aptara	BIOM	biom˙1014	**Dispatch:** March 6, 2008	**CE:** KJI
	Journal	**MSP No.**	**No. of pages:** 9	**PE:** Adrian

BIOMETRICS DOI: 10.1111/j.1541-0420.2008.01014.x

Evaluating Candidate Principal Surrogate Endpoints

Peter B. Gilbert[1,*] and Michael G. Hudgens[2]

[1]Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, Washington 98109, U.S.A.
[2]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
[*]*email:* pgilbert@scharp.org

SUMMARY. Frangakis and Rubin (2002, *Biometrics* **58,** 21–29) proposed a new definition of a surrogate endpoint (a "principal" surrogate) based on causal effects. We introduce an estimand for evaluating a principal surrogate, the *causal effect predictiveness (CEP) surface*, which quantifies how well causal treatment effects on the biomarker predict causal treatment effects on the clinical endpoint. Although the CEP surface is not identifiable due to missing potential outcomes, it can be identified by incorporating a baseline covariate(s) that predicts the biomarker. Given case–cohort sampling of such a baseline predictor and the biomarker in a large blinded randomized clinical trial, we develop an estimated likelihood method for estimating the CEP surface. This estimation assesses the "surrogate value" of the biomarker for reliably predicting clinical treatment effects for the same or similar setting as the trial. A CEP surface plot provides a way to compare the surrogate value of multiple biomarkers. The approach is illustrated by the problem of assessing an immune response to a vaccine as a surrogate endpoint for infection.

KEY WORDS: Case cohort; Causal inference; Clinical trial; HIV vaccine; Postrandomization selection bias; Structural model; Prentice criteria; Principal stratification.

1. Introduction

Identifying biomarkers that can be used as approximate surrogates for clinical endpoints in randomized trials is useful for many reasons including shortening studies, reducing costs, sparing study participants discomfort, and elucidating treatment effect mechanisms. As a motivating example, a central objective of placebo-controlled preventive HIV vaccine efficacy trials is the evaluation of vaccine-induced immune responses as surrogate endpoints for HIV infection. An immunological surrogate would be useful for several purposes including guiding iterative development of immunogens between basic and clinical research, informing regulatory decisions and immunization policies, and bridging efficacy of a vaccine observed in a trial to a new setting.

The surrogate evaluation field was catalyzed by Prentice's (1989) definition of a surrogate endpoint as a replacement endpoint that provides a valid test of the null hypothesis of no treatment effect on the clinical endpoint. The two main criteria for checking this definition are: (i) the distribution of the clinical endpoint conditional on the surrogate is the same as the distribution of the clinical endpoint conditional on the surrogate and treatment (i.e., all of the clinical treatment effect is "mediated" through the surrogate); and (ii) the surrogate and clinical endpoints are correlated. Frangakis and Rubin (2002) (henceforth FR) noted that this definition is based on observable random variables, and named a biomarker satisfying criterion (i) a "statistical surrogate." Since 1989, many surrogate-evaluation methods have been designed to check if a biomarker is a statistical surrogate, including methods for estimating the proportion of the treatment effect explained (Freedman, Graubard, and Schatzkin, 1992). Notably some approaches have not been based on (i); for example, the adjusted association estimand is designed for evaluating the correlation criterion (ii), and the relative effect estimand is based on average causal effects (Buyse and Molenberghs, 1998).

Treatment effects adjusted for a variable measured after randomization (called *net effects*) are susceptible to postrandomization selection bias. Because candidate surrogates are measured after randomization, criterion (i) defining a statistical surrogate is based on net effects. FR pointed out that this definition does not have a causal interpretation, and proposed a new surrogate definition based on principal causal effects. FR's definition of a "principal surrogate" is based on the potential outcomes framework for causal inference, which Robins (1995) also considered for studying treatment effects subject to postrandomization selection bias. To date, statistical methods for evaluating principal surrogates have not been elaborated. A recent review paper noted that FR "present a convincing case for the principal surrogate definition" and called for such elaborations (Weir and Walley, 2006).

The literature on statistical methods for evaluating surrogate endpoints contains approaches based on a single large clinical trial and on metaanalysis. Here we develop an approach for evaluating a principal surrogate within the former setting. Following Follmann (2006), our approach uses a baseline covariate(s) to predict missing potential biomarker

outcomes. After defining statistical and principal surrogates in Section 2, in Section 3 we introduce the *causal effect predictiveness (CEP) surface* and the *marginal CEP curve*, plus associated summary causal estimands, which quantify how well a biomarker predicts population-level causal effects of treatment. In Section 4, we develop an estimated-likelihood approach for estimating the causal estimands based on case–cohort sampling of the biomarker, and parametric or nonparametric marginal structural mean models. In Section 5, we evaluate the nonparametric method in simulations based on an HIV vaccine trial, and in Section 6 we conclude with discussion.

2. Statistical and Principal Surrogates

Throughout we consider a randomized trial with treatment assignment Z ($Z = 1$ or 0), a discrete or continuous biomarker S measured at fixed time t_0 after treatment assignment, and a binary clinical endpoint Y($Y = 1$ for disease, 0 otherwise) measured after t_0. Because S must be measured prior to disease to evaluate it as a candidate surrogate, the analysis is restricted to subjects disease free at t_0; denote this evaluability criterion by the indicator $V = 1$. The biomarker S is only measured in those with $V = 1$, and otherwise is undefined (denoted by $S = *$). We consider two phase outcome-dependent case–cohort sampling, wherein baseline covariates X are measured for everyone (phase 1) and in the second phase a baseline covariate(s) W is measured for all or almost all cases (those with $Y = 1$) and for a random "subcohort" of controls (those with $Y = 0$). Let δ indicate whether W is measured. For subjects with $V = 1$, S is measured for those with W measured. Case–cohort sampling is efficient when W or S is expensive (Prentice, 1986). For vaccine trials, W and S can be measured after the trial using stored specimens (Gilbert et al., 2005).

2.1 *Definition of a Statistical Surrogate*

Following FR, methods for evaluating statistical surrogates are based on comparing the risk distributions

$$\text{risk}(s \mid Z = 1) \equiv \Pr(Y = 1 \mid Z = 1, V = 1, S = s) \quad \text{and}$$

$$\text{risk}(s \mid Z = 0) \equiv \Pr(Y = 1 \mid Z = 0, V = 1, S = s).$$

If S is continuous then these definitions abuse notation; however, to avoid the distraction of technical details the formal definitions are placed in Web Appendix A. FR defined S to be a *statistical surrogate* if, for all values s of S, $\text{risk}(s \mid Z = 1) = \text{risk}(s \mid Z = 0)$.

Because S and V are measured after randomization, a comparison of $\text{risk}(s \mid Z = 1)$ and $\text{risk}(s \mid Z = 0)$ measures the net effect of treatment, i.e., differences due to a mixture of the causal treatment effect and any differences in characteristics between treatment 1 subjects who have response level s, $\{Z = 1, V = 1, S = s\}$, and treatment 0 subjects who have response level s, $\{Z = 0, V = 1, S = s\}$. Consequently, application of a method that evaluates a statistical surrogate may mislead about the capacity of a biomarker to reliably predict causal clinical treatment effects.

2.2 *Definition of a Principal Surrogate Endpoint*

Let $Y(Z)$ be the potential clinical endpoint after time t_0 under assignment to treatment Z. Similarly define potential outcomes $S(Z)$ for the biomarker endpoint measured at t_0, and let $V(Z)$ be the potential indicators of whether the subject is disease free at t_0. Note that $S(Z)$ and $Y(Z)$ are undefined if $V(Z) = 0$; in this case $S(Z) = Y(Z) = *$. We suppose that (Z_i, X_i, δ_i, $\delta_i W_i$, $V_i(1)$, $V_i(0)$, $S_i(1)$, $S_i(0)$, $Y_i(1)$, $Y_i(0)$), $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.), and for simplicity assume no drop-out. Further we assume A1–A3:

Assumption A1: *Stable unit treatment value assumption (SUTVA)*

Assumption A2: *Ignorable treatment assignments (Rubin, 1986)*: Conditional on X, Z is independent of (W, $V(1)$, $V(0)$, $S(1)$, $S(0)$, $Y(1)$, $Y(0)$)

Assumption A3: *Equal individual clinical risk up to time* $t_0 : V(1) = 1$ if and only if $V(0) = 1$

A1 implies that the potential outcomes ($V_i(1)$, $V_i(0)$, $S_i(1)$, $S_i(0)$, $Y_i(1)$, $Y_i(0)$) are independent of the treatment assignments of other subjects, which implies "consistency," $(V_i(Z_i), S_i(Z_i), Y_i(Z_i)) = (V_i, S_i, Y_i)$. A2 holds for blinded randomized trials, where the randomization may depend on X. A3 will be needed for identifying the causal estimand based on data from subjects observed to be at risk for disease at t_0. Inferences will be robust to A3 if t_0 is near baseline relative to the period of follow-up for clinical events and the vast majority of subjects are at risk at t_0, in which case $V_i(1) = V_i(0) = 1$ for almost all i.

