\dot{\Psi}_\beta(\beta_0, F_0^*, \pi_0) = \dot{\Psi}_\beta^{(\epsilon)}(\beta_0, F_0^*, \pi_0)$.

A.3.1.0.2 2. To show that $a_0^{(\epsilon)} = \epsilon a_0$

$a_0(\delta, Y, S, Z, X)^{(\epsilon)} = a_0(\delta, Y, S, Z, X)$ by definition of $\hat{\beta}^{(\epsilon)}$.

A.3.1.0.3 3. To show that $a_1^{(\epsilon)} = \epsilon a_1 + o_p(1)$

We first derive the general asymptotic properties of a perturbed MLE estimator:
$L(\beta) = \prod_{i=1}^n p_\beta(x_i)$, $l(\beta) = \sum_{i=1}^n logp_\beta(x_i)$, $\frac{\partial}{\partial\beta}l(\beta) \equiv \dot{l}(\beta) \equiv U(\beta)$, and the MLE $\hat{\beta}$ is obtained by solving the estimating equation $\sum_{i=1}^n U(\beta|X_i) = 0$,
$0 = \frac{1}{\sqrt{n}}\dot{l}_n(\hat{\beta}) = \frac{1}{\sqrt{n}}\dot{l}_n(\beta_0) - (-\frac{\ddot{l}_n(\beta_n^*)}{n})\sqrt{n}(\hat{\beta} - \beta_0)$, where $|\beta_n^* - \beta_0| \leq |\hat{\beta} - \beta_0|$. And $-\frac{1}{n}\ddot{l}_n(\beta_n^*) = -\frac{1}{n}\ddot{l}_n(\beta_0) + o_p(1) \underset{p}{\to} I(\beta_0)$.

Thus $\sqrt{n}(\hat{\beta} - \beta_0) = I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_n(\beta_0|X_i) + o_p(1) \underset{d}{\to} N(0, I^{-1}(\beta_0))$.

Now define the perturbed MLE $\hat{\beta}^{(\epsilon)}$ as the solution of
$\sum_{i=1}^n \dot{l}^{(\epsilon)}(\beta|X_i) = 0$, where $\dot{l}^{(\epsilon)}(\beta|X_i) = \epsilon_i\dot{l}(\beta|X_i)$.
Then $0 = \frac{1}{\sqrt{n}}\dot{l}_n^{(\epsilon)}(\hat{\beta}^{(\epsilon)}) = \frac{1}{\sqrt{n}}\dot{l}_n^{(\epsilon)}(\beta_0) - (-\frac{\ddot{l}_n^{(\epsilon)}(\beta_n^*)}{n})\sqrt{n}(\hat{\beta}^{(\epsilon)} - \beta_0)$, where $|\beta_n^* - \beta_0| \leq |\hat{\beta}^{(\epsilon)} - \beta_0|$.
Note that $\ddot{l}^{(\epsilon)}(\beta|X_i) = \epsilon_i\ddot{l}(\beta|X_i)$, thus $-E_0\left\{\ddot{l}_{jk}^{(\epsilon)}(\beta_0|X)\right\} = -E_0\left\{\ddot{l}_{jk}(\beta_0|X)\right\} = I(\beta_0)$, $-\frac{1}{n}\ddot{l}_n^{(\epsilon)}(\beta_n^*) = -\frac{1}{n}\ddot{l}_n^{(\epsilon)}(\beta_0) + o_p(1) \underset{p}{\to} I(\beta_0)$.

Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\beta}^{(\epsilon)} - \beta_0) &= I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_n^{(\epsilon)}(\beta_0|X_i) + o_p(1) \\ &= I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_n(\beta_0|X_i)\cdot\epsilon_i + o_p(1).\end{aligned}$$

Take the specific case of a Bernoulli distribution. If $X_i's$ are i.i.d. Bernoulli(p), then the MLE $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$. The perturbed version of $\hat{p}$ takes the form $\hat{p}^{(\epsilon)} = \frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} \epsilon_i}$.

Notice that a_1 represents an adjustment due to estimating $F^*_{k_0}(s) = Pr(S \leq s | X = x_k, \delta = 1) = p_{k_0}$ by $F_{N_k} = \frac{\sum_i^N \delta_i I(X_i = x_k) I(S_i \leq s)}{\sum_i^N \delta_i I(X_i = x_k)}$. Conditional on $X = x_k$ and $\delta = 1$, $I(S_i \leq s) \sim Bernoulli(p_{k_0})$. Thus following the above arguments:
if $\sqrt{N}\left[F_{N_k}(s) - F^*_{K_0}(s)\right] = \frac{1}{N}\sum_{i=1}^{N} \phi_{F^*_{k_0}}(s)(\delta_i, S_i, X_i) + o_p(1)$,
then $\sqrt{N}\left[F^{(\epsilon)}_{N_k}(s) - F^*_{K_0}(s)\right] = \frac{1}{N}\sum_{i=1}^{N} \phi_{F^*_{k_0}}(s)(\delta_i, S_i, X_i) \cdot \epsilon_i + o_p(1)$, where $F^{(\epsilon)}_{N_k} = \frac{\sum_i^N \delta_i I(X_i = x_k) I(S_i \leq s) \epsilon_i}{\sum_i^N \delta_i I(X_i = x_k) \epsilon_i}$.

Following similar arguments for $\dot{\Psi}^{(\epsilon)}_{\beta} = \dot{\Psi}_{\beta}$, it is straightforward to show $\dot{\Psi}^{(\epsilon)}_{F^*_k}(\beta_0, F^*_0) = \dot{\Psi}_{F^*_k}(\beta_0, F^*_0)$.
Thus

$$\begin{aligned} a^{(\epsilon)}_1(\delta_i, S_i, X_i) &= \sum_1^k \dot{\Psi}^{(\epsilon)}_{F^*_k}(\beta_0, F^*_0)\left[F^{(\epsilon)}_{Nk}(s) - F^*_{K_0}(s)\right] \\ &= \sum_1^k \dot{\Psi}_{F^*_k}(\beta_0, F^*_0)\left[F_{Nk}(s) - F^*_{K_0}(s)\right] \cdot \epsilon_i + o_p(1) \\ &= a_1(\delta_i, S_i, X_i) \cdot \epsilon_i + o_p(1). \end{aligned}$$

A.3.1.0.4 4. To show that $a^{(\epsilon)}_2 = \epsilon a_2 + o_p(1)$

Let $\pi(Y, Z, X, \alpha)$ denote $P(\delta = 1 | Y, Z, X)$. We estimate α by maximizing $LogL(\alpha | Y, Z, X) = \sum_{i=1}^{n} \delta_i log\pi(Y_i, Z_i, X_i; \alpha) + (1 - \delta_i) log\left\{1 - \pi(Y_i, Z_i, X_i; \alpha)\right\}$. Therefore $\hat{\alpha}$ is a MLE estimator and following similar arguments in 3., we get that $\sqrt{N}\dot{\Psi}_{\alpha}(\beta_0, F^*_0, \pi(\alpha_0))[\hat{\alpha} - \alpha_0] = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_2(\delta_i, Y_i, Z_i, X_i) + o_p(1)$,
$\sqrt{N}\dot{\Psi}^{(\epsilon)}_{\alpha}(\beta_0, F^*_0, \pi(\alpha_0))[\hat{\alpha}^{(\epsilon)} - \alpha_0] = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a^{(\epsilon)}_2(\delta_i, Y_i, Z_i, X_i) + o_p(1)$,
and $a^{(\epsilon)}_2(\delta_i, Y_i, Z_i, X_i) = a_2(\delta_i, Y_i, Z_i, X_i) \cdot \epsilon_i + o_p(1)$.

