STATISTICS IN MEDICINE
Statist. Med. (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3122

Evaluating a surrogate endpoint at three levels, with application to vaccine development

Peter B. Gilbert*, †, Li Qin and Steven G. Self

Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, WA 98109, U.S.A.

SUMMARY

Identification of an immune response to vaccination that reliably predicts protection from clinically significant infection, i.e. an immunological surrogate endpoint, is a primary goal of vaccine research. Using this problem of evaluating an immunological surrogate as an illustration, we describe a hierarchy of three criteria for a valid surrogate endpoint and statistical analysis frameworks for evaluating them. Based on a placebo-controlled vaccine efficacy trial, the first level entails assessing the correlation of an immune response with a study endpoint in the study groups, and the second level entails evaluating an immune response as a surrogate for the study endpoint that can be used for predicting vaccine efficacy for a setting similar to that of the vaccine trial. We show that baseline covariates, innovative study design, and a potential outcomes formulation can be helpful for this assessment. The third level entails validation of a surrogate endpoint *via* meta-analysis, where the goal is to evaluate how well the immune response can be used to predict vaccine efficacy for new settings (building bridges). A simulated vaccine trial and two example vaccine trials are presented, one supporting that certain anti-influenza antibody levels are an excellent surrogate for influenza illness and another supporting that certain anti-HIV antibody levels are not useful as a surrogate for HIV infection. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: clinical trial; counterfactual; immune correlate; meta-analysis; potential outcomes; principal surrogate; statistical surrogate

1. INTRODUCTION

Identification of a vaccine-induced immune response that predicts whether vaccine recipients will be protected from disease with a pathogen (i.e. an 'immune correlate') is a primary goal of vaccine research [1–3]. Owing to resource limitations, large clinical trials that provide direct estimates

*Correspondence to: Peter B. Gilbert, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, U.S.A.
†E-mail: pgilbert@scharp.org

Contract/grant sponsor: NIH; contract/grant number: 2 R01 AI54165-04

Received 19 September 2007
Accepted 20 September 2007

of vaccine efficacy can be done only over a limited spectrum of settings (geographic regions, genetics of human populations and pathogen populations, vaccine formulations, etc.). Therefore, the availability of an immune correlate would allow prediction of vaccine efficacy for new settings where direct data on vaccine efficacy are not available.

Despite the long history of searching for immune correlates and the consensus on the importance of finding them (e.g. it is one of the 14 'Grand Challenges of Global Health' of the NIH and Gates Foundation), there is only a small statistical literature for their evaluation [4–6]. Furthermore, three different conceptual definitions of immune correlates have been implicitly used, which has led to confusion about what exactly is assessable in vaccine trials. The three definitions of immune correlate form a hierarchy in the amount of information that is needed to evaluate them and in the strength of empirical validation for forming a basis for reliably predicting vaccine efficacy in new settings. The extensive statistical literature on the evaluation of surrogate endpoints suggests ways to evaluate the three kinds of immune correlates (reviewed in [7]), and in a companion clinical paper we summarized an approach for 'technology transfer' of some of these methods to evaluating the three kinds of immune correlates [8]. This article describes in greater statistical detail our approach to applying certain surrogate endpoint evaluation methods to the assessment of immune correlates and expands on new methodology for evaluating a level 1 surrogate endpoint.

2. THREE TIERS OF AN IMMUNOLOGICAL SURROGATE ENDPOINT

For concreteness we assume a randomized placebo-controlled vaccine efficacy trial, with the immune response of interest S measured at fixed time t_0 after randomization. Let Y be the study endpoint (e.g. clinically significant disease) and Z be the vaccination status ($Z=1$, vaccine; $Z=0$, placebo).

Table I defines the three kinds of immune correlates, which we refer to as a correlate of risk (CoR), a surrogate of protection (SoP) at validation level 1, and a SoP at validation level 2. Level 1 validation entails evaluating the reliability of the biomarker for predicting vaccine efficacy for the same setting as the trial, while the higher level 2 validation evaluates reliability for predicting vaccine efficacy across different settings (building bridges).

2.1. Correlate of risk (CoR)

A CoR is simply an immune response that is correlated with the rate or level of a study endpoint that is relevant to pathogen-specific disease (i.e. S is a CoR if it correlates with Y). A CoR can be assessed in epidemiological studies or in preventive vaccine efficacy or proof-of-concept trials. In efficacy and proof-of-concept trials, interest centers on assessing CoRs for the primary study endpoint within each randomized study group. Examples of primary endpoints in vaccine trials are (i) acute hepatitis illness with detection of positive hepatitis B surface antigens [9]; (ii) severe diarrheal illness with rotavirus [10] or *Vibrio cholerae* [11] isolated in stool; and (iii) cervical infection with type 16 or 18 human papillomavirus [12]. Although identifying CoRs for true clinical endpoints may be of greatest interest, it is also sometimes of interest to identify CoRs for biomarker study endpoints; for example, an objective of current vaccine efficacy trials of a T cell-based HIV vaccine is the evaluation of potential CoRs for set-point viral load [13].

Substantial variability of an immunological measurement among sampled individuals is necessary to evaluate the measurement as a CoR. We distinguish two types of putative CoRs, those that vary in both the vaccine and placebo arms of the trial, and those that have no or very limited

DOI: 10.1002/sim

Table I. Definitions of three levels of an immunological correlate of protection.

