Inference on treatment effect modification by biomarker response in a three-phase sampling design

MICHAL JURASKA*

*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center,
1100 Fairview Ave N., Seattle, WA 98109, USA*
mjuraska@fredhutch.org

YING HUANG, PETER B. GILBERT

*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, 1100 Fairview
Ave N., Seattle, WA 98109, USA and Department of Biostatistics, University of Washington,
1705 NE Pacific Street, Seattle, WA 98195, USA*

SUMMARY

An objective in randomized clinical trials is the evaluation of “principal surrogates,” which consists of analyzing how the treatment effect on a clinical endpoint varies over principal strata subgroups defined by an intermediate response outcome under both or one of the treatment assignments. The latter effect modification estimand has been termed the marginal causal effect predictiveness (mCEP) curve. This objective was addressed in two randomized placebo-controlled Phase 3 dengue vaccine trials for an antibody response biomarker whose sampling design rendered previously developed inferential methods highly inefficient due to a three-phase sampling design. In this design, the biomarker was measured in a case-cohort sample and a key baseline auxiliary strongly associated with the biomarker (the “baseline surrogate measure”) was only measured in a further sub-sample. We propose a novel approach to estimation of the mCEP curve in such three-phase sampling designs that avoids the restrictive “placebo structural risk” modeling assumption common to past methods and that further improves robustness by the use of non-parametric kernel smoothing for biomarker density estimation. Additionally, we develop bootstrap-based procedures for pointwise and simultaneous confidence intervals and testing of four relevant hypotheses about the mCEP curve. We investigate the finite-sample properties of the proposed methods and compare them to those of an alternative method making the placebo structural risk assumption. Finally, we apply the novel and alternative procedures to the two dengue vaccine trial data sets.

Keywords: Biomarker; Dengue; Principal stratification; Principal surrogate endpoint; Three-phase sampling design; Treatment effect modification; Vaccine.

*To whom correspondence should be addressed.

1. INTRODUCTION

Over the past 20 years, a sizable body of work has accumulated on estimation and inference on principal stratification estimands with application to “principal surrogate” evaluation. The goal of principal surrogate evaluation is to study how a clinical treatment effect in a randomized trial varies over principal strata subgroups defined by the potential outcome biomarker response under both treatment assignments (Frangakis and Rubin, 2002) or by the potential outcome biomarker response under one of the treatment assignments (Gilbert and Hudgens, 2008). These effect modification estimands have been called the causal effect predictiveness surface and the marginal causal effect predictiveness (mCEP) curve, respectively. Both estimands are useful, and, in this work, we focus on the mCEP curve estimand because of its greater facilitation for decision-making and its easier identifiability. Gilbert and Hudgens (2008) studied fully parametric and fully non-parametric estimated maximum likelihood estimation (EML) methods for this estimand, Huang and Gilbert (2011) relaxed the fully parametric EML methods to semi-parametric EML methods, and Huang and others (2013) and Huang (2018) replaced EML with pseudo-score (PS) estimation, demonstrating improved efficiency compared to EML and providing analytic variance estimation that was not possible via EML. These methods considered a binary clinical endpoint, and Gabriel and Gilbert (2014) and Gabriel and others (2015) extended the work to accommodate a time-to-event clinical endpoint subject to right-censoring. Moreover, Li and others (2010) and other papers from the same group studied full likelihood Bayesian methods, as did Zigler and Belin (2011).

The present work is motivated by two randomized placebo-controlled Phase 3 trials of a dengue vaccine, where the primary clinical endpoint of interest was symptomatic virologically confirmed dengue (VCD) occurrence between the month 13 visit and the month 25 visit. Overall vaccine efficacy [one minus relative risk (vaccine/placebo) of VCD times 100%] to prevent VCD over this follow-up period was estimated at 56.5% (95% CI 43.8–66.4) in the trial in Asia (Capeding and others, 2014) and at 60.8% (95% CI 52.0–68.0) in the trial in Latin America (Villar and others, 2015). In these trials with harmonized study designs and protocols, the biomarker S of interest to study as a modifier of vaccine efficacy (in participants free of VCD through month 13) was a participant’s average $\log_{10}$ neutralizing antibody titer against the four dengue virus strains represented in the vaccine construct (one of each serotype), measured from a serum sample taken at the month 13 study visit (Moodie and others, 2018). A particular three-phase case-cohort sampling design was used for measuring S from month 13 samples that rendered the previously developed EML and PS methods highly inefficient for solving the problem, and also opened an opportunity for a new approach studied here that better takes advantage of the data structure. (This work does not consider a Bayesian approach, which could also be fruitful for this setting.) Specifically, baseline serum samples from a random sample of approximately 10% (20%) of all participants in the trial in Asia (Latin America) were collected, whereas month 13 serum samples were collected from all participants in each trial. With covariate S_b defined the same as S except measured at baseline, both S_b and S were measured in all participants randomly sampled at entry into the subcohort, and S was measured in all participants experiencing the VCD primary endpoint. Because S_b and S are the same variable measured at different times, they are highly correlated, making S_b an ideal “baseline immunogenicity predictor” (Follmann, 2006; Gilbert and Hudgens, 2008) [referred to as a “baseline surrogate measure (BSM)” in Gabriel and Gilbert (2014)]. Such a predictor is a key ingredient of all of the EML and PS methods to yield reasonably precise estimation of the mCEP curve. However, all of the EML and PS methods require that S_b be measured from all vaccine recipients with S measured, which means that the methods applied to the data would discard data from approximately 90% or 80% of VCD endpoint cases in the vaccine group for the two trials, respectively.

However, with $S(z)$ the potential outcome of S for the vaccine ($z = 1$) and placebo ($z = 0$) group, the new opportunity generated by the dengue Phase 3 trials is that S_b , baseline covariates, and $S(0)$ may plausibly contain all information for VCD risk in the placebo group without needing to also condition

on $S(1)$, because neutralizing antibody titers are the key known predictor of VCD in both vaccinated and unvaccinated individuals [e.g., analysis and references in [Katzelnick and others \(2017\)](#)]. With this novel assumption ((A4) below), this method departs from previous methods. While this assumption can be challenged, it beneficially removes the need to make the “placebo structural risk” modeling assumption common to all past EML and PS methods that links the conditional risk of VCD if assigned placebo to the candidate surrogate if assigned vaccine, a quantity not identifiable from the observed data unless close-out placebo vaccination is used ([Follmann, 2006](#)), which was not done in the dengue vaccine trials. The other novel assumption underlying this new approach takes advantage of the fact that the baseline immunogenicity predictor is a BSM, which allows making a time-constancy assumption about the distribution of $S(1)$ conditional on baseline covariates and S_b or $S(0)$ ((A5) below). By estimating the density of $S(1)$ conditional on baseline covariates and S_b with non-parametric kernel smoothing and not making the placebo structural risk modeling assumption, this new method gives more flexible estimation of the mCEP curve than the previous methods, with advantages in bias, efficiency, and confidence interval (CI) coverage as illustrated in Sections 4 and 5.

