Evaluating Causal Effect Predictiveness of Candidate 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.

[*] *e-mail:* pgilbert@scharp.org

SUMMARY. Frangakis and Rubin (2002, Biometrics) 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. While the CEP surface is not identifiable due to missing potential outcomes, it can be identified by incorporating a baseline covariate that predicts the biomarker. Given case-cohort sampling of such a baseline predictor and the biomarker in a single 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; Post-Randomization Selection Bias; Prentice Criteria; Principal Stratification; Principal Surrogate.

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 et al., 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 post-randomization selection bias. Since 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 post-randomization 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 meta-analysis. Here we develop an approach for evaluating a principal surrogate within the former setting. Following Follmann (2006), our approach uses a baseline covariate 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. In Section 5 we evaluate the 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 baseline covariate W, 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 $S = *$).

2.1 *Definition of a Statistical Surrogate*

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

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

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, $risk(s|Z=1) = risk(s|Z=0)$. The full mediation criterion (i) requires that a treatment effect on S is necessary and sufficient for a treatment effect on Y; 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 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 $(Z_i, W_i, V_i(1), V_i(0), S_i(1), S_i(0), Y_i(1), Y_i(0)), i = 1, \ldots, n$, are iid, and for simplicity assume no drop-out. Further we assume A1, A2 (Rubin 1986), and A3:

A1 *Stable Unit Treatment Value Assumption (SUTVA)*

A2 *Ignorable Treatment Assignments*: Z is independent of

$(V(1), V(0), S(1), S(0), Y(1), Y(0))$

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

A1 states that the potential outcomes $(V_i(1), V_i(0), S_i(1), S_i(0), Y_i(1), Y_i(0))$ 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. 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 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. FR defined a biomarker S to be a principal surrogate endpoint if the comparison between

$$risk_{(1)}(s_1, s_0) \equiv \Pr(Y(1) = 1 | V(1) = 1, V(0) = 1, S(1) = s_1, S(0) = s_0) \quad \text{and}$$
$$risk_{(0)}(s_1, s_0) \equiv \Pr(Y(0) = 1 | V(1) = 1, V(0) = 1, S(1) = s_1, S(0) = s_0)$$

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, since $(S(1), S(0))$ is only defined if $V(1) = V(0) = 1$. For notational simplicity henceforth all probability statements involving $S_i(Z)$ implicitly condition on $V_i(Z) = 1$. A contrast in $risk_{(1)}(s_1, s_0)$ and $risk_{(0)}(s_1, s_0)$ 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: $risk_{(1)}(s_1, s_0) = risk_{(0)}(s_1, s_0)$ 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 $risk_{(1)}(s_1, s_0)$ $\neq risk_{(0)}(s_1, s_0)$ for all $|s_1 - s_0| > C$.

For the 1-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 1-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. Heretofore 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

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

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 (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 $h(x, x) = 0$, 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 $\{S(1) = s_1, S(0) = s_0\}$. Note that $CEP^{risk}(s, s) = 0$ for all s is equivalent to Average Causal Necessity, whereas $CEP^{risk}(s_1, s_0) \neq 0$ for all $|s_1 - s_0| > C$ (or the 1-sided analog) 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 often have $|CEP^{risk}(s_1, s_0)|$ 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 $|CEP^{risk}(s_1, s_0)|$ to 0 for s_1 near s_0 and by the extent to which $|CEP^{risk}(s_1, s_0)|$ increases with $|s_1 - s_0|$,

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 $|CEP^{risk}(s_1, s_0)|$ 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 cdf of S. Note that $R(v) = risk(S = F^{-1}(v))$, 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, $CEP^R(v_1, v_0) \equiv h(R_{(1)}(v_1, v_0), R_{(0)}(v_1, v_0))$, where

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

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 1-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:

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

where $risk_{(Z)}(s_1) \equiv \Pr(Y(Z) = 1|S(1) = s_1)$. Similarly $mCEP^R(v_1)$ 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 $mCEP^R(\cdot)$ 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 $CEP^{risk}(c, c)$ and the associative effects by $CEP^{risk}(s_1, c)$ for $s_1 \neq c$. For example, with $c = L$ the lower bound of S, the nearer $CEP^{risk}(c, c)$ is to zero and the greater the increase of $|CEP^{risk}(s_1, c)|$ 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 $CEP(s_1, s_0)$. In general, however, $mCEP(s_1)$ does not measure the association between causal biomarker effects and causal clinical effects, and hence does not measure principal surrogate value. Nevertheless, under A1 and A2 $mCEP(s_1)$ 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/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 1-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)*: $EAE(w) \equiv$

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

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

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

We also define the *proportion associative effect* by

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

Values of $PAE(w) \in [0, 0.5]$ suggest the biomarker has no surrogate value, while values in $(0.5, 1]$ suggest some surrogate value.