Therefore, the proof for (A.10) is completed. Furthermore,

$$\sqrt{n}(\hat{\beta}^{(\epsilon)} - \hat{\beta}) = -[\dot{\Psi}_{\beta}]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_1(\delta_i, Y_i, S_i, Z_i, X_i)(\epsilon_i - 1) + o_p(1).$$

A.3.2

In the this section, we show that

$$\sqrt{N}(\hat{\gamma}^{(\epsilon)}(\hat{\alpha}^{(\epsilon)}) - \hat{\gamma}(\hat{\alpha})) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_2(\delta_i, Y_i, S_i(1), Z_i, X_i) \cdot (\epsilon_i - 1) + o_p(1).$$

Similar to the arguments used in Section C.1,

$$\sqrt{N}\begin{pmatrix} \hat{\gamma}^{(\epsilon)}(\alpha_0) - \gamma_0 \\ \hat{\alpha}^{(\epsilon)} - \alpha_0 \end{pmatrix} = \sqrt{N}\begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} \frac{\delta_i}{\pi_{0i}} \widetilde{l}_0(S_i) \cdot \epsilon_i \\ \frac{1}{N}\sum_{i=1}^{N} \widetilde{l}_0^{\alpha} \cdot \epsilon_i \end{pmatrix}. \tag{A.11}$$

Moreover

$$\frac{1}{N}\sum_{i=1}^{N}(\frac{\delta_i}{\hat{\pi}_i} - \frac{\delta_i}{\pi_0})\widetilde{l}_0(S_i) \cdot \epsilon_i = o_p(1) - \frac{1}{N}\sum_{i=1}^{N} \delta_i \widetilde{l}_0(S_i) \cdot \epsilon_i \left[\frac{\dot{\pi}_0^T}{\pi^2}(\hat{\alpha} - \alpha_0)\right].$$

Then

$$\sqrt{N}\left[\hat{\gamma}^{(\epsilon)}(\hat{\alpha}^{(\epsilon)}) - \gamma_0\right] = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_2(\delta_i, Y_i, S_i(1), Z_i, X_i) \cdot \epsilon_i + o_p(1).$$

Therefore

$$\sqrt{N}(\hat{\gamma}^{(\epsilon)}(\hat{\alpha}^{(\epsilon)}) - \hat{\gamma}(\hat{\alpha})) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_2(\delta_i, Y_i, S_i(1), Z_i, X_i) \cdot (\epsilon_i - 1) + o_p(1).$$

Conditional on the data and given that $E[\epsilon] = 1$, $Var[\epsilon] = 1$,$E[(\epsilon - 1)^2] = 1$, we have $E[\phi_0(\epsilon - 1)] = 0$, $Var[\phi_0(\epsilon - 1)] = Var[\phi_0]$,
$cov(\phi_1(\delta, Y, S, Z, X)(\epsilon - 1), \phi_2(\delta, Y, S, Z, X)(\epsilon - 1)) = cov(\phi_1(\delta, Y, S, Z, X), \phi_2(\delta, Y, S, Z, X))$.

Therefore, conditional on the data

$$\sqrt{n}\left(\begin{pmatrix} \hat{\beta}^{(\epsilon)} \\ \hat{\gamma}^{(\epsilon)} \end{pmatrix} - \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}\right) \underset{d}{\to} N\left(\mathbf{0}, \begin{bmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{bmatrix}\right). \tag{A.12}$$

Because $log\hat{RR}^{(\epsilon)}$ is a continuous function of $\hat{\beta}^{(\epsilon)}$ and $\hat{\gamma}^{(\epsilon)}$, the regular delta method applies and the unconditional distribution of $\sqrt{N}\left\{log\hat{RR} - logRR_0\right\}$ can be approximated by $\sqrt{N}\left\{log\hat{RR}^{(\epsilon)} - log\hat{RR}\right\}$ conditional on the observed data.

Appendix B

APPENDIX FOR CHAPTER 4

B.1 Three Useful Convenient Facts

$$If\ \ U \sim N(u, \sigma^2),\ then\ \ E[\Phi(a+U)] = \Phi\left[\frac{a+u}{\sqrt{1+\sigma^2}}\right] \tag{B.1}$$

$$\begin{aligned} If\ \ \begin{pmatrix} X \\ Y \end{pmatrix} &\sim BivariateNormal\left\{\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right\} \\ Then\ \ X|Y = y_0 &\sim N\left(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y_0 - \mu_y), \left(\sigma_x\sqrt{1-\rho^2}\right)^2\right) \end{aligned} \tag{B.2}$$

$$\begin{aligned} If B \sim & N(\mu_1, \sigma_1^2) \\ S|B = b \sim & N(\gamma_0 + \gamma_1 b, \sigma_2^2) \\ Then\ \ \begin{pmatrix} B \\ S \end{pmatrix} \sim & BivariateNormal\left\{\begin{pmatrix} \mu_B \\ \mu_S \end{pmatrix}, \begin{pmatrix} \sigma_B^2 & \rho\sigma_B\sigma_S \\ \rho\sigma_B\sigma_S & \sigma_S^2 \end{pmatrix}\right\} \end{aligned} \tag{B.3}$$

$$\begin{aligned} where\ \ \mu_B =& \mu_1;\ \ \mu_S = \gamma_0 + \gamma_1\mu_1; \\ \sigma_B^2 =& \sigma_1^2;\ \ \sigma_S^2 = \sigma_2^2 + \gamma_1^2\sigma_1^2;\ \ \rho^2 = \frac{\gamma_1^2\sigma_1^2}{\sigma_2^2 + \gamma_1^2\sigma_1^2} \end{aligned}$$

B.2 Derivation of Expression 4.4, 4.5, and 4.6

$$P(X|S) = \frac{f(S|X)P(X)}{\sum_{i=1}^{D} f(S|x_i)P(x_i)},$$

where

$$f(S|X) = \int f(S, b|X)dF(b|X) = \int f(S|b, X)f(b|X)db.$$

$$P(X|S, B > c) = \frac{f(S|B > c, X)P(X|B > c)}{\sum_{i=1}^{D} f(S|B > c, x_i)P(x_i|B > c)},$$

where

$$f(S|B > c, X) = \frac{f(S, B > c|X)}{f(B > c|X)} = \frac{\int_c^\infty f(S, b|X)db}{\int_c^\infty f(b|X)db} = \frac{\int_c^\infty f(S|b, X)f(b|X)db}{\int_c^\infty f(b|X)db}.$$

$$P(X|S, B = c) = \frac{f(S|B = c, X)P(X|B = c)}{\sum_{i=1}^{D} f(S|B = c, x_i)P(x_i|B = c)}.$$