With these preliminaries, we now define a principal surrogate endpoint. FR defined the *basic principal stratification* P_0 with respect to the postrandomization variable S as the partition of units $i = 1, \ldots, n$ such that within any set of P_0, all units have the same vector ($S_i(1)$, $S_i(0)$). A *principal stratification* is a partition of units whose sets are unions of sets in P_0. FR defined a biomarker S to be a principal surrogate endpoint if the comparison between

$$\begin{aligned}\text{risk}_{(1)}(s_1, s_0) \equiv \Pr(Y(1) = 1 \mid V(1) = 1, V(0) = 1,\\ S(1) = s_1, S(0) = s_0) \quad \text{and}\end{aligned}$$

$$\begin{aligned}\text{risk}_{(0)}(s_1, s_0) \equiv \Pr(Y(0) = 1 \mid V(1) = 1, V(0) = 1,\\ S(1) = s_1, S(0) = s_0),\end{aligned}$$

results in equality for all $s_1 = s_0$. FR did not explicitly condition on $V(1) = V(0) = 1$ in their definition; however, implicitly they must have, because ($S(1)$, $S(0)$) is only defined if $V(1) = V(0) = 1$. For notational simplicity henceforth all probability statements involving $S(Z)$ implicitly condition on $V(Z) = 1$. A contrast in $\text{risk}_{(1)}(s_1, s_0)$ and $\text{risk}_{(0)}(s_1, s_0)$ measures a population-level causal treatment effect on Y for subjects with $\{S_i(1) = s_1, S_i(0) = s_0\}$. Such a contrast is causal because it conditions on a principal stratification, which, by construction, is unaffected by treatment. Thus in FR's definition, S is a principal surrogate if groups of subjects with no causal effect on the biomarker have no causal effect on the clinical endpoint. We call this property average causal necessity.

Average causal necessity: $\text{risk}_{(1)}(s_1, s_0) = \text{risk}_{(0)}(s_1, s_0)$ for all $s_1 = s_0$.

Biomarkers with the greatest utility for predicting clinical treatment effects will not only be necessary for a clinical effect, but also sufficient. For example, knowing that an antibody

titer > 1000 is sufficient for a vaccine to protect individuals against HIV infection is exactly the information needed to use titer as a reliable predictor of protection. We define average causal sufficiency as

Average causal sufficiency: There exists a constant $C \geq 0$ such that $\text{risk}_{(1)}(s_1, s_0) \neq \text{risk}_{(0)}(s_1, s_0)$ for all $|s_1 - s_0| > C$.

For the one-sided situation where interest is in assessing if higher treatment 1 biomarker responses ($S(1) > S(0)$) predict clinical benefit of treatment 1 ($Y(1) = 0$ and $Y(0) = 1$) (e.g., a placebo-controlled trial), a one-sided version of average causal sufficiency may be more appropriate, defined as above with $\neq$ replaced with $<$ and $|s_1 - s_0|$ replaced with $s_1 - s_0$. In either case, we suggest a refined definition of a principal surrogate endpoint as a biomarker that satisfies both average causal necessity and average causal sufficiency. Henceforth we use this definition of a principal surrogate endpoint.

3. Causal Effect Predictiveness Estimands

3.1 *Quantitation of Associative and Dissociative Effects*

FR suggested that the quality of a surrogate be measured by its "associative effects" relative to its "dissociative effects." As defined in equations 5.3 and 5.4 of FR, an *associative effect* is a comparison between the ordered sets

$$\{Y_i(1) : S_i(1) \neq S_i(0)\} \quad \text{and} \quad \{Y_i(0) : S_i(1) \neq S_i(0)\},$$

and a *dissociative effect* is a comparison between the ordered sets

$$\{Y_i(1) : S_i(1) = S_i(0)\} \quad \text{and} \quad \{Y_i(0) : S_i(1) = S_i(0)\}.$$

For the purpose of quantifying these effects, we introduce a *causal effect predictiveness (CEP) surface*. Let $CE \equiv h(\Pr(Y(1) = 1), \Pr(Y(0) = 1))$ be the overall causal effect of treatment on the clinical endpoint, where $h(\cdot, \cdot)$ is a known contrast function satisfying $h(x, y) = 0$ if and only if $x = y$, for example $h(x, y) = x - y$ or $\log(x/y)$. Let

$$\text{CEP}^{\text{risk}}(s_1, s_0) \equiv h(\text{risk}_{(1)}(s_1, s_0), \text{risk}_{(0)}(s_1, s_0)),$$

be this contrast conditional on $\{S(1) = s_1, S(0) = s_0\}$. Note that $\text{CEP}^{\text{risk}}(s, s) = 0$ for all s is equivalent to average causal necessity, whereas $\text{CEP}^{\text{risk}}(s_1, s_0) \neq 0$ for all $|s_1 - s_0| > C$ (or the one-sided analog) is equivalent to average causal sufficiency. Therefore, the criteria for a principal surrogate can be checked through inference on the CEP surface. Moreover, biomarkers with capacity to predict clinical treatment effects will often have $|\text{CEP}^{\text{risk}}(s_1, s_0)|$ increasing in $|s_1 - s_0|$, reflecting the situation that on average groups of persons with a greater causal effect on the marker have a greater causal effect on the clinical endpoint. We refer to the capacity of a biomarker to reliably predict the population level causal effect of treatment on the clinical endpoint as the biomarkers' *surrogate value*. This value can be quantified both by the nearness of $|\text{CEP}^{\text{risk}}(s_1, s_0)|$ to 0 for s_1 near s_0 and by the extent to which $|\text{CEP}^{\text{risk}}(s_1, s_0)|$ increases with $|s_1 - s_0|$, with a greater increase reflecting greater associative effects. Note that even if one or both of average causal necessity or sufficiency fail, a biomarker can still have surrogate value if $|\text{CEP}^{\text{risk}}(s_1, s_0)|$ increases with $|s_1 - s_0|$; Figure 2 (dashed line) will illustrate this. Moreover, two principal surrogates can have different surrogate values as reflected by different CEP surfaces.

If S is continuous, then the CEP surface can alternatively be defined in terms of percentiles of the marker S. To formulate this, consider Huang, Pepe, and Feng's (2007) proposal to judge the value of a continuous marker S for predicting disease Y by the *predictiveness curve*, $R(v) \equiv \Pr(Y = 1 \,|\, S = F^{-1}(v))$, $v \in [0, 1]$, where F is the cumulative distributive function (cdf) of S. Note that $R(v) = \text{risk}(S = F^{-1}(v))$, i.e., $R(v)$ is risk as a function of the quantiles of S, which provides a common scale for comparing multiple markers. The predictiveness curve $R(v)$ usefully informs about both absolute risks at different marker quantiles and the frequency of these risks in the population. A predictive marker is one with $R(v)$ monotone (or approximately so) in v with large $|R(1) - R(0)|$.

Applying these ideas, we propose a scale-independent version of the causal effect predictiveness surface, $\text{CEP}^R(v_1, v_0) \equiv h(R_{(1)}(v_1, v_0), R_{(0)}(v_1, v_0))$, where

$$R_{(Z)}(v_1, v_0) \equiv \Pr\big(Y(Z) = 1 \,|\, S(1) = F^{-1}_{(1)}(v_1), \\ S(0) = F^{-1}_{(1)}(v_0)\big) \quad \text{for } Z = 0, 1.$$

In this definition, $S(1)$ and $S(0)$ are standardized relative to the distribution $F_{(1)}$ of $S(1)$. Figure 1 illustrates two CEP surfaces for the one-sided setting where interest is in predicting clinical benefit of treatment 1 from higher treatment 1 biomarker responses.