A weight function is included in $EAE(w)$ to reflect the idea that a biomarker with high surrogate value should have large $|CEP^{risk}(s_1, s_0)|$ for large $s_1 - s_0 > 0$. 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: $PAE(w) = |CEP^{risk}(U, L)|/\{|EDE| + |CEP^{risk}(U, L)|\}$.

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 $PAE(w = 1)$ equals the *proportion associative (PA)*, defined by

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

(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 $PAE(w)$ 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.

Biomarkers satisfying Average Causal Necessity have $EDE = 0$ and thus $PAE(w) = 1$, in which case $EAE(w)$ contributes no information to the $PAE(w)$. Therefore additional measures are needed for summarizing the magnitude of associative effects. One such measure is the *associative span (AS)*, defined by $AS \equiv |CEP^{risk}(U, L)| - |EDE|$. Figure 2 illustrates $PAE(w = 1)$ and AS. While 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.

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

$$\begin{aligned} risk_{(1)}(s_1, s_0) &= \Pr(Y = 1 | Z = 1, V = 1, S = s_1, S(0) = s_0) \quad \text{and} \\ risk_{(0)}(s_1, s_0) &= \Pr(Y = 1 | Z = 0, V = 1, S(1) = s_1, S = s_0), \end{aligned}$$

demonstrating that $risk_{(Z)}(s_1, s_0)$ would be identified if we knew the potential outcomes

$S_i(1-Z_i)$'s for the treatment not assigned. Estimating the CEP surface will therefore require a way to predict the missing potential biomarkers. This challenge is relatively easy in Case CB, under which $risk_{(1)}(s_1, c) = risk(s_1|Z = 1)$, i.e., $risk_{(1)}(s_1, c)$ is identified by the observed data in arm $Z = 1$. However, A1-A3 do not identify $risk_{(0)}(s_1, c)$, and the remaining task to identify the CEP surface entails determining values $S_i(1)$ for arm $Z_i = 0$ subjects. This same task will suffice to identify the marginal CEP curve in the general case, because A1-A3 identify $risk_{(1)}(s_1)$ by $risk(s_1|Z = 1)$.

4.2 Baseline Irrelevant Predictor Study Design and Likelihood

We consider estimation of the CEP surface under A1-A3 for the special Case CB, with the constant value c for $S_i(0)$ equal to the realized lower bound L of the biomarker $S(1)$, $c = L = min\{S_i(1)\} = min\{S_i|Z_i = 1\}$. The identical approach will provide estimates of the marginal CEP curve in the general case. The method is based on one of the augmented vaccine trial designs proposed by Follmann (2006), wherein a baseline covariate W that is predictive of $S(1)$ is measured in subjects in both treatment arms. A model predicting $S(1)$ from 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 $S(1)|Z = 1, W =^d S(1)|Z = 0, W$, where $=^d$ denotes equality in distribution.

We assume the baseline predictor W is not a risk factor after accounting for $S(1)$:

A4 *Baseline Irrelevant Predictor*: $Y(Z)|W, S(1) =^d Y(Z)|S(1) \qquad$ for $Z = 0, 1.$

In the discussion we consider an approach to relax A4.

Whereas Follmann (2006) considered complete sampling, we consider outcome-dependent case-cohort sampling, wherein W is measured for all or almost all cases (those with $Y = 1$) and for a random "sub-cohort" of controls (with $Y = 0$). The biomarker S is measured for all arm $Z = 1$ subjects with W measured. Case-cohort sampling is efficient when W or S is expensive (Prentice, 1986). For vaccine trials, W

and $S(1)$ can be measured after the trial using stored specimens (Gilbert et al., 2005).

Let δ_i indicate whether W_i is measured. We observe iid data $O_i \equiv (Z_i, V_i, Y_i, \delta_i, \delta_i W_i, \delta_i Z_i S_i), i = 1, \ldots, n$. Only subjects i with $V_i = 1$ contribute terms to the likelihood. Subjects with $Z_i\delta_i = 1$ contribute $risk_{(1)}(S_i, c; \beta)^{Y_i}(1 - risk_{(1)}(S_i, c; \beta))^{1-Y_i}$, where $risk_{(1)}(\cdot, c; \beta)$ 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 $risk_{(0)}(\cdot, c; \beta)$ over the conditional cdf $F_{(1)}^{S|W}$ of $S(1)|W$, i.e., $\int risk_{(0)}(s_1, c; \beta) dF_{(1)}^{S|W}(s_1|W_i)$ or one minus this integral; note that A4 is used here. Subjects with $\delta_i = 0$ contribute $\int risk_{(Z_i)}(s_1, c; \beta) dF_{(1)}(s_1)$ or one minus this integral. Thus, with $\nu \equiv (F_{(1)}^{S|W}, F_{(1)})$, the conditional likelihood is $L(\beta, \nu) \equiv \prod_{i=1}^n f(Y_i|Z_i, V_i, \delta_i, \delta_i W_i, \delta_i Z_i S_i)^{V_i}$, where

