Yingying Zhuang[1] / Ying Huang[2] / Peter B. Gilbert[3]

Simultaneous Inference of Treatment Effect Modification by Intermediate Response Endpoint Principal Strata with Application to Vaccine Trials

[1] Department of Biostatistics, University of Washington, Seattle, WA, USA, E-mail: yyzhuang@uw.edu
[2] Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[3] Fred Hutchinson Cancer Research Center & University of Washington, Seattle, WA, USA

Abstract:
In randomized clinical trials, researchers are often interested in identifying an inexpensive intermediate study endpoint (typically a biomarker) that is a strong effect modifier of the treatment effect on a longer-term clinical endpoint of interest. Motivated by randomized placebo-controlled preventive vaccine efficacy trials, within the principal stratification framework a pseudo-score type estimator has been proposed to estimate disease risks conditional on the counter-factual biomarker of interest under each treatment assignment to vaccine or placebo, yielding an estimator of biomarker conditional vaccine efficacy. This method can be used for trial designs that use baseline predictors of the biomarker and/or designs that vaccinate disease-free placebo recipients at the end of the trial. In this article, we utilize the pseudo-score estimator to estimate the biomarker conditional vaccine efficacy adjusting for baseline covariates. We also propose a perturbation resampling method for making simultaneous inference on conditional vaccine efficacy over the values of the biomarker. We illustrate our method with datasets from two phase 3 dengue vaccine efficacy trials.

Keywords: dengue vaccine, estimated likelihood, principal stratification, pseudo-score, resampling method
DOI: 10.1515/ijb-2018-0058
Received: June 9, 2018; **Revised:** February 9, 2019; **Accepted:** June 10, 2019

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. The research in this manuscript 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. Such post-randomization effect modification research has been referred to as principal surrogate evaluation by Frangakis and Rubin [1].

Various research has been conducted for evaluating surrogates under the principal stratification framework [2–5]. For evaluating continuous immune response biomarkers as principal surrogates in vaccine trials, Gilbert and Hudgens [2] proposed the marginal Causal Effect Predictiveness (mCEP) curve as an estimand for principal surrogacy. The mCEP curve is defined in terms of clinical risks conditional on the potential biomarker if assigned to vaccine and it quantifies how well causal treatment effects on the biomarker predict causal treatment effects on the clinical endpoint. Based on the mCEP curve, a useful biomarker is one that is a strong effect modifier, where the mCEP varies over subgroups defined by the biomarker levels if receiving vaccine [6].

To estimate the mCEP curve in the presence of missing counter-factual biomarker values if receiving vaccine in an efficacy trial, Huang, Gilbert, and Wolfson [7] proposed a pseudo-score type estimator (inspired by Chatterjee, Chen, and Breslow [8]). This estimator utilizes a baseline predictor associated with the biomarker if receiving vaccine as an auxiliary variable to estimate the parameters in the assumed parametric risk model. Huang et al. [7] 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 biomarker if receiving vaccine. This condition may be reasonable in some settings. For example, in a HIV vaccine efficacy trial where the biomarker

Yingying Zhuang is the corresponding author.

of interest is an immune response to HIV, the trial can be designed to inoculate everyone with an irrelevant vaccine, say rabies, prior to randomization and the immune response to the rabies vaccine at baseline is a good choice for the baseline predictor [9]. In general, however, baseline predictors could still contribute to infection risk in addition to biomarker if receiving vaccine, and their effects need to be properly accounted for valid inference. In this article, we develop methodology to estimate the mCEP curve allowing the risk model to be dependent on the baseline predictor after conditioning on the biomarker if receiving vaccine. Moreover, we develop a new procedure for making simultaneous inference of the mCEP curve over a range of biomarker values, which utilizes the perturbation resampling method to approximate the asymptotic distribution of our estimator. 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. All existing work [2, 7, 10, 11] covered inference at individual points of the mCEP curve only.

The research for this paper was motivated by two phase 3 dengue vaccine efficacy trials: the CYD14 trial conducted in Asia (Capeding et al. [12]) and the CYD15 trial conducted in Latin America (Villar et al. [13]) where the biomarker of interest is neutralizing antibody titer to the vaccine measured at 13 months post randomization. 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 for 25 months post first vaccination with active surveillance for occurrence of the primary study endpoint of symptomatic virologically confirmed dengue disease (VCD). Out of the 10,275 (20,869) participants in CYD14 (CYD15), there were 117 (176) VCD cases in the vaccine group and 133 (221) VCD cases in the control group that took place after 13 months post randomization. We demonstrate application of our proposed method using data from CYD14 and CYD15 in Section 5.

The remainder of this article is organized as follows. In Section 2, we introduce the problem setting and propose an estimator for the mCEP curve that accounts for effects of the baseline predictor together with the biomarker if receiving vaccine in the risk model. In Section 3, we introduce a perturbation resampling method to approximate the distribution of our estimator for the mCEP curve and make simultaneous inference. In Section 4, we evaluate the finite-sample performance of the proposed estimator through extensive numerical studies; we also evaluate the performance of the proposed resampling procedure for constructing confidence bands and compare it with alternative bootstrap procedures. In Section 5, we present an analysis of the two phase 3 dengue vaccine efficacy trials, CYD14 and CYD15, using our proposed methods. Finally, in Section 6 we end the article by discussing some potential limitations of our methods and making suggestions for future research.

2 Methods

We consider data from a two-arm vaccine efficacy trial that randomizes n participants to either the vaccine arm or the placebo arm, with Z being the binary indicator of assignment to the vaccine arm. Let W be the baseline covariates measured in everyone such as demographics, S be the intermediate response endpoint measured at a fixed time τ post randomization, Y be the indicator of clinical endpoint occurrence (e. g. infection with a pathogen) after τ over a fixed follow-up period, and Y^τ be the indicator that a subject has the clinical endpoint before τ. The variable S is only measured among those who remain free of Y through time τ and is thus undefined if $Y^\tau = 1$. 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 $S(z) = *$.