For some studies, the *marginal CEP curve* is a related causal estimand of interest:

$$m\text{CEP}^{\text{risk}}(s_1) \equiv h(\text{risk}_{(1)}(s_1), \text{risk}_{(0)}(s_1)),$$

where $\text{risk}_{(Z)}(s_1) \equiv \Pr(Y(Z) = 1 \,|\, S(1) = s_1)$. Similarly $m\,\text{CEP}^R(v_1)$ is defined as $h(R_{(1)}(v_1), R_{(0)}(v_1))$ with $R_{(Z)}(v_1) \equiv \Pr(Y(Z) = 1 \,|\, S(1) = F^{-1}_{(1)}(v_1))$. With $h(x, y) = x - y$, if S is continuous and strictly increasing then the area between $m\,\text{CEP}^R(\cdot)$ and the zero-line equals $CE = \Pr(Y(1) = 1) - \Pr(Y(0) = 1)$ (proof in Web Appendix A).

If $S_i(0)$ is constant across subjects, then the CEP surface (trivially) equals the marginal CEP curve. We refer to this special case as case CB:

Case CB. *Constant Biomarkers*: $S_i(0) = c$ for all i for some constant c

HIV vaccine trials fit case CB, with (almost) all subjects having no immune response under placebo ($Z = 0$). This occurs because S is an HIV-specific immune response, so that vaccine antigens must be presented to the immune system to induce a response (Gilbert et al., 2005). The dissociative effect can be measured by $\text{CEP}^{\text{risk}}(c, c)$ and the associative effects by $\text{CEP}^{\text{risk}}(s_1, c)$ for $s_1 \neq c$. For example, with $c = L$ the lower bound of S, the nearer $\text{CEP}^{\text{risk}}(c, c)$ is to zero and the greater the increase of $|\text{CEP}^{\text{risk}}(s_1, c)|$ with $s_1 > c$, the greater the surrogate value (Figure 2).

For placebo-controlled trials for which case CB fails yet $S_i(0)$ has much less variability than $S_i(1)$, the marginal CEP curve has interpretation approximately equal to that of $\text{CEP}(s_1, s_0)$. In general, however, $m\text{CEP}(s_1)$ does not measure the association between causal biomarker effects and causal

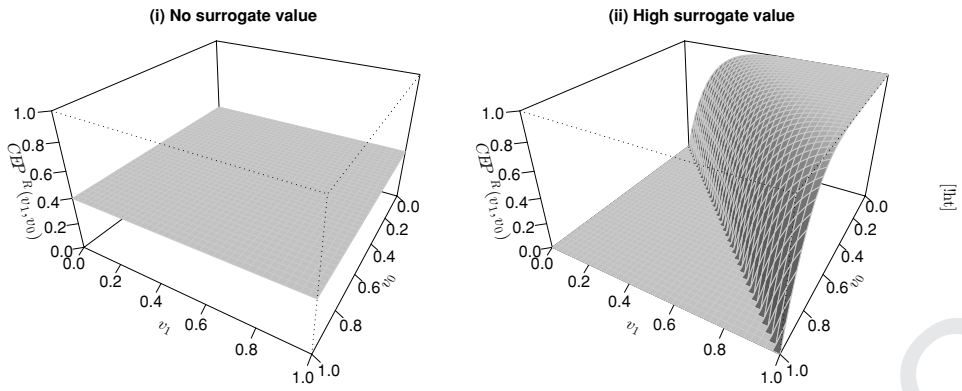

Figure 1. Example $\text{CEP}^R(v_1, v_0) = h(R_{(1)}(v_1, v_0), R_{(0)}(v_1, v_0))$ surfaces, with $h(x, y) = 1 - x/y$. The surface in (i) reflects a biomarker with no surrogate value, wherein the clinical treatment effect is the same for all treatment effects on the biomarker. The surface in (ii) reflects a biomarker with high surrogate value, wherein the average causal effect on the clinical endpoint is zero for $v_1 = v_0$ and has a large increase in $v_1 - v_0$ for $v_1 > v_0$. Because $\text{CEP}^R(v_1, v_0) = 0$ for $v_1 = v_0$, both biomarkers satisfy average causal necessity. Furthermore, because $\text{CEP}^R(v_1, v_0) > 0$ for all $v_1 > v_0$, the biomarker in (ii) satisfies one-sided average causal sufficiency. This figure appears in color in the electronic version of this article.

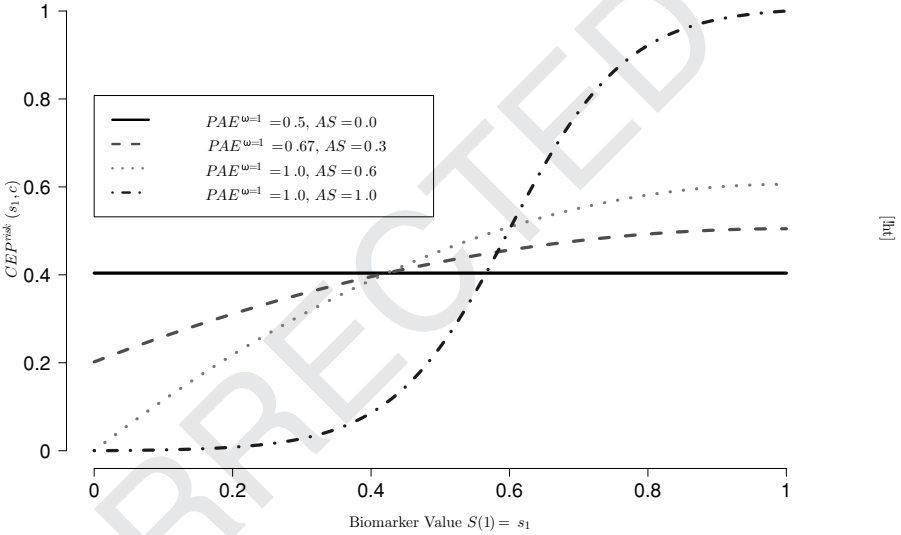

Figure 2. For case CB with $S_i(0) = c$ for all i with $c = L$ the lower bound of S, biomarkers S that have no (horizontal solid line), modest (dashed line), moderate (dotted line), and high (hatched line) surrogate value. Here $\text{CEP}^{\text{risk}}(s_1, c) = h(\text{risk}_{(1)}(s_1, c), \text{risk}_{(0)}(s_1, c))$ with $h(x, y) = 1 - x/y$. Because $\text{CEP}^{\text{risk}}(c, c) = 0$ and $\text{CEP}^{\text{risk}}(s_1, c) > 0$ for all $s_1 > c$, the latter two S's satisfy average causal necessity and average causal sufficiency, and hence are principal surrogates. This figure appears in color in the electronic version of this article.

clinical effects, and hence does not measure principal surrogate value. Nevertheless, under A1 and A2 $m\text{CEP}(s_1)$ has a different but useful interpretation as the population level causal treatment effect on Y for subjects with $S(1) = s_1$, where conditioning on $S(1)$ is equivalent to conditioning on a baseline covariate. As such, the marginal CEP curve can be used for predicting how clinical efficacy varies with the biomarker $S = S(1)$ observed in persons attending a treatment or vaccine clinic.

3.2 *Estimands for Summarizing Surrogate Value*

We suggest functionals of the CEP surface that summarize the surrogate value of a biomarker. We again consider the one-sided setting where interest is in assessing whether $S(1) > S(0)$ predicts clinical benefit of treatment 1 ($Y(1) = 0$ and $Y(0) = 1$). To summarize the associative and dissociative effects, we consider the *expected associative effect (EAE)* and the *expected dissociative effect (EDE)*:

$$\text{EAE}^{\omega} \equiv E[\omega(S(1), S(0))\text{CEP}^{\text{risk}}(S(1), S(0)) \,|\, S(1) > S(0)]/ E[\omega(S(1), S(0)) \,|\, S(1) > S(0)], \quad (1)$$

$$\text{EDE} \equiv E[\text{CEP}^{\text{risk}}(S(1), S(0)) \,|\, S(1) = S(0)], \quad (2)$$

where $\omega(\cdot, \cdot)$ is a nonnegative weight function. For case CB with $c = L, \text{EAE}^{\omega} = \{\int_{s_1 > c} \omega(s_1, c)\, dF_{(1)}(s_1)\}^{-1} \int_{s_1 > c} \times \omega(s_1, c)\text{CEP}^{\text{risk}}(s_1, c)\, dF_{(1)}(s_1)$ and $\text{EDE} = \text{CEP}^{\text{risk}}(c, c)$.

