UW Biostatistics Working Paper Series

7-6-2006

Evaluating Causal Effect Predictiveness of Candidate Surrogate Endpoints

Peter B. Gilbert
Fred Hutchinson Cancer Research Center & University of Washington, pgilbert@scharp.org

Michael Hudgens
University of North Carolina, Chapel Hill, mhudgens@bios.unc.edu

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 specific example motivating this work, an objective of placebo-controlled preventive HIV vaccine efficacy trials is the evaluation of various vaccine-induced immune responses as surrogate endpoints for HIV infection. Development of such a surrogate is a central goal of vaccine research; for example the Foundation of the NIH and the Gates Foundation list it as one of the 14 "Grand Challenges in Global Health." An immunological surrogate would be useful for several purposes including guiding iterative development of immunogens between basic and clinical research, guiding regulatory decisions and public immunization policy, and bridging efficacy of a vaccine observed in a trial to a new setting.

Statistical methods for evaluating surrogate endpoints has emerged as an important research area (Weir and Walley, 2006). This 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) observed that this definition is based on observable random variables, and named a biomarker satisfying criterion (i) a "statistical surrogate." Since 1989, most of the surrogate-evaluation methods have been designed to check if a biomarker is a statistical surrogate. These methods include estimation of the proportion of the treatment effect explained (Freedman et al., 1992; Lin, Fleming, and DeGruttola, 1997; Wang and Taylor, 2002) and of the relative effect and adjusted association (Buyse and Molenberghs, 1998), as well as meta-analysis (Daniels and Hughes, 1997; Buyse et al., 2000; Gail, 2000).

Treatment effects adjusted for a variable measured after randomization (called *net effects*) are susceptible to post-randomization selection bias (e.g., Rosenbaum, 1984; Robins and Greenland, 1992; Hudgens, Hoering, and Self, 2003). Since potential 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 introduced a potential outcomes framework for evaluating "principal surrogates," but statistical methods for doing so 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 only work in this area of which we are aware is Taylor et al.'s (2005) summary measure of surrogate quality.

Here we develop an approach for evaluating a principal surrogate from a single large clinical trial, which to our knowledge constitutes the first such method. Following Follmann (2006), our approach uses baseline covariates to predict missing potential biomarker outcomes. After defining and comparing statistical and principal surrogates in Section 2, in Section 3 we introduce a *causal effect predictiveness surface*, plus associated summary parameters, which serve as appropriate estimands for quantifying how well a biomarker predicts population level causal clinical treatment effects. Motivated by the problem of assessing an immune response to an HIV vaccine as a surrogate endpoint for HIV infection, in Section 4 we consider the important special case where the biomarker has no variability in one of the treatment arms. For this setting we develop an estimated likelihood-based method for estimating the causal estimands based on case-cohort sampling of the biomarkers. In Section 5 we evaluate the method in simulations based on a vaccine trial, and in Section 6 we conclude with discussion.

2. Comparison of Statistical and Principal Surrogates

2.1 *Statistical Surrogates are Based on Net Effects, not Causal Effects*

Throughout we consider a randomized trial with treatment assignment Z ($Z = 1$ or 0), a biomarker endpoint 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 potential 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 $S = *$). Following FR, methods for evaluating statistical surrogates are based on comparing the risk distributions

$$\begin{aligned} risk(s|Z=1) &\equiv Pr(Y^{obs}=1|Z=1, V^{obs}=1, S^{obs}=s) \quad \text{and} \\ risk(s|Z=0) &\equiv Pr(Y^{obs}=1|Z=0, V^{obs}=1, S^{obs}=s), \end{aligned}$$

where *obs* indicates the variable is observed. FR defined S to be a *statistical surrogate* if, for all fixed values s of S, $risk(s|Z = 1) = risk(s|Z = 0)$. The full mediation criterion (i) requires that a treatment effect on S^{obs} is necessary and sufficient for a treatment effect on Y^{obs}; statistical surrogacy is the necessity part of (i).

Because S and V are measured after randomization, a comparison of $risk(s|Z = 1)$ and $risk(s|Z = 0)$ measures treatment differences due to a mixture of the causal treatment effect and any differences in characteristics between treatment 1 subjects who have response level s, $\{Z = 1, V^{obs} = 1, S^{obs} = s\}$, and treatment 0 subjects who have response level s, $\{Z = 0, V^{obs} = 1, S^{obs} = s\}$ (i.e., the net effect). If there is no treatment effect on S, then the net effect may approximate the causal effect.

Biomarkers of interest are usually affected by treatment, however, and the greater the treatment effect on S, the greater the anticipated discrepancy between the net effect and the causal effect of interest. FR concluded that because the statistical surrogate definition is based on net effects, employing it for evaluating a surrogate may mislead about the biomarker's capacity for reliably predicting clinical treatment effects.

2.2 Evaluating Statistical Surrogates Based on the PTE

Freedman et al. (1992) introduced the *proportion of treatment effect explained (PTE)* as a quantitative measure of the quality of a biomarker as a statistical surrogate, and several methods for evaluating surrogates have been developed based on the PTE (Lin, Fleming, and DeGruttola, 1997; Wang and Taylor, 2002). To define the PTE, consider two generalized linear models:

$$\begin{aligned} g_Y\{E(Y_i^{obs}|Z_i, V_i^{obs}=1)\} &= \beta_0 + \beta_1 Z_i, \\ g_Y\{E(Y_i^{obs}|Z_i, V_i^{obs}=1, S_i^{obs})\} &= \theta_0 + \theta_1 Z_i + \theta_2 S_i^{obs}, \end{aligned}$$

where $g_Y\{\cdot\}$ is a known link function. For a binary clinical endpoint Y, Freedman et al. (1992) defined the PTE as $PTE \equiv 1 - \theta_1/\beta_1$, which equals

$$1 - \frac{g_Y\{E(Y_i^{obs}|Z_i=1, V_i^{obs}=1, S_i^{obs})\} - g_Y\{E(Y_i^{obs}|Z_i=0, V_i^{obs}=1, S_i^{obs})\}}{g_Y\{E(Y_i^{obs}|Z_i=1, V_i^{obs}=1)\} - g_Y\{E(Y_i^{obs}|Z_i=0, V_i^{obs}=1)\}}.$$

A perfect statistical surrogate has $PTE = 1$, which means there is a treatment effect on the clinical endpoint ($\beta_1 \neq 0$) and no net treatment effect after adjusting for the observed surrogate ($\theta_1 = 0$). The latter condition is implied by $risk(s|Z=1) = risk(s|Z=0)$ for all fixed values s of S, showing that FR's definition of a statistical surrogate implies $PTE = 1$. The numerator θ_1 of the PTE is a net effect, whereas under standard assumptions A1 and A2 made for a randomized trial given below, the denominator β_1 is a causal effect. Consequently, the common description of the PTE as a measure of the amount of the clinical treatment effect mediated through the surrogate seems misleading, because "mediation" should only reflect a causal effect. The post-randomization bias inherent in the PTE suggests that alternative summary measures, based solely on causal effects, should be considered.

2.3 Definition of a Principal Surrogate Endpoint

We introduce the potential outcomes notation (Rubin, 1986) and assumptions that will be used for defining and identifying the causal estimands of interest. For subject i, let $Y_i(Z)$ be the potential clinical endpoint (i.e., disease) under assignment to treatment Z, $Z = 0, 1$. Similarly define potential outcomes $S_i(Z)$ for the biomarker endpoint, which is measured at time t_0 after treatment assignment, and let $V_i(Z)$ be the potential indicators of whether the ith subject is disease free at t_0. Note that $S_i(Z)$ is undefined if $V_i(Z) = 0$; in this case $S_i(Z) = *$. We suppose that $(V_i(1), V_i(0), S_i(1), S_i(0), Y_i(1), Y_i(0)), i = 1, \ldots, n$, are iid, and for simplicity assume no drop-out. We also make the following assumptions A1, A2 (Rubin 1986), and A3.

A1 *Stable Unit Treatment Value Assumption (SUTVA)*

A2 *Ignorable Treatment Assignments*: Z_i is independent of $(V_i(1), V_i(0), S_i(1), S_i(0), Y_i(1), Y_i(0))$ for all i

A3 *Equal Individual Clinical Risk Up to Time* t_0: $V_i(1) = 1$ if and only if $V_i(0) = 1$.

A1 states that the potential outcomes $(V_i(1), V_i(0), S_i(1), S_i(0), Y_i(1), Y_i(0))$ for each subject are independent of the treatment assignments of other subjects, which implies so-called "consistency", $(V_i(Z_i), S_i(Z_i), Y_i(Z_i)) = (V_i^{obs}, S_i^{obs}, Y_i^{obs})$. A2 holds for randomized and blinded trials. A3 will be needed for identifying the causal estimand based on data from subjects observed to be at risk at t_0. This assumption will approximately hold if the risk of disease is the same in the two arms up to t_0, or if most subjects are at risk for disease at t_0, e.g., if t_0 is near baseline. A1-A3 often hold in our motivating application. In particular, A3 should approximately attain in the two ongoing HIV vaccine efficacy trials, since the candidate immunological surrogate endpoints are measured near baseline (at $t_0 = 8$ weeks) (Mehrotra, Li, and Gilbert, 2006).