B.3 Specific Parameterization

This section provides a detailed estimation procedure of $mCEP^{risk}(S)$, $mCEP^{risk}(S, B > c)$, and $mCEP^{risk}(S, B = c)$ for the case where $F^{B|X}$ is assumed censored normal, $F^{S|B,X}$ is assume censored normal, and the risk functions take the form $risk_z\{S, B, X\} = g\{\beta; S, B, Z, X\} = \Phi(\beta_0 + \beta_1 Z + \beta_2 S + \beta_3 Z \cdot S + \beta_4 B + \beta_5 Z \cdot B + \beta_6 X)$, where Φ denotes the standard normal cdf. With $S = max(S^*, c)$ and $B = max(B^*, c)$ where c is the limit of detection, we assume the following models for $\left(F^{B^*|X}, F^{S^*|B^*,X}\right)$

(B1) $B^*|X \sim N(\mu_1, \sigma_1^2)$ where $\mu_1 = \alpha_0 + \alpha_1 X$

(B2) $S^*|B^*, X \sim N(\mu_2, \sigma_2^2)$ where $\mu_2 = \gamma_0 + \gamma_1 B^* + \gamma_2 X$

Then by Convenient Fact B.2 and B.3 from Appendix B.1:

$$(B^*, S^*)|X \sim BivariateNormal\left\{\begin{pmatrix}\mu_B \\ \mu_S\end{pmatrix}, \begin{pmatrix}\sigma_B^2 & \rho\sigma_B\sigma_S \\ \rho\sigma_B\sigma_S & \sigma_S^2\end{pmatrix}\right\}$$

$$where\ \ \mu_B = \mu_1,\ \ \mu_S = \gamma_0 + \gamma_1\mu_1 + \gamma_2 X$$

$$\sigma_B^2 = \sigma_1^2,\ \ \sigma_S^2 = \sigma_2^2 + \gamma_1^2\sigma_1^2,\ \ \rho^2 = \frac{\gamma_1^2\sigma_1^2}{\sigma_2^2 + \gamma_1^2\sigma_1^2},$$

and $B^*|S^*, X \sim N(\mu_3, \sigma_3^2)$ where $\mu_3 = \mu_B + \rho\frac{\sigma_B}{\sigma_S}[S^* - \mu_S]$ and $\sigma_3^2 = \sigma_B^2(1 - \rho^2)$. To make inference, first the nuisance parameter ν is estimated by sample estimators $\widehat{\nu} = \left(\widehat{\mu}_1, \widehat{\sigma}_1^2, \widehat{\mu}_2, \widehat{\sigma}_2^2,\right)^T$.

B.3.1 Estimating the Nuisance Parameter

Without loss of generality, we consider the estimation of the nuisance parameters in the case where $\{i : \delta_{Bi} = 1\}$ is a random sample of all study subjects (Phase-II sample), and $\{i : \delta_i = 1\}$ is the union of $\{i : Z_i = 1, \delta_{Bi} = 1\}$ and $\{i : Z_i = 1, Y_i = 1\}$.

1. To estimate μ_1 and σ_1:

$$L_1 = \prod_{i:\delta_{Bi=1}} f(B_i|X_i)$$

where

$$\begin{aligned} f(B|X) &= I(B = c)P(B^* \leq c|X) + I(B > c)f(B^*|X) \\ &= I(B = c)\Phi(c; \alpha_0 + \alpha_1 X, \sigma_1) + I(B > c)\phi(B; \alpha_0 + \alpha_1 X, \sigma_1) \end{aligned}$$

$(\hat{\alpha_0}, \hat{\alpha_1}, \hat{\sigma_1})$ are the solutions that maximize the likelihood function L_1. And $\hat{\mu_1} = \hat{\alpha_0} + \hat{\alpha_1} X$.

2. To estimate μ_2 and σ_2:

$$L_2 = \prod_{i:\delta_{Bi=1,Z_i=1}} f(S_i|B_i, X_i)$$

where

$$f(S|B, X) = I(S = c)P(S^* \leq c|B, X) + I(S > c)f(S^*|B, X)$$

$$\begin{aligned} P(S^* \leq c|B, X) &= I(B = c)P(S^* \leq c|B^* \leq c, X) + I(B > c)P(S^* \leq c|B^*, X) \\ &= I(B = c)\frac{\int_{-\infty}^{c} P(S^* \leq c|b^*, X)f(b^*|X)db^*}{P(B^* \leq c, X)} + I(B > c)P(S^* \leq c|B^*, X) \\ &= I(B = c)\frac{\int_{-\infty}^{c} \Phi(c; \gamma_0 + \gamma_1 b^* + \gamma_2 X, \sigma_2)\phi(b^*; \mu_1, \sigma_1)db^*}{\Phi(c; \mu_1, \sigma_1)} \\ &+ I(B > c)\Phi(c; \gamma_0 + \gamma_1 B + \gamma_2 X, \sigma_2) \end{aligned}$$

and

$$\begin{aligned} f(S^*|B,X) &= I(B=c)f(S^*|B^* \le c, X) + I(B>c)f(S^*|B^*,X) \\ &= I(B=c)\frac{\int_{-\infty}^{c} f(S^*|b^*,X)f(b^*|X)db^*}{P(B^* \le c|X)} + I(B>c)f(S^*|B^*,X) \\ &= I(B=c)\frac{\int_{-\infty}^{c} \phi(S^*;\gamma_0+\gamma_1 b^*+\gamma_2 X,\sigma_2)\phi(b^*;\mu_1,\sigma_1)db^*}{\Phi(c;\mu_1,\sigma_1)} \\ &+ I(B>c)\phi(S^*;\gamma_0+\gamma_1 B^*+\gamma_2 X,\sigma_2) \end{aligned}$$

$(\hat{\gamma_0},\hat{\gamma_1},\hat{\gamma_2},\hat{\sigma_2})$ are the solutions that maximize the likelihood function L_2. And $\hat{\mu_2} = \hat{\gamma_0} + \hat{\gamma_1}B + \hat{\gamma_2}X$.

Notice that because both set $\{i : \delta_{Bi} = 1\}$ and set $\{i : \delta_{Bi} = 1, \delta_i = 1\}$ are random samples of the entire study subjects, inverse probability weighting of S (IPW) is not needed here.

B.3.2 Likelihood Contribution for each subject

B.3.2.1 $\delta_{Bi} = \delta_i = 1$

Subjects with $\delta_{Bi} = \delta_i = 1$ contribute to likelihood $risk_{Z_i}(S_i,B_i,X_i;\beta)^{Y_i}(1 - risk_{Z_i}(S_i,B_i,X_i;\beta))^{1-Y_i}$.