The remainder of this article is organized as follows. In Sections 1.1–1.4, we introduce notation, define the estimand of interest, state identifiability assumptions, establish identifiability based on these assumptions, and discuss plausibility of the assumptions and potential for their violations. We describe the estimation method, under a modeling assumption, including the construction of simultaneous CIs, in Section 2 and characterize procedures for testing of hypotheses of interest in Section 3. The design and findings from the simulation experiment are presented in Section 4. We apply the proposed estimation and inference procedures to data from the two dengue vaccine trials in Section 5.

1.1. Notation

We consider a randomized placebo-controlled trial with treatment assignment Z ($Z = 1$, treatment; $Z = 0$, placebo), and a discrete or continuous univariate biomarker S measured at fixed time τ after randomization. In many vaccine trials, placebo recipients have variable levels of S reflecting pre-existing immunity arising from past exposure (e.g., natural infection and/or prior vaccination) to the disease-causing pathogen (e.g., dengue virus, *Plasmodium falciparum*, or influenza virus). It is of interest to evaluate S as a modifier of the treatment effect on a binary clinical endpoint Y ($Y = 1$, disease; $Y = 0$ no disease) measured after τ . To this end, S needs to be measured prior to Y ; therefore, we restrict the analysis to trial participants who are observed to be endpoint-free at τ and denote this status as $Y^\tau = 0$. If $Y^\tau = 1$, S is undefined, and we set $S = *$.

We consider a three-phase outcome-dependent case-cohort sampling design as follows: Z , Y^τ , Y , and a baseline covariate vector $\mathbf{X} = (X_1, \dots, X_k)$ are measured in all randomized participants (phase 1). Next, at baseline, Bernoulli sampling of all randomized participants is used to determine a subcohort $\mathcal{S}$, and the biomarker S is measured at time τ in the subset of this subcohort with $Y^\tau = 0$ as well as in all or almost all cases (those with $Y = 1$) with $Y^\tau = 0$, regardless of their membership in $\mathcal{S}$ (phase 2) [this is a classic case-cohort sampling design ([Prentice, 1986](#))]. Finally, the biomarker at baseline S_b is measured only in the subcohort $\mathcal{S}$ with $Y^\tau = 0$ (phase 3); S_b is the BSM in the [Gabriel and Gilbert \(2014\)](#) design. Let ϵ and δ indicate, respectively, that S and S_b are measured in phases 2 and 3. The sampling design is graphically illustrated in [Supplementary Figure 1](#) available at *Biostatistics* online and may be expressed as δ sampled from the same Bernoulli distribution for all participants and then, conditional on (δ, Y^τ, Y) , ϵ sampled according to the sampling probability

$$P(\epsilon = 1 \mid \delta, Y^\tau, Y) = (1 - Y^\tau)\{\delta + \psi(1 - \delta)Y\}$$

for a fixed constant $\psi \in (0, 1]$.

Our causal estimand of interest is the mCEP curve that has been previously studied in several papers. To define the mCEP curve, let $Y^\tau(z)$, $\epsilon(z)$, $S(z)$, and $Y(z)$ be the potential outcomes of Y^τ , ϵ , S , and Y under treatment assignment z . If $Y^\tau(z) = 1$, $\epsilon(z) = 0$ and $S(z)$ is undefined, therefore we set $S(z) = *$.

1.2. Estimand of interest

Let $CE = h(P\{Y(1) = 1\}, P\{Y(0) = 1\})$ measure the overall causal effect of treatment on the clinical endpoint Y , where $h(\cdot, \cdot)$ is a known contrast function such that $h(x, y) = 0$ if and only if $x = y$. Denote $p_z(s_1) = P\{Y(z) = 1 | S(1) = s_1, Y^\tau(1) = Y^\tau(0) = 0\}$ for $z = 0, 1$. We define the causal estimand of interest as

$$\text{mCEP}(s_1) = h\{p_1(s_1), p_0(s_1)\}; \quad (1.1)$$

this estimand is termed the marginal CEP curve in the literature. If $S(1)$ is continuous (assumed henceforth), then (1.1) abuses notation for expositional simplicity, with the formal definition provided in the [supplementary material](#) available at *Biostatistics* online. The contrast $h(x, y) = 1 - x/y$ gives $\text{mCEP}(s_1)$ an interpretation as the percent reduction in clinical endpoint risk among treatment recipients with biomarker level $S(1) = s_1$ compared to if they had been assigned placebo, whereas the additive-difference contrast $h(x, y) = y - x$ gives an attributable risk interpretation for treatment recipients with biomarker level $S(1) = s_1$.

1.3. Identifiability assumptions

We suppose that $(Z_i, \mathbf{X}_i, \delta_i, \delta_i S_{b,i}, Y_i^\tau(0), Y_i^\tau(1), \epsilon_i(0), \epsilon_i(0)S_i(0), \epsilon_i(1), \epsilon_i(1)S_i(1), Y_i(0), Y_i(1))$, $i = 1, \dots, n$, are independent and identically distributed and assume no drop-out for simplicity. We consider commonly made identifiability assumptions ([Gilbert and Hudgens, 2008](#); [Huang and Gilbert, 2011](#); [Huang and others, 2013](#); [Huang, 2018](#)):

- (A1) Stable Unit Treatment Value Assumption (SUTVA): $(Y_i^\tau(0), Y_i^\tau(1), \epsilon_i(0), \epsilon_i(0)S_i(0), \epsilon_i(1), \epsilon_i(1)S_i(1), Y_i(0), Y_i(1))$ of the i -th subject is independent of $Z_j, j \neq i$.
- (A2) Ignorable treatment assignment: Z_i is conditionally independent of $(\delta_i, \delta_i S_{b,i}, Y_i^\tau(0), Y_i^\tau(1), \epsilon_i(0), \epsilon_i(0)S_i(0), \epsilon_i(1), \epsilon_i(1)S_i(1), Y_i(0), Y_i(1))$ given $\mathbf{X}_i$.
- (A3) Equal early clinical risk: $P\{Y_i^\tau(0) \neq Y_i^\tau(1)\} = 0$.

(A1) implies “consistency,” i.e., $(Y_i^\tau(Z_i), \epsilon_i(Z_i)S_i(Z_i), Y_i(Z_i)) = (Y_i^\tau, \epsilon_i S_i, Y_i)$. SUTVA may be violated in vaccine trials due to herd immunity and other factors, but may be approximately valid if trial participants do not interact with each other and the study sites are in large geographically dispersed regions. (A2) holds by randomization that may be stratified by $\mathbf{X}$. (A3) is more credible when τ is near baseline relative to the length of follow-up and it takes time for the effect of the intervention to occur. For simplicity of exposition, henceforth all conditional and unconditional probabilities of $Y(z) = 1$ and densities of $S(1)$ implicitly condition on $Y^\tau(1) = Y^\tau(0) = 0$.