Gilbert and Hudgens [2] defined the mCEP curve as

$$mCEP^{risk}(s_1) \equiv h(risk_1(s_1), risk_0(s_1)), \tag{1}$$

where

$$risk_z(s_1) \equiv P\left(Y(z) = 1 | S(1) = s_1, Y^\tau(1) = Y^\tau(0) = 0\right), \tag{2}$$

for $z = 0$, 1 and the function $h(x, y)$ is a known contrast function satisfying $h(x, y) = 0$ if and only if $x = y$. A common choice of the contrast function h is the vaccine efficacy function:

$$\mathrm{VE}(s_1) \equiv 1 - \frac{risk_1(s_1)}{risk_0(s_1)}, \tag{3}$$

which we refer to as the VE curve and it constitutes the estimand of interest in this paper. 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 it would have been had they been assigned to the placebo arm. Large variation 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 mCEP curves are not identifiable from the standard assumptions made in randomized vaccine efficacy trials due to missing potential outcomes. To address this problem, two augmented trial designs have been proposed by Follmann [9]. 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. For example, in dengue vaccine trials, study participants are vaccinated either with dengue vaccine ($z = 1$) or placebo ($z = 0$) and immune response to the dengue vaccine (S) is measured at month 13 post randomization in the vaccine group. Therefore the potential outcome $S(1)$ is observable for the vaccine group but missing for the placebo group. If we can identify a baseline covariate(s) W that is predictive of $S(1)$, and W is measured in subjects in both the vaccine and the placebo groups, then a model predicting $S(1)$ from W fit from the vaccine group can be used to predict $S(1)$ for the placebo group. 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 $S(1)$ as if they originally had been assigned to the vaccine arm. For example, in dengue vaccine trials, if a CPV component were to be added, a subset of placebo patients who are free of the dengue disease primary endpoint at the end of the trial follow-up period will receive the dengue vaccine at study closeout and their immune response will be measured at month 13 post vaccination. Follmann named the second augmented design "closeout placebo vaccination" (CPV). A major advantage of including a CPV component is that it makes it possible to evaluate 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.

Frequently a two-phase sampling design is used. In the first phase, baseline covariates W and the clinical outcome data Y and Y^τ are measured for everyone, and in the second phase, the potential outcome $S(1)$ (and S_c if a CPV component is added) is measured in a subcohort of study participants sampled according to a random mechanism. Sampling of $S(1)$ can depend on other phase-I variables such as Y. For example, in our motivating dengue studies, a two-phase case-control sampling using a BIP-only design could carry out in a way such that in the first phase, baseline covariates W such as age, gender, and country along with the clinical endpoint of dengue disease Y and Y^τ are measured for everyone while in the second phase, all cases in the vaccine arm (defined by $Y = 1$, $Y^\tau = 0$, and $Z = 1$) have $S(1)$ measured and a random subset of controls in the vaccine arm (defined by $Y = 0$ and $Z = 1$) are sampled for $S(1)$ measurement. On the other hand, a two-phase case-control sampling using a BIP+CPV design could carry out in a way such that in the first phase, W, Y and Y^τ are measured for everyone while in the second phase, all cases in the vaccine arm and a random subset of controls in the vaccine arm have $S(1)$ measured, and a random subset of placebo recipients who are uninfected of dengue at the end of the trial (defined by $Y = 0$ and $Z = 0$) are sampled for S_c measurement.

When Gilbert et al. [14] examined the power for detecting vaccine efficacy modification using a parametric estimated likelihood method, the results were seemingly counterintuitive. In some scenarios where W was strongly correlated with $S(1)$ (and S_c if a CPV component is added), the BIP-only design was more powerful than the BIP+CPV design. Huang et al. [7] investigated this result and concluded 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 but not in the step of estimating the conditional distribution of the biomarker. To address the inconsistent use of the CPV component, Huang et al. [7] 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 $risk_Z\{S(1), W\} \equiv P(Y(Z) = 1 | S(1), W, Y^\tau(1) = Y^\tau(0) = 0) = g\{\beta; S(1), Z, W\}$. Furthermore, in order to estimate the mCEP curve, specifically the VE curve as a function of $S(1)$, Huang et al. [7] considered the scenario that the risk model is independent of W after conditioning on $S(1)$: $risk_Z\{S(1), W\} = g\{\beta; S(1), Z\}$. Then it follows that the VE curve estimator can be represented as a function of the risk model directly:

$$\mathrm{VE}(s_1) = 1 - \frac{g\{\beta; S(1) = s_1, Z = 1\}}{g\{\beta; S(1) = s_1, Z = 0\}}. \tag{4}$$

In general, however, clinical risks under Z might not be independent of W after conditioning on $S(1)$. 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 W conditional on the immune response biomarker if receiving vaccine $S(1)$. 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 the pseudo-score estimator by Huang et al. [7] but allows the risk model to depend on W after conditioning on $S(1)$.

2.1 Identifiability assumptions

Throughout we adopt assumptions A1–A7 made by Huang et al. [7] to help identify the pseudo-score estimators from the observed data.

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

(A2) Ignorable Treatment Assignment [15]:

Conditional on W, Z is independent of $(Y^\tau(1), Y^\tau(0), S(1), Y(1), Y(0))$;

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

A1–A2 are standard assumptions in individual-randomized clinical trials. A1 implies that the potential outcomes $(Y_i^\tau(1), Y_i^\tau(0), S_i(1), Y_i(1), Y_i(0))$ are independent of the treatment assignments of other subjects, which implies $(Y_i^\tau(Z_i), S_i(Z_i), Y_i(Zi)) = (Y_i^\tau, S_i, Y_i)$. A2 holds for individual-randomized clinical trials, where the randomization may depend on W. A3 is plausible if the vaccine confers no protective immunity, nor any unintended harmful effects, until after month 13 post first vaccination; this could occur if the body takes more than a year to make enough protective immune cells in response to vaccination. A1–A3 imply that

$$\begin{aligned} risk_z\{s_1, w\} &\equiv P\left(Y(z) = 1 | S(1) = s_1, W = w, Y^\tau(1) = Y^\tau(0) = 0\right) \\ &= P\left(Y = 1 | Z = z, S(1) = s_1, W = w, Y^\tau = 0\right). \end{aligned}$$

Therefore, with additional identifiability assumptions as given below, $risk_z$ can be identified based on subjects who are observed to be at risk at time τ (i. e. $Y^\tau = 0$). Henceforth, we drop the notation of $Y^\tau(1) = Y^\tau(0) = 0$ and tacitly assume all probabilities condition on $Y^\tau(1) = Y^\tau(0) = 0$.