With these preliminaries, we now define a principal surrogate endpoint. FR suggested that a surrogate S should satisfy the following property:

Causal Necessity: S is necessary for the effect of treatment on the outcome Y in the sense that an effect of treatment on Y can occur only if an effect of treatment on S has occurred. At the individual level, this means that $S_i(1) = S_i(0)$ implies $Y_i(1) = Y_i(0)$.

A population level definition of Causal Necessity, which is used in our approach to surrogate evaluation, is given below.

FR defined the *basic principal stratification* P_0 with respect to the post-randomization 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. Estimands that condition on a principal stratification are causal because, by construction, the stratification is unaffected by treatment. FR defined a biomarker S to be a principal surrogate endpoint if, for all fixed $s_1 = s_0$, the comparison between

$$\begin{aligned} risk_{(1)}(s_1, s_0) &\equiv \Pr(Y_i(1) = 1 | V_i(1) = 1, V_i(0) = 1, S_i(1) = s_1, S_i(0) = s_0) \quad \text{and} \\ risk_{(0)}(s_1, s_0) &\equiv \Pr(Y_i(0) = 1 | V_i(1) = 1, V_i(0) = 1, S_i(1) = s_1, S_i(0) = s_0) \end{aligned}$$

results in equality. FR did not explicitly condition on $V_i(1) = V_i(0) = 1$ in their definition; however implicitly they must have, since $(S_i(1), S_i(0))$ is only defined if $V_i(1) = V_i(0) = 1$. Henceforth, for brevity all probability statements that involve $S_i(1)$ and $S_i(0)$ are implicitly intersected with $\{V_i(1) = V_i(0) = 1\}$. A contrast in $risk_{(1)}(s_1, s_0)$ and $risk_{(0)}(s_1, s_0)$ measures a population level or average causal effect on

Y for subjects with $\{S_i(1) = s_1, S_i(0) = s_0\}$. Thus with FR's definition S is a principal surrogate if subjects with no causal effect on the biomarker have no average causal effect on the clinical endpoint. This is a population version of Causal Necessity, which we call Average Causal Necessity. For reference, we define this property as follows.

Average Causal Necessity: $risk_{(1)}(s_1, s_0) = risk_{(0)}(s_1, s_0)$ for all fixed $s_1 = s_0$.

Biomarkers with the greatest utility for predicting clinical treatment effects will not only be necessary for a clinical effect, but also sufficient. Causal Sufficiency can be defined as follows:

Causal Sufficiency: S is sufficient for the effect of treatment on the outcome Y in the sense that an effect of treatment on S implies an effect of treatment on Y. At the individual level, this means that $S_i(1) \neq S_i(0)$ implies $Y_i(1) \neq Y_i(0)$.

Often Causal Sufficiency is at least as important scientifically as Causal Necessity. For example, knowing that an antibody titer > 1000 is sufficient for a vaccine to protect an individual 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: $risk_{(1)}(s_1, s_0) = risk_{(0)}(s_1, s_0)$ for all fixed $s_1 \neq s_0$,

and suggest a refined definition of a principal surrogate endpoint:

Principal Surrogate Endpoint: A biomarker S that satisfies both Average Causal Necessity and Average Causal Sufficiency as defined above.

Heretofore we use this definition of a principal surrogate endpoint.

Evaluating a principal surrogate is challenging due to the missing data, which results from observing only one of $(V_i(1), S_i(1), Y_i(1))$ or $(V_i(0), S_i(0), Y_i(0))$ from each subject (Holland, 1986). At present, it is not clear when this missing data problem can be satisfactorily overcome to provide practically useful inferential tools for evaluating principal surrogates. Inaccurate modeling of the missing data could lead to bias in assessing whether a biomarker is a principal surrogate, and it is unclear when this bias will exceed that inherent in the definition of a statistical surrogate. However, there are particular settings where it is auspicious to solve the missing data problem under assumptions that are all plausible or testable, in which case it is relatively easy to evaluate a principal surrogate. In Sections 3 and 4, we develop an evaluation method for one such setting: where the biomarker has no variation for one treatment arm.

2.4 Illustration of Statistical versus Principal Surrogates

To illustrate the more useful scientific interpretation of a principal than statistical surrogate, we consider a placebo-controlled vaccine trial where Y is infection and S is a binary, taking values positive or negative immune response (vaccine "take" or not). We suppose $S_i(0) = 0$ for all i. The top half of Table 1 presents a perfect principal surrogate, wherein subjects in the "not take" principal stratum have a 30% chance of becoming infected under either assignment vaccine or placebo (0% protection), and subjects in the "take" stratum have a 0% chance of becoming infected under vaccine assignment and a 15% chance under placebo assignment (100% protection). Therefore the vaccine effect on the immune response predicts perfectly whether a subject is protected, and S is a perfect principal surrogate. However, S is not a statistical surrogate, because for subjects with $S_i^{obs} = 0$, the probabilities of infection $\Pr(Y_i^{obs} = 1|S_i^{obs} = 0, Z_i = z)$ for vaccine and placebo recipients are unequal (0.3 for $Z = 1$ and 0.2 for $Z = 0$). Thus the definition of a statistical surrogate misses the predictive capacity of S (a "false negative"). The bottom half of Table 1 presents an immune response that does not predict whether a subject is protected at all yet is a statistical surrogate (a "false positive"). The statistical surrogate definition fails in these examples because of the causal vaccine effect on S, with 67% versus 0% responders in the vaccine and placebo arms, and the large amount of selection bias that is reflected in the net effect. This bias could arise because vaccine recipients who fail to mount an immune response have relatively weak immune systems, which places them at high risk for infection.

3. Causal Effect Predictiveness Surface

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", with a 'good' surrogate having large associative effects and small dissociative effects. As defined in equations 5.3 and 5.4 of FR, an *associative effect* is a comparison between the ordered sets

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

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

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

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

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

be this contrast conditional on $\{S_i(1) = s_1, S_i(0) = s_0\}$. Note that $CEP^{risk}(s, s) = 0$ for all s is a population version of no dissociative effects, and is equivalent to Average Causal Necessity, whereas $CEP^{risk}(s_1, s_0) \neq 0$ for all $s_1 \neq s_0$ is a population version of 100% associative effects, and is equivalent to Average Causal Sufficiency. Therefore the criteria for a principal surrogate can be checked by estimating the CEP surface. Moreover, biomarkers with capacity to predict clinical treatment effects will usually have $|CEP^{risk}(s_1, s_0)|$ monotone in $|s_1 - s_0|$, reflecting the situation that on average 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*, which can be quantified by both the nearness of $|CEP^{risk}(s_1, s_0)|$ to 0 for s_1 near s_0, and by the extent to which $|CEP^{risk}(s_1, s_0)|$ increases with $|s_1 - s_0|$.

The CEP surface can alternatively be defined in terms of percentiles of the marker S. To formulate this, consider Huang et al.'s (2006) proposal to judge the value of a marker S for predicting disease Y by the *predictiveness curve*, $R(v) \equiv Pr(Y^{obs} = 1|F(S^{obs}) = v), v \in [0, 1]$, where F is the cdf of S^{obs}. If F^{-1} exists, then

$$R(v) = Pr(Y^{obs} = 1|S^{obs} = F^{-1}(v)) = risk(S^{obs} = F^{-1}(v)),$$

i.e., $R(v)$ is risk as a function of the quantiles of S^{obs}, which provides a common scale for comparing multiple markers. If we assume $R(v)$ is a monotone increasing function of v, then $R(v) = p$ implies v percent of the population have risk less than or equal to p. The predictiveness curve $R(v)$ usefully informs about both absolute risks at different marker quantiles and the frequency of these risks in the population. Huang, Pepe, and Feng (2006) proposed plotting an estimate of $R(v)$ versus v as a graphical tool for assessing and comparing the predictiveness of markers. 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, $CEP^R(v_1, v_0) \equiv h(R_{(1)}(v_1, v_0), R_{(0)}(v_1, v_0))$, where

$$\begin{aligned} R_{(1)}(v_1, v_0) &\equiv Pr(Y(1) = 1|S(1) = F_{(1)}^{-1}(v_1), S(0) = F_{(1)}^{-1}(v_0)) \quad \text{and} \\ R_{(0)}(v_1, v_0) &\equiv Pr(Y(0) = 1|S(1) = F_{(1)}^{-1}(v_1), S(0) = F_{(1)}^{-1}(v_0)). \end{aligned}$$

In this definition, $S(1)$ and $S(0)$ are standardized relative to the distribution $F_{(1)}$ of

$S(1)$. Using the same reference distribution for $S(1)$ and $S(0)$ makes the marker values under assignment to the two arms comparable, ensuring $S(1) = S(0)$ if and only if $v_1 = v_0$. With $h(x, y) = x - y$, the volume between $CEP^R(\cdot, \cdot)$ and the zero-plane equals $CE = Pr(Y(1) = 1) - Pr(Y(0) = 1)$. The nearer $|CEP^R(v_1, v_0)|$ is to zero for $|v_1 - v_0|$ near zero and the larger $|CEP^R(v_1, v_0)|$ is for large $|v_1 - v_0|$, the greater the causal treatment effect on S is predictive of the average causal treatment effect on Y.