We additionally consider the following identifiability assumptions:

- (A4) $P\{Y(0) = 1 | \mathbf{X}, S(0), S(1)\} = P\{Y(0) = 1 | \mathbf{X}, S(0)\}$, i.e., the risk of $Y(0)$ is conditionally independent of $S(1)$ given $(\mathbf{X}, S(0))$.
- (A5) Time constancy: $f(s_1 | S(0) = s_0, \mathbf{X} = \mathbf{x}) = \tilde{f}(s_1 | S_b = s_0, \mathbf{X} = \mathbf{x})$ for all $(s_1, s_0, \mathbf{x})$, where f and $\tilde{f}$ are conditional density functions of $S(1)$.

In a standard trial design for biomarker effect modification evaluation, $S(1)$ is unobserved in all placebo recipients, and thus the validity of (A4) cannot be tested in general. (A4) may be violated if the occurrence of the event $Y(0) = 1$ is correlated with $S(1)$ within subgroups defined by $(\mathbf{X}, S(0))$, stemming from unmeasured factors, and we discuss its plausibility in the analysis of the dengue vaccine trials in Section 5. (A5) may be plausible as both $S(0)$ and S_b measure pre-existing natural and/or vaccine-induced immunity, only $S(0)$ is measured τ time units later than S_b . The estimation section includes a technique that relaxes this assumption.

1.4. Establishing identifiability

Each of the two conditional risks $p_1(s_1)$ and $p_0(s_1)$ in $\text{mCEP}(s_1)$, defined in (1.1), is identified by the observed data and assumptions (A1)–(A5), as sketched here. Bayes' theorem and (A4) yield

$$\begin{aligned} p_0(s_1) &= \int \cdots \int P\{Y(0) = 1 \mid \mathbf{X} = \mathbf{x}, S(0) = s_0, S(1) = s_1\} d^{k+1}F_0(s_0, \mathbf{x} \mid s_1) \\ &= \int \cdots \int P\{Y(0) = 1 \mid \mathbf{X} = \mathbf{x}, S(0) = s_0\} \frac{f(s_1 \mid s_0, \mathbf{x})g(s_0 \mid \mathbf{x})h(\mathbf{x})}{m(s_1)} ds_0 d\mathbf{x}_1 \cdots d\mathbf{x}_k, \end{aligned} \quad (1.2)$$

where $m(s_1) = \int \cdots \int f(s_1 \mid s_0, \mathbf{x})g(s_0 \mid \mathbf{x})h(\mathbf{x}) ds_0 d\mathbf{x}_1 \cdots d\mathbf{x}_k$, $F_0(s_0, \mathbf{x} \mid s_1)$ is a conditional joint cumulative distribution function of $(S(0), \mathbf{X})$ given $S(1)$, $f(s_1 \mid s_0, \mathbf{x})$ a conditional density of $S(1)$ given $(S(0), \mathbf{X})$, $g(s_0 \mid \mathbf{x})$ a conditional density of $S(0)$ given $\mathbf{X}$, $h(\mathbf{x})$ a marginal density/probability function of $\mathbf{X}$, and $m(s_1)$ a marginal density of $S(1)$. The decomposition in (1.2) is advantageous because it enables us to identify (and estimate) $p_0(s_1)$ by separately identifying each component in (1.2).

First, $g(s_0 \mid \mathbf{x})$ is identified from placebo recipients with $\epsilon = 1$ and the sampling design, and $h(\mathbf{x})$, which involves phase 1 covariates only, is identified from all randomized participants. Second, for identifying $f(s_1 \mid s_0, \mathbf{x})$, note that, under the sampling design,

$$\begin{aligned} \tilde{f}(s_1 \mid S_b = s_0, \mathbf{X} = \mathbf{x}, \delta = 1, \epsilon = 1) &= \tilde{f}(s_1 \mid S_b = s_0, \mathbf{X} = \mathbf{x}, \delta = 1) \\ &= \tilde{f}(s_1 \mid S_b = s_0, \mathbf{X} = \mathbf{x}) \\ &= f(s_1 \mid S(0) = s_0, \mathbf{X} = \mathbf{x}), \end{aligned}$$

where the first equality holds because $\delta = 1$ implies $\epsilon = 1$, the second equality holds because participants with $\delta = 1$ are a random sample of all randomized participants, and the third equality holds by (A5). Because $\tilde{f}(s_1 \mid S_b = s_0, \mathbf{X} = \mathbf{x}, \delta = 1, \epsilon = 1)$ is identified from the observed data, the needed term $f(s_1 \mid S_0 = s_0, \mathbf{X} = \mathbf{x})$ is identified. Lastly, $P\{Y(0) = 1 \mid \mathbf{X} = \mathbf{x}, S(0) = s_0\}$ in (1.2) is identified from placebo recipients with $\epsilon = 1$ and the sampling design. Similar to this last term, the conditional risk $p_1(s_1)$ in (1.1) is identified from treatment recipients with $\epsilon = 1$ and the sampling design; this conditional risk is straightforward to identify because both $Y(1)$ and $S(1)$ are observable from the same treatment recipient.

2. ESTIMATION METHOD

2.1. Modeling assumption

We develop an estimation and inference method under the following modeling assumption:

(A6) The risk of $Y(z)$ conditional on $S(z)$ and $\mathbf{X}$ follows a generalized linear model (GLM) for $z = 0, 1$.

(A6) for $z = 1$ has been made in previous papers and constitutes a standard regression problem without identifiability problems. (A6) for $z = 0$ is novel, replacing the untestable “placebo structural risk” assumption from previous papers that $Y(0)$ conditional on $S(1)$ and $\mathbf{X}$ follows a GLM. The GLMs specified in (A6) are estimated in participants with $\epsilon = 1$ using methods for case-cohort designs.

2.2. Estimation of the causal estimand

We propose a plug-in estimator for $\text{mCEP}(s_1)$ by separately estimating $p_1(s_1)$ and $p_0(s_1)$. We estimate $p_1(s_1)$ by fitting the GLM specified in (A6) accounting for the case-cohort sampling of S . To estimate $p_0(s_1)$, leveraging the identifiability results, we estimate $f(s_1 | S_0 = s_0, \mathbf{X} = \mathbf{x})$ by an estimate of $\tilde{f}(s_1 | S_b = s_0, \mathbf{X} = \mathbf{x}, \delta = 1, \epsilon = 1)$, which we obtain via non-parametric kernel smoothing. Because participants with $\delta = 1$ constitute a random sample from all randomized participants, $\tilde{f}(s_1 | S_b = s_0, \mathbf{X} = \mathbf{x})$ is estimated by an estimate of $\tilde{f}(s_1 | S_b = s_0, \mathbf{X} = \mathbf{x}, \delta = 1, \epsilon = 1)$ generated using data from treatment recipients with $\delta = 1$.