(A4) Risk functions have a generalized linear model form:

$risk_Z\{S(1), W\} = g\{\beta; S(1), Z, W\}$ for some known link function $g(\cdot)$;

For the BIP+CPV design only:

(A5) 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 and independent of other variables including Y, W, and S^{true};

(A6) No placebo subjects event-free at closeout experience the endpoint over the next τ time-units

In A5, S^{true} is the true time constant immune response if receiving vaccine, which is observed subject to measurement errors U. If a CPV component is incorporated in the trial design, all or a fraction of placebo recipients who remain free of the clinical endpoint at the closeout of the trial are vaccinated and the immune response biomarker S_c is measured at time τ after vaccination. A6 assumes no infections occur between closeout of the trial and when S_c is measured; therefore all placebo recipients who remain free of the clinical endpoint at closeout could potentially have S_c measured. Under (A5) and (A6), S_c is substituted for $S(1)$ for subjects selected for CPV. Henceforth we consolidate the notation of $S(1)$ and S_c, and let S be the immune response measurement if receiving vaccine obtained either during the standard trial follow-up or during the CPV and let δ be the indicator for the availability of S.

(A7) $\int\int P(\delta = 1 | y, z, W)\mathrm{d}y\mathrm{d}z > \varepsilon$ for some constant $\varepsilon > 0$ for W almost everywhere.

A7 is needed for the pseudo-score estimators. Note that for A7 to hold, the probability of observing S does not have to be non-zero within each of the four subgroups defined by treatment arm Z and case/control status Y, thus the pseudo-score estimator is applicable to both the BIP only and the BIP+CPV designs.

2.2 The pseudo-score estimator

Assuming the study participants make up a random sample from a large population of interest, the observed data $O_i \equiv (Z_i, W_i, Y_i^\tau, Y_i, \delta_i, S_i\delta_i)'$, $i = 1,\ldots, n$, are i.i.d. copies of a random vector $O \equiv (Z, W, Y^\tau, Y, \delta, S\delta)'$. Based on the missing at random assumption, we consider the maximization of the observed likelihood

$$\begin{aligned} L &= \prod_{\delta_i=1} P(Y_i | Z_i, W_i, S_i) \prod_{\delta_i=0} P(Y_i | Z_i, W_i) \\ &= \frac{\prod_{\delta_i=1} P(Y_i | Z_i, W_i, S_i) \prod_{\delta_i=0} P(Y_i | Z_i, W_i)}{\prod_{\delta_i=1} P(Y_i | Z_i, W_i, S_i) \prod_{\delta_i=0} \int P(Y_i | Z_i, W_i, s) dF(s | Z_i, W_i)}, \end{aligned} \tag{5}$$

where $F(S|Z,W)$ is the CDF of S conditional on Z and W. Earlier work for identifying risk model parameters involves two steps. In the first step, based on (A2), the Ignorable Treatment Assignment assumption, $dF(S|Z,W) = dF(S|W)$. Thus $F(S|Z,W)$ is estimated using vaccine recipients with S measured. When sampling of S depends on variables such as Y, inverse probability weighting (IPW) can be implemented to correct for biased sampling. In the second step, the estimator of $F(S|Z,W)$ is substituted into (5) and risk model parameters are estimated as the maximizer of the resulting estimated likelihood [2, 11, 16]. In a BIP+CPV design, data from the CPV component are used in the second step but not the first step due to the fact that all infected placebo recipients have zero sampling probability for S, thus IPW cannot be applied to the whole S sample in estimating $F(S|W)$. This inconsistency could cause decreased efficiency as observed in Gilbert et al. [14]. Huang et al. [7] proposed a pseudo-score type estimator to address this issue. The score equation of the observed likelihood L is

$$\frac{\partial l}{\partial \beta} = \sum_{\delta_i=1} U_\beta(Y_i|Z_i,W_i,S_i) + \sum_{\delta_i=0} \frac{\int U_\beta(Y_i|Z_i,W_i,s)P(Y_i|Z_i,W_i,s)dF(s|Z_i,W_i)}{\int P(Y_i|Z_i,W_i,s)dF(s|Z_i,W_i)} = 0, \tag{6}$$

where $U_\beta(Y|Z,W,S) = \frac{\partial log P(Y|Z,W,S)}{\partial \beta}$. Based on (A2), $dF(S|Z,W) = dF(S|W)$. Furthermore, according to Bayes' theorem, $dF(S|W) = \frac{dF(S|W,\delta=1)P(\delta=1|W)}{P(\delta=1|S,W)}$. Thus, the pseudo-score type estimator proposed by Huang et al. [7] is defined as the solution to the pseudo-score estimation equation

$$U(\beta,F_0,\pi_0) = \sum_{\delta_i=1} U_\beta(Y_i|S_i,Z_i,W_i) + \sum_{\delta_i=0} \frac{\int U_\beta(Y_i|s,Z_i,W_i)\frac{P(Y_i|s,Z_i,W_i)}{P(\delta=1)|s,W_i)}dF(s|W_i,\delta=1)}{\int \frac{P(Y_i|s,Z_i,W_i)}{P(\delta=1)|s,W_i)}dF(s|W_i,\delta=1)}. \tag{7}$$

The name "pseudo-score type estimator" was first introduced by Chatterjee et al. [8] and adopted by Huang et al. [7]. The key component of the pseudo-score estimation equation is that instead of directing estimating the term $F(S|Z,W)$ in eq. (5), which introduces the problem of inconsistent usage of the CPV component, the pseudo-score method forms an estimating function depending on $F(S|W,\delta=1)$. For the BIP-only design, the cumulative distribution function $F_0 \equiv F(s|w,\delta=1)$ is estimated using S data from the vaccine recipients while for the BIP+CPV design it is estimated using S data from both the vaccine recipients and the CPV component. For discrete W, we estimate F_0 empirically by $F_N \equiv \frac{\sum_{\delta_i=1} I(S_i\le s,W_i=w)}{\sum_{\delta_i=1} I(W_i=w)}$. Note that in both BIP-only and BIP+CPV designs, whether S is available is independent of the values of S after conditioning on Y, Z, and W, thus $P(\delta=1|y,z,S,W) = P(\delta=1|y,z,W)$. Therefore, the probability $\pi_0 \equiv P(\delta=1|S,W) = \int\int P(\delta=1|y,z,S,W)P(y,z|S,W)dydz = \int\int P(\delta=1|y,z,W)P(y|S,z,W)P(z)dydz$ can be estimated by $\hat{\pi} \equiv \hat{P}(\delta=1|S,W) = \sum_{z=0}^{1}\sum_{y=0}^{1}\hat{P}(\delta=1|y,z,W)\hat{P}(y|S,z,W)P(z)$. By substituting F_0 and π_0 in expression (7) 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 $risk_Z\{S,W\} = g\{\beta;S,Z,W\}$ is then obtained by solving the equation $U(\beta,F_N,\hat{\pi}) = 0$.