We also define the *proportion associative (PA) effect* by

$$\mathrm{PAE}^{\omega} \equiv |\mathrm{EAE}^{\omega}|/\{|\mathrm{EDE}| + |\mathrm{EAE}^{\omega}|\}. \quad (3)$$

Values of $\mathrm{PAE}^{\omega} \in [0, 0.5]$ suggest the biomarker has no surrogate value, whereas values in (0.5, 1] suggest some surrogate value.

A weight function is included in EAE^{ω} to reflect the idea that a biomarker with high surrogate value should have large $|\mathrm{CEP}^{\mathrm{risk}}(s_1, s_0)|$ for large $s_1 - s_0 > 0$. For example, weights $\omega(s_1, s_0) = s_1 - s_0$ or $I(s_1 = U, s_0 = L)$ may be used, where $L(U)$ is the lower (upper) bound of S. With the latter weight, PAE^{ω} compares the clinical effect among groups with the maximum surrogate effect and with no surrogate effect:

$$\mathrm{PAE}^{\omega} = |\mathrm{CEP}^{\mathrm{risk}}(U, L)|/\{|\mathrm{EDE}| + |\mathrm{CEP}^{\mathrm{risk}}(U, L)|\}.$$

If $h(x, y) = x - y$, $\Pr(S(1) > S(0)) = 0.5$, and an additional monotonicity assumption is made (that $Y_i(1) \leq Y_i(0)$ for all i, i.e., no one is harmed by treatment 1), then $\mathrm{PAE}^{\omega=1}$ equals the PA, defined by

$$\mathrm{PA} \equiv \Pr(S(1) > S(0), Y(1) = 0, \\ Y(0) = 1)/\Pr(Y(1) = 0, Y(0) = 1)$$

(proof in Web Appendix A). This summary measure, proposed by Taylor, Wang, and Thiebaut (2005), is interpreted as the proportion of the study population with a beneficial causal clinical effect that also has a positive causal surrogate effect. The PA depends on the underlying principal strata distribution $F_{(1),(0)}(s_1, s_0) \equiv \Pr(S(1) \leq s_1, S(0) \leq s_0)$; if $\Pr(S(1) > S(0))$ is small (large) then the PA will tend to be small (large), irrespective of the biomarker's surrogate value. By conditioning on $(S(1), S(0))$, the PAE^{ω} is designed to be robust to $F_{(1),(0)}(\cdot, \cdot)$; the PAE^{ω} reflects the relative magnitude of clinical effects for those with and without surrogate effects.

Biomarkers satisfying average causal necessity have $\mathrm{EDE} = 0$ and thus $\mathrm{PAE}^{\omega} = 1$, in which case EAE^{ω} contributes no information to the PAE^{ω}. Therefore, additional measures are needed for summarizing the magnitude of associative effects. One such measure is the *associative span (AS)*, defined by $\mathrm{AS} \equiv |\mathrm{CEP}^{\mathrm{risk}}(U, L)| - |\mathrm{EDE}|$. Figure 2 illustrates $\mathrm{PAE}^{\omega=1}$ and AS. Although the summary parameters may be useful, it is important to estimate the CEP estimands over the range of marker values or quantiles to provide a full picture of the associative and disssociative effects.

Below we also consider estimands defined as above except they condition on X and/or W; for example $\mathrm{risk}_{(Z)}(s_1, s_0, x, w) \equiv \Pr(Y(Z) = 1 \,|\, S(1) = s_1, S(0) = s_0, X = x, W = w)$ and $\mathrm{CEP}^{\mathrm{risk}}(s_1, s_0, x, w) \equiv h(\mathrm{risk}_{(1)}(s_1, s_0, x, w), \mathrm{risk}_{(0)}(s_1, s_0, x, w))$. The conditional estimands reflect baseline covariate-specific surrogate value.

4. Estimating the CEP Surface and Marginal CEP Curve

We consider one approach to identifying and estimating the CEP surface in the practically important special case CB. The same approach identifies and estimates the marginal CEP curve in the general case that $S_i(0)$ has arbitrary variability. In case CB it is difficult to evaluate a statistical surrogate, because it is not possible to study the correlation of S with Y in arm $Z = 0$ subjects, and it is conceptually difficult to evaluate whether S fully mediates clinical treatment effects (Chan et al., 2002).

4.1 *Identifiability of the Causal Estimands*

Due to missing potential outcomes the CEP surface and marginal CEP curve are not identified without further assumptions. A1–A3 imply

$$\mathrm{risk}_{(1)}(s_1, s_0, x, w) = \Pr(Y = 1 \,|\, Z = 1, S = s_1, \\ S(0) = s_0, X = x, W = w),$$

$$\mathrm{risk}_{(0)}(s_1, s_0, x, w) = \Pr(Y = 1 \,|\, Z = 0, S(1) = s_1, \\ S = s_0, X = x, W = w),$$

demonstrating that $\mathrm{risk}_{(Z)}(s_1, s_0, x, w)$ would be identified if we knew the potential outcomes $S_i(Z)$ of subjects assigned the opposite treatment $1 - Z$. Estimating the CEP surface will therefore require a way to predict the missing potential biomarkers. This challenge is relatively easy in case CB, for which $\mathrm{risk}_{(1)}(s_1, c, x, w)$ is identified by the observed data in arm $Z = 1$. However, A1–A3 do not identify $\mathrm{risk}_{(0)}(s_1, c, x, w)$, and the remaining task to identify the CEP surface entails determining values $S_i(1)$ for arm $Z_i = 0$ subjects.

4.2 *Baseline Predictor Study Design and Likelihood*

Our method of inference is based on one of the augmented vaccine trial designs proposed by Follmann (2006), wherein a baseline covariate(s) W that is predictive of $S(1)$ is measured in subjects in both treatment arms. A model predicting $S(1)$ from X and W fit from arm $Z = 1$ subjects is used to predict $S(1)$ for arm $Z = 0$ subjects. The predictions are unbiased because A1–A3 imply $S(1) \,|\, Z = 1, X, W =^d S(1) \,|\, Z = 0, X, W$, where $=^d$ denotes equality in distribution.

We observe iid data $O_i \equiv (Z_i, X_i, V_i, Y_i, \delta_i, \delta_i W_i, \delta_i Z_i S_i)$, $i = 1, \ldots, n$. Only subjects with $V_i = 1$ contribute to the likelihood. Subjects with $Z_i\delta_i = 1$ contribute $\mathrm{risk}_{(1)}(S_i, c, X_i, W_i; \beta)^{Y_i}(1 - \mathrm{risk}_{(1)}(S_i, c, X_i, W_i; \beta))^{1-Y_i}$, where $\mathrm{risk}_{(1)}(\cdot, c, \cdot, \cdot; \beta)$ is modeled as a function of unknown parameters β. The likelihood contribution for subjects with $(1 - Z_i)\delta_i = 1$ is obtained by integrating $\mathrm{risk}_{(0)}(\cdot, c, X_i, W_i; \beta)$ over the conditional cdf $F_{(1)}^{S|X,W}$ of $S(1) \,|\, X, W$. The contribution for subjects with $\delta_i = 0$ is obtained by integrating $\mathrm{risk}_{(Z_i)}(\cdot, c, X_i; \beta)$ over the conditional cdf of $S(1)$, $W \,|\, X$, which is $F_{(1)}^{S|X,W} \times F^{W|X}$, where $F^{W|X}$ is the conditional cdf of $W \,|\, X$. Thus, with $\nu \equiv (F_{(1)}^{S|X,W}, F^{W|X})$, the conditional likelihood is $L(\beta, \nu) \equiv \prod_{i=1}^{n} f(Y_i \,|\, Z_i, X_i, V_i, \delta_i, \delta_i W_i, \delta_i Z_i S_i)^{V_i}$, where