$$
\begin{aligned}
f(Y|Z, V, \delta, \delta W, \delta ZS) &= \left\{risk_{(1)}(S, c; \beta)^Y (1 - risk_{(1)}(S, c; \beta))^{1-Y}\right\}^{Z\delta} \\
&\times \left\{\left(\int risk_{(0)}(s_1, c; \beta) dF_{(1)}^{S|W}(s_1|W)\right)^Y \left(1 - \int risk_{(0)}(s_1, c; \beta) dF_{(1)}^{S|W}(s_1|W)\right)^{1-Y}\right\}^{(1-Z)\delta} \\
&\times \left\{\left(\int risk_{(Z)}(s_1, c; \beta) dF_{(1)}(s_1)\right)^Y \left(1 - \int risk_{(Z)}(s_1, c; \beta) dF_{(1)}(s_1)\right)^{1-Y}\right\}^{(1-\delta)}. \qquad (4)
\end{aligned}
$$

Since $CEP^{risk}(\cdot, c; \beta)$ 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 $F_{(1)}^{S|W}$ and $F_{(1)}$. We use estimated likelihood (Pepe and Fleming, 1991), also called pseudo-likelihood, wherein consistent estimates of ν are obtained based on treatment arm 1 data, and then $L(\beta, \widehat{\nu})$ is maximized in β. The bootstrap is used to get standard errors for $\widehat{\beta}$. A re-sampling approach seems to be required because there is no analytic expression for the asymptotic variance of $\widehat{\beta}$ that accounts for the variations in $\widehat{\nu}$, 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 is observed (e.g., Pepe and Fleming 1991).

4.3 Models for $risk_{(Z)}(\cdot, c)$ and $\nu = (F_{(1)}^{S|W}, F_{(1)})$

The estimated likelihood approach can be used for a variety of models for $risk_{(Z)}(s_1, c)$ and the nuisance parameters ν. Here we consider two types of models. The first is fully parametric, where we assume $(W, S(1))$ follows a particular distribution and the risk functions have a generalized linear model form

$$risk_{(1)}(s_1, c; \beta_1) = g(\beta_{10} + \beta_{11} s_1) \quad \text{and} \quad risk_{(0)}(s_1, c; \beta_0) = g(\beta_{00} + \beta_{01} s_1), \tag{5}$$

for $s_1 \geq c$ and some known link function $g(\cdot)$. For example, similar to Follmann (2006), we might assume $(W, S(1))$ is bivariate normal and (5) holds with g equal to the standard normal cdf Φ. With $h(x, y) = g^{-1}(x) - g^{-1}(y)$, model (5) implies

$$CEP^{risk}(s_1, c) = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01}) s_1.$$

Simple calculations yield $EDE = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})L$, $AS = |(\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})U| - |EDE|$, and $EAE(w = 1) = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})E[S(1)|S(1) > c]$. The parametric approach can accommodate a general parametric model for $(W, S(1))$. If W is high-dimensional, then assumptions alternative to A4 are desirable (see Discussion).

We also 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. Nonparametric models are specified by $\theta_{jk} \equiv \Pr(S(1) = j, W = k)$, $f_{(1)}(j) \equiv \Pr(S(1) = j) = \sum_{k=1}^{K} \theta_{jk} \equiv \theta_j$, $f_{(1)}^{S|W}(j|k) \equiv \Pr(S(1) = j|W = k) = \theta_{jk} / \sum_{l=1}^{J} \theta_{lk}$, and $risk_{(Z)}(j, 1; \beta) = \beta_{Zj}$, for $Z = 0, 1; j = 1, \ldots, J; k = 1, \ldots, K$. Then $CEP^{risk}(j, 1; \beta) = h(\beta_{1j}, \beta_{0j})$, $AS = |h(\beta_{1J}, \beta_{0J})| - |h(\beta_{11}, \beta_{01})|$, $EDE = h(\beta_{11}, \beta_{01})$, and $EAE(w) = \sum_{j=2}^{J} \tilde{w}(j, 1) h(\beta_{1j}, \beta_{0j})$ with $\tilde{w}(j, 1) = w(j, 1)\theta_j / \sum_{l=2}^{J} w(l, 1)\theta_l$.

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

4.4 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 maximum estimated likelihood estimator (MELE)

$\widehat{PAE}(w)$ minus 0.5 divided by its bootstrap standard error. Similarly Wald tests of $AS = 0$ can be implemented based on $\widehat{AS}$. For categorical $(W, S(1))$ 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, 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 mean 0.41, standard deviation 0.55, and correlation $\rho = 0.5$, 0.7, and 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_i 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 $S_i(1)$, $Y_i(Z)$ was generated from Bernoulli(β_{Zj}), with β_{Zj} set to achieve the infection rate $\Pr(Y(1) = 1) = 0.067$ 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 only for all infected vaccine recipients and a sub-cohort 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 sub-cohort of uninfected placebo recipients. For each arm, the ratio of controls to cases was 3:1. The simulated data sets satisfied A1-A4.