To illustrate the interpretation of $CEP^R(v_1, v_0)$, we consider the unidirectional situation where interest is in assessing if higher responses of S if assigned treatment 1 ($S_i(1) > S_i(0)$) predict clinical benefit of treatment 1. For example, this situation might occur in trials of active treatment 1 versus placebo 0. In Figure 1(i), the fact that $CEP^R(v_1, v_0) = CE$ for all (v_1, v_0) indicates the biomarker has no surrogate value. In contrast, in Figure 1(ii) $CEP^R(v_1, v_0) = 0$ for all $v_1 \leq v_0$ and $|CEP^R(v_1, v_0)|$ is monotone in $v_1 - v_0$ with large amount of increase, reflecting a biomarker with high surrogate value.

Next, we consider the interpretation of the CEP surface in the special case where $S_i(0)$ is constant. We refer to this case as A4:

A4 *Uniform Biomarkers*: $S_i(0) = c$ for all i for some constant c

HIV vaccine trials fit case A4, because S is an HIV-specific immune response, which will be 0 for all subjects in the placebo arm $Z = 0$, since vaccine antigens must be presented to the immune system to induce a response (Gilbert et al., 2005). Under A4 the $CEP^{risk}(s_1, c)$ surface is a curve in s_1 and the $CEP^R(v_1, F_{(1)}(c))$ surface is a curve in v_1. The dissociative effect can be measured by $CEP^R(F_{(1)}(c), F_{(1)}(c))$, and the associative effects by $CEP^R(v_1, F_{(1)}(c))$ for $v_1 \neq F_{(1)}(c)$. For example, with $c = L$ the lower bound of S (e.g., an assay detection limit), the nearer $CEP^R(F_{(1)}(c), F_{(1)}(c))$ is to zero and the greater the increase of $|CEP^R(v_1, F_{(1)}(c))|$ with $v_1 > F_{(1)}(c)$, the greater the surrogate value of the biomarker (Figure 2). This kind of plot provides an interpretable way to compare the surrogate value of multiple biomarkers.

3.2 Summary Measures of Surrogate Value

We suggest parameters that summarize the surrogate value of a biomarker, which are functionals of the CEP surface. Again we consider the situation where interest is in assessing whether $S_i(1) > S_i(0)$ predicts clinical benefit of treatment 1 ($Y_i(1) = 0$ and $Y_i(0) = 1$). To summarize the asociative and dissociative effects, we consider the *expected associative effect (EAE)* and the *expected dissociative effect (EDE)*:

$$\begin{aligned} EAE(w) &\equiv E[w(S(1), S(0))CEP^{risk}(S(1), S(0))|S(1) > S(0)] \quad (1) \\ EDE &\equiv E[CEP^{risk}(S(1), S(0))|S(1) \leq S(0)], \quad (2) \end{aligned}$$

where $w(\cdot, \cdot)$ is a known nonnegative weight function. The $EAE(w)$ can equivalently be written as $EAE(w) = \int_{v_1 > v_0} w(v_1, v_0)CEP^R(v_1, v_0)dv_1 dv_0/Pr(S_i(1) > S_i(0))$ and similarly for EDE. Thus $EAE(w = 1)$ is the volume between $CEP^R(v_1, v_0)$ and the zero-plane in the $v_1 > v_0$ quadrant divided by $Pr(S_i(1) > S_i(0))$, and EDE is the volume between $CEP^R(v_1, v_0)$ and the zero-plane in the $v_1 \leq v_0$ quadrant divided by $Pr(S_i(1) \leq S_i(0))$.

We also define the *proportion associative effect* by

$$PAE(w) \equiv \frac{|EAE(w)|}{|EDE| + |EAE(w)|}. \quad (3)$$

The $PAE(w)$ is the magnitude of the expected associative effect relative to the combined magnitude of the expected associative effect and the expected dissociative effect. Values $PAE(w) \leq 0.5$ suggest the biomarker has no surrogate value, while values in $(0.5, 1]$ suggest some surrogate value.

A weight function is included in $EAE(w)$, and thus $PAE(w)$, to allow the parameters to reflect the idea that a biomarker with high surrogate value should have large $|CEP^{risk}(s_1, s_0)|$ for large $|s_1 - s_0|$. For example, weights $w(s_1, s_0) = |s_1 - s_0|$ 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, $PAE(w)$ compares the clinical effect among groups with the maximum surrogate effect and with no surrogate effect: $PAE(w) = |CEP^R(1, 0)|/ [|EDE| + |CEP^R(1, 0)|]$.

If $h(x, y) = x - y$, $Pr(S_1(1) > S_i(0)) = 0.5$, 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 $PAE(w = 1)$ equals the *proportion associative (PA)*, defined by

$$PA \equiv \frac{Pr(S_i(1) > S_i(0), Y_i(1) = 0, Y_i(0) = 1)}{Pr(Y_i(1) = 0, Y_i(0) = 1)}.$$

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. Note that the PA depends on the underlying principal strata distribution $F_{(1),(0)}(s_1, s_0) = Pr(S(1) \leq s_1, S(0) \leq s_0)$; if $Pr(S_i(1) > S_i(0))$ is small (large) then the PA will tend to be small (large), irrespective of the surrogate value of the biomarker. By conditioning on $(S_i(1), S_i(0))$, the $PAE(w)$ is designed to be robust to $F_{(1),(0)}(\cdot, \cdot)$; the $PAE(w)$ reflects the relative magnitude of clinical effects for those with and without surrogate effects.

Note that biomarkers satisfying Average Causal Necessity have $EDE = 0$ and thus $PAE(w) = 1$, in which case $EAE(w)$ contributes no information in the $PAE(w)$. Therefore other summary measures are needed to compare multiple biomarkers satisyfing Average Causal Necessity, and more generally for better summarizing the magnitude of associative effects. The $EAE(w)$ itself may be useful for this purpose, as may contrasts of $|EAE(w)|$ with $|EDE|$ other than the $PAE(w)$. For example, with $w(s_1, s_0) = I(s_1 = U, s_0 = L)$, the difference $|EAE(w)| - |EDE|$ equals $AS \equiv |CEP^R(1, 0)| - |EDE|$, which we refer to as the *associative span (AS)*. Table 1 illustrates $EAE(w=1)$, EDE, $PAE(w=1)$, and AS for two hypothetical biomarkers S. The first has high (in fact perfect) surrogate value, with $PAE(w=1) = AS = 1$, and the second has no surrogate value, with $PAE(w=1) = 0.5$ and $AS = 0$.

While the summary parameters may be useful, in general it is important to estimate the CEP surface over the whole range of marker values or quantiles, to provide a full picture of the associative and disssociative effects.

4. Estimating the Causal Effect Predictiveness Surface

4.1 Identifiability of the Causal Effect Predictiveness Surface

Due to missing potential outcomes the CEP surface is not identified without further assumptions. A1-A3 imply

$$\begin{aligned} risk_{(1)}(s_1, s_0) &= \Pr\{Y_i^{obs} = 1 | Z_i = 1, V_i^{obs} = 1, S_i^{obs} = s_1, S_i(0) = s_0\} \quad \text{and} \\ risk_{(0)}(s_1, s_0) &= \Pr\{Y_i^{obs} = 1 | Z_i = 0, V_i^{obs} = 1, S_i(1) = s_1, S_i^{obs} = s_0\}, \end{aligned}$$

demonstrating that $risk_{(1)}(s_1, s_0)$ would be identified if the $S_i(0)$'s of arm $Z = 1$ subjects were known, and similarly $risk_{(0)}(s_1, s_0)$ would be identified if the $S_i(1)$'s of arm $Z = 0$ subjects were known. Estimating the CEP surface will therefore require a study design and plausible assumptions that provide a way to predict the missing potential biomarker outcomes. While generally challenging, these requirements are attainable in the important special case A4. Under A4 the joint values $(S_i(1), S_i(0))$ are observed or known for all subjects in arm $Z = 1$, so that $risk_{(1)}(s_1, c) = risk(s_1 | Z = 1)$, i.e., $risk_{(1)}(s_1, c)$ is identified by the observed data in arm $Z = 1$. However, $risk_{(0)}(s_1, c)$ is still not identified, and the remaining task to identify the CEP surface entails determining values $S_i(1)$ for arm $Z = 0$ subjects.

An additional advantage under A4 is that the Average Causal Necessity criterion is greatly simplified, to $CEP^{risk}(c, c) = 0$. Thus the CEP surface only has to be estimated at a single biomarker value to check this property. Furthermore, in case A4 it is difficult to evaluate a statistical surrogate, because it is not possible to study the correlation of S_i^{obs} with Y_i^{obs} in arm $Z = 0$ subjects, and it is conceptually difficult to evaluate whether S fully mediates clinical treatment effects (Chan et al., 2002).

4.2 Baseline Predictor Study Design and Likelihood

Under A1-A4 and a standard clinical trial design with S binary, recently developed methods (for a different application) provide estimators of $CEP^{risk}(0,0)$ and $CEP^{risk}(1,0)$, as well as of $PAE(w)$ and AS (Hudgens and Halloran, 2006; Shepherd et al., 2006). These sensitivity analysis methods posit a class of non-identified models for the post-randomization selection bias, and repeat the estimation under each model. Alternatively, in the current work, for S continuous or categorical we leverage an innovative trial design to develop a non-sensitivity analysis approach for estimating the CEP surface. Throughout we assume A1-A4 and that the constant value c for $S_i(0)$ is the realized lower bound L of the biomarker $S(1)$, $c = L = min\{S_i(1)\} = min\{S_i^{obs}|Z_i = 1\}$.