B.3.2.2 $\delta_{Bi} = 1, \delta_i = 0$

For subjects with $\delta_{Bi} = 1, \delta_i = 0$, B_i is observed and S_i is missing. Contribution to likelihood is obtained by integrating $risk_{Z_i}(\cdot,B_i,X_i;\beta)$ over the conditional cdf $F^{S|B,X}$:

$$\begin{aligned} &\left\{ \left(\int risk_{Z_i}(s_1,B_i,X_i;\beta) dF^{S|B,X}(s_1|B,X) \right)^{Y_i} \right. \\ &\qquad \left. \times \left(1 - \int risk_{Z_i}(s_1,B_i,X_i;\beta) dF^{S|B,X}(s_1|B,X) \right)^{1-Y_i} \right\} \end{aligned}$$

1. If the observed $B_i > c$, then $B_i = B_i^*$, thus by Convenient Fact B.1

$$\begin{aligned}
&\int_{c-}^{\infty} risk_{Z_i}(s_1, B_i, X_i; \beta) dF^{S|B,X}(s_1|B, X) \\
=&\int_{c-}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1 + \beta_3 s_1 Z_i + \beta_4 B_i + \beta_5 B_i Z_i + \beta_6 X_i) dF^{S|B^*,X}(s_1|B^*, X) \\
=&P(S_i^* \le c|B_i^*, X_i) \cdot \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1 + \beta_3 s_1 Z_i + \beta_4 B_i + \beta_5 B_i Z_i + \beta_6 X_i) \\
&+ \int_{c+}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1 + \beta_3 s_1 Z_i + \beta_4 B_i + \beta_5 B_i Z_i + \beta_6 X_i) dF^{S^*|B^*,X}(s_1|B^*, X) \\
=&\Phi\left(\frac{c - \mu_2}{\sigma_2}\right) \cdot \Phi(\beta_0 + \beta_1 Z_i + \beta_2 c + \beta_3 c Z_i + \beta_4 B_i + \beta_5 B_i Z_i + \beta_6 X_i) \\
&+ \int_{c+}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1 + \beta_3 s_1 Z_i + \beta_4 B_i + \beta_5 B_i Z_i + \beta_6 X_i) dF^{S^*|B^*,X}(s_1|B^*, X) \\
\equiv&A(Z_i, B_i, X_i; \beta)
\end{aligned}$$

where $\mu_2 = \gamma_0 + \gamma_1 B_i + \gamma_2 X_i$. Note that $S^*|B^*, X \sim N(\mu_2, \sigma_2^2)$, thus the second part of the integral can be calculated by

$$\int_{c}^{\infty} \Phi(\beta_0 + \beta_2 s_1 + \beta_4 B_i + \beta_6 X_i) f_{N(\mu_2, \sigma_2^2)}(s_1) ds_1$$

where $f_{N(\mu_2,\sigma_2^2)}$ is the pdf for Normal distribution with mean μ_2 and variance σ_2^2.

2. If the observed $B_i = c$, then we need to integrate $risk_{Z_i}(\cdot, B_i, X_i; \beta)$ over unknown s_1 w.r.t. distribution of $S|B_i = c, X$

$$\begin{aligned}
P(S = c|B = c, X) = P(S^* \le c|B^* \le c, X) &= \frac{P(S^* \le c, B^* \le c|X)}{P(B^* \le c|X)} \\
&= \frac{\int_{-\infty}^{c} \int_{-\infty}^{c} f_{BiN}(b, s_1) db ds_1}{\Phi\left(\frac{c - \mu_B}{\sigma_B}\right)} \equiv p_1
\end{aligned} \tag{B.4}$$

where $f_{BiN}(b, s_1)$ are the pdf for Bivariate Normal distribution $BiN\left\{\binom{\mu_B}{\mu_S}, \begin{pmatrix} \sigma_B^2 & \rho\sigma_B\sigma_S \\ \rho\sigma_B\sigma_S & \sigma_S^2 \end{pmatrix}\right\}$.

The pdf for distribution $S^*|B=c,X$ is:

$$f(s^*|B=c,X) = f(S^*|B^* \le c,X) = \frac{f(S^*, B^* \le c|X)}{f(B^* \le c|X)}$$
$$= \frac{f(B^* \le c|S^*,X)\cdot f(S^*|X)}{f(B^* \le c|X)} = \frac{\Phi\left(\frac{c-\mu_3}{\sigma_3}\right)\cdot f_{N(\mu_S,\sigma_S^2)}(s^*}{\Phi\left(\frac{c-\mu_B}{\sigma_B}\right)} \tag{B.5}$$

where $f_{N(\mu_S,\sigma_S^2)}$ are the pdf for Normal distribution with mean μ_S and variance σ_S^2.

thus

$$\begin{aligned}
&\int_{c-}^{\infty} risk_{Z_i}(s_1, B_i, X_i;\beta) dF^{S|B,X}(s_1|B,X) \\
=&\int_{c-}^{\infty} \Phi(\beta_0+\beta_1 Z_i+\beta_2 s_1+\beta_3 s_1 Z_i+\beta_4 c+\beta_5 c Z_i+\beta_6 X_i) dF^{S|B_i=c,X}(s_1|B_i=c,X) \\
=&P(S=c|B=c,X)\cdot\Phi(\beta_0+\beta_1 Z_i+\beta_2 c+\beta_3 c Z_i+\beta_4 c+\beta_5 c Z_i+\beta_6 X_i) \\
+&\int_{c+}^{\infty} \Phi(\beta_0+\beta_1 Z_i+\beta_2 s_1+\beta_3 s_1 Z_i+\beta_4 c+\beta_5 c Z_i+\beta_6 X_i) dF^{S^*|B_i=c,X}(s_1|B_i=c,X) \\
\equiv&B(Z_i,c,X_i;\beta)
\end{aligned}$$

which can be calculated using results from B.4 and B.5.

Therefore, contribution for subjects with $\delta_{Bi}=1,\delta_i=0$ are

$$\left\{\left(A(Z_i,B_i,X_i;\beta)^{I(B_i>c)}B(Z_i,c,X_i;\beta)^{I(B_i=c)}\right)^{Y_i} \times\left(1-A(Z_i,B_i,X_i;\beta)^{I(B_i>c)}B(Z_i,c,X_i;\beta)^{I(B_i=c)}\right)^{1-Y_i}\right\}$$

B.3.2.3 $\delta_i=1,\delta_{Bi}=0$

For subjects with $\delta_i=1,\delta_{Bi}=0$, B_i is missing and S_i is observed. Likelihood contribution equals

$$\left\{\left(\int risk_{Z_i}(S_i,b,X_i;\beta)dF^{B|S,X}(b|S,X)\right)^{Y_i} \times\left(1-\int risk_{Z_i}(S_i,b,X_i;\beta)dF^{B|S,X}(b|S,X)\right)^{1-Y_i}\right\}$$