For each of 1000 simulated data sets the MELE $\widehat{\beta}$ was computed using the nonparametric approach described in Section 4.3. Then, with $h(x,y) = \log(x/y)$, $\widehat{\beta}$ was used to compute the MELEs of $CEP^{risk}(j,1)$, AS, and $PAE(w)$ for $w(j,1) = 1, j$, and $I(j = J = 4)$. Wald tests (with bootstrap standard errors) based on $\widehat{PAE}(w) - 0.5$, $\widehat{AS}$, and T were used to test for any surrogate value.

The MELEs of $CEP^{risk}(j,1)$, $PAE(w)$ and AS performed well (Tables 1 and 2). The MELEs exhibited negligible bias and the confidence intervals (CIs) about them had nominal coverage, with coverage improving for higher ρ. The tests for any surrogate value had approximately nominal size and showed high power to detect surrogate value; the nonparametric trend test had power 0.88, 0.99, and 1.00 for $\rho = 0.5$, 0.7, and 0.9.

Additional simulations were conducted to evaluate the performance of the MELE method with binned (quartilized) covariates when the data are 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 probit model (5) for $risk_{(1)}(s_1, c; \beta)$

with $\beta_{11} = -0.73$ and $\beta_{10} = -1.23$, 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 probit model (5) with either (i) $\beta_{01} = \beta_{11}$ or (ii) $\beta_{01} = 0$. In (i) $CEP^{risk}(s_1, c; \beta) = \beta_{10} - \beta_{00} = -1.23 - (-0.825) = -0.405$, so that S has no surrogate value (with $AS = 0$); in (ii) $CEP^{risk}(s_1, c; \beta) = -0.405 - 0.73 s_1$ so that S has high surrogate value (with $AS = 0.73$). Using $h(x, y) = \Phi^{-1}(x) - \Phi^{-1}(y)$, the MELEs and CIs for $CEP^{risk}(j, 1)$ performed well (results not shown), as did the MELEs and CIs for $PAE(w)$ and AS (Table 3). Tests for any surrogate value had approximately nominal size, with power only slightly lower than in the previous set of simulations. This simulation study provides 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 causal effect predictiveness (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; such inferences apply for measuring surrogate predictiveness for the same or similar setting as the trial. Thus the inferences do not form an empirical basis for bridging information about clinical efficacy to a new setting (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 meta-analysis are needed.

Since 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-A4, a reasonably

good model predicting S from W in treatment arm 1, and models for $risk_{(Z)}(s_1, c)$ or its marginal counterpart $risk_{(Z)}(s_1)$, 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. While A3 is untestable, violations of it will not significantly influence the results if the vast majority of participants are at-risk when the biomarker is measured, since in this case A3 will hold for almost all subjects. Otherwise it will be important to extend the methods to facilitate sensitivity analyses to departures from A3. Models for S given W can be directly checked using arm $Z = 1$ data, and under A1-A4 parametric modeling assumptions placed on $risk_{(Z)}(s_1, c)$ can be tested.

Assumption A4 is testable for treatment arm $Z = 1$ but not for $Z = 0$. A4 can be very strong, because it requires that after accounting for $S(1)$ the baseline predictor W of $S(1)$ does not predict clinical risk. If W is a vector containing multiple clinical risk factors then A4 will likely be implausible, so that in practice a well-chosen low dimensional W is desired. For example, a study that vaccinated 75 individuals simultaneously with hepatitis A and B vaccines showed a linear correlation of 0.85 among A- and B-specific antibody titers (Czeschinski, Binding, and Witting, 2000). Given there is little cross-reactivity among the hepatitis A and B proteins, W = hepatitis A titer may be an excellent baseline predictor for $S(1)$ = hepatitis B titer that satisfies A4. For HIV vaccine trials, two available scalar W's may plausibly satisfy A4. First, Follmann (2006) considered as W the antibody titer to a rabies glycoprotein vaccine. Because rabies is not acquired sexually, it is plausible that anti-rabies antibodies are independent of risk of HIV infection given $S(1)$. Second, in the ongoing HIV vaccine efficacy trials, a current leading candidate W is the titer of antibodies that neutralize

the Adenovirus serotype 5 vector that carries the HIV genes in the vaccine. This W has been shown to inversely correlate with the $S(1)$ of primary interest (T cell response levels measured by ELISpot) (Catanzaro et al., 2006), and since Adenovirus 5 is a respiratory infection virus, A4 may plausibly hold. In general detailed biological knowledge may help identify a suitable W.