Huang et al. [7] considered estimating the VE curve under the scenario that the risk model is independent of W after conditioning on immune response if receiving vaccine S, which we refer to as (A8):

(A8) $risk_Z\{S,W\} = risk_Z\{S\} = g\{\beta;S,Z\}$.

Follmann [9] made exactly the same assumption that W has no effect on $Y(Z)$ once S and Z are in the model in his eq. (1). Follmann also discussed that this assumption can be relaxed when using a maximum likelihood approach with generalizations to the risk model that includes an additional main effect of W, or even allows for an interaction term between W 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 W and Z. Furthermore, (A8) is similar to a standard assumption made in many "surrogate" measurement error statistical methods that the conditional distribution of Y given (S,Z,W) depends only on (S,Z), in which case W is said to be a *surrogate* [17].

With (A8), it follows that

$$\mathrm{VE}(s_1) = 1 - \frac{g\{\beta;S=s_1,Z=1\}}{g\{\beta;S=s_1,Z=0\}}. \tag{8}$$

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 (A8)-based VE estimator $\widehat{\mathrm{VE}}^{(A8)}(s_1)$. Asymptotic normality of $\widehat{\mathrm{VE}}^{(A8)}(s_1)$ follows by applying the delta method. We investigate the effects of assumption (A8) on estimation in further detail in Section 4.1.

2.3 Estimating the marginal CEP curve for general settings

In this section, we consider estimating $\mathrm{VE}(s_1)$ with a two-phase sampling design, without assuming (A8), and call this estimator $\widehat{\mathrm{VE}}^{new}(s_1)$. We consider situations where W is categorical with D levels: $w_1, w_2, ..., w_D$. Then $\mathrm{VE}(s_1)$ can be represented as

$$\mathrm{VE}(s_1) = 1 - \frac{\sum_{j=1}^{D} g\left\{\beta; S = s_1, Z = 1, W = w_j\right\} \cdot P(W = w_j | S = s_1)}{\sum_{j=1}^{D} g\left\{\beta; S = s_1, Z = 0, W = w_j\right\} \cdot P(W = w_j | S = s_1)}. \tag{9}$$

The parameter β can be estimated by the pseudo-score type estimators following the approach in Huang et al. [7] as summarized in Section 2.2. We propose to model $P(W|S)$ with a multinomial logistic function. Denote $P(W = w_j|S = s_1) \equiv \mu_j\{\gamma; s_1\}$, $j = 1,2, ...D$ for some pre-specified parametric functions μ_j. Suppose the multinomial logistic function can be represented as:

$$\begin{aligned}
ln\left(\frac{P(W = w_1|S = s_1)}{P(W = w_D|S = s_1)}\right) &= \gamma_{10} + \gamma_{11}h_1(s_1) + \gamma_{12}h_2(s_1) + ... + \gamma_{1T}h_T(s_1) = \boldsymbol{\gamma_1}^T\boldsymbol{h}(s_1) \\
ln\left(\frac{P(W = w_2|S = s_1)}{P(W = w_D|S = s_1)}\right) &= \boldsymbol{\gamma_2}^T\boldsymbol{h}(s_1) \\
&\vdots \\
ln\left(\frac{P(W = w_{D-1}|S = s_1)}{P(W = w_D|S = s_1)}\right) &= \boldsymbol{\gamma_{D-1}}^T\boldsymbol{h}(s_1),
\end{aligned}$$

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

$$\begin{aligned}
P(W = w_1|S = s_1) &= \frac{e^{\boldsymbol{\gamma_1^T h}(s_1)}}{1 + \sum_{d=1}^{D-1} e^{\boldsymbol{\gamma_d^T h}(s_1)}} \equiv p_1(s_1) \\
P(W = w_2|S = s_1) &= \frac{e^{\boldsymbol{\gamma_2^T h}(s_1)}}{1 + \sum_{d=1}^{D-1} e^{\boldsymbol{\gamma_d^T h}(s_1)}} \equiv p_2(s_1) \\
&\vdots \\
P(W = w_D|S = s_1) &= \frac{1}{1 + \sum_{d=1}^{D-1} e^{\boldsymbol{\gamma_d^T h}(s_1)}} \equiv p_D(s_1).
\end{aligned}$$

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 $\pi_{0i} \equiv P(\delta_i = 1|y_i, z_i, w_i)$ and $p_{ji} \equiv p_j(s_{1i}) = P(W = w_j|S = s_{1i})$ where $i = 1,2,..., n$ and $j = 1,2,..., D$. Given that W 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 W: $G_1, G_2, ..., G_D$. If $W = w_d$, then $G_d = 1$ and all other $G's = 0$. Then the likelihood function for $P(W|S)$ can be derived as $L(\gamma) = \prod_{i=1}^{n} p_{1i}^{g_{1i}} \cdot p_{2i}^{g_{2i}} \cdots p_{Di}^{g_{Di}}$. In the Supplemental Material Section A, we show that the estimating function for γ can be found by taking the first partial derivatives of the log-likelihood function $l(\gamma)$ with respect to each of the $D(T+1)$ unknown parameters: $\frac{\partial l(\gamma)}{\partial \gamma_{jt}} = \sum_{i=1}^{n} h_t(s_{1i})(g_{ji} - p_{ji})$ 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

$$\Psi(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi_{0i}} h_t(s_{1i})(g_{ji} - p_{ji}) = 0. \tag{10}$$

When π_0 is unknown and needs to be estimated with a consistent estimator $\hat{\pi}_\alpha$, α is substituted with its maximum likelihood estimator (MLE) from the Phase-I observations and π_0 is be substituted with $\hat{\pi}_\alpha$ in eq. (10) to obtain $\hat{\gamma}$. Based on the estimators $\hat{\beta}$ and $\hat{\gamma}$, we estimate $\mathrm{VE}(s_1)$ with

$$\widehat{\mathrm{VE}}^{(new)}(s_1) = 1 - \frac{\sum_{j=1}^{D} g\left\{\hat{\beta}; S = s_1, Z = 1, W = w_j\right\} \cdot \mu_j\{\hat{\gamma}; s_1\}}{\sum_{j=1}^{D} g\left\{\hat{\beta}; S = s_1, Z = 0, W = w_j\right\} \cdot \mu_j\{\hat{\gamma}; s_1\}}. \tag{11}$$

Under regularity conditions specified in the Supplemental Material Section B.1, estimators $\hat{\beta}$, $\hat{\gamma}$, and $\widehat{\mathrm{VE}}^{(new)}(s_1)$ can be shown to be consistent and asymptotically normally distributed. Theorem 1 in the Supplemental Material Section B.2 describes the asymptotic distribution of $\hat{\beta}$, $\hat{\gamma}$, and $\widehat{\mathrm{VE}}^{(new)}(s_1)$ with a proof sketched.