The estimation approach is based on Follmann (2006), who proposed augmented vaccine trial designs for discerning if an immunological correlate of HIV infection risk causatively impacts infection risk. Follmann did not develop this work as a method for evaluating a principal surrogate, and here we show how it can be built upon to provide a technique for estimating the CEP surface. Follmann proposed two techniques for predicting $S(1)$ for arm $Z = 0$ subjects, of which we consider the first, wherein a baseline covariate vector W that is predictive of $S(1)$ is measured in subjects in both treatment arms. The correlation of W and $S(1)$ observed in subjects assigned arm $Z = 1$ is used to predict $S(1)$ for subjects in arm $Z = 0$. A1-A3 imply $S(1)|Z = 1, V^{obs} = 1, W =^d S(1)|Z = 0, V^{obs} = 1, W$, ensuring validity of this procedure. Several potential baseline predictors are being collected in the ongoing HIV vaccine efficacy trials (Mehrotra, Li, and Gilbert, 2006).

To develop an estimation procedure using the baseline predictor W, we assume W does not predict clinical risk after accounting for $S(1)$:

A5: $Y(Z)|W, S(1) =^d Y(Z)|S(1), \qquad Z = 0, 1.$

We consider a case-cohort sampling design, in which a sub-sample of trial participants is selected for measurement of W_i, which includes all cases and a "sub-cohort" of controls. The biomarker S_i^{obs} is measured for all arm $Z_i = 1$ subjects for whom W_i is measured. Case-cohort sampling is efficient when W_i or S_i is an expensive covariate (Prentice, 1986). For HIV vaccine trials, $S_i(1)$ (and likely components of W_i) can be measured after the trial using stored blood samples (Gilbert et al., 2005).

Let δ_i indicate whether W_i is measured. We observe iid data $O_i \equiv (Z_i, V_i^{obs}, Y_i^{obs}, \delta_i, \delta_i W_i, \delta_i Z_i S_i^{obs}), i = 1, \ldots, n$. Subjects i with $V_i^{obs} = 1$ contribute terms to the likelihood. For subjects with $Z_i \delta_i = 1$, $Pr(Y_i^{obs} = 1|Z_i = 1, V_i^{obs} = 1, S_i^{obs}) = risk_{(1)}(S_i^{obs}, 0; \beta)$, where $risk_{(1)}(S_i^{obs}, 0; \beta)$ 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 $risk_{(0)}(S_i(1), 0; \beta)$ over the conditional cdf $G^{S|W}$ of $S(1)|W$, $Pr(Y_i^{obs} = 1|Z_i = 0, V_i^{obs} = 1, W_i) = \int risk_{(0)}(s_1, 0; \beta) dG^{S|W}(s_1|W_i)$; note that A5 is used here. Subjects with $\delta_i = 0$ contribute $Pr(Y_i^{obs} = 1|Z_i, V_i^{obs} = 1) = \int risk_{(Z_i)}(s_1, 0; \beta) dG^S(s_1)$, where G^S is the cdf of $S(1)$. Thus the likelihood is $L(\beta, G^{S|W}, G^S) \equiv \prod_{i=1}^n f(O_i)^{V_i^{obs}}$, where $f(O)$ equals

$$\left\{risk_{(1)}(S^{obs}, 0; \beta)^{Y^{obs}} (1 - risk_{(1)}(S^{obs}, 0; \beta))^{1-Y^{obs}}\right\}^{Z\delta}$$
$$\times \left\{ \left(\int risk_{(0)}(s_1, 0; \beta) dG^{S|W}(s_1|W) \right)^{Y^{obs}} \left(1 - \int risk_{(0)}(s_1, 0; \beta) dG^{S|W}(s_1|W) \right)^{1-Y^{obs}} \right\}^{(1-Z)\delta}$$
$$\times \left\{ \left(\int risk_{(Z)}(s_1, 0; \beta) dG^{S}(s_1) \right)^{Y^{obs}} \left(1 - \int risk_{(Z)}(s_1, 0; \beta) dG^{S}(s_1) \right)^{1-Y^{obs}} \right\}^{(1-\delta)}.$$

Since $CEP^{risk}(s_1, 0; \beta)$ depends on β but not $G^{S|W}$ and G^S, these cdfs are nuisance parameters. Although profile likelihood is thus a natural approach to pursue, it is difficult to implement because the likelihood integrates over $G^{S|W}$ and G^S. An alternative approach would estimate $(\beta, G^{S|W}, G^S)$ by full maximum likelihood; however this would require specification of the joint distribution of $(W, S(1))$ and complex numerical integration. We use estimated likelihood (Pepe and Fleming, 1991), wherein consistent estimates of $G^{S|W}$ and G^S are obtained based on treatment arm 1 data, and then $L(\beta, \widehat{G}^{S|W}, \widehat{G}^S)$ is maximized in β. The bootstrap can be used to get standard errors for $\widehat{\beta}$. A re-sampling approach seems to be required because in general there is no analytic expression for the asymptotic variance of $\widehat{\beta}$ that accounts for the variations in $\widehat{G}^{S|W}$ and $\widehat{G}^S$, and previously developed techniques for deriving the asymptotic variance of $\widehat{\beta}$ 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 $risk_{(Z)}$, $G^{S|W}$, and G^S

An advantage of the estimated likelihood approach is that it can be used generally for a variety of models for $risk_{(Z)}(\cdot,\cdot)$, $G^{S|W}$, and G^S. The dimensionality of W and S determine whether parametric modeling assumptions are needed for stably estimating $G^{S|W}$ and G^S. For the case that $\delta_i = 1$ for all i, Follmann considered a fully parametric model, with $(W, S(1))$ assumed bivariate normal and $Pr(Y(Z) = 1|S(1) = s_1)$ assumed to follow a probit model, for $Z = 0, 1$.

We allow case-cohort sampling and focus on the setting that S has J categories and W has K categories. In this case nonparametric models can be used: with $\theta_{jk} \equiv Pr(S(1) = j, W = k|V^{obs} = 1)$, $g^S(j) = Pr(S(1) = j|V^{obs} = 1) = \sum_{k=1}^{K} \theta_{jk} \equiv \theta_j$, $g^{S|W}(j|k) = \theta_{jk}/\sum_{l=1}^{J} \theta_{lk}$, and $risk_{(Z)}(j, 0; \beta) = \beta_{Zj}$, for $Z = 0, 1; j = 1, \ldots, J; k = 1, \ldots, K$. Then for any $h(\cdot,\cdot)$ contrast function $CEP^{risk}(j, 0; \beta) = h(\beta_{1j}, \beta_{0j}), AS = |h(\beta_{1J}, \beta_{0J})| - |h(\beta_{11}, \beta_{01})|$, $EAE(w) = (1 - \theta_1)^{-1} \sum_{j=2}^{J} w(j, 1)h(\beta_{1j}, \beta_{0j})Pr(S^{obs} = j|Z = 1)$, and $EDE = \theta_1^{-1} h(\beta_{11}, \beta_{01})Pr(S^{obs} = 1|Z = 1)$.

4.4 Nonparametric Maximum Estimated Likelihood Estimation (MELE)

For estimating $G^{S|W}$ and G^S, a consistent estimator of θ_{jk} based on treatment $Z=1$ data is given by

$$\widehat{\theta}_{jk} = \widehat{\theta}_j^{-1}\left\{(n_1(j,k)/n_1)AR + (n_0(j,k)/n_0)(1-AR)\right\},$$

where $\widehat{\theta}_j = (n_1(j)/n_1)AR + (n_0(j)/n_0)(1-AR)$, $AR = \sum_{i=1}^n Z_i V_i^{obs} I[Y_i^{obs} = 1]/\sum_{i=1}^n Z_i V_i^{obs}$, $n_y(j,k) = \sum_{i=1}^n Z_i V_i^{obs} \delta_i I[Y_i^{obs} = y, S_i^{obs} = j, W_i = k]$, $n_y(j) = \sum_{i=1}^n Z_i V_i^{obs} \delta_i I[Y_i^{obs} = y, S_i^{obs} = j]$, and $n_y = \sum_{i=1}^n Z_i V_i^{obs} \delta_i I[Y_i^{obs} = y]$, for $y = 0, 1$.