Additionally, A4 may be replaced with other assumptions that may be more plausible. For example, consider the fully parametric approach, expanded to allow a vector-valued W and a vector X of baseline covariates measured on all subjects. Substitute for A4 and the risk models (5) the modeling assumption

$$risk_{(Z)}(s_1, c, w, x; \beta_Z) = g(\beta_{Z0} + \beta_{Z1} s_1 + \beta'_{Z2} w + \beta'_{Z3} x) \text{ for } Z = 0, 1, \tag{6}$$

where $risk_{(Z)}(s_1, c, w, x; \beta_Z) \equiv \Pr(Y(Z) = 1 | S(1) = s_1, S(0) = c, W = w, X = x)$ and $g(\cdot)$ is a known link function. With $h(x, y) = g^{-1}(x) - g^{-1}(y)$, (6) implies $(\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01}) s_1$ is the CEP^{risk} surface controlling for W and X. For the case that $g = \Phi$ and $(S(1), W', X')'$ has a multivariate normal distribution, Web Appendix C provides a proof, adapted from a proof by Dean Follmann, that $\beta_1 = (\beta_{10}, \beta_{11}, \beta'_{12}, \beta'_{13})'$ and $\beta_0 = (\beta_{00}, \beta_{01}, \beta'_{02}, \beta'_{03})'$ are identified under the constraint that one of the components of $(\beta'_{12}, \beta'_{13})'$ is supposed equal to the corresponding component of $(\beta'_{02}, \beta'_{03})'$. This approach allows clinical risk to depend on W and X, alleviating the "irrelevancy" condition in A4. The MELE $(\widehat{\beta}, \widehat{\theta})$ may be calculated as described in Section 4 and Web Appendix B, with minor modifications to the estimated likelihood (4).

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 estimands describe how the average or population level causal effect on Y depends on the causal effect on S. R code for the nonparametric methods with Y binary is available at the second author's website.

Supplementary Materials

Web Appendices referenced in Sections 2.1, 3.1, 3.2, 4.3, and 6 are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgements

The authors thank Dean Follmann, Margaret Pepe, Ross Prentice, and Steve Self for helpful comments, and Allan DeCamp for producing Figure 1. 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.

Catanzaro, A., Koup, R. and Roederer, M. (2006). Safety and immunogenicity evaluation of a multiclade HIV-1 candidate vaccine delivered by a replication-defective recombinant adenovirus vector. *Journal of Infectious Diseases* **194**, 1638–1649.

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.

Czeschinski, P., Binding, N. and Witting, U. (2000). Hepatitis A and hepatitis B vaccinations: immunogenicity of combined vaccine and of simultaneously or separately applied single vaccines. *Vaccine* **18**, 1074–1080.

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. and 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.

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**.

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.

Table 1

Bernoulli model simulation results for the nonparametric MELEs $\widehat{CEP}^{risk}(j,1) = \log(\widehat{\beta}_{1\mathrm{j}}/\widehat{\beta}_{0\mathrm{j}})$ *for* $j = 1, \cdots, 4$[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	$CEP^{risk}(1,1) = -0.69$	-0.03	0.39	0.38	0.98	0.48	$CEP^{risk}(1,1) = -0.22$	0.04	0.54	0.62	0.98	0.05
	$CEP^{risk}(2,1) = -0.69$	0.09	0.51	0.62	0.98	0.14	$CEP^{risk}(2,1) = -0.51$	0.01	0.54	0.64	0.98	0.10
	$CEP^{risk}(3,1) = -0.69$	0.09	0.54	0.74	1.00	0.08	$CEP^{risk}(3,1) = -0.92$	0.02	0.52	0.65	0.99	0.23
	$CEP^{risk}(4,1) = -0.69$	-0.10	1.23	1.24	0.97	0.14	$CEP^{risk}(4,1) = -1.61$	0.01	0.60	0.63	0.98	0.64
0.7	$CEP^{risk}(1,1) = -0.69$	0.01	0.27	0.27	0.97	0.65	$CEP^{risk}(1,1) = -0.22$	-0.01	0.35	0.39	0.97	0.11
	$CEP^{risk}(2,1) = -0.69$	0.00	0.41	0.48	0.98	0.27	$CEP^{risk}(2,1) = -0.51$	0.02	0.45	0.52	0.98	0.14
	$CEP^{risk}(3,1) = -0.69$	0.05	0.45	0.55	0.99	0.19	$CEP^{risk}(3,1) = -0.92$	0.01	0.46	0.52	0.98	0.42
	$CEP^{risk}(4,1) = -0.69$	-0.01	0.89	1.03	0.97	0.17	$CEP^{risk}(4,1) = -1.61$	0.02	0.43	0.44	0.98	0.82
0.9	$CEP^{risk}(1,1) = -0.69$	0.00	0.18	0.18	0.96	0.93	$CEP^{risk}(1,1) = -0.22$	0.00	0.23	0.23	0.97	0.18
	$CEP^{risk}(2,1) = -0.69$	-0.01	0.37	0.39	0.97	0.45	$CEP^{risk}(2,1) = -0.51$	0.00	0.45	0.45	0.97	0.26
	$CEP^{risk}(3,1) = -0.69$	0.03	0.49	0.53	0.96	0.32	$CEP^{risk}(3,1) = -0.92$	0.00	0.47	0.47	0.95	0.51
	$CEP^{risk}(4,1) = -0.69$	-0.02	0.56	0.54	0.96	0.32	$CEP^{risk}(4,1) = -1.61$	0.02	0.31	0.30	0.96	0.98

[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 $\widehat{CEP}^{risk}(j,1)$. SEE is the median of the bootstrap standard error estimates based on 200 bootstrap replicates. CP is the empirical coverage of standard normal 95% confidence intervals for $\widehat{CEP}^{risk}(j,1)$ using bootstrap standard error estimates. Power refers to power of the Wald test to reject $H_0 : CEP^{risk}(j,1) = 0$. 1000 simulations were done to compute the table elements for each model.