The above approach to estimation of $\tilde{f}(s_1 | S_b = s_0, \mathbf{X} = \mathbf{x})$ assumes (A5), whose veracity may be supported by regression modeling that indicates an association of S_b and $S(0)$ close to identity among placebo recipients with measured $(S_b, S(0))$. Otherwise, (A5) could be violated, and an estimated regression model, e.g., $E\{S_b | S(0) = s_0\} = \hat{a} + \hat{b}s_0$, could be employed for calibration to estimate $\tilde{f}(s_1 | S_b = \hat{a} + \hat{b}s_0, \mathbf{X} = \mathbf{x}, \delta = 1, \epsilon = 1)$.

We estimate $g(s_0 | \mathbf{X} = \mathbf{x})$ by estimating $g(s_0 | \mathbf{X} = \mathbf{x}, \epsilon = 1)$ via non-parametric kernel smoothing after random proportional deletion of a subset of cases among participants with measured $S(0)$ to attain the same case:control ratio as in the target population that is represented by randomized participants observed to be endpoint-free at τ (i.e., $Y^\tau = 0$). Another, more powerful, approach would use inverse probability weighting; however, we conjecture that minimal efficiency gain would be achieved in a rare endpoint setting common in vaccine trials. The multivariate density/probability function $h(\mathbf{x})$ is estimated using phase 1 covariate data from all randomized participants. In Section 4, we also consider an alternative parametric estimator for density functions in (1.2) based on maximum likelihood estimation in the Gaussian family of distributions.

Pointwise and simultaneous Wald-type CIs for $\text{mCEP}(s_1)$ are obtained by assessing $\text{mCEP}(s_1)$ in B bootstrap samples, with cases and controls sampled separately yielding a fixed number of cases and controls in each bootstrap sample. We construct the CIs in two steps. First, CIs for the η -transformed estimand (defined below) are constructed, and then the inverse transformation is applied to the confidence bounds.

2.3. Simultaneous confidence interval for the mCEP curve

Let $\eta(s_1) := \eta\{\text{mCEP}(s_1)\}$ be a “symmetrizing” transformation of the mCEP curve estimand that helps make Wald-type CIs more accurate in finite samples; e.g., if $h(x, y) = 1 - x/y$, we consider $\eta\{h(x, y)\} = \log\{1 - h(x, y)\}$, or, if $h(x, y) = y - x$, we may consider the identity transformation or $\eta\{h(x, y)\} = \text{logit}\{[h(x, y) + 1]/2\}$. For an arbitrary subset $\mathbb{S}$ of the support of $S(1)$, denote

$$U := \sup_{s_1 \in \mathbb{S}} |\hat{\eta}(s_1) - \eta(s_1)| / \text{SE}\{\hat{\eta}(s_1)\},$$

where $\hat{\eta}(s_1) = \eta\{\widehat{\text{mCEP}}(s_1)\}$ and $\text{SE}(\cdot)$ denotes a standard error. For a fixed $\alpha \in (0, 1)$, we define c_α as the solution to the equation $P(U \leq c_\alpha) = 1 - \alpha$.

Further, let $\hat{\eta}^{(b)}(s_1)$ be the estimate of $\eta(s_1)$ based on the b -th bootstrap sample, and $\text{SE}\{\hat{\eta}(s_1)\}$ be the sample standard deviation of the bootstrap estimates $\hat{\eta}^{(b)}(s_1)$, $b = 1, \dots, B$. Let $U^{(b)} := \sup_{s_1 \in \mathbb{S}} |\hat{\eta}^{(b)}(s_1) -$

$\widehat{\eta}(s_1)|/\text{SE}\{\widehat{\eta}(s_1)\}$. Because the distributions of U and $U^{(b)}$ are asymptotically equivalent, we estimate c_α by c_α^* defined as the empirical quantile in the bootstrap sample $U^{(b)}$, $b = 1, \dots, B$, at the probability level $1 - \alpha$. Finally, the simultaneous Wald-type bootstrap $(1 - \alpha) \times 100\%$ CI for the $\{\text{mCEP}(s_1), s_1 \in \mathbb{S}\}$ curve is obtained by the $\eta^{-1}(\cdot)$ transformation of the bounds

$$(l_\alpha^n(s_1), u_\alpha^n(s_1)) = \widehat{\eta}(s_1) \mp c_\alpha^* \text{SE}^*\{\widehat{\eta}(s_1)\}.$$

It is of note that the CI width depends on the size of $\mathbb{S}$, which may be chosen based on statistical and biological considerations.

3. HYPOTHESIS TESTING

It is of interest to evaluate, separately, the following four null hypotheses, each against a general alternative hypothesis:

- $H_0^1 : \text{mCEP}(s_1) \equiv \text{CE}$ for all $s_1 \in \mathbb{S}$,
- $H_0^2 : \text{mCEP}(s_1) \equiv c$ for all $s_1 \in \mathbb{S}_1 \subseteq \mathbb{S}$ and a known constant $c \in \mathbb{R}$,
- $H_0^3 : \text{mCEP}_1(s_1) = \text{mCEP}_2(s_1)$ for all $s_1 \in \mathbb{S}_1 \subseteq \mathbb{S}$, where mCEP_1 and mCEP_2 are each associated with either a different biomarker (measured in the same units) or a different endpoint or both, and
- $H_0^4 : \text{mCEP}(s_1|X = 1) = \text{mCEP}(s_1|X = 0)$ for all $s_1 \in \mathbb{S}_1 \subseteq \mathbb{S}$, where X is a baseline dichotomous phase 1 covariate of interest included in $\mathbf{X}$.

A test of H_0^1 assesses $S(1)$ as a modifier of the clinical treatment effect. It is commonly of interest to test H_0^2 with c representing the absence of a treatment effect, to assess if there exists a subgroup of treatment recipients defined by biomarker levels in $\mathbb{S}_1$ with some treatment effect. A test of H_0^3 allows comparisons of two biomarkers and/or two endpoints, while a test of H_0^4 allows baseline subgroup or between-trial comparisons.

We follow Roy and Bose (1953) to construct tests of hypotheses H_0^i , $i = 1, \dots, 4$. Let $U_\eta^{(b)}(\mathbb{S}, a) := \sup_{s_1 \in \mathbb{S}} |\widehat{\eta}^{(b)}(s_1) - \eta(a)|/\text{SE}^*\{\widehat{\eta}(s_1)\}$, where $a \in \mathbb{R}$. For testing H_0^1 and H_0^2 at significance level α , we use as the regions of rejection

$$U_1 := \sup_{s_1 \in \mathbb{S}} |\widehat{\eta}(s_1) - \eta(\widehat{\text{CE}})|/\text{SE}^*\{\widehat{\eta}(s_1)\} > c_{1\alpha}^*$$

and

$$U_2 := \sup_{s_1 \in \mathbb{S}_1} |\widehat{\eta}(s_1) - \eta(c)|/\text{SE}^*\{\widehat{\eta}(s_1)\} > c_{2\alpha}^*,$$

respectively, where $\widehat{\text{CE}}$ is an estimator for CE, and $c_{1\alpha}^*$ and $c_{2\alpha}^*$ are empirical quantiles in bootstrap samples $U_\eta^{(b)}(\mathbb{S}, \widehat{\text{CE}})$ and $U_\eta^{(b)}(\mathbb{S}_1, c)$, $b = 1, \dots, B$, respectively, at the probability level $1 - \alpha$. We obtain the two-sided p-values as the empirical probabilities that $U_\eta^{(b)}(\mathbb{S}, \widehat{\text{CE}}) > U_1$ and $U_\eta^{(b)}(\mathbb{S}_1, c) > U_2$.