3 Simultaneous inference

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: $\mathrm{VE}(s_1) = ve$ 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: $\mathrm{VE}(S = s_1) = \mathrm{VE}(S' = s_1)$ 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: $\mathrm{VE}(s_1|W_1 = 1) = \mathrm{VE}(s_1|W_1 = 0)$ for $s_1 \in [s_l, s_u]$, where W_1 is the indicator of trial 1 (i. e. $W_1 = 1$ for trial 1 and $W_1 = 0$ 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. In this article, we propose a perturbation resampling method Parzen, Wei, and Ying inspired by Parzen, Wei, and Ying [18] 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 $\mathscr{E} \equiv \{\epsilon_i, i = 1, 2, ..., n\}$.

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

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

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

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

when π_0 is known, or

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

when π_0 is unknown, where $\hat{\pi}^{(\epsilon)}(y_i, z_i, w_i) = \hat{P}^{(\epsilon)}(\delta = 1|y_i, z_i, w_i)$, 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:

$$\begin{aligned} log\widehat{RR}^{(\epsilon)}(s_1) &= log\left\{\frac{risk_1^{(\epsilon)}(s_1)}{risk_0^{(\epsilon)}(s_1)}\right\} \\ &= log\left\{\frac{\sum_{j=1}^{D} g\left\{\hat{\beta}^{(\epsilon)};S=s_1,Z=1,W=w_j\right\}\cdot\mu_j\{\hat{\gamma}^{(\epsilon)};s_1\}}{\sum_{j=1}^{D} g\left\{\hat{\beta}^{(\epsilon)};S=s_1,Z=0,W=w_j\right\}\cdot\mu_j\{\hat{\gamma}^{(\epsilon)};s_1\}}\right\}. \end{aligned}$$

Repeat Steps 1–4 B_0 times to obtain B_0 realizations of $log\widehat{RR}^{(\epsilon)}(s_1)$, denoted by $\left\{log\widehat{RR}^{(b)}(s_1), b = 1, 2, ..., B_0\right\}$. The empirical distribution of $\widehat{W}_{logRR}(s_1)^{(b)} \equiv \sqrt{n}\left\{log\widehat{RR}^{(b)}(s_1) - log\widehat{RR}(s_1)\right\}$ conditional on the observed data can be used to approximate the distribution of $\widehat{W}_{logRR}(s_1) \equiv \sqrt{n}\left\{log\widehat{RR}(s_1) - logRR_0(s_1)\right\}$. In the Supplemental Material Section C, we provide theoretical justification for why the distribution of $\widehat{W}_{logRR}(s_1)^{(b)}$ given the observed data $O_i = (Z_i, W_i, Y_i^{\tau}, Y_i, \delta_i, S_i\delta_i)'$, $i = 1,..., n$ 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 $\left\{log\widehat{RR}^{(b)}(s_1), b = 1, 2, ..., B_0\right\}$. A $100(1-\alpha)\%$ pointwise confidence interval and simultaneous confidence band for $\{logRR(s_1), s_1 \in \zeta\}$ may be constructed as $log\widehat{RR}(s_1) \pm \mathcal{Z}_{1-\alpha/2}\hat{\sigma}_{logRR}(s_1)$ and $log\widehat{RR}(s_1) \pm \mathcal{Q}_{1-\alpha}\hat{\sigma}_{logRR}(s_1)$, respectively, where $\mathcal{Z}_{\eta}$ is the 100ηth percentile of the standard normal distribution $N(0, 1)$ 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 $\widehat{\text{VE}}(s_1)$ are obtained by transformation of the symmetric bounds from the *logRR* scale back to the VE scale.

4 Simulation studies

We conducted simulation studies to examine the finite-sample performance of our proposed estimator $\widehat{\text{VE}}^{(new)}(s_1)$ and the perturbation resampling procedure for inference. For comparison, we also studied $\widehat{\text{VE}}^{(A8)}(s_1)$, the estimator that makes the extra assumption (A8) that after accounting for Z and S the conditional risk does not depend on W, as well as a traditional bootstrap procedure for making pointwise and simultaneous inference. Specifically, we assessed (1) the bias of $\widehat{\text{VE}}^{(new)}(s_1)$ compared to $\widehat{\text{VE}}^{(A8)}(s_1)$, (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. W was generated to have four categories, 1, 2, 3, 4, from the following multinomial model conditional on S: $ln\left(\frac{P(W=1|S)}{P(W=4|S)}\right) = -1.99 + 0.89S$; $ln\left(\frac{P(W=2|S)}{P(W=4|S)}\right) = -4.80 + 1.84S$; $ln\left(\frac{P(W=3|S)}{P(W=4|S)}\right) = -9.79 + 3.29S$. 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 correlation between W and S was 0.5. The risk model $P(Y = 1|S, Z, W)$ was assumed to take the form $P(Y = 1|S, Z, W) = \Phi(\beta_0 + \beta_1 Z + \beta_2 S + \beta_3 SZ + \beta_4 W)$ 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, CYD15, and of HIV infection in RV144, a phase 3 HIV vaccine efficacy trial conducted in Thailand (Rerks-Ngarm et al. [19]), 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. We show the results from the BIP+CPV design in this section. Analyses from the BIP-only design yielded similar conclusions and corresponding results have been included in the Supplemental Material Section D.

4.1 Bias of $\widehat{\mathrm{VE}}^{(A8)}(s_1)$ and $\widehat{\mathrm{VE}}^{(new)}(s_1)$

For each of 1000 simulated data sets, the estimates $\widehat{\mathrm{VE}}^{(A8)}(s_1)$ and $\widehat{\mathrm{VE}}^{(new)}(s_1)$ were computed. Figure 1 displays the true VE curve, the average $\widehat{\mathrm{VE}}^{(A8)}(s_1)$, and average $\widehat{\mathrm{VE}}^{(new)}(s_1)$ over the 1000 simulations for different sampling ratios in the Non-Rare, Med-Rare, Rare case, respectively. It also reports the average sampling proportions of S for each treatment arm $\left(\overline{\pi}_v, \overline{\pi}_p\right)$, 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 1 that the $\widehat{\mathrm{VE}}^{(A8)}(s_1)$ estimator could significantly underestimate the true VE for small values of S. For example, the bias in $\widehat{\mathrm{VE}}^{(A8)}(s_1)$ 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 $\widehat{\mathrm{VE}}^{(new)}(s_1)$ is generally negligible across all scenarios. These results confirmed the superior performance of our proposed estimator over the estimator making the assumption of (A8), $risk_Z\{S, W\} = risk_Z\{S\} = g\{\beta; S, Z\}$, when the assumption is violated.