Table 2

Bernoulli model simulation results for the nonparametric MELEs $\widehat{PAE}(w)$ and $\widehat{AS}$[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	$PAE(w1) = 0.50$	-0.03	0.18	0.20	0.94	0.03	$PAE(w1) = 0.82$	-0.07	0.18	0.21	0.96	0.29
	$PAE(w2) = 0.50$	-0.02	0.18	0.20	0.97	0.02	$PAE(w2) = 0.85$	-0.08	0.17	0.19	0.96	0.39
	$PAE(w3) = 0.50$	0.06	0.20	0.20	0.96	0.07	$PAE(w3) = 0.88$	-0.06	0.17	0.18	0.95	0.63
	$AS = 0.00$	0.21	0.93	0.91	0.97	0.05	$AS = 1.39$	-0.14	0.66	0.74	0.99	0.53
0.7	$PAE(w1) = 0.50$	-0.02	0.15	0.17	0.97	0.02	$PAE(w1) = 0.82$	-0.05	0.14	0.17	0.97	0.46
	$PAE(w2) = 0.50$	-0.02	0.15	0.17	0.98	0.01	$PAE(w2) = 0.85$	-0.06	0.13	0.15	0.96	0.61
	$PAE(w3) = 0.50$	0.04	0.19	0.19	0.97	0.05	$PAE(w3) = 0.88$	-0.04	0.12	0.13	0.96	0.83
	$AS = 0.00$	0.09	0.63	0.75	0.98	0.05	$AS = 1.39$	-0.11	0.45	0.50	0.98	0.79
0.9	$PAE(w1) = 0.50$	-0.02	0.12	0.13	0.97	0.01	$PAE(w1) = 0.82$	-0.01	0.12	0.13	0.98	0.71
	$PAE(w2) = 0.50$	-0.02	0.12	0.14	0.98	0.01	$PAE(w2) = 0.85$	-0.02	0.10	0.11	0.97	0.86
	$PAE(w3) = 0.50$	0.02	0.17	0.17	0.93	0.06	$PAE(w3) = 0.88$	-0.01	0.09	0.09	0.98	0.96
	$AS = 0.00$	0.04	0.46	0.46	0.96	0.06	$AS = 1.39$	-0.05	0.37	0.37	0.95	0.96

[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 $\widehat{PAE}(w)$ and $\widehat{AS}$. SEE is the median of the bootstrap standard error estimates based on 200 bootstrap replicates. CP is the empirical coverage of standard normal 95% confidence intervals for $PAE(w)$ and AS using bootstrap standard error estimates. Power is for 1-sided tests of $H_0 : PAE(w) = 0.5$ versus $H_1 : PAE(w) > 0.5$ or $H_0 : AS = 0$ versus $H_1 : AS > 0$ at level $\alpha = 0.05$. For the PAE weights, $w_1(j, 1) = 1$, $w_2(j, 1) = j$, and $w_3(j, 1) = I[j = J = 4]$. 1000 simulations were done to compute the table elements for each model.

Table 3

Probit model simulation results for the nonparametric MELEs $\widehat{PAE}(w)$ and $\widehat{AS}$[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	$PAE(w1) = 0.50$	-0.06	0.19	0.19	0.92	0.04	$PAE(w1) = 0.82$	-0.09	0.17	0.20	0.96	0.29
	$PAE(w2) = 0.50$	-0.05	0.18	0.19	0.95	0.04	$PAE(w2) = 0.85$	-0.10	0.16	0.19	0.95	0.37
	$PAE(w3) = 0.50$	0.00	0.20	0.20	0.96	0.03	$PAE(w3) = 0.87$	-0.08	0.17	0.19	0.94	0.53
	$AS = 0.00$	0.01	0.35	0.36	0.98	0.03	$AS = 0.73$	-0.13	0.34	0.37	0.97	0.48
0.7	$PAE(w1) = 0.50$	-0.03	0.15	0.17	0.95	0.03	$PAE(w1) = 0.82$	-0.04	0.15	0.17	0.96	0.49
	$PAE(w2) = 0.50$	-0.03	0.15	0.16	0.96	0.02	$PAE(w2) = 0.85$	-0.05	0.13	0.15	0.95	0.60
	$PAE(w3) = 0.50$	-0.02	0.18	0.19	0.96	0.03	$PAE(w3) = 0.87$	-0.04	0.12	0.14	0.96	0.78
	$AS = 0.00$	-0.03	0.28	0.29	0.99	0.03	$AS = 0.73$	-0.09	0.24	0.27	0.98	0.75
0.9	$PAE(w1) = 0.50$	-0.02	0.12	0.13	0.96	0.02	$PAE(w1) = 0.82$	-0.02	0.13	0.13	0.97	0.69
	$PAE(w2) = 0.50$	-0.02	0.12	0.13	0.97	0.02	$PAE(w2) = 0.85$	-0.02	0.11	0.11	0.95	0.82
	$PAE(w3) = 0.50$	-0.02	0.17	0.16	0.93	0.04	$PAE(w3) = 0.87$	-0.02	0.10	0.10	0.97	0.93
	$AS = 0.00$	-0.04	0.23	0.23	0.96	0.05	$AS = 0.73$	-0.04	0.20	0.20	0.95	0.94