1. If the observed $S_i > c$, then $S_i = S_i = S_i^*$, thus by Convenient Fact B.1

$$
\begin{aligned}
&\int_{c-}^{\infty} risk_{Z_i}(S_i, b, X_i; \beta) dF^{B|S,X}(b|S, X) \\
=&\int_{c-}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 S_i + \beta_3 S_i Z_i + \beta_4 b + \beta_5 b Z_i + \beta_6 X_i) dF^{B|S^*,X}(b|S^*, X) \\
=&P(B_i^* \leq c|S_i^*, X_i) \cdot \Phi(\beta_0 + \beta_1 Z_i + \beta_2 S_i + \beta_3 S_i Z_i + \beta_4 b + \beta_5 b Z_i + \beta_6 X_i) \\
&+ \int_{c+}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 S_i + \beta_3 S_i Z_i + \beta_4 b + \beta_5 b Z_i + \beta_6 X_i) dF^{B^*|S^*,X}(b^*|S^*, X) \\
=&\Phi\left(\frac{c - \mu_3}{\sigma_3}\right) \cdot \Phi(\beta_0 + \beta_1 Z_i + \beta_2 S_i + \beta_3 S_i Z_i + \beta_4 c + \beta_5 c Z_i + \beta_6 X_i) \\
&+ \int_{c+}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 S_i + \beta_3 S_i Z_i + \beta_4 b + \beta_5 b Z_i + \beta_6 X_i) dF^{B^*|S^*,X}(b^*|S^*, X) \\
\equiv&C(Z_i, S_i, X_i; \beta)
\end{aligned}
$$

where $\mu_3 = \alpha_0 + \alpha_1 X_i + \rho \frac{\sigma_1}{\sqrt{\sigma_2^2 + \gamma_1^2 \sigma_1^2}} [S_i - (\gamma_0 + \gamma_1(\alpha_0 + \alpha_1 X_i) + \gamma_2 X_i)]$.

2. If the observed $S_i = c$, then we need to integrate $risk_{Z_i}(S_i, b, X_i; \beta)$ over unknown b w.r.t. distribution of $B|S = c, X$.

$$
\begin{aligned}
P(B = c|S = c, X) = P(B^* \leq c|S^* \leq c, X) &= \frac{P(B^* \leq c, S^* \leq c|X)}{P(S^* \leq c|X)} \\
&= \frac{\int_{-\infty}^{c} \int_{-\infty}^{c} f_{BiN}(b, s_1) dbds_1}{\Phi\left(\frac{c - \mu_S}{\sigma_S}\right)} \equiv p_2
\end{aligned} \tag{B.6}
$$

where $f_{BiN}(b, s_1)$ are defined the same as before: the pdf for Bivariate Normal distribution $BiN\left\{\binom{\mu_B}{\mu_S}, \begin{pmatrix} \sigma_B^2 & \rho\sigma_B\sigma_S \\ \rho\sigma_B\sigma_S & \sigma_S^2 \end{pmatrix}\right\}$

The pdf for distribution $B^*|S = c, X$ is:

$$
\begin{aligned}
f(B^*|S = c, X) = f(B^*|S^* \leq c, X) &= \frac{f(B^*, S^* \leq c|X)}{f(S^* \leq c|X)} \\
&= \frac{f(S^* \leq c|B^*, X) \cdot f(B^*|X)}{f(S^* \leq c|X)} = \frac{\Phi\left(\frac{c - \mu_2}{\sigma_2}\right) \cdot f_{N(\mu_B, \sigma_B^2)}(B^*)}{\Phi\left(\frac{c - \mu_S}{\sigma_S}\right)}
\end{aligned} \tag{B.7}
$$

where $f_{N(\mu_B, \sigma_B^2)}$ are the pdf for Normal distribution with mean μ_B and variance σ_B^2.

Thus

$$
\begin{aligned}
&\int_{c-}^{\infty} risk_{Z_i}(S_i, b, X_i; \beta) dF^{B|S,X}(b|S, X) \\
=&\int_{c-}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 c + \beta_3 Z_i c + \beta_4 b + \beta_5 Z_i b + \beta_6 X_i) dF^{B|S=c,X}(b|S = c, X) \\
=&P(B = c|S = c, X) \cdot \Phi(\beta_0 + \beta_1 Z_i + \beta_2 c + \beta_3 c Z_i + \beta_4 c + \beta_5 c Z_i + \beta_6 X_i) \\
&+ \int_{c+}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 c + \beta_3 Z_i c + \beta_4 b^* + \beta_5 Z_i b^* + \beta_6 X_i) dF^{B^*|S=c,X}(b^*|S = c, X) \\
\equiv&D(Z_i, c, X_i; \beta)
\end{aligned}
$$

which can be calculated using results B.6 and B.7.

Therefore, contribution for subjects with $\delta_i = 1, \delta_{Bi} = 0$ are

$$
\begin{aligned}
&\left\{ \left(C(Z_i, S_i, X_i; \beta)^{I(S_i > c)} D(Z_i, c, X_i; \beta)^{I(S_i = c)} \right)^{Y_i} \right. \\
&\qquad \left. \times \left(C(Z_i, S_i, X_i; \beta)^{I(S_i > c)} D(Z_i, c, X_i; \beta)^{I(S_i = c)} \right)^{1 - Y_i} \right\}
\end{aligned}
$$

B.3.2.4 $\delta_i = 0, \delta_{Bi} = 0$

For subjects with $\delta_i = 0$ and $\delta_{Bi} = 0$, both B_i and S_i are missing. Likelihood contribution equals

$$
\begin{aligned}
&\left\{ \left(\int \int risk_{Z_i}(s_1, b, X_i; \beta) dF^{S|B,X}(s_1|b, X) dF^{B|X}(b|X) \right)^{Y_i} \right. \\
&\left. \times \left(1 - \int \int risk_{Z_i}(s_1, b, X_i; \beta) dF^{S|B,X}(s_1|b, X) dF^{B|X}(b|X) \right)^{1 - Y_i} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\int\int risk_{Z_i}(s_1, b, X_i; \beta) dF^{S|B,X}(s_1|b, X) dF^{B|X}(b|X) \\
=&\int_{c-}^{\infty}\int_{c-}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1 + \beta_3 Z_i \cdot s_1 + \beta_4 b + \beta_5 Z_i \cdot b + \beta_6 X_i) dF^{S|B,X}(s_1|b, X) dF^{B|X}(b|X) \\
=&P(S = c, B = c|X) \times \Phi(\beta_0 + \beta_1 Z_i + \beta_2 c + \beta_3 Z_i \cdot c + \beta_4 c + \beta_5 Z_i \cdot c + \beta_6 X_i) \\
+&\int_{c+}^{\infty} P(S = c|b^*, X) \times \Phi(\beta_0 + \beta_1 Z_i + \beta_2 c + \beta_3 Z_i \cdot c + \beta_4 b^* + \beta_5 Z_i \cdot b^* + \beta_6 X_i) dF^{B^*|X}(b^*|X) \\
+&\int_{c+}^{\infty} P(B = c|s_1^*, X) \times \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1^* + \beta_3 Z_i \cdot s_1^* + \beta_4 c + \beta_5 Z_i \cdot c + \beta_6 X_i) dF^{S^*|X}(s_1^*|X) \\
+&\int_{c+}^{\infty}\int_{c+}^{\infty} \Phi(\beta_0 + \beta_1 Z_i + \beta_2 s_1^* + \beta_3 Z_i \cdot s_1^* + \beta_4 b^* + \beta_5 Z_i \cdot b^* + \beta_6 X_i) dF^{S^*|B^*,X}(s_1^*|b^*, X) \\
&dF^{B^*|X}(b^*|X) \\
\equiv&E(Z_i, X_i; \beta)
\end{aligned} \tag{B.8}
$$