To maximize $L(\beta, \widehat{G}^{S|W}, \widehat{G}^S)$ it is convenient to partition $\beta = (\beta_1, \beta_0)'$, where $risk_{(Z)}(\cdot,\cdot)$ depends on $\beta_Z \equiv (\beta_{Z1}, \ldots, \beta_{ZJ})'$ only, $Z = 0, 1$. Then the estimated likelihood factors as $L(\beta, \widehat{G}^{S|W}, \widehat{G}^S) = L_1(\beta_1, \widehat{G}^{S|W}, \widehat{G}^S) \times L_0(\beta_0, \widehat{G}^{S|W}, \widehat{G}^S)$, where $L_1 = \prod_{i=1}^n f(O_i)^{Z_i V_i^{obs}}$ and $L_0 = \prod_{i=1}^n f(O_i)^{(1-Z_i)V_i^{obs}}$. Based on L_1 a closed form MELE of $\beta_{1j}, j = 1, \ldots, J$ can be derived as $\widehat{\beta}_{1j} = (n_1(j)/n_1) \times (AR/\widehat{\theta}_j)$. An EM algorithm can be used to find the MELE of the β_{0j}. The E step entails computing the expectation of $I_{ij} \equiv I[S_i^{obs} = j]$ given the observed data. For $\delta_i = 0$, $E[I_{ij}|Z_i = 0, V_i^{obs} = 1, Y_i^{obs} = y, \delta_i = 0] = \{\beta_{0j}^{Y_i^{obs}}(1-\beta_{0j})^{1-Y_i^{obs}}\widehat{\theta}_j\}/\{\sum_{l=1}^J \beta_{0l}^{Y_i^{obs}}(1-\beta_{0l})^{1-Y_i^{obs}}\widehat{\theta}_l\}$, and for $\delta_i = 1$, $E[I_{ij}|Z_i = 0, V_i^{obs} = 1, Y_i^{obs} = y, \delta_i = 1, W_i = k] = \{\beta_{0j}^{Y_i^{obs}}(1-\beta_{0j})^{1-Y_i^{obs}}\widehat{\theta}_{jk}\}/\{\sum_{l=1}^J \beta_{0l}^{Y_i^{obs}}(1-\beta_{0l})^{1-Y_i^{obs}}\widehat{\theta}_{lk}\}$. The M step entails replacing I_{ij} with $\mu_{ij} \equiv E[I_{ij}|Z_i = 0, V_i^{obs} = 1, Y_i^{obs} = y, \delta_i, \delta_i W_i]$ in the complete data likelihood, which when maximized yields $\widehat{\beta}_{0j} = \{\sum_{i=1}^n (1-Z_i)V_i^{obs} I[Y_i^{obs} = y]\mu_{ij}\}/\{\sum_{i=1}^n (1-Z_i)V_i^{obs}\mu_{ij}\}$.

4.5 Tests for Whether a Biomarker has Any Surrogate Value

Since $PAE(w) = 0.5$ supports that S has no surrogate value, Wald tests for any surrogate value can be based on the MELE $\widehat{PAE}(w)$ minus 0.5 divided by its bootstrap standard error. Similarly Wald tests of $AS = 0$ can be implemented based on $\widehat{AS}$. We also consider a test statistic $T = \sum_{j=2}^J (j-1)\{\widehat{\beta}_{0j} - (\widehat{\beta}_{0j} + \widehat{\beta}_{1j})(\widehat{\mu}_0/(\widehat{\mu}_0 + \widehat{\mu}_1))\}$ divided by its bootstrap standard error, where $\widehat{\mu}_Z = \frac{1}{J}\sum_{j=1}^J \widehat{\beta}_{Zj}$. This test evaluates $H_0 : CEP^{risk}(j,1) = CE$ for all j versus the monotone alternative that $CEP^{risk}(j,1)$

increases in j, and is similar to the Breslow-Day test for trend (Breslow and Day, 1980).

5. Simulation Study

We are preparing to apply the methods to forthcoming datasets (Mehrotra, Li, and Gilbert, 2006). In the interim, 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 nonparametric MELE method. The vaccine trial was double-blind with 2:1 randomization to vaccine:placebo. A biomarker of interest S was the percentage of antibody blocking of the binding of the HIV GNE8 recombinant gp120 molecule to recombinant soluble CD4 (the "CD4 blocking level") measured from a serum sample drawn at the month 6.5 visit after randomization, and Y was HIV infection during the 3 year follow-up period. The lower quantification limit of the CD4 blocking assay was 0.084, and all 46 placebo recipients with S measured had $S_i^{obs} < 0.084$; thus the data fit case A4. The range of S_i^{obs} was $[0.084, 0.92]$, which we rescaled to $[0, 1]$, so that A4 holds with $c = L = 0$. In vaccine recipients S_i^{obs} was approximately normally distributed, with average 0.576 and variance 0.0238.

We simulated vaccine trials with the following steps. **Step 1:** For all 3330 (1691) subjects in the vaccine (placebo) arm, $(W_i, S_i(1))$ was generated from a bivariate normal distribution with means 0.576, variances 0.0238, and correlation $\rho = 0.5, 0.7$, or 0.9. Then W_i and $S_i(1)$ were binned into quartiles. **Step 2:** For subjects i with quartile j value of $S_i(1)$, $Y_i(Z)$ was generated from Bernoulli(β_{Zj}), with the β_{Zj} set to achieve the infection rate $Pr(Y(1) = 1) = 0.057$ that was observed in the vaccine arm of the trial and overall vaccine efficacy of 50% ($Pr(Y(0) = 1) = 2 \times Pr(Y(1) = 1)$), and to reflect a biomarker with either (i) no or (ii) high surrogate value. In scenario (i) $CEP^{risk}(j, 1; \beta) \equiv log(risk_{(1)}(j, 1; \beta_1)/risk_{(0)}(j, 1; \beta_0)) = -0.69$ for $j = 1, \ldots, 4$, and in scenario (ii) $CEP^{risk}(j, 1; \beta) = -0.22, -0.51, -0.92, -1.61$ for $j = 1, \ldots, 4$. With vaccine efficacy $VE(j,1) \equiv 1 - exp(CEP^{risk}(j,1;\beta))$, scenario (i) specifies constant $VE(j,1) = 0.5$ and scenario (ii) specifies $VE(j,1) = 0.2, 0.4, 0.6, 0.8$ for $j = 1, \ldots, 4$. **Step 3:** To achieve case-cohort sampling, $(W_i, S_i(1))$ was retained for all infected vaccine recipients and for the 406 uninfected vaccine recipients who had immunological assays performed, and was set to missing for all other vaccine recipients. For the placebo arm $S_i(1)$ was set to missing for everyone and W_i was retained for all infected placebo recipients and for a random sample of 406 uninfected placebo recipients.

For each of 1000 simulated data sets the MELE $\widehat{\beta}$ was computed, which was then used to compute the MELEs of $CEP^{risk}(j,1)$, AS, and $P\widehat{AE}(w)$ for $w(s_1, s_0) = 1, |s_1 - s_0|$, and $I(s_1 = 4, s_0 = 1)$. Wald tests (with bootstrap standard errors) based on $\widehat{PAE}(w) - 0.5\ \widehat{AS}$, and on T were used to test for any surrogate value.

Performance of the MELE $\widehat{\beta}$ was excellent (Table 2). The MELEs of $CEP^{risk}(j,1)$ also performed well, though with some bias. This bias results from the facts that $\widehat{CEP}^{risk}(j,1)$ involves the ratio $\widehat{\beta}_{1j}/\widehat{\beta}_{0j}$, and estimators defined by the ratio of two estimators, each of which is unbiased for its estimand, may be biased in moderate samples (Chick, Barth-Jones, and Koopman, 2001). In contrast the MELEs of $CEP^{risk}(j,1)$ with $h(x,y) = x - y$ were unbiased (results not shown).

The MELEs of $PAE(w)$ and AS were unbiased and the confidence intervals about them had nominal coverage. The tests for any surrogate value had approximately nominal size and showed high power to detect surrogate value when ρ was 0.7 or higher; the trend test had power 0.73, 0.91, and 0.99 for $\rho = 0.5$, 0.7, and 0.9. These results demonstrate "proof-of-principle" that the methods can reliably estimate the CEP surface when a reasonably good baseline predictor of the biomarker is used, and can distinguish between biomarkers S 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 causal effect predictiveness (CEP) surface as an appropriate estimand for measuring the predictive capacity of a candidate surrogate. The CEP surface is not identified from the data collected in a standard trial design, however. On the other hand, the net effect estimand used in the alternative framework for evaluating surrogates (statistical surrogates) is identified, but is not causal, which may make it less useful for measuring predictive capacity. Therefore when applied to a single trial, each framework for defining and evaluating surrogates has a serious but different limitation. As such, both approaches may be useful for generating preliminary evidence about approximate surrogacy, which will require further validation. In fact, based on a single efficacy trial neither approach is suitable for evaluating whether a biomarker can be reliably used for bridging information about clinical efficacy to a new setting (e.g., bridge to a new human population or treatment formulation); for this additional experiments (such as mechanistic studies and studies that deliberately manipulate the biomarker) and meta-analysis are needed.

Since the definition of the CEP surface involves counterfactuals, strong untestable assumptions may be needed to identify it, which may preclude its reliable estimation. While we think this critique will sometimes hold, a thesis of this work is that given innovative data collection and a particular kind of biomarker S, the CEP surface can be identified and estimated under plausible assumptions. The estimation method we developed requires A1-A5, a reasonably good baseline predictor W, a model predicting S^{obs} from W in treatment arm 1, and models for $risk_{(Z)}(s_1, c) \equiv Pr(Y(Z) = 1|S(1) = s_1, S(0) = c)$, for $Z = 0, 1$. A1-A2 are standard in randomized trials, and A4 is easy to check. A1 (SUTVA) is a potentially dubious assumption 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. While untestable, A3 will not significantly influence the results if S is measured near baseline or if most randomized subjects are disease free at its measurement time, but otherwise it will be important to conduct sensitivity analyses to violations of A3. A5 is testable for treatment arm $Z = 1$ but not for $Z = 0$. Thus it is important to evaluate plausibility of $Y(0)|W, S(1) =^d Y(0)|S(1)$ from biological knowledge. Under A1-A5 any parametric modeling assumptions placed on $risk_{(1)}(s_1, c)$ and $risk_{(0)}(s_1, c)$ can be tested. Finally, models for S^{obs} given W can be directly checked using arm $Z = 1$ data.