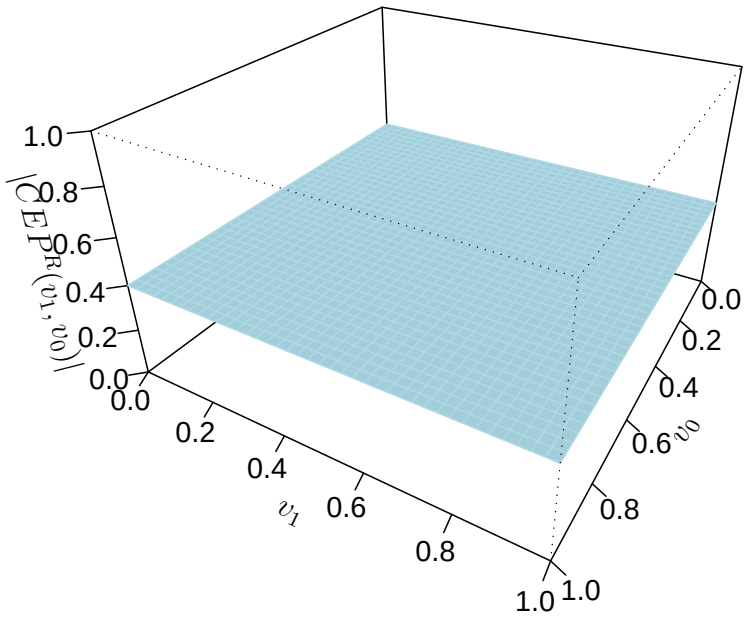
[a] ρ is the linear correlation of the simulated bivariate normal variables W and $S(1)$. Bias is the median bias. SE is the empirical standard error of $\widehat{PAE}(w)$ and $\widehat{AS}$. SEE is the median of the bootstrap standard error estimates based on 200 bootstrap replicates. CP is the empirical coverage of standard normal 95% confidence intervals for $PAE(w)$ and AS using bootstrap standard error estimates. Power is for 1-sided tests of $H_0 : PAE(w) = 0.5$ versus $H_1 : PAE(w) > 0.5$ or $H_0 : AS = 0$ versus $H_1 : AS > 0$ at level $\alpha = 0.05$. For the PAE weights, $w_1(j,1) = 1$, $w_2(j,1) = j$, and $w_3(j,1) = I[j = J = 4]$. 1000 simulations were done to compute the table elements for each model.

Figure Legends

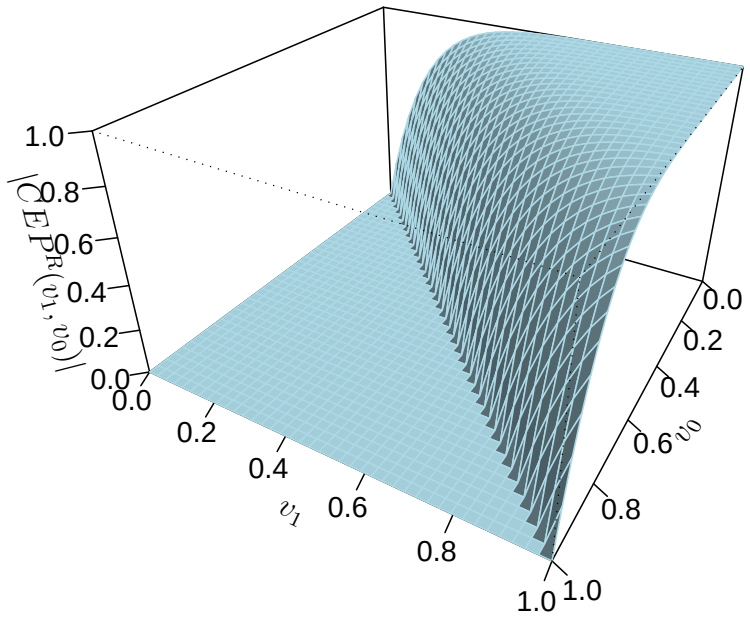
Figure 1. Example $CEP^R(v_1, v_0) = h(R_{(1)}(v_1, v_0), R_{(0)}(v_1, v_0))$ surfaces, with $h(x, y) = x - y$ or $1 - x/y$. The surface in (i) reflects a biomarker with no surrogate value, wherein the clinical treatment effect is the same for all treatment effects on the biomarker. The surface in (ii) 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 $CEP^R(v_1, v_0) = 0$ for $v_1 = v_0$, both biomarkers satisfy Average Causal Necessity. Furthermore, because $CEP^R(v_1, v_0) > 0$ for all $v_1 > v_0$, the biomarker in (ii) satisfies 1-sided Average Causal Sufficiency.