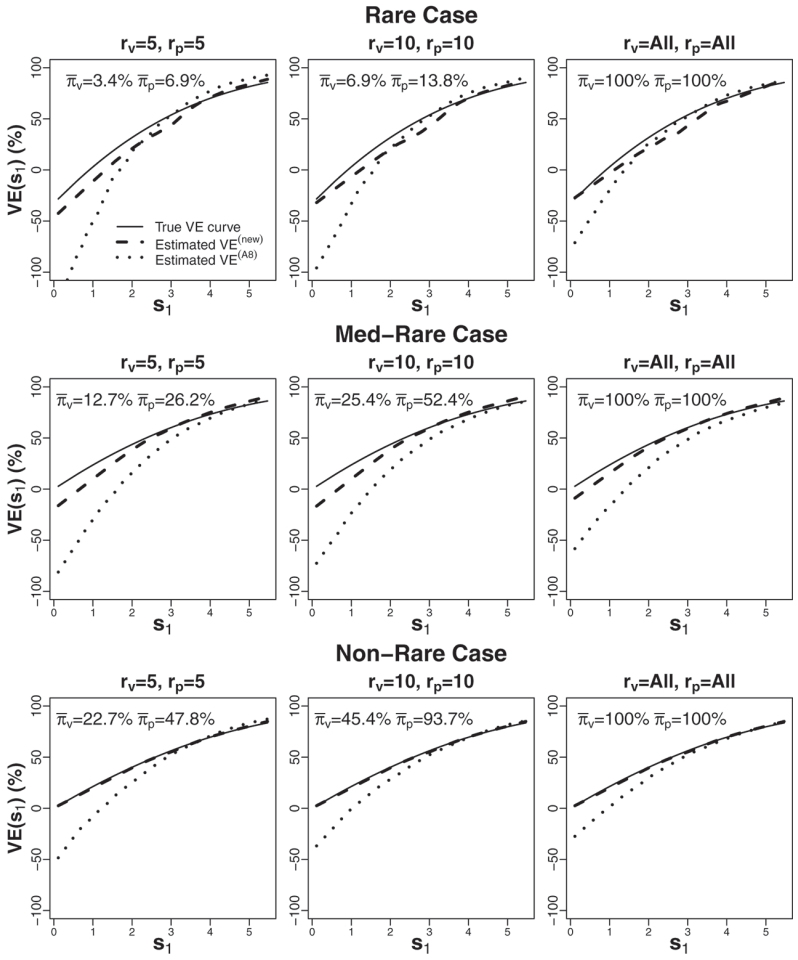

Figure 1: Estimated vaccine efficacy curve using our proposed method without assumption A8 and using the [7] method with assumption A8, compared to the true VE curve for checking the bias of these two estimators based on 500 simulated datasets for the Rare case where the probability of $Y = 1$ for $Z = 0$ (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.

4.2 Standard error estimator

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

on 500 nonparametric bootstrap samples. For comparison, we calculated the Monte Carlo standard errors as a benchmark for the correct standard errors. The results are summarized in Figure 2. Generally the perturbation resampling procedure yielded standard error estimates closer to the Monte Carlo standard errors than the nonparametric bootstrap procedure. When the disease endpoint is rare, perturbation-based standard errors were remarkably smaller than the bootstrap-based standard errors.

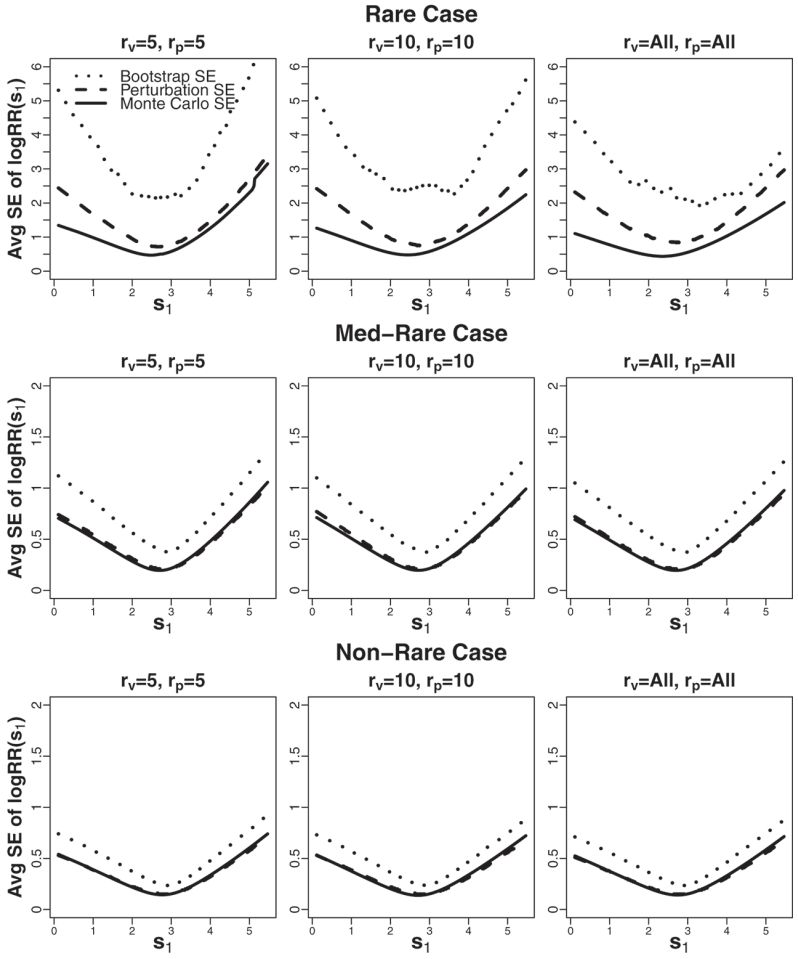

Figure 2: Estimated standard errors of $log\widehat{RR}(s_1)$, solid for the perturbation resampling approach and dashed for the bootstrap approach, for the Rare case, the Med-Rare case, and the Non-Rare case with a BIP+CPV design.