$$f(Y \,|\, Z, X, V, \delta, \delta W, \delta ZS) \\ = \left\{\mathrm{risk}_{(1)}(S, c, X, W; \beta)^{Y}(1 - \mathrm{risk}_{(1)}(S, c, X, W; \beta))^{1-Y}\right\}^{Z\delta} \\ \times \left\{\left(\int \mathrm{risk}_{(0)}(s_1, c, X, W; \beta)\, dF_{(1)}^{S|X,W}(s_1 \,|\, X, W)\right)^{Y}\right. \\ \left. \times \left(1 - \int \mathrm{risk}_{(0)}(s_1, c, X, W; \beta)\, dF_{(1)}^{S|X,W}\right.\right.$$

$$\times (s_1 \mid X, W)\Big)\Big)^{1-Y}\Big\}^{(1-Z)\delta}$$
$$\times \Big\{\Big(\iint \text{risk}_{(Z)}(s_1, c, X, w; \beta)\, dF_{(1)}^{S|X,W} \times (s_1 \mid w, X)\, dF^{W\mid X}(w \mid X)\Big)^{Y} \times \Big(1 - \iint \text{risk}_{(Z)}(s_1, c, X, w; \beta)\, dF_{(1)}^{S\mid X,W} \times (s_1 \mid w, X)\, dF^{W\mid X}(w \mid X)\Big)^{1-Y}\Big\}^{(1-\delta)}.$$

because $\text{CEP}^{\text{risk}}(\cdot, c, X, W; \beta)$ depends on β but not ν, the ν are nuisance parameters. Although profile likelihood is a natural approach to pursue, it is difficult to implement because the likelihood integrates over $F_{(1)}^{S|X,W}$ and $F^{W|X}$. We use estimated likelihood (Pepe and Fleming, 1991), also called pseudolikelihood, wherein consistent estimates of ν are obtained based on treatment arm 1 data, and then $L(\beta, \hat{\nu})$ is maximized in β. The bootstrap is used to estimate standard errors for $\hat{\beta}$. A re-sampling approach is used because there does not appear to be an analytical expression for the asymptotic variance of $\hat{\beta}$ that accounts for the variations in $\hat{\nu}$, and previously developed techniques for deriving the variance do not apply because they would assume that all subjects have a non-zero probability that $S(1)$ is observed (e.g., Pepe and Fleming, 1991).

4.3 *Models for* $\text{Risk}_{(z)}(\cdot, c, \cdot, \cdot)$ *and* $\nu = (F_{(1)}^{S|X,W}, F^{W|X})$

The estimated likelihood approach can be used for a variety of structural models for $\text{risk}_{(z)}(s_1, c, x, w)$ and the nuisance parameters ν. Here we consider two types of models for case CB. The first is fully parametric, where we assume $F_{(1)}^{S|X,W}$ and $F^{W|X}$ have particular parametric distributions, and $S(1)$ is continuous subject to "limit of detection" left-censoring: $S(1) \equiv \max(S^*(1), c)$, where $S^*(1)$ has a continuous cdf with $\Pr(S^*(1) \le c) > 0$. We also assume the risk functions have a generalized linear model form

A4-P: $\text{risk}_{(z)}(s_1, c, x, w; \beta_z) = g(\beta_{z0} + \beta_{z1}s_1 + \beta_{z2}^T x + \beta_{z3}^T w)$ for $z = 0, 1$,

for $s_1 \ge c$ and some known link function $g(\cdot)$. For example, we might assume $F^{W|X}$ is normal and $F_{(1)}^{S|X,W}$ is censored normal, with left-censoring of values below c, and A4-P holds with g equal to the standard normal cdf Φ. This set-up extends Follmann (2006) to account for censoring. With $h(x, y) = g^{-1}(x) - g^{-1}(y)$, A4-P then implies

$$\text{CEP}^{\text{risk}}(s_1, c, x, w; \beta) = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})s_1 + (\beta_{12} - \beta_{02})^T x + (\beta_{13} - \beta_{03})^T w.$$

Simple calculations yield $\text{EDE}(x, w) = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})L + (\beta_{12} - \beta_{02})^T x + (\beta_{13} - \beta_{03})^T w$, $\text{AS}(x, w) = |(\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})U + (\beta_{12} - \beta_{02})^T x + (\beta_{13} - \beta_{03})^T w| - |\text{EDE} \times (x, w)|$, and $\text{EAE}^{\omega=1}(x, w) = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})E[S(1)|S(1) > c, x, w] + (\beta_{12} - \beta_{02})^T x + (\beta_{13} - \beta_{03})^T w$. For the case that $g = \Phi$, Web Appendix C provides a proof, adapted from a proof by Dean Follmann, that β is identified under the untestable imposed constraint that one of the components of $(\beta_{12}^T, \beta_{13}^T)$ equals the corresponding component of $(\beta_{02}^T, \beta_{03}^T)$. Therefore identifiability requires assuming the absence of one interaction, but otherwise if and how the CEP curve varies with X and W can be evaluated. If no interactions between treatment and X or W are assumed, then $\text{CEP}^{\text{risk}}(s_1, c; \beta) = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})s_1$ is interpreted as the covariate-adjusted CEP curve.

Secondly, we consider a nonparametric approach wherein S and W are treated as categorical variables with J and K levels, which may be discretized versions of continuous measurements. Here we assume the lowest category $j = 1$ corresponds to the constant c in case CB. With W the only baseline covariate, nonparametric models are specified by $\nu_{jk} \equiv \Pr(S(1) = j, W = k)$ and

A4-NP: $\text{risk}_{(z)}(j, 1, k; \beta) = \beta_{zj} + \beta'_k$

for $j = 1, \ldots, J$; $k = 1, \ldots, K$; and $z = 0, 1$. The parameters $\beta \equiv (\beta_{zj}, \beta'_k : z = 0, 1; j = 1, \ldots, J; k = 1, \ldots K)$ are constrained such that $0 \le \beta_{zj} + \beta'_k \le 1$ for all z, j, k and $\sum_{k=1}^K \beta'_k = 0$ for identifiability. Under model A4-NP, W has the same effect on risk for the two study arms. This no-interaction assumption identifies the model, and the expanded model with β'_k replaced with β'_{zk} is not identified (see Web Appendix C).

In the simulations we consider the CEP curve estimand $\text{CEP}^{\text{risk}}(j, 1; \beta) = \log(\text{risk}_{(1)}(j, 1; \beta_{1j})/\text{risk}_{(0)}(j, 1; \beta_{0j}))$ based on average risks $\text{risk}_{(z)}(j, 1; \beta_{zj}) \equiv K^{-1}\sum_{k=1}^K \text{risk}_{(z)} \times (j, 1, k; \beta) = \beta_{zj}, z = 0, 1$. It follows that $\text{CEP}^{\text{risk}}(j, 1; \beta) = \log(\beta_{1j}/\beta_{0j})$, $\text{AS} = |\log(\beta_{1J}/\beta_{0J})| - |\log(\beta_{11}/\beta_{01})|$, $\text{EDE} = \log(\beta_{11}/\beta_{01})$, and $EAE^{\omega} = \sum_{j=2}^J \tilde{\omega}(j, 1)\log(\beta_{1j}/\beta_{0j})$ with $\tilde{\omega}(j, 1) = \omega(j, 1)\nu_j / \sum_{l=2}^J \omega(l, 1)\nu_l$ and $\nu_j = \sum_{k=1}^K \nu_{jk}$.

For both the parametric and nonparametric approaches, Web Appendix B describes consistent estimators of ν and procedures for maximizing $L(\beta, \hat{\nu})$ in β.

4.4 *Tests for Whether a Biomarker has Any Surrogate Value*

Because $\text{PAE}^{\omega} = 0.5$ supports that S has no surrogate value, Wald tests for any surrogate value can be based on the maximum estimated likelihood estimator (MELE) $\widehat{\text{PAE}}^{\omega}$ minus 0.5 divided by its bootstrap standard error. Similarly Wald tests of $\text{AS} = 0$ can be implemented based on $\widehat{\text{AS}}$. For the nonparametric approach assuming model A4-NP, we also consider a test statistic $T = \sum_{j=2}^J (j-1)\{\hat{\beta}_{0j} - (\hat{\beta}_{0j} + \hat{\beta}_{1j})(\hat{\mu}_0/(\hat{\mu}_0 + \hat{\mu}_1))\}$ divided by its bootstrap standard error, where $\hat{\mu}_z = \frac{1}{J}\sum_{j=1}^J \hat{\beta}_{zj}$. This test evaluates $H_0 : \text{CEP}^{\text{risk}}(j, 1) = CE$ for all j versus the monotone alternative that $\text{CEP}^{\text{risk}}(j, 1)$ increases in j, similar to the Breslow–Day trend test (Breslow and Day, 1980). The null and alternative hypotheses indicate that average causal sufficiency does not and does hold, respectively.