For testing H_0^3 , let $\theta(s_1) := \eta\{\text{mCEP}_1(s_1)\} - \eta\{\text{mCEP}_2(s_1)\}$ be a contrast pertaining to the transformation η . Let $\text{SE}^*\{\widehat{\theta}(s_1)\}$ denote the sample standard deviation of the bootstrap estimates $\widehat{\theta}^{(b)}(s_1)$, $b = 1, \dots, B$, and let $U_\theta^{(b)} := \sup_{s_1 \in \mathbb{S}_1} |\widehat{\theta}^{(b)}(s_1)|/\text{SE}^*\{\widehat{\theta}(s_1)\}$. We define $c_{3\alpha}^*$ as the empirical quantile in the

bootstrap sample $U_\theta^{(b)}$, $b = 1, \dots, B$, at the probability level $1 - \alpha$. Subsequently, we test H_0^3 at significance level α by using as the region of rejection

$$U_3 := \sup_{s_1 \in \mathbb{S}_1} |\widehat{\theta}(s_1)| / \text{SE}^* \{\widehat{\theta}(s_1)\} > c_{3\alpha}^*,$$

and we obtain the two-sided p-value as the empirical probability that $U_\theta^{(b)} > U_3$. For testing H_0^4 , we proceed as for the test of H_0^3 except, due to independence of the baseline subgroups defined by X , we obtain $\text{SE}^* \{\widehat{\theta}(s_1)\}$ as $\left\{ \text{SE}^{*2} [\eta \{\widehat{\text{mCEP}}(s_1 | X = 1)\}] + \text{SE}^{*2} [\eta \{\widehat{\text{mCEP}}(s_1 | X = 0)\}] \right\}^{1/2}$.

4. SIMULATION STUDY

Consider the treatment efficacy estimand $\text{TE}(s_1) := \text{mCEP}(s_1)$ defined by the contrast function $h(x, y) = 1 - x/y$. The simulation study aims to evaluate and compare finite-sample performance of the proposed estimator for $\text{TE}(s_1)$ with an alternative pseudo-score estimator (PSN) of Huang (2018) and examine size/power of the proposed tests of H_0^1 , H_0^2 , and H_0^4 . Note that the test of H_0^3 is identical to that of H_0^4 except that it additionally accounts for correlation of the contrasted estimators for $\eta \{\text{mCEP}(s_1)\}$. Two approaches to probability density estimation in (1.2)—non-parametric kernel smoothing and Gaussian maximum likelihood estimation—are considered throughout the simulation, resulting in two variants of the proposed estimator for $\text{TE}(s_1)$ (denoted NP-TE and MLE-TE, respectively). More specifically, the NP-TE estimator employs the generalized product kernel density estimation method of Hall and others (2004) with optimal bandwidths selected by likelihood cross-validation.

The simulation design mimics characteristics of the dengue vaccine trials introduced in Section 1. We generate $(S_b, S(0), S(1))$ as follows. First, we generate i.i.d. vectors from $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (2, 2, 3)^T$, $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^3$ with $\sigma_{ii} = 1$ for all i , and the correlation matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}$ is chosen to emulate relationships between biomarker measurements at baseline and month 13 observed in the dengue trials:

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{pmatrix}.$$

Then each component of the generated triplets is left-censored at the value of 1.5, which represents, e.g., the biomarker assay's lower limit of quantitation.

Furthermore, using conditional independence in (A4), we posit a probit model for the association of the biomarkers $(S(0), S(1))$ with the endpoint indicator in each treatment group:

$$P\{Y(z) = 1 | S(0) = s_0, S(1) = s_1\} = \Phi\{\beta_0 + \beta_1 z + \beta_2(1 - z)s_0 + \beta_3 z s_1\}, \quad z = 0, 1, \quad (4.1)$$

where Φ is the cumulative distribution function of $\mathbf{N}(0, 1)$. Model (4.1) yields

$$p_0(s_1) = \int_{-\infty}^{+\infty} \Phi(\beta_0 + \beta_2 s_0) v(s_0 | s_1) ds_0, \quad (4.2)$$

where v is the conditional probability density of $S(0)$ given $S(1) = s_1$. For $z = 0, 1$, denote the marginal mean risk $\lambda_z = E[p_z\{S(1)\}]$ and the risk “gradient” $\delta_z = p_z(s_{1,0.9}) / p_z(s_{1,0.1})$, where $s_{1,q}$ is the quantile of the marginal distribution of $S(1)$ at probability q . The values of the probit model coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ are chosen such that $\lambda_0 = 0.1$, $\lambda_1 = 0.05$, $\delta_0 = 0.96$, and $\delta_1 = 0.30$. We consider $\delta_0 < 1$ to reflect the

Fig. 1. The true $TE(s_1)$ curves in the simulation design, each satisfying $\lambda_0 = 0.1$ and $\lambda_1 = 0.05$. In addition, the solid curve reflects $\delta_0 = 0.96$ and $\delta_1 = 0.30$. The dot-dashed curve, used only to assess power of the test of H_0^4 , reflects the same δ_0 but $\delta_1 = 0.52$. The dashed line, used only to assess size of the tests of H_0^1 and H_0^2 , reflects $\delta_0 = \delta_1 = 1$.

assumed positive correlation between $S(0)$ and $S(1)$. Using (4.2), the resultant $TE(s_1) = 1 - p_1(s_1)/p_0(s_1)$ for $s_1 \in [1.5, s_{1,0.95}]$, representing the truth, is shown in Figure 1 (solid curve).

To estimate $TE(s_1)$, the proposed NP-TE and MLE-TE estimators assume the logit link function in model (4.1). The PSN estimator of Huang (2018) for $TE(s_1)$ models the endpoint risk as a function of Z and $S(1)$ and utilizes a discrete baseline covariate for predicting a missing $S(1)$ value. To implement the PSN estimator, we discretize S_b by quartiles to arrive at S_b^* , which is used as the auxiliary variable for predicting $S(1)$. We model $P\{Y(z) = 1|S(1) = s_1\} = \Phi\{\gamma_0 + \gamma_1 z + \gamma_2 s_1 + \gamma_3 s_1 z\}$ using the PSN approach and construct the corresponding pointwise Wald-type CI for $TE(s_1)$ using the analytical variance estimator developed by Huang (2018). We examine and compare finite-sample relative bias and mean squared error (MSE) of the proposed and the PSN estimators for $TE(s_1)$ and coverage probabilities of pointwise and simultaneous 95% Wald-type bootstrap CIs for $TE(s_1)$.