A vector of baseline covariates X measured on all subjects could easily be incorporated into the developed estimation methods. This would allow modifying A5 to the more defensible assumption A5′: $Y(Z)|X, W, S(1) =^d Y(Z)|X, S(1)$, $Z = 0, 1$. It would also allow addressing the interaction question of how the CEP surface depends on X, and could increase the precision for estimating the CEP surface. Furthermore, to accommodate study drop-out that leads to missing Y's, including covariates could help justify a missing at random assumption, facilitating making unbiased inferences. To allow for drop-out A3 must be modified to $A3'$: Equal Individual Clinical Risk and Drop-out Up to Time t_0: $V_i(1) = 1$ if and only if $V_i(0) = 1$, where now $V_i(Z)$ is the potential indicator of whether the ith subject is at risk for disease at t_0.

The estimands and estimation techniques developed here for a binary clinical endpoint Y also apply for a quantitative clinical endpoint Y, with all expressions $Pr(Y(Z) = 1|\cdot)$ replaced with $E(Y(Z)|\cdot)$. In either case the CEP surface describes how the average or population level causal effect on Y depends on the causal effect on S. It is beyond the scope of this article to address the statistical generalizability of S, that is, how reliably it can be used for predicting Y^{obs} in a new setting. We note that the estimated CEP surface may be useful for this purpose, by providing a prediction of the overall clinical effect CE in the new setting based on measurements of S, which could be compared to an estimate of CE computed ignoring S.

Acknowledgements

The authors thank Margaret Pepe, Ross Prentice, and Steve Self 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.* International Agency for Research on Cancer, Lyon, France.

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

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). The validation on surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–67.

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.

Chick, S., Barth-Jones, D. and Koopman, J. (2001). Bias reduction for risk ratio and vaccine effect estimators. *Statistics in Medicine* **20**, 1609–1624.

Daniels, M. and Hughes, M. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16**, 1965–1982.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics* **in press**.

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.

Gail, M., Pfeiffer, R., Van Houwelingen, H. and Carroll, R. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246.

Gilbert, P., Peterson, M., Follmann, D., Hudgens, M., Francis, D., Gurwith, M., Heyward, W., Jobes, D., Popovic, V., Self, S., Sinangil, F., Burke, D. and Berman, P. (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.

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

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–961.

Huang, Y., Pepe, M. and Feng, Z. (2006). *Evaluating the predictiveness of a continuous marker.* Department of Biostatistics, University of Washington.

Hudgens, M. and Halloran, M. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association* **101**, 51–64.

Hudgens, M., Hoering, A. and Self, S. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* **22**, 2281–2298.

Lin, D., Fleming, T. and De Gruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16**, 1515–1527.

Mehrotra, D., Li, X. and Gilbert, P. (2006). Dual-endpoint evaluation of vaccine efficacy: Application to a proof-of-concept clinical trial of a cell mediated immunity-based HIV vaccine. *Biometrics* **in press**.

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. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology* **3**, 143–155.

Rosenbaum, P. (1984). The consequence of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* **147**, 656–666.

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

Shepherd, B., Gilbert, P., Jemiai, Y. and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62**, 332–342.

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.

Wang, Y. and Taylor, J. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **58**, 803–812.

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

Table 1

Examples illustrating a principal surrogate compared to a statistical surrogate, S binary with $S_i(0) = 0$ for all i, with $h(x, y) = 1 - x/y$

Perfect Principal Surrogate but Not a Statistical Surrogate[a]

					S_i^{obs}/	
Principal		Fraction	$Pr(Y_i(1) = 1\|$	$Pr(Y_i(0) = 1\|$	$Pr(Y_i^{obs} = 1\|S_i^{obs}, Z_i)$	
Stratum (PS)	($S_i(1)$,$S_i(0)$)	in PS	$S_i(1), S_i(0))$	$S_i(1), S_i(0))$	$Z_i = 1$	$Z_i = 0$
Vacc. not take	(0,0)	1/3	0.3	0.3	0/0.3	0/0.2
Vacc. take	(1,0)	2/3	0.0	0.15	1/0.0	0/0.2

No Value as a Principal Surrogate but a Statistical Surrogate[b]

					S_i^{obs}/	
Principal		Fraction	$Pr(Y_i(1) = 1\|$	$Pr(Y_i(0) = 1\|$	$Pr(Y_i^{obs} = 1\|S_i^{obs}, Z_i)$	
Stratum (PS)	($S_i(1)$,$S_i(0)$)	in PS	$S_i(1), S_i(0))$	$S_i(1), S_i(0))$	$Z_i = 1$	$Z_i = 0$
Vacc. not take	(0,0)	1/3	0.2	0.4	0/0.2	0/0.2
Vacc. take	(1,0)	2/3	0.05	0.1	1/0.0	0/0.2

[a] $CE = 1 - [(1/3) \times 0.3 + (2/3) \times 0.0]/[(1/3) \times 0.3 + (2/3) \times 0.15] = 0.5$;
$Pr(Y_i^{obs} = 1|S_i^{obs} = 0, Z_i = 1) = (1) \times 0.3 = 0.3$;
$Pr(Y_i^{obs} = 1|S_i^{obs} = 0, Z_i = 0) = (1/3) \times 0.3 + (2/3) \times 0.15 = 0.2$;
$CEP^{risk}(0, 0) = 1 - 0.3/0.3 = 0.0$; $CEP^{risk}(1, 0) = 1 - 0.0/0.15 = 1.0$;
$EAE(w = 1) = (2/3) \times 1.0/(2/3) = 1.0$, $EDE = (1/3) \times 0.0/(1/3) = 0.0$;
$PAE(w = 1) = 1.0/(0.0 + 1.0) = 1.0$; $AS = 1.0 - 0.0 = 1.0$
(the parameters $EAE(w)$, EDE, and $PAE(w)$ are defined at (1)-(3)).

[b] $CE = 1 - [(1/3) \times 0.2 + (2/3) \times 0.05]/[(1/3) \times 0.4 + (2/3) \times 0.1] = 0.5$;
$Pr(Y_i^{obs} = 1|S_i^{obs} = 0, Z_i = 1) = (1) \times 0.2 = 0.2$;
$Pr(Y_i^{obs} = 1|S_i^{obs} = 0, Z_i = 0) = (1/3) \times 0.4 + (2/3) \times 0.1 = 0.2$;
$CEP^{risk}(0, 0) = 1 - 0.2/0.4 = 0.5$; $CEP^{risk}(1, 0) = 1 - 0.05/0.1 = 0.5$;
$EAE(w = 1) = (2/3) \times 0.5/(2/3) = 0.5$; $EDE = (1/3) \times 0.5/(1/3) = 0.5$;
$PAE(w = 1) = 0.5/(0.5 + 0.5) = 0.5$; $AS = 0.5 - 0.5 = 0.0$

Table 2

Simulation results for the nonparametric MELE $\widehat{\beta}^a$

Cor.	No Surrogate Value Scenario							High Surrogate Value Scenario						
		Bias			SEE				Bias			SEE		
ρ	Parameter	Mean	Med	SE	Mean	Med	CP	Parameter	Mean	Med	SE	Mean	Med	CP
0.5	$\beta_{01} = .181$	-.001	-.003	.060	.062	.061	.95	$\beta_{01} = .113$	-.001	-.002	.050	.051	.050	.93
	$\beta_{02} = .136$	.000	-.013	.076	.079	.078	.94	$\beta_{02} = .113$	.004	-.005	.062	.070	.068	.96
	$\beta_{03} = .090$	-.002	-.006	.049	.056	.055	.96	$\beta_{03} = .113$	.002	-.006	.061	.071	.069	.96
	$\beta_{04} = .045$	.006	.002	.036	.035	.035	.91	$\beta_{04} = .113$	-.002	-.002	.049	.051	.049	.94
	$\beta_{11} = .090$	.001	.001	.012	.012	.012	.95	$\beta_{11} = .090$	.000	-.001	.012	.012	.012	.95
	$\beta_{12} = .068$	.000	.000	.010	.010	.010	.94	$\beta_{12} = .068$	.001	.000	.010	.010	.010	.96
	$\beta_{13} = .045$	.000	.000	.008	.008	.008	.94	$\beta_{13} = .045$	.000	.000	.008	.008	.008	.95
	$\beta_{14} = .023$	.000	.000	.005	.006	.005	.95	$\beta_{14} = .023$	.000	.000	.005	.005	.005	.94
0.7	$\beta_{01} = .181$	.001	.003	.049	.049	.048	.94	$\beta_{01} = .113$	.000	.000	.038	.041	.040	.94
	$\beta_{02} = .136$	-.001	-.008	.060	.066	.064	.94	$\beta_{02} = .113$	.002	-.003	.052	.058	.057	.96
	$\beta_{03} = .090$	-.003	-.005	.039	.046	.045	.95	$\beta_{03} = .113$	.002	-.002	.051	.058	.056	.96
	$\beta_{04} = .045$	.003	.001	.027	.028	.028	.93	$\beta_{04} = .113$	-.001	-.001	.038	.041	.040	.94
	$\beta_{11} = .090$	-.001	-.001	.012	.012	.012	.94	$\beta_{11} = .090$	.000	-.001	.012	.012	.012	.95
	$\beta_{12} = .068$	.000	.000	.010	.010	.010	.95	$\beta_{12} = .068$	.000	-.001	.010	.010	.010	.95
	$\beta_{13} = .045$	.000	.000	.008	.008	.008	.94	$\beta_{13} = .045$	.000	.000	.008	.008	.008	.95
	$\beta_{14} = .023$	.000	.000	.005	.005	.005	.94	$\beta_{14} = .023$	.000	.000	.006	.005	.005	.94
	NA													
0.9	$\beta_{01} = .181$	.002	.001	.033	.034	.033	.96	$\beta_{01} = .113$	.000	-.001	.028	.027	.027	.93
	$\beta_{02} = .136$	-.001	-.002	.050	.049	.048	.93	$\beta_{02} = .113$	.000	-.002	.046	.046	.045	.92
	$\beta_{03} = .090$	.002	-.001	.042	.040	.040	.89	$\beta_{03} = .113$	.000	-.002	.046	.046	.045	.93
	$\beta_{04} = .045$	-.001	-.001	.019	.019	.019	.91	$\beta_{04} = .113$	-.001	-.002	.027	.028	.027	.95
	$\beta_{11} = .090$	.001	.001	.012	.012	.012	.95	$\beta_{11} = .090$	.000	.000	.012	.012	.012	.94
	$\beta_{12} = .068$	.000	.000	.010	.010	.010	.94	$\beta_{12} = .068$	.001	-.001	.011	.010	.010	.94
	$\beta_{13} = .045$	.000	.000	.008	.008	.008	.95	$\beta_{13} = .045$	.000	.000	.008	.008	.008	.94
	$\beta_{14} = .023$	.000	.000	.005	.005	.005	.95	$\beta_{14} = .023$	.000	.000	.006	.006	.005	.93