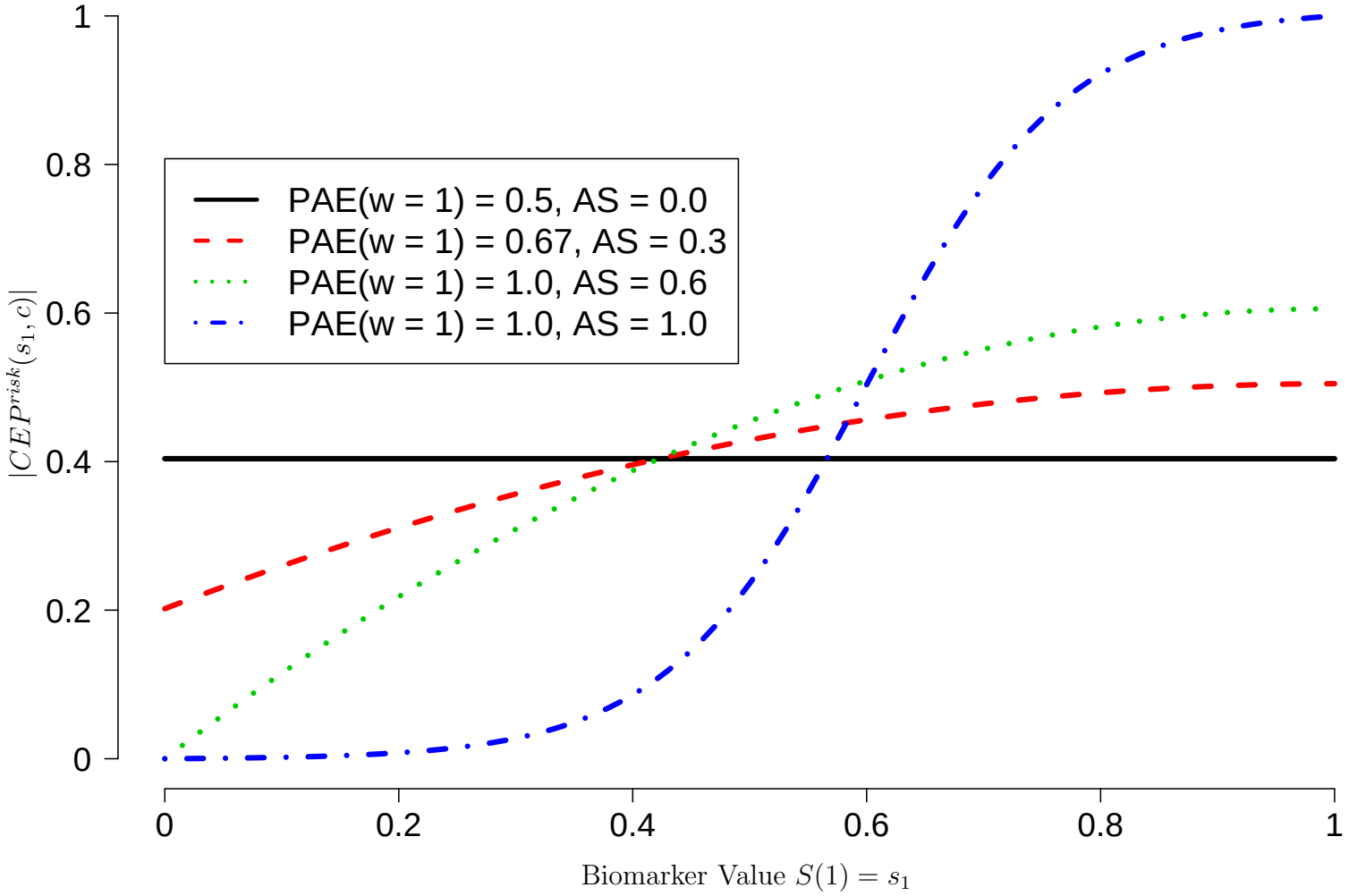
Figure 2. For Case CB with $S_i(0) = c$ for all i with $c = L$ the lower bound of S, biomarkers S that have no (horizontal solid line), modest (dashed line), moderate (dotted line), and high (hatched line) surrogate value. Because $CEP^{risk}(c, c) = 0$ and $CEP^{risk}(s_1, c) > 0$ for all $s_1 > c$, the latter two S's satisfy Average Causal Necessity and Average Causal Sufficiency, and hence are principal surrogates.

(i) No surrogate value



(ii) High surrogate value




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
$|CEP^{risk}(s_1, c)|$
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1
Biomarker Value $S(1) = s_1$