where

$$P(S = c, B = c|X) = P(S^* \leq c, B^* \leq c|X) = \int_{-\infty}^{c}\int_{-\infty}^{c} f_{BiN}(b, s_1) dbds_1$$

where $f_{BiN}(b, s_1)$ are defined the same as before: the pdf for Bivariate Normal distribution $BiN\left\{\binom{\mu_B}{\mu_S}, \begin{pmatrix} \sigma_B^2 & \rho\sigma_B\sigma_S \\ \rho\sigma_B\sigma_S & \sigma_S^2 \end{pmatrix}\right\}$.

$$P(S = c|b^*, X) = P(S^* \leq c|b^*, X) = \Phi\left(\frac{c - \mu_2}{\sigma_2}\right),$$

and

$$P(B = c|s_1^*, X) = P(B^* \leq c|s_1^*, X) = \Phi\left(\frac{c - \mu_3}{\sigma_3}\right).$$

Therefore, contribution for subjects with $\delta_i = 0, \delta_{Bi} = 0$ are

$$\{E(Z_i, X_i; \beta)\}^{Y_i} \times \{1 - E(Z_i, X_i; \beta)\}^{1-Y_i}.$$

Thus, the conditional likelihood can be expressed as

$$
\begin{aligned}
L(\beta,\nu) = \prod_{i=1}^{n} & \left\{ risk_{Z_1}(S_i(1), B_i, X_i; \beta)^{Y_i} (1 - risk_{Z_i}(S_i(1), B_i, X_i; \beta))^{1-Y_i} \right\}^{\delta_{Bi}\delta_i} \\
\times & \left\{ \left(A(Z_i, B_i, X_i; \beta)^{I(B_i>c)} B(Z_i, c, X_i; \beta)^{I(B_i=c)} \right)^{Y_i} \times \right. \\
& \left. \left(1 - A(Z_i, B_i, X_i; \beta)^{I(B_i>c)} B(Z_i, c, X_i; \beta)^{I(B_i=c)} \right)^{1-Y_i} \right\}^{\delta_{Bi}(1-\delta_i)} \\
\times & \left\{ \left(C(Z_i, S_i(1), X_i; \beta)^{I(S_i(1)>c)} D(Z_i, c, X_i; \beta)^{I(S_i(1)=c)} \right)^{Y_i} \times \right. \\
& \left. \left(1 - C(Z_i, S_i(1), X_i; \beta)^{I(S_i(1)>c)} D(Z_i, c, X_i; \beta)^{I(S_i(1)=c)} \right)^{1-Y_i} \right\}^{(1-\delta_{Bi})\delta_i} \\
\times & \left\{ E(Z_i, X_i; \beta)^{Y_i} \times (1 - E(Z_i, X_i; \beta))^{1-Y_i} \right\}^{(1-\delta_{Bi})(1-\delta_i)},
\end{aligned}
$$

which takes into account the left-censoring of the censored normal random variables $S(1)$ and B. The estimated conditional likelihood $L(\beta, \widehat{\nu})$ is maximized in β using standard numerical techniques.

Appendix C

APPENDIX FOR CHAPTER 5

C.1 Technical Details for 5.3

Under $\mathcal{M}_a$,

$$\begin{aligned}
&P(Y=1|Y^\tau=0,S=s_1,Z=1,X=x)\\
=&P(Y(1)=1|Y^\tau(1)=0,S=s_1,Z=1,X=x) && \text{by (3)}\\
=&P(Y(1)=1|Y^\tau(1)=0,S=s_1,X=x) && \text{by (4)}\\
=&g_v(s_1,x;\eta_v^a) && \text{by (M5a)}
\end{aligned}$$

Under $\mathcal{M}_b$,

$$\begin{aligned}
&P(Y=1|Y^\tau=0,S=s_1,Z=1,X=x)\\
=&P(Y(1)=1|Y^\tau(1)=0,S=s_1,Z=1,X=x) && \text{by (3)}\\
=&P(Y(1)=1|Y^\tau(1)=0,S=s_1,X=x) && \text{by (4)}\\
=&\frac{\frac{w^{-1}(s_1,x,y=1;\beta,\alpha)w(s_1,x,y=1;\beta,\alpha)P(Y(1)=1|Y^\tau(1)=0,S=s_1,X=x)}{P(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)}}{\sum_{y=0}^1 \frac{w^{-1}(s_1,x,y;\beta,\alpha)w(s_1,x,y;\beta,\alpha)P_{Y(1)|Y^\tau(1),S,X}(y|Y^\tau(1)=0,S=s_1,X=x)}{P(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)}} && \text{if w}(\cdot)>0\\
=&\frac{w^{-1}(s_1,x,y=1;\beta,\alpha)P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)}{\sum_{y=0}^1 w^{-1}(s_1,x,y;\beta,\alpha)P(Y(1)=y|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)} && \text{by (A4), (M2)}\\
=&\frac{w^{-1}(s_1,x,y=1;\beta,\alpha)g_v^{EAAR}(s_1,x;\eta_v^b)}{w^{-1}(s_1,x,y=1;\beta,\alpha)g_v^{EAAR}(s_1,x;\eta_v^b)+w^{-1}(s_1,x,y=0;\beta,\alpha)(1-g_v^{EAAR}(s_1,x;\eta_v^b))} && \text{by (M5b)}\\
\equiv&g_v^*(s_1,x;\beta,\alpha,\eta_v^b).
\end{aligned}$$

The third last equality is derived based on that $w(s_1,x,y=1;\beta,\alpha)P(Y(1)=1|Y^\tau(1)=0,S=s_1,X=x)=P(Y(1)=1,Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)=P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)P(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)$.