4.3 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. 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 standard error estimates were remarkably large, yielding wide confidence intervals and high bootstrap CI coverage, while the perturbation-based standard error estimates continued to perform well as shown in Section 4.2. The empirical coverage levels of the 95 % simultaneous confidence bands from the perturbation/bootstrap methods are also reported in Figure 3 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 standard errors and therefore narrower pointwise confidence intervals and simultaneous confidence bands while maintaining nominal coverage.

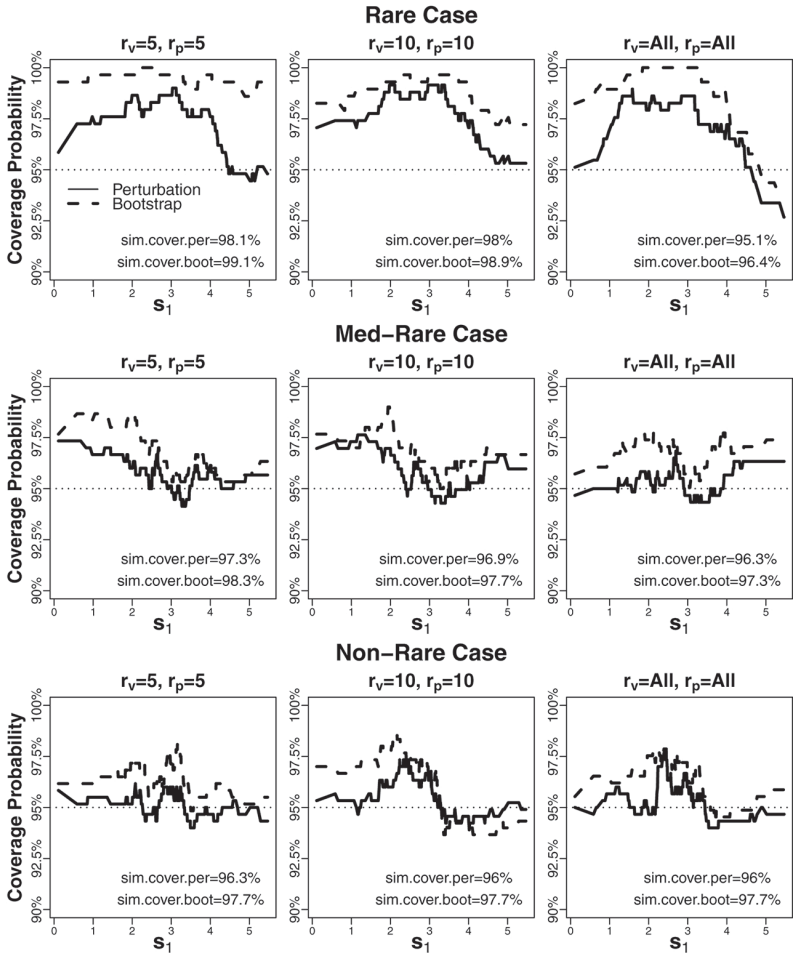

Figure 3: Empirical coverage probabilities of 95 % pointwise confidence intervals and simultaneous confidence bands about $\mathrm{VE}(s_1)$, for the Rare case, the Med-Rare case, and the Non-Rare case with a BIP+CPV design.

Lastly, we compare efficiency of pseudo-score type estimators for β allowing $\beta_4 \neq 0$ (denoted by $\hat{\beta}^{(new)}$) to the pseudo-score type estimators that set $\beta_4 = 0$ (denoted by $\hat{\beta}^{(A8)}$) when the true risk model $risk_Z\{S, W\} = \Phi(\beta_0 + \beta_1 Z + \beta_2 S + \beta_3 SZ + 0W)$, therefore the conditional risk does not depend on W after accounting for Z and S, and the assumption (A8) made in $\widehat{\mathrm{VE}}^{(A8)}$ is correct. The relative efficiency, defined as the ratio of sampling variance of $\hat{\beta}^{(new)}$ and sampling variance of $\hat{\beta}^{(A8)}$, was 1.25, 1.08, 1.28, and 1.07 for the Med-Rare Case with $r_v = 10$ and $r_p = 10$ for β_0, β_1, β_2, β_3, respectively. Relative efficiency for a BIP-only design is included in the Supplemental Material Section D. There is some efficiency loss by using the proposed approach allowing $\beta_4 \neq 0$ when assumption (A8) holds, but the efficiency loss is minor especially for estimating the interaction between S and Z.

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 standard error estimators in general were smaller than the bootstrap-based standard error estimators especially when the disease endpoint is rare.

5 Dengue example

We demonstrate application of our proposed method using data from CYD14 and CYD15. 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 strains at Month 13 (time τ) in case-control samples. For illustration of our methods, 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 W, and Y is the indicator of the same dengue disease endpoint noted above. See Moodie et al. [20] for the reporting of the full analysis to a clinical audience.

A probit risk model was assumed conditional on Z, S, and W: $risk_Z\{S, W\} = \Phi(\beta_0 + \beta_1 Z + \beta_2 S + \beta_3 SZ + \beta_4 W)$. Our proposed methods were conducted on the data set of 9–16 year olds pooling across both CYD14 and CYD15 to assess how VE varied with Month 13 neutralizing antibody titers if assigned to receive the vaccine. Justification for pooling the data across the trials is provided in Moodie et al. [20], with main justification that the protocols for these two trials were essentially the same. In Figure 4 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.

Furthermore, let $LB(s_1)$ denote the lower bound and $UB(s_1)$ denote the upper bound of the simultaneous band for $VE(s_1)$. Then a test of the hypothesis $H_0 : VE(s_1) = ve_0$ for all $s_1 \in \zeta$ with size α is obtained by rejecting H_0 when and only when, at least for one s_1, the statement

$$LB(s_1) \leq ve_0 \leq UB(s_1) \tag{12}$$

is false. It is evident that the size of the test is α, since by the definition of the simultaneous bands constructed in Section 3, $1 - \alpha$ is the chance for the statement (12) to be simultaneously true for all $s_1 \in \zeta$. The fact that a horizonal line cannot be drawn between the $100(1-\alpha)\%$ simultaneous confidence bands without intersecting both the lower and the upper limits in Figure 4 indicates at least for one s_1, the statement (12) is false. Therefore, we reject H_0 and conclude that $VE(s_1)$ significantly varies with s_1.