Modifications of the described mechanism are needed for generating data under H_0^1 , H_0^2 , H_0^4 , and under respective alternative hypotheses. To evaluate size of the tests of H_0^1 and H_0^2 , we set $\beta_2 = \beta_3 = 0$ in (4.1) and recompute β_0 and β_1 in order to maintain the marginal probabilities $\lambda_0 = 0.1$ and $\lambda_1 = 0.05$. The respective constant $TE(s_1)$ curve is shown in Figure 1 (dashed line). We evaluate power of the tests of H_0^1 and H_0^2 for the setting described above. This setting also serves to evaluate size of the test of H_0^4 by drawing two independent samples, each used for estimating $TE(s_1)$ separately, representing distinct populations. To evaluate power of the test of H_0^4 , the first sample considers the setting specified above, while the second sample considers the same setting except $\delta_1 = 0.52$. The resultant $TE(s_1)$ curve for this modified setting is also shown in Figure 1 (dot-dashed curve).

We consider a three-phase case-cohort sampling design. In phase 1, 5000 subjects are randomized at a 1:1 ratio to receive treatment ($Z = 1$) or placebo ($Z = 0$) and followed for the binary outcome Y . Although not required, for simplicity, we assume that all endpoints occur after time t_0 at which $S(Z)$ is measured (i.e., $V = 1$ for all subjects in the absence of drop-out). In phase 2, we measure $S(Z)$ in a Bernoulli sample S , drawn at baseline with sampling probability π , and in all cases ($Y = 1$) whether or not they were in this subcohort. In phase 3, we measure S_b in subcohort S only, i.e., S_b is missing in cases not included

Fig. 2. Top row: Estimated relative bias of the proposed NP-TE and MLE-TE estimators for $TE(s_1)$, compared with that of the PSN estimator of Huang (2018), as a function of the probability π of sampling into the phase 2/3 subcohort $\mathcal{S}$. Bottom row: Estimated MSE of the proposed NP-TE and MLE-TE estimators for $TE(s_1)$, compared with that of the PSN estimator of Huang (2018), as a function of probability π of sampling into the phase 2/3 subcohort $\mathcal{S}$.

in $\mathcal{S}$ and the proportion of such cases varies with π . We evaluate the performance of the estimation and inferential procedures for $TE(s_1)$ as a function of π , setting $\pi = 0.1, 0.25$, and 0.5 (note that π was set to 0.1 and 0.2 in the two dengue trials). The results are based on 10^3 replicated data sets with 500 bootstrap samples drawn in each data set.

For all values of π , the NP-TE and MLE-TE estimators exhibit minimal bias for $S(1) \geq 2.0$, with an increase in bias in the left-censored tail, whereas the PSN estimator is heavily biased in both tails for $\pi = 0.1$, and its bias becomes comparable to that of the NP-TE and MLE-TE estimators as π increases to 0.5 (Figure 2, top row). Additionally, for $\pi = 0.1$, the NP-TE and MLE-TE estimators substantially reduce the MSE in both tails compared with the PSN estimator, with comparable MSE across all three estimators achieved for $\pi = 0.5$ (Figure 2, bottom row). Coverage probabilities of pointwise 95% Wald-type bootstrap CIs induced by the NP-TE and MLE-TE estimators are within the Monte-Carlo (MC) error band except with slight undercoverage in the central region by the MLE-TE estimator using $\pi = 0.1$

Table 1. *Estimated coverage probabilities of the simultaneous 95% Wald-type bootstrap CI for $\{\text{TE}(s_1), s_1 \in \mathbb{S}\}$ based on the NP-TE and MLE-TE estimators, where $\mathbb{S}$ spans the observed values of $S(z)$, as a function of the probability π of sampling into the phase 2/3 subcohort $\mathcal{S}$*

π	NP-TE	MLE-TE
0.1	0.959	0.943
0.25	0.956	0.944
0.5	0.959	0.954

Table 2. *Size and power of the two-sided tests, based on the NP-TE and MLE-TE estimators, of H_0^1 , H_0^2 with $c = 0.5$, and H_0^4 against general alternative hypotheses, as a function of the probability π of sampling into the phase 2/3 subcohort $\mathcal{S}$. The nominal significance level is taken to be 0.05*

π	Test of H_0^1		Test of H_0^2		Test of H_0^4	
	Size	Power	Size	Power	Size	Power
NP-TE						
0.1	0.01	0.73	0.04	0.83	0.04	0.12
0.25	0.01	0.84	0.05	0.89	0.04	0.15
0.5	0.01	0.89	0.05	0.93	0.04	0.18
MLE-TE						
0.1	0.01	0.87	0.06	0.92	0.05	0.17
0.25	0.01	0.91	0.05	0.95	0.05	0.20
0.5	0.01	0.92	0.06	0.96	0.05	0.20

(see Section 3 in the [supplementary material](#) available at *Biostatistics* online). In contrast, pointwise 95% Wald-type CIs induced by the PSN estimator uniformly overcover for $\pi = 0.1$, with coverage probabilities within the MC error band attained for $\pi = 0.25$ and 0.5. The simultaneous 95% Wald-type bootstrap CIs for $\{\text{TE}(s_1), s_1 \in \mathbb{S}\}$ induced by both the NP-TE and MLE-TE estimators exhibit adequate coverage for all values of π (Table 1). Overall, the NP-TE and MLE-TE estimators perform comparably, with considerable precision gain and improved CI coverage over the PSN estimator for small values of π .

Table 2 summarizes observed sizes and powers of the tests of H_0^1 , H_0^2 , and H_0^4 described in Section 3, based on both the NP-TE and MLE-TE estimators. For all values of π , sizes of the tests of H_0^2 and H_0^4 are in good agreement with the nominal significance level, whereas the test of H_0^1 is markedly conservative. Nevertheless, for each π , the power of the test of H_0^1 is only slightly smaller than that of the test of H_0^2 suggesting that the former test is also useful in practice. Powers of the test of H_0^4 are small for the given comparison (see Figure 1) indicating that larger contrasts are needed to be detected with sufficient power. Finally, the tests based on the MLE-TE estimator are slightly more powerful than those based on the NP-TE estimator.

5. APPLICATION

The two harmonized Phase 3 dengue trials, introduced in Section 1, randomized 31 144 healthy children (aged 2–14 and 9–16 years in Asia and Latin America, respectively) at a 2:1 ratio to receive either three doses of Sanofi Pasteur’s recombinant, live, attenuated, tetravalent dengue vaccine (Dengvaxia/CYD-TDV), or placebo at months 0, 6, and 12 after randomization (Capeding and others, 2014; Villar and others, 2015). Participants were followed with active surveillance for 25 months for the primary clinical endpoint of VCD. The current indication of CYD-TDV requires the minimum age of 9 years, and therefore we choose to present herein a trial-pooled analysis in the 25 824 children aged ≥ 9 years.