Under $\mathcal{M}_a$,

$$
\begin{aligned}
&P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)\\
=&\frac{P(Y(1)=1,Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)}{P(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)}\\
=&\frac{P(Y^\tau(0)=0|Y(1)=1,Y^\tau(1)=0,S=s_1,X=x)P(Y(1)=1|Y^\tau(1)=0,S=s_1,X=x)}{\sum_{y=0}^{1}P(Y^\tau(0)=0|Y(1)=y,Y^\tau(1)=0,S=s_1,X=x)P(Y(1)=y|Y^\tau(1)=0,S=s_1,X=x)}\\
=&\frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x;\eta_v^a)}{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x;\eta_v^a)+w(s_1,x,y=0;\beta,\alpha)(1-g_v(s_1,x;\eta_v^a))}\\
\equiv&g_v^{*EAAR}(s_1,x;\beta,\alpha,\eta_v^a)
\end{aligned}
$$

Under $\mathcal{M}_b$,

$$
\begin{aligned}
&P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)\\
=&g_v^{EAAR}(s_1,x;\eta_v^b)
\end{aligned}
$$

Under $\mathcal{M}_a$ or $\mathcal{M}_b$,

$$
\begin{aligned}
&P(Y=1|Y^\tau=0,S=s_1,Z=0,X=x)\\
=&P(Y(0)=1|Y^\tau(0)=0,S=s_1,Z=0,X=x) && \text{by (3)}\\
=&P(Y(0)=1|Y^\tau(0)=0,S=s_1,X=x) && \text{by (4)}\\
=&g_p(s_1,x;\eta_p) && \text{by (M4)}
\end{aligned}
$$

$$
\begin{aligned}
&f_{S|Y^\tau,Z,X}(s_1|Y^\tau=0,Z=1,X=x)\\
=&f_{S|Y^\tau(1),Z,X}(s_1|Y^\tau(1)=0,Z=1,X=x)\\
=&f_{S|Y^\tau(1),X}(s_1|Y^\tau(1)=0,X=x)\\
\equiv&f(s_1|x;\gamma)
\end{aligned}
$$

$$
\begin{aligned}
& f_{S|Y^\tau,Z,X}(s_1|Y^\tau=0,Z=0,X=x) \\
=& f_{S|Y^\tau(0),Z,X}(s_1|Y^\tau(0)=0,Z=0,X=x) \\
=& f_{S|Y^\tau(0),X}(s_1|Y^\tau(0)=0,X=x) \\
=& f_{S|Y^\tau(0),Y^\tau(1),X}(s_1|Y^\tau(0)=Y^\tau(1)=0,X=x) \\
=& \frac{f(S=s_1,Y^\tau(0)=0|Y^\tau(1)=0,X=x)}{P(Y^\tau(0)=0|Y^\tau(1)=0,X=x)} \\
=& \frac{f(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)}{P(Y^\tau(0)=0|Y^\tau(1)=0,X=x)} f(S=s_1|Y^\tau(1)=0,X=x) \\
=& \frac{\frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x)}{g_v^{EAAR}(s_1,x)} f_p(s_1|x;\gamma)}{\int \frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x)}{g_v^{EAAR}(s_1,x)} f_p(s_1|x;\gamma)ds_1}
\end{aligned}
$$

based on the following two results:

$$
\begin{aligned}
& f(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x) \\
=& \frac{f(Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)}{P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)} \\
=& \frac{P(Y(1)=1,Y^\tau(0)=0|Y^\tau(1)=0,S=s_1,X=x)}{P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)} \\
=& \frac{P(Y^\tau(0)=0|Y^\tau(1)=0,Y(1)=1,S=s_1,X=x)P(Y(1)=1|Y^\tau(1)=0,S=s_1,X=x)}{P(Y(1)=1|Y^\tau(0)=Y^\tau(1)=0,S=s_1,X=x)} \\
=& \frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x)}{g_v^{EAAR}(s_1,x)}
\end{aligned}
$$

and

$$
\begin{aligned}
& P(Y^\tau(0)=0|Y^\tau(1)=0,X=x) \\
=& \int f(S=s,Y^\tau(0)=0|Y^\tau(1)=0,X=x)dF(s|Y^\tau(1)=0,X=x) \\
=& \int \frac{w(s_1,x,y=1;\beta,\alpha)g_v(s_1,x)}{g_v^{EAAR}(s_1,x)} f(s_1|x;\gamma)ds_1
\end{aligned}
$$

$$
\begin{aligned}
& P(Y^\tau = 0 | Z = 1, X = x) \\
=& P(Y^\tau(1) = 0 | X = x) \\
=& \theta_v(x; \mu)
\end{aligned}
$$

$$
\begin{aligned}
& P(Y^\tau = 0 | Z = 0, X = x) \\
=& P(Y^\tau(0) = 0 | X = x) \\
=& P(Y^\tau(0) = Y^\tau(1) = 0 | X = x) \\
=& P(Y^\tau(1) = 0 | X = x) P(Y^\tau(0) = 0 | Y^\tau(1) = 0, X = x) \\
=& \theta_v(x; \mu) \int \frac{w(s_1, x, y = 1; \beta, \alpha) g_v(s_1, x)}{g_v^{EAAR}(s_1, x)} f(s_1 | x; \gamma) ds_1 \\
=& \begin{cases} \theta_v(x; \mu) \int \frac{w(s_1, x, y=1; \beta, \alpha) g_v(s_1, x; \eta_v^a)}{g_v^{*EAAR}(s_1, x; \beta, \alpha, \eta_v^a)} f(s_1 | x; \gamma) ds_1 & \text{under } \mathcal{M}_\text{a} \\ \theta_v(x; \mu) \int \frac{w(s_1, x, y=1; \beta, \alpha) g_v^*(s_1, x; \beta, \alpha, \eta_v^b)}{g_v^{EAAR}(s_1, x; \eta_v^a)} f(s_1 | x; \gamma) ds_1 & \text{under } \mathcal{M}_\text{b} \end{cases} \\
\equiv& \theta_p(x; \mu, \beta, \alpha, \eta_v, \gamma)
\end{aligned}
$$

C.2 Constructing the Likelihood

C.2.1 $Y_i^\tau = 0,\ Z_i = 1,\ \delta_i = 1$

For subjects with $Y_i^\tau = 0,\ Z_i = 1,\ \delta_i = 1,\ Y_i = y,\ S_i(1) = s_1,\ X_i = x$.

$$
\begin{aligned}
f_O(O) \propto & P_{\delta | Y, Y^\tau, S, Z, X}(\delta = 1 | Y^\tau = 0, Y = y, S = s_1, Z = 1, X = x) && \text{(C.1)} \\
& P_{Y | Y^\tau, S, Z, X}(y | Y^\tau = 0, S = s_1, Z = 1, X = x) && \text{(C.2)} \\
& f_{S | Y^\tau, Z, X}(s_1 | Y^\tau = 0, Z = 1, X = x) && \text{(C.3)} \\
& P_{Y^\tau | Z, X}(Y^\tau = 0 | Z = 1, X = x) && \text{(C.4)}
\end{aligned}
$$

Expression (C.1)$= P(\delta_i = 1 | Y^\tau = 0, Y = y, S = s_1, Z = z, X = x) = P(\delta_i = 1 | Y^\tau = 0, Y = y, Z = z, X = x) \equiv \pi_0(y, z, x)$. When π_0 is unknown, we can substitute π_0 with a consistent estimator $\hat{\pi}_\lambda$ where λ is a parameter to be estimated by ML from observations $\{i : Y_i^\tau = 0\}$.