5. Simulation Study

Based on data from the first preventive HIV vaccine efficacy trial (Gilbert et al., 2005), we conducted a simulation study to evaluate performance of the MELE methods. The vaccine trial was double blind with 2:1 randomization to vaccine:placebo. A biomarker of interest S was the 50% neutralization titers against the HIV recombinant gp120 molecule measured from a serum sample drawn at the month 1.5 visit, and Y was

HIV infection during the time period $t_0 = 1.5$ months to 36 months. The lower quantification limit of the neutralization assay was 1.65, and 44 of 47 placebo recipients with S measured at 1.5 months had left-censored values; thus the data essentially fit case CB. The range of S_i was [1.65, 4.09], which we rescaled to [0, 1], so that $c = L = 0$.

We simulated vaccine trials with the following steps. Step 1: For all 3598 (1805) subjects in the vaccine (placebo) arm, $(W_i, S_i(1))$ was generated from a bivariate normal distribution with means 0.41, standard deviation 0.55, and correlation $\rho = 0.5$, 0.7, or 0.9; the standard deviation was chosen such that 23% of $S_i(1)$ values were less than 0 on average. Simulated values of $S_i(1)$ and W less than 0 (greater than 1) were set equal to 0 (1). Step 2: The W_i and $S_i(1)$ were binned into quartiles. For subjects i with quartile j value of $S_i(1)$ and quartile k value of W_i, $Y_i(Z)$ was generated from a Bernoulli distribution with success probability determined by model A4-NP with values of β_{zj} and β'_k set as follows. First, the β_{zj} were set to achieve the infection rate $\Pr(Y(1) = 1) = 0.067$ that was observed in the vaccine arm of the trial, an overall vaccine efficacy of 50% (i.e., $\Pr(Y(0) = 1) = 2 \times \Pr(Y(1) = 1)$) and to reflect a biomarker with either (i) no or (ii) high surrogate value. Based on risk averaged over W described in Section 4.3, in scenario (i) $\text{CEP}^{\text{risk}}(j,\ 1;\ \beta) \equiv \log\left(\text{risk}_{(1)}(j,\ 1;\ \beta_{1j})/\text{risk}_{(0)}(j,\ 1;\ \beta_{0j})\right) = -0.69$ for $j = 1, \ldots, 4$ and in scenario (ii) $\text{CEP}^{\text{risk}}(j,\ 1;\ \beta) = -0.22,\ -0.51,\ -0.92,\ -1.61$ for $j = 1, \ldots, 4$. With vaccine efficacy $\text{VE}(j,\ 1) \equiv 1 - \exp(\text{CEP}^{\text{risk}}(j,\ 1;\ \beta))$, scenario (i) specifies constant $\text{VE}(j,\ 1) = 0.5$ and scenario (ii) specifies $\text{VE}(j,\ 1) = 0.2,\ 0.4,\ 0.6,\ 0.8$ for $j = 1, \ldots, 4$. For both scenarios (i) and (ii), we let $\beta'_k = 0.015 - 0.01(k-1)$ for $k = 1, \ldots, 4$. Step 3: To achieve case–cohort sampling, $(W_i, S_i(1))$ was retained for all infected vaccine recipients and a subcohort of uninfected vaccine recipients. For the placebo arm $S_i(1)$ was set to missing for everyone and W_i was retained only for all infected placebo recipients and for a subcohort of uninfected placebo recipients. For each arm, the ratio of controls to cases was 3:1. The simulated data sets satisfied A1–A3 and A4-NP.

For each of 1000 simulated data sets the MELE $\hat{\beta}$ was computed using the nonparametric approach described in Section 4.3. Then, with $h(x, y) = \log(x/y)$, $\hat{\beta}$ was used to compute the MELEs of $\text{CEP}^{\text{risk}}(j;\ 1)$, AS, and PAE^{ω} for $\omega(j;\ 1) = 1$, j, and $I(j = J = 4)$. Wald tests (with bootstrap standard errors) based on $\widehat{\text{PAE}}^{\omega} - 0.5, \widehat{\text{AS}}$, and T were used to test for any surrogate value. The MELEs of $\text{CEP}^{\text{risk}}(j;\ 1)$, PAE^{ω} and AS performed well (Tables 1 and 2). Bootstrap percentile confidence intervals (CIs) had approximately nominal coverage and for higher values of ρ the MELEs exhibited negligible bias. The tests for any surrogate value had approximately nominal size and showed adequate power to detect surrogate value; the nonparametric trend test had power 0.83, 0.99, and >0.99 for $\rho = 0.5$, 0.7, and 0.9 under scenario (ii).

Additional simulations were conducted to evaluate the performance of the MELE method with binned covariates when the data were generated from a continuous model. Specifically, Step 2 described above was replaced with Step 2′: For vaccine arm subjects, $Y_i(1)$ was generated using model A4-P with $g = \Phi$ and $(\beta_{10},\ \beta_{11},\ \beta_{12},\ \beta_{13}) = (-1.21,\ -0.67,\ 0,\ -0.1)$ set to fit the real vaccine arm data with infection rate 0.067. For the placebo arm, we supposed overall vaccine efficacy of 50% and generated $Y_i(0)$ assuming model A4-P with $\beta_{02} = 0$, $\beta_{03} = \beta_{13}$ and either (i) $\beta_{01} = \beta_{11}$ or (ii) $\beta_{01} = 0$. For $h(x,\ y) = g^{-1}(x) - g^{-1}(y)$, under (i) $\text{CEP}^{\text{risk}}(s_1,\ c;\ \beta) = \beta_{10} - \beta_{00} = -1.21 - (-1.1) = -0.11$ is constant, so that S has no surrogate value (with AS = 0 and $\text{PAE}^{\omega} = 0.5$ for any weight function ω); under (ii) $\text{CEP}^{\text{risk}}(s_1,\ c;\ \beta) = -0.11 - 0.67\ s_1$, so that S has high surrogate value (with AS = 0.67 and $\text{PAE}^{\omega} = 0.88$ for $\omega(j;\ 1) = I(j = J = 4)$). Using $h(x,\ y) = g^{-1}(x) - g^{-1}(y)$, the MELEs and CIs of PAE^{ω} and AS performed well (Table 3). Relative to Table 2, there was a slight increase in bias and in standard errors. Tests for any surrogate value had approximately nominal size, with power only slightly lower than in the previous set of simulations.

Table 1

Model A4-NP simulation results for the nonparametric MELEs $\widehat{CEP}^{\text{risk}}(j, 1; \beta) = \log(\hat{\beta}_{1j}/\hat{\beta}_{0j})$ *for* $j = 1, \ldots, 4$[a]