We aim to study modification of the effect of CYD-TDV vs. placebo on VCD risk by a biomarker $S(1)$ under assignment to CYD-TDV that is measured by the PRNT₅₀ assay from a serum sample collected at the month 13 visit; S is the average of $\log_{10}$ neutralizing antibody titers to the four parental dengue strains of the vaccine constructs, one for each dengue serotype (Moodie and others, 2018). The following three-phase case-cohort sampling design was employed: the phase 1 covariate vector $\mathbf{X}$ —age category (≤ 11 vs. > 11 years) and country—was observed in all participants. Baseline serum samples for measuring S were collected from a Bernoulli sample $\mathcal{S}$ of all participants, whereas month 13 serum samples were collected from all participants. Subsequently, S was measured at month 13 in subcohort $\mathcal{S}$ and in all VCD cases (phase 2), whereas the biomarker’s baseline value, S_b , was only measured in subcohort $\mathcal{S}$ (phase 3). Because S is measured at month 13, we restrict the analysis to the 24 768 children aged ≥ 9 years who were at risk and free of VCD at month 13; this cohort includes 503 VCD cases. Of the 2766 controls with S measured, all but 7 also had S_b measured. In contrast, of the 502 cases with S measured, only 55 (11.0%) also had S_b measured.

We consider two $\text{mCEP}(s_1)$ estimands of interest defined by contrasts $h_1(x, y) = \log(x/y)$ and $h_2(x, y) = y - x$ in (1.1), with identity and $\text{logit}\{[h_2(x, y) + 1]/2\}$ as their respective $\eta(\cdot)$ transformations (for the latter estimand, one could alternatively also use the identity transformation; simulations showed similar coverage of CIs under the two transformations). We employ the proposed estimator for $\text{mCEP}(s_1)$, with probability densities estimated by the generalized kernel method of Hall and others (2004). We model the conditional VCD risk in (A6) by specifying inverse probability-weighted logistic regression models. In addition, a hinge model (Fong and others, 2017) is used for modeling the effect of $S(z)$ because a likelihood ratio test supports a better fit ($p < 0.05$) and the number of VCD cases is sufficient to support this more flexible model. Moreover, a hinge model is used because it specifies that variability in $S(z)$ near the lower limit of quantitation (LLOQ) does not associate with VCD risk, which is a desirable (biologically credible) model feature given that such values largely reflect PRNT₅₀ technical measurement error.

For the PSN estimator of Huang (2018), we model $P\{Y(z) = 1 | \mathbf{X}, S(1)\}$ using a probit model adjusted for the main effects of $\mathbf{X}$, and the main effects and interaction of Z and $S(1)$ with a hinge. The marginal risk conditional on $S(1)$ is then estimated by integrating $P\{Y(z) = 1 | \mathbf{X}, S(1)\}$ over the distribution of $\mathbf{X} | S(1)$, which is estimated through a multinomial model of $\mathbf{X}$ on natural cubic spline basis of $S(1)$ with nodes at its quantiles at probabilities 0.25, 0.5, and 0.75.

In the cohort of participants at risk at month 13, logistic regression adjusted for treatment assignment as the sole covariate yields an overall estimated VCD risk of 0.013 and 0.035 in vaccine and placebo recipients, respectively, resulting in an overall log relative risk estimate of -0.998 and an overall additive risk difference estimate of 0.022. We report and compare inference about the $\text{mCEP}(s_1)$ estimands defined by contrasts h_1 and h_2 using the proposed estimator (employing non-parametric kernel smoothing) and, where it is applicable, the PSN estimator. Point estimates and pointwise and simultaneous 95% Wald-type bootstrap CIs for both $\text{mCEP}(s_1)$ estimands are shown in Figure 3. The proposed estimator yields substantially narrower CIs than the PSN estimator in this setting. The test of $H_0^1 : \text{mCEP}(s_1) \equiv \text{CE}$ for all $s_1 \in [\log_{10}(5), \log_{10}(22127)]$ proposed in Section 3 yields p-values < 0.001 and 0.16 for h_1 and h_2 , respectively, indicating that $S(1)$ is a significant modifier of relative VCD risk but not of VCD risk

Fig. 3. Point and 95% Wald-type bootstrap CI estimates of the log relative risk (vaccine/placebo) of VCD and the additive VCD risk difference (placebo–vaccine) in the trial-pooled analysis of 9–16 year olds using the proposed estimator with non-parametric kernel smoothing vs. the PSN estimator of Huang (2018).

difference. The test of $H_0^2 : \text{mCEP}(s_1) \equiv 0$ for all $s_1 \in [\log_{10}(5), \log_{10}(57)]$, where 57 is the estimated hinge-point titer, yields p-values < 0.001 for both h_1 and h_2 , respectively, indicating a significant vaccine effect on both the relative risk and the risk difference in a subgroup of vaccine recipients with $S(1)$ near the assay's LLOQ. This supports that the biomarker $S(1)$ does not satisfy the average causal necessity property (Gilbert and Hudgens, 2008) which if true implies that this biomarker does not fully mediate the CYD-TDV vaccine effect on VCD risk (VanderWeele, 2008).

Because (A3) is evidently violated in the dengue trials, we developed a sensitivity analysis method that relaxes (A3) to the weaker monotonicity assumption of “no early harm by treatment” (A3’), i.e., $P\{Y^\tau(0) = 0, Y^\tau(1) = 1\} = 0$, which is plausible for the dengue trials. The [supplementary material](#) available at *Biostatistics* online describes the method and applied it to the data, which indicates that violations of (A3) leads to underestimation of the $\text{mCEP}(s_1)$ curve for both the log relative risk and risk difference contrast. (A4) states that the neutralizing antibody response to vaccination does not affect placebo-group VCD risk after accounting for the existing natural neutralizing antibody response (due to prior dengue infections) and for measured baseline covariates. It could be violated if the ability to mount a strong neutralizing

antibody response to vaccination is correlated with an intrinsic VCD susceptibility factor that is not fully captured by the other variables. The fact that the natural response $S(0)$ and the vaccine response $S(1)$ measure a similar quantity using the identical laboratory assay may limit the degree of violation of (A4). Furthermore, in placebo recipients with measured $(S_b, S(0))$, robust linear regression (Yohai, 1987) and locally weighted polynomial regression indicate an association of S_b and $S(0)$ close to identity, supporting validity of (A5) (see the [supplementary material](#) available at *Biostatistics* online for details).

6. DISCUSSION

Driven by data constraints of a BSM three-phase sampling design, this article develops semi-parametric inferential procedures for a principal stratification estimand, the mCEP curve, designed to assess modification of a clinical treatment effect by the potential outcome marker response under assignment to treatment. We demonstrate that current alternative approaches may be biased and highly inefficient in this setting. We show that, under a different set of assumptions, we can remove the untestable placebo structural risk modeling assumption common to all past EML and PS methods and employ non-parametric kernel smoothing to estimate the placebo group's clinical endpoint risk conditional on the potential outcome $S(1)$ under treatment assignment. We use a bootstrap procedure to construct pointwise and simultaneous CIs and tests of relevant hypotheses about $\text{mCEP}(s_1)$ estimands.