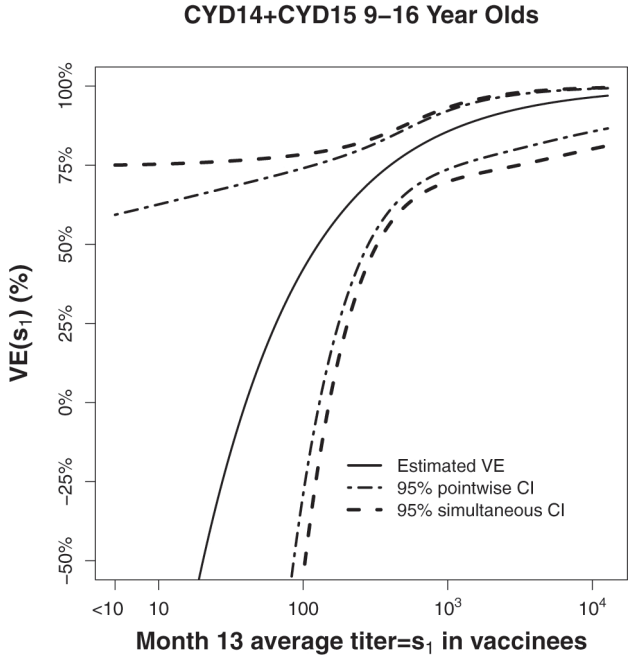

Figure 4: Estimated vaccine efficacy against virologically confirmed dengue disease of any serotype through Month 25 by Month 13 average titer if receiving vaccine with 95 % pointwise confidence intervals and simultaneous confidence bands in 9–16 year olds in the two Phase 3 dengue vaccine efficacy trials combined (CYD14 and CYD15).

6 Discussion

The identification and evaluation of response markers as surrogate endpoints and as effect modifiers are important goals 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, exploration of biological mechanisms of clinical treatment efficacy, and for selecting biomarker study endpoints for evaluating treatments/vaccines in new trials.

In this paper 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 W 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. Theoretically, if sample sizes are large enough, and there are a large number of events, bootstrap and perturbation resampling procedures should both yield correct inference.

However, through simulation we discovered that 500 perturbation resamples approximate standard errors better than 500 bootstrap resamples. When the number of events is small, and some events are associated with extreme S values, the bootstrap resampling procedure will sometimes delete these high influential points, and sometime repeat them. Therefore, the influence from these points will be wiped out in some bootstrap samples while compounded in others, yielding unstable estimates for different samples. On the other hand, the perturbation resampling procedure keeps the same data points but 'perturbs' their influence through continuously generated weights, thus yielding more stable estimates. Our simulation results also showed that the precision gain from perturbation compared to bootstrap is more dramatic when the disease endpoint is rare. Therefore, the perturbation procedure is favorable given finite sample size and/or rare events but the relative advantage of perturbation should diminish with a large number of events as the relative influence of the most influential points should be diminished.

While we have defined the baseline covariates W to be categorical such that we estimated $F(S|W, \delta = 1)$ nonparametrically and employed a multinomial model for $P(W|S)$, our proposed definition and estimation procedure can be extended to scenarios where W is continuous by adopting a nonparametric smoothing technique to estimate $F(S|W, \delta = 1)$ and using a parametric, semiparametric, or nonparametric model to estimate $P(W|S)$. Furthermore, this article assumed the baseline covariates W are phase-I variables that are available in all subjects. Extending our estimator to settings where some Ws are only available from a subset of participants is of interest for future research. The R code implementing the proposed methods is available upon request.

Acknowledgements

The authors thank the participants, investigators, and sponsors of the CYD14 and CYD15 trials. Research reported in this publication was supported by Sanofi Pasteur and the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services, under award number R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or Sanofi Pasteur.

References

[1] Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58:21–9.
[2] Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. Biometrics. 2008;64:1146–54.
[3] Joffe MM, Greene T. Related causal frameworks for surrogate outcomes. Biometrics. 2009;65:530–8.
[4] Li Y, Taylor JM, Elliott MR. A bayesian approach to surrogacy assessment using principal stratification in clinical trials. Biometrics. 2010;66:523–31.
[5] Wolfson J, Gilbert P. Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. Biometrics. 2010;66:1153–61.
[6] Gilbert PB, Gabriel EE, Huang Y, Chan IS. Surrogate endpoint evaluation: Principal stratification criteria and the prentice definition. J Causal Inference. 2015;3:157–75.
[7] Huang Y, Gilbert PB, Wolfson J. Design and estimation for evaluating principal surrogate markers in vaccine trials. Biometrics. 2013;69:301–9.
[8] Chatterjee N, Chen Y-H, Breslow NE. A pseudoscore estimator for regression problems with two-phase sampling. J Am Stat Assoc. 2003;98:158–68.
[9] Follmann, D. Augmented designs to assess immune response in vaccine trials. Biometrics. 2006;62:1161–9.
[10] Gabriel EE, Gilbert PB. Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. Biostatistics. 2013;15:251–65.
[11] Huang Y, Gilbert PB. Comparing biomarkers as principal surrogate endpoints. Biometrics. 2011;67:1442–51.
[12] Capeding MR, Tran NH, Hadinegoro SR, Ismail HI, Chotpitayasunondh T, Chua MN, 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. Lancet. 2014;384:1358–65.
[13] Villar L, Dayan GH, Arredondo-García JL, Rivera DM, Cunha R, Deseda C. Efficacy of a tetravalent dengue vaccine in children in Latin America. N Engl J Med. 2015;372:113–23.
[14] Gilbert PB, Grove D, Gabriel E, Huang Y, Gray G, Hammer SM. A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple hiv vaccine regimens. Stat Commun Infect Dis. 2011;3. DOI: 10.2202/1948-4690.1037
[15] Rubin DB. Comment: Which ifs have causal answers. J Am Stat Assoc 1986;81:961–2.
[16] Qin L, Gilbert PB, Follmann D, Li D. Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the cox model. Ann Appl Stat. 2008;2:386.
[17] Carroll RJ., Ruppert D, Crainiceanu CM, Stefanski LA. Measurement error in nonlinear models: a modern perspective. Boca Raton, Florida: Chapman & Hall/CRC, 2006.
[18] Parzen M, Wei L, Ying Z. A resampling method based on pivotal estimating functions. Biometrika. 1994;81:341–50.

[19] Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R. Vaccination with alvac and aidsvax to prevent hiv-1 infection in Thailand. N Engl J Med. 2009;361:2209–20.
[20] Moodie Z, Juraska M, Huang Y, Zhuang Y, Fong Y, Carpp LN, et al. Neutralizing antibody correlates analysis of tetravalent dengue Vvaccine efficacy trials in Asia and Latin America. J Infect Dis. 2017;217:742–53. DOI: 10.1093/infdis/jix609.

Supplementary Material: The online version of this article offers supplementary material (DOI:https://doi.org/10.1515/ijb-2018-0058).