[a] ρ is the linear correlation of the simulated bivariate normal variables latent to the quartilized variables W and $S(1)$. SE is the empirical standard error of $\hat{\beta}_{Zj}$. SEE is the mean or median (Med) of the bootstrap standard error estimates based on 200 bootstrap replicates. CP is the empirical coverage of standard normal 95% confidence intervals for $\hat{\beta}_{Zj}$ using bootstrap standard error estimates. 1000 simulations were done to compute the table elements for each model.

Table 3

Simulation results for the nonparametric MELEs $\widehat{CEP}^{risk}(j,1) = \log(\widehat{\beta}_{1j}/\widehat{\beta}_{0j})$ *for* $j = 1, \cdots, 4$, $\widehat{PAE}(w)$, *and* $\widehat{AS}$[a]

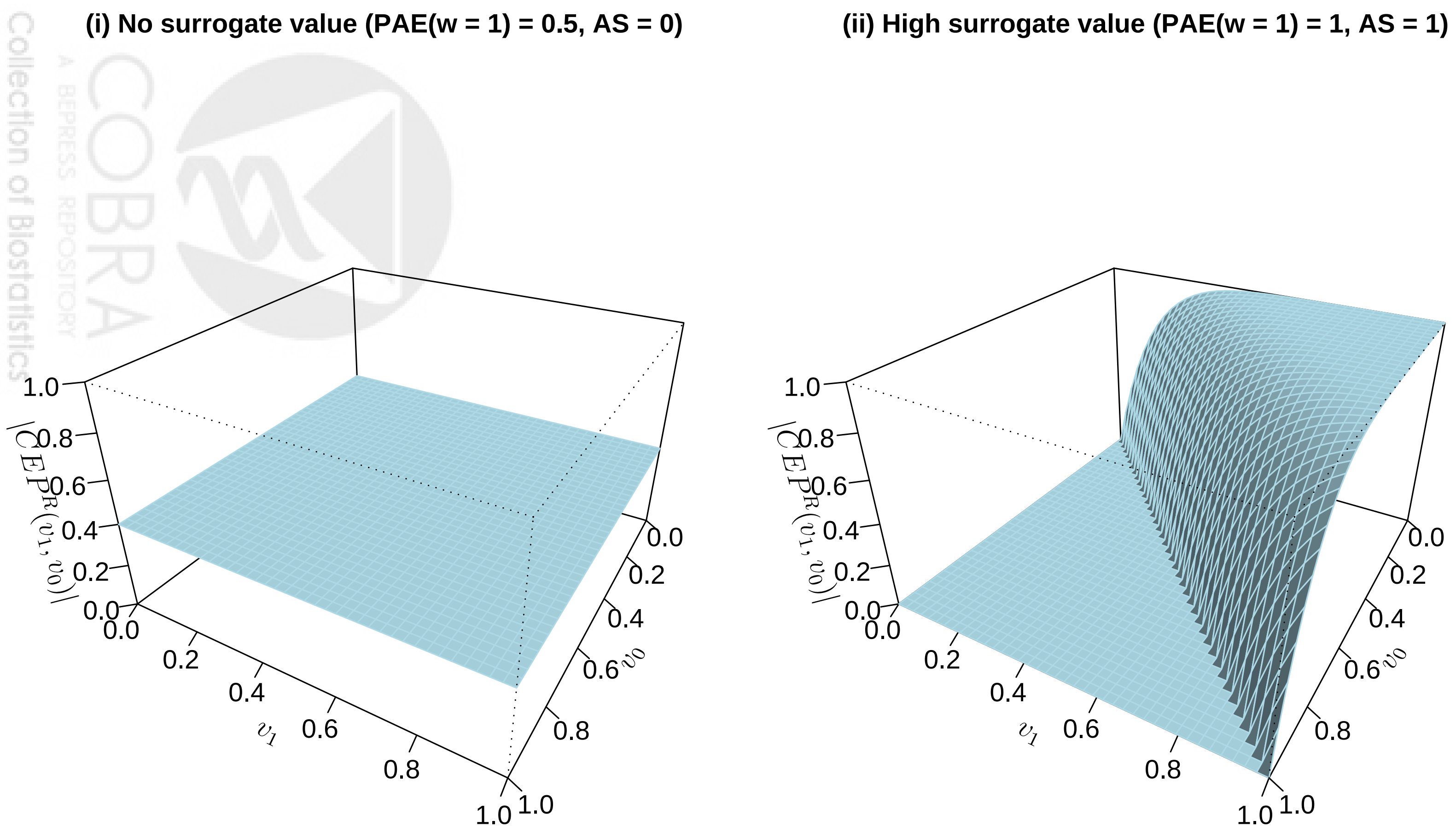
Cor.	No Surrogate Value Scenario								High Surrogate Value Scenario							
		Bias			SEE					Bias			SEE			
ρ	Parameter	Mean	Med	SE	Mean	Med	CP	Pow	Parameter	Mean	Med	SE	Mean	Med	CP	Pow
0.5	$CEP^{risk}(1,1) = -.69$	.09	.00	.44	.54	.48	.98	.32	$CEP^{risk}(1,1) = -02$	.15	.01	.63	.83	.74	.99	.05
	$CEP^{risk}(2,1) = -.69$	.15	.11	.65	.80	.74	.98	.12	$CEP^{risk}(2,1) = -.51$	.13	.04	.70	.86	.78	.99	.06
	$CEP^{risk}(3,1) = -.69$	.19	.07	.75	.98	.88	.99	.05	$CEP^{risk}(3,1) = -.92$	.12	.05	.62	.86	.79	.99	.15
	$CEP^{risk}(4,1) = -.69$	.39	-.05	1.52	1.60	1.45	.97	.11	$CEP^{risk}(4,1) = -1.61$	.14	.01	.67	.87	.75	.99	.55
	$PAE(w_1) = .50$	-.03	-.03	.21	.22	.22	.95	.04	$PAE(w_1) = .82$	-.14	-.12	.20	.22	.22	.93	.23
	$PAE(w_2) = .50$	.00	.00	.21	.22	.22	.97	.04	$PAE(w_2) = .84$	-.13	-.11	.19	.21	.21	.94	.30
	$PAE(w_3) = .50$	.07	.09	.21	.21	.22	.96	.09	$PAE(w_3) = .88$	-.12	-.08	.18	.19	.19	.94	.53
	$AS = 0.00$	.47	.32	1.11	1.32	1.07	.97	.05	$AS = 1.39$	-.33	-.18	.74	.99	.85	1.00	.40
0.7	$CEP^{risk}(1,1) = -.69$	.02	-.02	.34	.38	.33	.97	.56	$CEP^{risk}(1,1) = -02$	.06	-.01	.45	.60	.50	.98	.09
	$CEP^{risk}(2,1) = -.69$	.12	.06	.55	.65	.61	.98	.18	$CEP^{risk}(2,1) = -.51$	.09	.03	.53	.68	.62	.98	.10
	$CEP^{risk}(3,1) = -.69$	.15	.06	.60	.75	.69	.99	.10	$CEP^{risk}(3,1) = -.92$	.09	.02	.55	.69	.64	.98	.27
	$CEP^{risk}(4,1) = -.69$	.19	-.04	1.11	1.40	1.19	.97	.14	$CEP^{risk}(4,1) = -1.61$	.06	.02	.50	.66	.55	.98	.71
	$PAE(w_1) = .50$	-.04	-.04	.18	.19	.19	.95	.02	$PAE(w_1) = .82$	-.09	-.08	.16	.20	.20	.97	.34
	$PAE(w_2) = .50$	-.03	-.03	.18	.19	.19	.98	.02	$PAE(w_2) = .84$	-.07	-.07	.14	.18	.17	.97	.48
	$PAE(w_3) = .50$	.02	.05	.19	.20	.20	.95	.07	$PAE(w_3) = .88$	-.07	-.05	.13	.16	.15	.97	.73
	$AS = 0.00$	.23	.16	.85	1.14	.87	.98	.05	$AS = 1.39$	-.19	-.13	.52	.74	.62	.99	.62
0.9	$CEP^{risk}(1,1) = -.69$	.00	-.01	.21	.23	.21	.98	.84	$CEP^{risk}(1,1) = -02$	.03	.01	.28	.32	.28	.96	.16
	$CEP^{risk}(2,1) = -.69$	.08	.02	.47	.52	.47	.96	.37	$CEP^{risk}(2,1) = -.51$	.11	.03	.54	.58	.55	.96	.22
	$CEP^{risk}(3,1) = -.69$	.12	-.01	.62	.63	.60	.95	.28	$CEP^{risk}(3,1) = -.92$	.09	.02	.54	.58	.55	.95	.41
	$CEP^{risk}(4,1) = -.69$	.14	.03	.72	.84	.69	.97	.24	$CEP^{risk}(4,1) = -1.61$	.02	.01	.36	.40	.36	.98	.94
	$PAE(w_1) = .50$	-.04	-.03	.14	.16	.16	.95	.01	$PAE(w_1) = .82$	-.04	-.03	.14	.16	.15	.97	.56
	$PAE(w_2) = .50$	-.04	-.02	.15	.16	.16	.95	.01	$PAE(w_2) = .84$	-.03	-.03	.11	.13	.13	.97	.77
	$PAE(w_3) = .50$	-.01	.01	.19	.19	.19	.93	.06	$PAE(w_3) = .88$	-.03	-.02	.10	.11	.10	.97	.90
	$AS = 0.00$	.08	.02	.53	.70	.54	.96	.06	$AS = 1.39$	-.08	-.06	.41	.46	.44	.97	.90