Expression (C.2)$= \{g_v(s_1, x; \eta_v^a)\}^y \{1 - g_v(s_1, x; \eta_v^a)\}^{1-y}$.

Expression (C.3)= $f_v(s_1|x;\gamma)$.

Expression (C.4)= $\theta_v(x;\mu)$.

C.2.2 $Y_i^\tau = 0,\ Z_i = 1,\ \delta_i = 0$

For subjects with $Y_i^\tau = 0,\ Z_i = 1,\ \delta_i = 0,\ Y_i = 0$(implied by $\delta_i = 0$), $X_i = x$, the observed data $O_i = (Z_i, X_i, Y_i^\tau, Y_i, \delta_i)$.

$$f_O(O) \propto P_{\delta|Y,Y^\tau,Z,X}(\delta = 0|Y^\tau = 0, Y = 0, Z = 1, X = x) \tag{C.5}$$

$$P_{Y|Y^\tau,Z,X}(y = 0|Y^\tau = 0, Z = 1, X = x) \tag{C.6}$$

$$P_{Y^\tau|Z,X}(Y^\tau = 0|Z = 1, X = x) \tag{C.7}$$

Expression (C.5) $= \pi_0(y = 0, z = 1, x)$.

$$\begin{aligned}\text{Expression(C.6)} &= \int f(y = 0, s_1|Y^\tau = 0, Z = 1, X = x)dF(s_1|Y^\tau = 0, Z = 1, X = x) \\ &= \int P(y = 0|Y^\tau = 0, Z = 1, S = s_1, X = x)f(s_1|Y^\tau = 0, Z = 1, X = x)ds_1 \\ &= \int \{1 - g_v(s_1, x; \eta_v^a))\}\, f_v(s_1|x;\gamma)ds_1.\end{aligned}$$

Expression (C.7)= $\theta_v(x;\mu)$.

C.2.3 $Y_i^\tau = 0,\ Z_i = 0$

For subjects with $Y_i^\tau = 0,\ Z_i = 0,\ Y_i = y,\ X_i = x$, the observed data $O_i = (Z_i, X_i, Y_i^\tau, Y_i, \delta_i)$.

$$f_O(O) \propto P_{\delta|Y,Y^\tau,Z,X}(\delta = 0|Y^\tau = 0, Y = y, Z = 0, X = x) \tag{C.8}$$

$$P_{Y|Y^\tau,Z,X}(y|Y^\tau = 0, Z = 0, X = x) \tag{C.9}$$

$$P_{Y^\tau|Z,X}(Y^\tau = 0|Z = 0, X = x) \tag{C.10}$$

Expression (C.8)= 1.

$$\begin{aligned}\text{Expression(C.9)} &= \int f(y, s_1|Y^\tau = 0, Z = 0, X = x)dF(s_1|Y^\tau = 0, Z = 0, X = x) \\ &= \int P(y|Y^\tau = 0, Z = 0, S = s_1, X = x)f(s_1|Y^\tau = 0, Z = 0, X = x)ds_1 \\ &= \int \{g_p(s_1, x; \eta_p)\}^y \{1 - g_p(s_1, x; \eta_p)\}^{1-y} f_p^a(s_1|x; \beta, \alpha, \eta_v^a, \gamma)ds_1.\end{aligned}$$

Expression (C.10)= $\theta_p^a(x; \mu, \beta, \alpha, \eta_v^a, \gamma)$.

C.2.4 $Y_i^\tau = 1$

For subjects with $Y_i^\tau = 1$, $Y_i = 1$(implied by $Y_i^\tau = 1$), S_i and δ_i are undefined. The observed data $O_i = (Z_i, X_i, Y_i^\tau = 1, Y_i = 1)$.

$$f_O(O) \propto P_{Y|Y^\tau,Z,X}(y = 1|Y^\tau = 1, Z = z, X = x) \tag{C.11}$$

$$P_{Y^\tau|Z,X}(Y^\tau = 1|Z = z, X = x) \tag{C.12}$$

Expression (C.11)= 1.

Expression (C.12)= $\{1 - \theta_v(x; \mu)\}^z \left\{1 - \theta_p^a(x; \mu, \beta, \alpha, \eta_v^a, \gamma)\right\}^{1-z}$.

C.3 Estimating distribution $X|S, Y^\tau(0) = Y^\tau(1) = 0$

$$\begin{aligned}&P(X = x|S = s_1, Y^\tau(0) = Y^\tau(1) = 0) \\ =&\frac{P(X = x, Y^\tau(0) = 0|S = s_1, Y^\tau(1) = 0)}{P(Y^\tau(0) = 0|S = s_1, Y^\tau(1) = 0)} \\ =&\frac{P(X = x, Y^\tau(0) = 0|S = s_1, Y^\tau(1) = 0)}{\sum_{all\ x} P(X = x, Y^\tau(0) = 0|S = s_1, Y^\tau(1) = 0)}.\end{aligned}$$

where the numerator

$$\begin{aligned}&P(X = x, Y^\tau(0) = 0|S = s_1, Y^\tau(1) = 0) \\ =&P(Y^\tau(0) = 0|X = x, S = s_1, Y^\tau(1) = 0) \cdot P(X = x|S = s_1, Y^\tau(1) = 0)\end{aligned}$$

$P(X = x|S = s_1, Y^\tau(1) = 0) = P(X = x|S = s_1, Y^\tau = 0, Z = 1)$ can either be estimated using a multinomial distribution by vaccine recipients with S measured with inverse prob. weighting, or

it can be estimated through:

$$P(X|S, Y^{\tau}(1) = 0) = \frac{P\left(X, S|Y^{\tau}(1) = 0\right)}{P\left(S|Y^{\tau}(1) = 0\right)} = \frac{P\left(X, S|Y^{\tau}(1) = 0\right)}{\sum_{all\ x} P\left(X, S|Y^{\tau}(1) = 0\right)}$$

and the numerator

$$P\left(X, S|Y^{\tau}(1) = 0\right) = P(S|X, Y^{\tau}(1) = 0) \cdot P(X|Y^{\tau}(1) = 0).$$

$P(S|X, Y^{\tau}(1) = 0)$ can be estimated by $\widehat{\gamma}$ and $P(X|Y^{\tau}(1) = 0)$ on subjects with $Y_i^{\tau}(1) = 0$.

And

$$\begin{aligned}
&P(Y^{\tau}(0) = 0|X = x, S = s_1, Y^{\tau}(1) = 0) \\
=&\sum_{y=0}^{1} P(Y^{\tau}(0) = 0|Y(1) = y, Y^{\tau}(1) = 0, S = s_1, X = x) \cdot \\
&P(Y(1) = y|Y^{\tau}(1) = 0, S = s_1, X = x) \\
=&w(s_1, x, y = 1)g_v(s_1, x) + w(s_1, x, y = 0)(1 - g_v(s_1, x))
\end{aligned}$$

where $g_v(s_1, x)$ takes some specific forms under $\mathcal{M}_a$ or $\mathcal{M}_b$.