Cor.		No surrogate value scenario						High surrogate value scenario-				
ρ	Parameter	Bias	SE	SEE	CP	Power	Parameter	Bias	SE	SEE	CP	Power
0.5	$\text{CEP}^{\text{risk}}(1,\ 1) = -0.69$	−0.04	0.42	0.41	0.98	0.45	$\text{CEP}^{\text{risk}}(1,\ 1) = -0.22$	−0.06	0.67	0.65	0.98	0.12
	$\text{CEP}^{\text{risk}}(2,\ 1) = -0.69$	0.11	0.91	0.90	0.99	0.09	$\text{CEP}^{\text{risk}}(2,\ 1) = -0.51$	0.09	0.96	0.93	1.00	0.04
	$\text{CEP}^{\text{risk}}(3,\ 1) = -0.69$	0.13	0.88	0.87	0.99	0.06	$\text{CEP}^{\text{risk}}(3,\ 1) = -0.92$	0.15	0.94	0.93	1.00	0.09
	$\text{CEP}^{\text{risk}}(4,\ 1) = -0.69$	0.09	0.80	0.72	0.98	0.18	$\text{CEP}^{\text{risk}}(4,\ 1) = -1.61$	−0.03	0.65	0.66	0.98	0.66
0.7	$\text{CEP}^{\text{risk}}(1,\ 1) = -0.69$	−0.03	0.30	0.29	0.96	0.62	$\text{CEP}^{\text{risk}}(1,\ 1) = -0.22$	−0.03	0.45	0.47	0.97	0.13
	$\text{CEP}^{\text{risk}}(2,\ 1) = -0.69$	0.09	0.80	0.77	0.99	0.17	$\text{CEP}^{\text{risk}}(2,\ 1) = -0.51$	0.06	0.87	0.84	0.99	0.08
	$\text{CEP}^{\text{risk}}(3,\ 1) = -0.69$	−0.02	0.82	0.79	1.00	0.11	$\text{CEP}^{\text{risk}}(3,\ 1) = -0.92$	−0.02	0.83	0.83	0.99	0.17
	$\text{CEP}^{\text{risk}}(4,\ 1) = -0.69$	0.06	0.73	0.64	0.97	0.22	$\text{CEP}^{\text{risk}}(4,\ 1) = -1.61$	0.00	0.47	0.48	0.96	0.82
0.9	$\text{CEP}^{\text{risk}}(1,\ 1) = -0.69$	0.00	0.19	0.19	0.95	0.90	$\text{CEP}^{\text{risk}}(1,\ 1) = -0.22$	−0.01	0.28	0.27	0.94	0.18
	$\text{CEP}^{\text{risk}}(2,\ 1) = -0.69$	0.02	0.48	0.48	0.96	0.37	$\text{CEP}^{\text{risk}}(2,\ 1) = -0.51$	0.01	0.66	0.59	0.95	0.26
	$\text{CEP}^{\text{risk}}(3,\ 1) = -0.69$	−0.02	0.68	0.63	0.96	0.27	$\text{CEP}^{\text{risk}}(3,\ 1) = -0.92$	0.00	0.62	0.58	0.95	0.40
	$\text{CEP}^{\text{risk}}(4,\ 1) = -0.69$	−0.01	0.53	0.50	0.96	0.32	$\text{CEP}^{\text{risk}}(4,\ 1) = -1.61$	−0.03	0.39	0.36	0.95	0.99

[a] ρ is the linear correlation of the simulated bivariate normal variables latent to the quartilized variables W and $S(1)$. Bias is the median bias. SE is the empirical standard error of $\widehat{\text{CEP}}^{\text{risk}}(j, 1)$. SEE is the median of the bootstrap standard error estimates based on 500 bootstrap replicates. CP is the empirical coverage of bootstrap percentile 95% confidence intervals for $\widehat{\text{CEP}}^{\text{risk}}(j, 1)$. Power refers to power of the Wald test to reject $H_0 : \text{CEP}^{\text{risk}}(j,\ 1) = 0$. 1000 simulations were done to compute the table elements for each model.

Table 2
Model A4-NP simulation results for the nonparametric MELEs $\widehat{\text{PAE}}^{\omega}$ *and* $\widehat{\text{AS}}$*, with* $h(x, y) = \log(x/y)$[a]

Cor.		No surrogate value scenario						High surrogate value scenario				
ρ	Parameter	Bias	SE	SEE	CP	Power	Parameter	Bias	SE	SEE	CP	Power
0.5	$\text{PAE}^{\omega_1} = 0.50$	−0.13	0.22	0.21	0.95	0.03	$\text{PAE}^{\omega_1} = 0.82$	−0.21	0.23	0.23	0.98	0.15
	$\text{PAE}^{\omega_2} = 0.50$	−0.12	0.21	0.20	0.96	0.02	$\text{PAE}^{\omega_2} = 0.84$	−0.18	0.19	0.20	0.97	0.21
	$\text{PAE}^{\omega_3} = 0.50$	0.03	0.21	0.20	0.99	0.04	$\text{PAE}^{\omega_3} = 0.88$	−0.11	0.17	0.19	0.99	0.51
	$\text{AS} = 0.00$	0.07	0.53	0.55	0.99	0.04	$\text{AS} = 1.39$	−0.22	0.70	0.71	0.98	0.51
0.7	$\text{PAE}^{\omega_1} = 0.50$	−0.09	0.19	0.19	0.94	0.02	$\text{PAE}^{\omega_1} = 0.82$	−0.12	0.18	0.20	0.97	0.27
	$\text{PAE}^{\omega_2} = 0.50$	−0.08	0.17	0.17	0.94	0.02	$\text{PAE}^{\omega_2} = 0.84$	−0.10	0.15	0.17	0.97	0.39
	$\text{PAE}^{\omega_3} = 0.50$	0.02	0.20	0.19	0.99	0.04	$\text{PAE}^{\omega_3} = 0.88$	−0.06	0.12	0.14	0.98	0.75
	$\text{AS} = 0.00$	0.04	0.50	0.49	0.99	0.05	$\text{AS} = 1.39$	−0.14	0.51	0.55	0.96	0.70
0.9	$\text{PAE}^{\omega_1} = 0.50$	−0.03	0.13	0.14	0.96	0.02	$\text{PAE}^{\omega_1} = 0.82$	−0.04	0.14	0.15	0.96	0.56
	$\text{PAE}^{\omega_2} = 0.50$	−0.02	0.13	0.14	0.96	0.02	$\text{PAE}^{\omega_2} = 0.84$	−0.04	0.11	0.12	0.96	0.75
	$\text{PAE}^{\omega_3} = 0.50$	0.01	0.19	0.17	0.98	0.08	$\text{PAE}^{\omega_3} = 0.88$	−0.02	0.09	0.10	0.97	0.94
	$\text{AS} = 0.00$	0.02	0.50	0.46	0.98	0.08	$\text{AS} = 1.39$	−0.03	0.45	0.43	0.96	0.94

[a] ρ is the linear correlation of the simulated bivariate normal variables latent to the quartilized variables W and $S(1)$. Bias is the median bias. SE is the empirical standard error of $\widehat{\text{PAE}}^{\omega}$ and $\widehat{\text{AS}}$. SEE is the median of the bootstrap standard error estimates based on 500 bootstrap replicates. CP is the empirical coverage of bootstrap percentile 95% confidence intervals for PAE^{ω} and AS. Power is for one-sided tests of $H_0 : \text{PAE}^{\omega} = 0.5$ versus $H_1 : \text{PAE}^{\omega} > 0.5$ or $H_0 : \text{AS} = 0$ versus $H_1 : \text{AS} > 0$ at level $\alpha = 0.05$. For the PAE weights, $\omega_1(j, 1) = 1$, $\omega_2(j, 1) = j$, and $\omega_3(j, 1) = I[j = J = 4]$. A total of 1000 simulations were done to compute the table elements for each model.

Table 3
Model A4-P (probit) model simulation results for the nonparametric MELEs $\widehat{\text{PAE}}^{\omega}$ *and* $\widehat{\text{AS}}$*, with* $h(x, y) = \Phi^{-1}(x) - \Phi^{-1}(y)$[a]

Cor.		No surrogate value scenario						High surrogate value scenario				
ρ	Parameter	Bias	SE	SEE	CP	Power	Parameter	Bias	SE	SEE	CP	Power
0.5	$\text{PAE}^{\omega_1} = 0.50$	−0.20	0.25	0.23	0.94	0.03	$\text{PAE}^{\omega_1} = 0.82$	−0.25	0.24	0.23	0.96	0.12
	$\text{PAE}^{\omega_2} = 0.50$	−0.19	0.23	0.22	0.94	0.03	$\text{PAE}^{\omega_2} = 0.85$	−0.24	0.22	0.22	0.94	0.15
	$\text{PAE}^{\omega_3} = 0.50$	0.01	0.21	0.21	1.00	0.05	$\text{PAE}^{\omega_3} = 0.88$	−0.17	0.20	0.20	0.97	0.31
	$\text{AS} = 0.00$	0.01	0.29	0.31	1.00	0.03	$\text{AS} = 0.67$	−0.26	0.39	0.36	0.93	0.30
0.7	$\text{PAE}^{\omega_1} = 0.50$	−0.14	0.21	0.21	0.92	0.02	$\text{PAE}^{\omega_1} = 0.82$	−0.14	0.20	0.21	0.96	0.21
	$\text{PAE}^{\omega_2} = 0.50$	−0.14	0.20	0.19	0.92	0.02	$\text{PAE}^{\omega_2} = 0.85$	−0.15	0.17	0.19	0.96	0.28
	$\text{PAE}^{\omega_3} = 0.50$	−0.02	0.21	0.20	0.99	0.04	$\text{PAE}^{\omega_3} = 0.88$	−0.11	0.17	0.17	0.97	0.50
	$\text{AS} = 0.00$	−0.03	0.27	0.26	0.99	0.04	$\text{AS} = 0.67$	−0.22	0.29	0.29	0.91	0.47
0.9	$\text{PAE}^{\omega_1} = 0.50$	−0.06	0.16	0.16	0.92	0.03	$\text{PAE}^{\omega_1} = 0.82$	−0.07	0.16	0.17	0.97	0.45
	$\text{PAE}^{\omega_2} = 0.50$	−0.07	0.15	0.16	0.91	0.02	$\text{PAE}^{\omega_2} = 0.85$	−0.08	0.14	0.15	0.96	0.55
	$\text{PAE}^{\omega_3} = 0.50$	−0.05	0.20	0.18	0.98	0.04	$\text{PAE}^{\omega_3} = 0.88$	−0.05	0.13	0.12	0.96	0.75
	$\text{AS} = 0.00$	−0.08	0.24	0.22	0.98	0.05	$\text{AS} = 0.67$	−0.16	0.22	0.21	0.85	0.76