[a] ρ is the linear correlation of the simulated bivariate normal variables latent to the quartilized variables W and $S(1)$. SE is empirical standard error of $\widehat{CEP}^{risk}(j,1)$, $\widehat{PAE}(w)$, and $\widehat{AS}$. SEE is the mean or median (Med) of the bootstrap standard error estimates based on 200 bootstrap replicates. CP is the empirical coverage of standard normal 95% confidence intervals for $CEP^{risk}(j,1)$, $PAE(w)$, and AS using bootstrap standard error estimates. Pow (Power) is for testing $H_0 : CEP^{risk}(j,1) = 0$, $H_0 : PAE(w) = 0.5$, and $H_0 : AS = 0$ at level $\alpha = 0.05$. For the PAE weights, $w_1(s_1, s_0) = 1$, $w_2(s_1, s_0) = |s_1 - s_0|$, and $w_3(s_1, s_0) = I[s_1 = 4, s_0 = 1]$. 1000 simulations were done to compute the table elements for each model.

Figure Legends

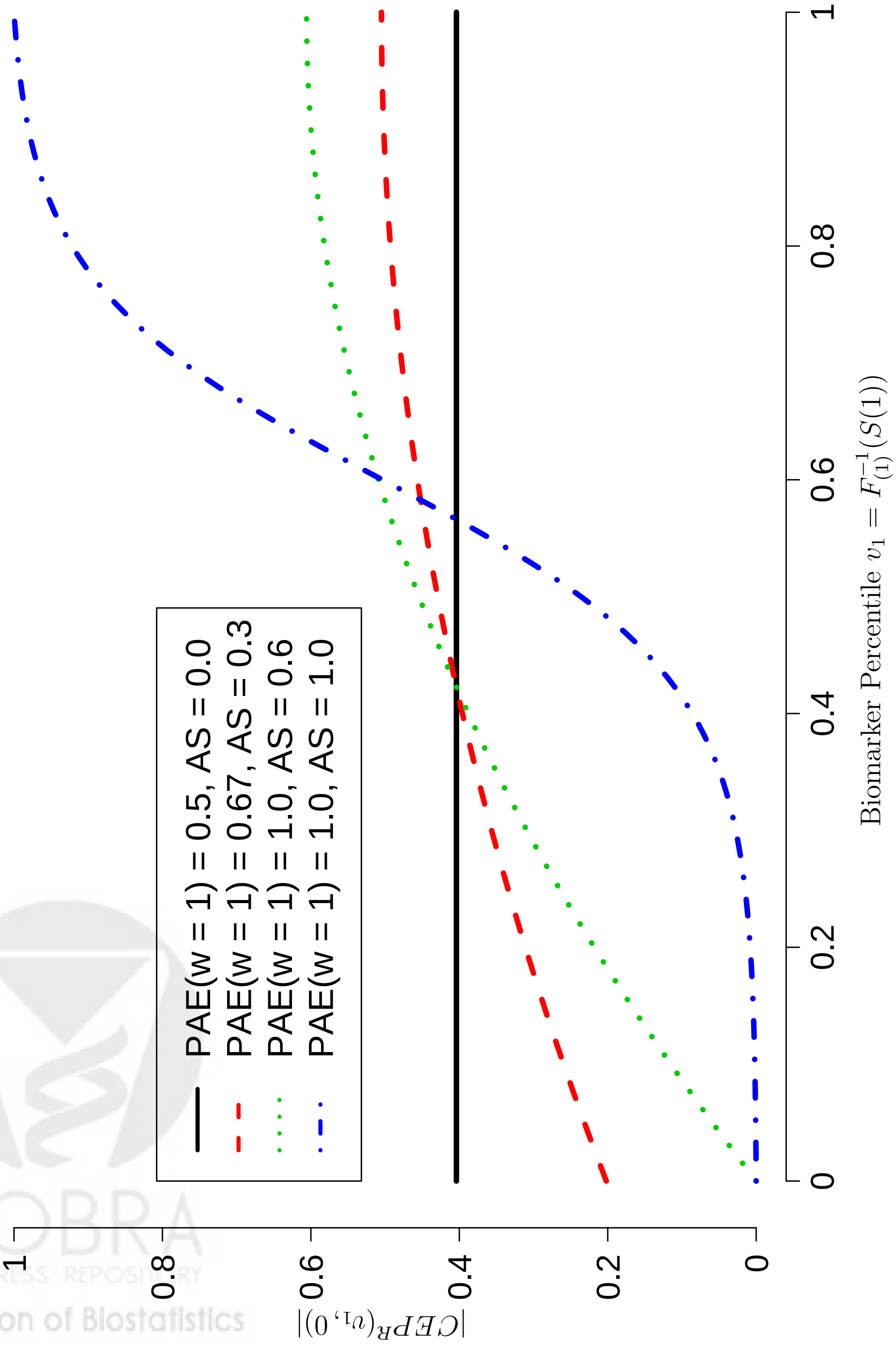
Figure 1. Example $CEP^R(v_1, v_0) = h(R_{(1)}(v_1, v_0), R_{(0)}(v_1, v_0))$ surfaces, with $h(x, y) = x - y$ or $1 - x/y$. The surface in (i) reflects a biomarker with no surrogate value ($PAE(w = 1) = 0.5, AS = 0$), wherein the clinical treatment effect is the same for all treatment effects on the biomarker. The surface in (ii) reflects a biomarker with high surrogate value ($PAE(w = 1) = 1, AS = 1$), wherein the average causal effect on the clinical endpoint is zero for all $v_1 \leq v_0$ and has a large increase in $v_1 - v_0$ for $v_1 > v_0$. Because $CEP^R(v_1, v_0) = 0$ for all $v_1 \leq v_0$ and $CEP^R(v_1, v_0) > 0$ for all $v_1 > v_0$, the biomarker evaluated in (ii) satisfies Average Causal Necessity and Average Causal Sufficiency, and hence is a principal surrogate.

Figure 2. For the situation A4 for which $S_i(0) = c$ 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. With $h(x, y) = x - y$, the area between each $|CEP^R(v_1, F_{(1)}(c))|$ curve and the zero-line equals the overall clinical treatment effect $|CE| = 0.4$. Because $CEP^R(F_{(1)}(c), F_{(1)}(c)) = 0$ and $CEP^R(v_1, F_{(1)}(c)) > 0$ for all $v_1 > F_{(1)}(c)$, the latter two S's satisfy Average Causal Necessity and Average Causal Sufficiency, and hence are principal surrogates.



(i) No surrogate value (PAE(w = 1) = 0.5, AS = 0)

(ii) High surrogate value (PAE(w = 1) = 1, AS = 1)


PAE(w = 1) = 0.5, AS = 0.0
PAE(w = 1) = 0.67, AS = 0.3
PAE(w = 1) = 1.0, AS = 0.6
PAE(w = 1) = 1.0, AS = 1.0
$|CEP_R(v_1, 0)|$
0
0.2
0.4
0.6
0.8
1
Biomarker Percentile $v_1 = F_{(1)}^{-1}(S(1))$
0
0.2
0.4
0.6
0.8
1