The proposed estimation method particularly offers an alternative to the PS estimation method of Huang and others (2013) and Huang (2018), which estimate the same $\text{mCEP}(s_1)$ estimand. The following deliberations shall provide guidance for method selection. First, the proposed method makes three assumptions not made by the PS method: (i) $Y(0)$ is conditionally independent of $S(1)$ given $S(0)$ and $\mathbf{X}$ ((A4) above); (ii) the distribution of $S(1)$ conditional on $\mathbf{X}$ and $S(0)$ or S_b remains identical ((A5) above); and (iii) under treatment assignment $Z = 0$, the clinical endpoint risk of $Y(0)$ conditional on the biomarker response $S(0)$ and baseline covariates $\mathbf{X}$ follows a GLM, which replaces the placebo structural risk assumption of the PS method that $Y(0)$ conditional on $S(1)$ and $\mathbf{X}$ follows a GLM (a part of (A6) above). Since (A4) is not testable without close-out placebo vaccination, if data support validity of (A5) and (A6) in a standard trial with a BSM three-phase sampling design, the present method may be preferable for its potentially substantial improvement in bias, efficiency, and CI coverage. Because (A4) is untestable in a standard trial and could easily be violated, future research is warranted for sensitivity and for relaxing the assumption. Second, the present method enables use of non-parametric kernel smoothing, which more flexibly estimates the $\text{mCEP}(s_1)$ curve, albeit involves the issue of bandwidth selection. For this reason, maximum likelihood estimation, as an alternative to kernel smoothing, may be employed with the present method as demonstrated in Section 4. Third, to the best of our knowledge, the present method is the first method providing formal tests of the null hypotheses that the $\text{mCEP}(s_1)$ curve equals to a constant and that the $\text{mCEP}(s_1)$ curve is identical in two baseline covariate subgroups.

Recognizing the importance of assumption (A3) and a possibly frequent occurrence of its violation, we developed a method of sensitivity analysis that relaxes (A3) to the weaker monotonicity assumption of “no early harm by treatment” (A3’), i.e., $P\{Y^\tau(0) = 0, Y^\tau(1) = 1\} = 0$, which, in practice, is far more easily justified than (A3) (the [supplementary material](#) available at *Biostatistics* online describes the method).

In using non-parametric smoothing, the newly proposed method is not well suited to studying multivariate response biomarkers as effect modifiers. However, for settings where many or even high-dimensional response biomarkers are measured, the approach may still be useful by first defining a univariate score biomarker as the fitted values of a model selected by statistical/machine learning to provide the best prediction of the clinical outcome (Price and others, 2018) and then to study the score biomarker as an effect modifier.

7. SOFTWARE

All proposed methods are implemented in the R package `pssmooth` (Juraska, 2018) available on the Comprehensive R Archive Network. The R code used for computing and plotting results reported in Sections 4 and 5 is available at <https://github.com/mjuraska/mCEPcurve-three-phase>.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

We thank the participants and investigators of the CYD14 and CYD15 dengue vaccine efficacy trials, our Sanofi Pasteur colleagues for collaboration and sharing of the data, Ted Holzman (Fred Hutchinson Cancer Research Center) for assistance with computational analysis, and Lindsay N. Carpp (Fred Hutchinson Cancer Research Center) for technical editing. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: None declared.

FUNDING

Sanofi Pasteur and National Institute of Allergy and Infectious Diseases of the National Institutes of Health (Award Number R37AI054165).

REFERENCES

- CAPEDING, M. R., TRAN, N. H., HADINEGORO, S. R. S., ISMAIL, H. I. H. M., CHOTPITAYASUNONDH, T., CHUA, M. N., LUONG, C. Q., RUSMIL, K., WIRAWAN, D. N., NALLUSAMY, R. *and others.* (2014). 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**, 1358–1365.
- FOLLMANN, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics* **62**, 1161–1169.
- FONG, Y., HUANG, Y., GILBERT, P. B. AND PERMAR, S. R. (2017). chngpt: threshold regression model estimation and inference. *BMC Bioinformatics* **18**(1): 454.
- FRANGAKIS, C. AND RUBIN, D. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.
- GABRIEL, E. AND GILBERT, P. (2014). Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics* **15**, 251–265.
- GABRIEL, E. E., SACHS, M. C. AND GILBERT, P. B. (2015). Comparing and combining biomarkers as principal surrogates for time-to-event clinical endpoints. *Statistics in Medicine* **34**, 381–395.
- GILBERT, P. B. AND HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146–1154.
- HALL, P., RACINE, J. AND LI, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99**, 1015–1026.
- HUANG, Y. (2018). Evaluating principal surrogate markers in vaccine trials in the presence of multiphase sampling. *Biometrics* **74**, 27–39.
- HUANG, Y. AND GILBERT, P. B. (2011). Comparing biomarkers as principal surrogate endpoints. *Biometrics* **67**, 1442–1451.

- HUANG, Y., GILBERT, P. B. AND WOLFSON, J. (2013). Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics* **69**, 301–309.
- JURASKA, M. (2018). *pssmooth: Flexible and Efficient Evaluation of Principal Surrogates/Treatment Effect Modifiers*. R package version 1.0.1, Comprehensive R Archive Network.
- KATZELNICK, L. C., GRESH, L., HALLORAN, M. E., MERCADO, J. C., KUAN, G., GORDON, A., BALMASEDA, A. AND HARRIS, E. (2017). Antibody-dependent enhancement of severe dengue disease in humans. *Science* **358**, 929–932.
- LI, Y., TAYLOR, J. AND ELLIOTT, M. (2010). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66**, 523–531.
- MOODIE, Z., JURASKA, M., HUANG, Y., ZHUANG, Y., FONG, Y., CARPP, L., SELF, S., CHAMBONNEAU, L., SMALL, R., JACKSON, N., NORIEGA, F. and others. (2018). Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *Journal of Infectious Diseases* **217**, 742–753.
- PRENTICE, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.
- PRICE, B., GILBERT, P. AND VAN DER LAAN, M. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics*. doi: 10.1111/biom.12879. [Epub ahead of print].
- ROY, S. N. AND BOSE, R. C. (1953). Simultaneous confidence interval estimation. *The Annals of Mathematical Statistics* **24**, 513–536.
- VANDERWEELE, T. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters* **78**, 2957–2962.
- VILLAR, L., DAYAN, G. H., ARREDONDO-GARCÍA, J. L., RIVERA, D. M., CUNHA, R., DESEDA, C., REYNALES, H., COSTA, M. S., MORALES-RAMÍREZ, J. O., CARRASQUILLA, G. and others. (2015). Efficacy of a tetravalent dengue vaccine in children in Latin America. *New England Journal of Medicine* **372**, 113–123.
- YOHAI, V. (1987). High breakdown-point and high-efficiency robust estimates for regression. *Annals of Statistics* **15**, 642–656.
- ZIGLER, C. M. AND BELIN, T. R. (2011). The potential for bias in principal causal effect estimation when treatment received depends on a key covariate. *Annals of Applied Statistics* **5**, 1876–1892.

[Received May 29, 2018; revised October 15, 2018; accepted for publication October 21, 2018]