[a] ρ is the linear correlation of the simulated bivariate normal variables W and $S(1)$. Bias is the median bias. SE is the empirical standard error of $\widehat{\text{PAE}}^{\omega}$ and $\widehat{\text{AS}}$. SEE is the median of the bootstrap standard error estimates based on 500 bootstrap replicates. CP is the empirical coverage of bootstrap percentile 95% confidence intervals for PAE^{ω} and AS. Power is for one-sided tests of $H_0 : \text{PAE}^{\omega} = 0.5$ versus $H_1 : \text{PAE}^{\omega} > 0.5$ or $H_0 : \text{AS} = 0$ versus $H_1 : \text{AS} > 0$ at level $\alpha = 0.05$. For the PAE weights, $\omega_1(j, 1) = 1$, $\omega_2(j, 1) = j$, and $\omega_3(j, 1) = I[j = J = 4]$. A total of 1000 simulations were done to compute the table elements for each model.

These simulation studies provide a "proof of principle" that the proposed methods can reliably estimate the CEP surface and distinguish biomarkers with no or high surrogate value.

6. Discussion

A main use of a surrogate endpoint is predicting treatment effects on a clinical endpoint. Within the principal surrogate framework, we have introduced the CEP surface and the marginal CEP curve as appropriate estimands for measuring the predictive capacity of a candidate surrogate. We developed estimation and testing methods under case–cohort sampling from a single large clinical trial (or multiple similar trials); such inferences apply for measuring surrogate predictiveness for the same or similar setting as the trial. The inferences do not form an empirical basis for bridging information about clinical efficacy to a new setting not represented in the trial(s) (e.g., to a new human population or treatment formulation); for this additional experiments (such as mechanistic studies and studies that deliberately manipulate the biomarker) and metaanalysis of heterogeneous studies are needed.

Because the definition of the CEP surface involves unobservable potential outcomes, strong untestable assumptions may be needed to identify it, possibly precluding its reliable

estimation. The estimation method we developed requires A1–A3, a reasonably good model predicting S from baseline covariates X and W in treatment arm 1, and models for $\text{risk}_{(z)}(s_1, c, x, w)$ or its marginal counterpart $\text{risk}_{(z)}(s_1, x, w)$, for $z = 0, 1$. A1–A2 are standard in blinded randomized trials. A1 (SUTVA) is potentially dubious in the infectious disease setting where dependent happenings are possible (Halloran and Struchiner, 1995), but should approximately hold in trials with a small study population relative to the total population of at risk individuals. A3 can be assessed by testing $H_0^x : \Pr(V = 1 \mid Z = 1, X = x) = \Pr(V = 1 \mid Z = 0, X = x)$ at each of multiple fixed baseline covariate levels x, where rejecting H_0^x for any x rejects A3. It is difficult to fully verify A3, however, due to the curse of dimensionality. The method is expected to be robust to violations of A3 if the vast majority of clinical events happen after the biomarker measurement time. Otherwise it will be important to extend the methods to facilitate sensitivity analyses to departures from A3.

Models for the conditional distribution of S given X and W can be directly checked using arm $Z = 1$ data, and under A1–A3 models for $\text{risk}_{(1)}(s_1, c, x, w)$ can be tested. The model for $\text{risk}_{(0)}(s_1, c, x, w)$ specified by A4-P or A4-NP is not testable; however, with extra data collection Follmann's (2006) "close-out placebo vaccination" approach would provide one way to test it. Given the challenge in verifying this assumption, sensitivity analysis and the use of multiple surrogate evaluation approaches is warranted.

Within the principal surrogate framework considered here, internal validity of the putative surrogate can be checked by comparing the estimated overall clinical treatment effect, $\widehat{\text{CE}} \equiv h(\widehat{\Pr}(Y(1) = 1), \widehat{\Pr}(Y(0) = 1))$, to the CE predicted from the biomarker. Under A1–A3 and case CB, CE can be predicted by $\text{Pred(CE)} = \int \widehat{\text{CEP}}^{\text{risk}}(s_1, c)\, d\hat{F}_{(1)}(s_1)$, which averages the predicted clinical treatment effect over the distribution of observed marker values of subjects assigned arm $Z = 1$. Furthermore, the estimated CEP curve can be used to check projective validity, that is, the utility of S for bridging efficacy predictions across populations. For example, suppose treatments $Z = 1$ and $Z = 0$ are compared within two subroups of a large trial. The CEP surface can be estimated from subgroup 1 data, and Pred(CE) calculated by estimating $F_{(1)}(\cdot)$ from the observed biomarker values S of subgroup 2 subjects in arm $Z = 1$. Then projective validity would be supported by Pred(CE) near $\widehat{\text{CE}}$ for subgroup 2.

The estimands and estimation techniques developed here for a binary clinical endpoint Y also apply for a quantitative clinical endpoint Y, with all expressions $\Pr(Y(Z) = 1 \mid \cdot)$ replaced with $E(Y(Z) \mid \cdot)$. In either case the CEP estimands describe how the average or population level causal effect on Y depends on the causal effect on S.

7. Supplementary Materials

Web Appendices referenced in Sections 2.1, 3.1, 3.2, and 4.3 are available under the Paper Information link at the *Biometrics* website `http://www.tibs.org/biometrics`. `R` code for the nonparametric method is also available at the *Biometrics* website.

Acknowledgements

The authors thank Dean Follmann, Margaret Pepe, Ross Prentice, Steve Self, and the associate editor for helpful comments. This work was supported by NIH grants 2 R01 AI54165-04 and 5 R37 AI029168-16.

References

Breslow, N. and Day, N. (1980). *Statistical Methods in Cancer Research, Volume 1*. Lyon, France: International Agency for Research on Cancer.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54,** 1014–1029.

Chan, I., Shu, L., Matthews, H., Chan, C., Vessey, R., Sadoff, J., and Heyse, J. (2002). Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine* **21,** 3411–3430.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics* **62,** 1161–1169.

Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Freedman, L., Graubard, B., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11,** 167–178.

Gilbert, P., Peterson, M., Follmann, D., et al. (2005). Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases* **191,** 666–677. **Q1**

Halloran, M. and Struchiner, C. (1995). Causal inferences in infectious diseases. *Epidemiology* **6,** 142–151.

Huang, Y., Pepe, M., and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics*, in press. **Q2**

Pepe, M. and Fleming, T. (1991). A non-parametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86,** 108–113.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

Prentice, R. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* **8,** 431–440.

Robins, J. (1995). An analytic method for randomized trials with informative censoring: Part I. *Lifetime Data Analysis* **1,** 241–254.

Rubin, D. (1986). Statistics and causal inference: Which ifs have causal answers. *Journal of the American Statistical Association* **81,** 961–962.

Taylor, J., Wang, Y., and Thibaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61,** 1102–1111.

Weir, C. and Walley, R. (2006). Statistical evaluation of biomarkers as surrogate endpoints: A literature review. *Statistics in Medicine* **25,** 183–203.

Received July 2006. *Revised December* 2007.
Accepted December 2007.

Queries

Q1 Author: As per journal style all author names need to be provided (if less than or equal to 12). Please confirm that this rule has been followed in reference Gilbert et al. (2005).

Q2 Author: Please update reference Huang et al. (2007).