Term	Definition	Framework for assessment	Analytic method
CoR (Correlate of risk)	An immunological measurement S that correlates with the study endpoint Y measuring vaccine efficacy in a defined population	Vaccine trial (efficacy or proof of concept) or epidemiological study	Regression models
Level 1 (**specific**) SoP (Surrogate of protection for the same setting)	An immunological measurement that is a CoR within a defined population of vaccine recipients and satisfies either:		
SoP^S (Statistical surrogate of protection for the same setting)	The relationship between the immunological measurement S and endpoint Y is the same in the vaccine and placebo groups ($Y\|S=s, Z=1 =^d Y\|S=s, Z=0$ for all s)	Single large efficacy trial	Statistical surrogate framework
SoP^P (Principal surrogate of protection for the same setting)	The immune response S satisfies average causal necessity and average causal sufficiency defined in Section 2.2	Single large efficacy trial	Principal surrogate framework
Level 2 (**general**) SoP (Surrogate of protection for new setting)	An immunological measurement that is predictive of vaccine efficacy in different settings (e.g. across human populations, viral populations, vaccine formulations)	Multiple trials (efficacy or proof of concept) and/or post-licensure studies	Meta-analysis

variability in the placebo arm. If many of the trial participants have been infected with the pathogen under investigation prior to enrolling into the trial (as for influenza vaccine trials), then the former case likely prevails. However, if trial participants have never been infected with the pathogen (as in HIV and human papillomavirus vaccine trials), then most or all placebo recipients will have no immune response, because the immune response is pathogen specific. If there is variability of the immune response in the placebo arm, then a potentially useful measure of the strength of CoR is the adjusted association [14], which measures the correlation between the biomarker and the clinical endpoint, adjusting for vaccination status. If there is no variability of S in the placebo arm, then standard measures of correlation within the vaccine arm can be used, such as the relative risk or parameters that account for the distribution of the biomarker in the population [15]. The type of immunological biomarker influences the pros and cons of the frameworks considered below for evaluating SoPs.

Identifying biomarkers as CoRs is a ubiquitous objective in epidemiological studies and clinical trials for many diseases, and for many surrogate evaluation methods the first step is to evaluate a biomarker as a CoR. Discovering that a biomarker is a CoR may raise the hypothesis that it has some value as a surrogate endpoint.

2.2. Level 1 SoP

An SoP is a CoR such that contrasts in the immune response in the vaccine and placebo groups of an efficacy trial reliably predict the level of vaccine efficacy against a study endpoint. Level 1 or 2 refers to what can be reliably predicted—level 1 refers to reliable prediction for the same setting

DOI: 10.1002/sim

as studied in the efficacy trial, while level 2 refers to reliable prediction across different settings. As emphasized in the statistical literature, there are a number of ways in which a CoR can fail to be a level 1 SoP; for example, if the vaccine impacts the study endpoint through a mechanism that does not involve the CoR [16–18].

Two main frameworks have been employed in the statistical literature for evaluating a level 1 surrogate endpoint based on one large clinical trial, which directly apply for evaluating a level 1 SoP. The first approach evaluates whether the CoR approximately satisfies the Prentice [19] definition of a surrogate endpoint; a variety of criteria have been used to evaluate consistency with the Prentice definition. These methods base the assessment on the statistical associations observable in the trial, and hence we refer to this type of SoP as a *statistical* SoP. The main criterion of several methods taking this approach is that the observed vaccine efficacy is 'completely explained' in a statistical model by the immunological measurements, or in other words, the effect of the vaccine on the study endpoint is fully mediated through the biomarker.

A second approach for assessing a level 1 SoP, still in its early development, is based on the potential outcomes framework of causal inference [20, 21]. Within this framework, proposed by Frangakis and Rubin [22] with related ideas found in Robins [23], each trial participant has a potential immune response $S_i(Z)$ if assigned vaccine ($Z=1$) and if assigned placebo ($Z=0$) as well as potential study endpoints $Y_i(Z)$, for $Z=0,1$. Furthermore, let $V_i(Z)$ be the potential indicator of whether the ith subject has not yet had the endpoint $Y_i(Z)=1$ by the time t_0 that the immune response is measured. For subjects in arm Z, $S=S(Z)$ and $Y=Y(Z)$ are observed, whereas $S(1-Z)$ and $Y(1-Z)$ are counterfactuals. With this notation we have implicitly assumed the following.

A1: *Stable Unit Treatment Values* (*SUTVA*) [24]. SUTVA implies that for each subject i the potential outcomes $(S_i(1), S_i(0), Y_i(1), Y_i(0))$ are unaffected by the treatment assignments Z_j of other subjects. We also assume the following throughout.

A2: *Ignorable treatment assignments*. This states that the $(V_i(1), V_i(0), S_i(1), S_i(0), Y_i(1), Y_i(0))$ are independent of the treatment assignment Z_i. A2 holds in randomized and placebo-controlled trials with integrity of randomization and blinding. Note that in the absence of measurement error in the immunological assay, $S_i(1) \geqslant S_i(0)$ will often hold.

The causal vaccine efficacy to prevent disease $Y=1$ can be defined as [25, 26]

$$\mathrm{VE} \equiv 1 - \frac{\Pr(Y(1)=1)}{\Pr(Y(0)=1)}$$

Under A1 and A2, $\mathrm{VE}=1-\Pr(Y=1|Z=1)/\Pr(Y=1|Z=0)$, identifying VE from the observed data. A causal estimand for measuring the surrogate value of a biomarker can be defined similar to VE by conditioning on the joint potential outcomes $(S(1), S(0))$, as we now describe.

Frangakis and Rubin [22] gave a definition of a principal surrogate endpoint, and Gilbert and Hudgens [27] suggested a modified definition that we employ here. For fixed levels s_1 of $S(1)$ and s_0 of $S(0)$, define

$$\mathrm{risk}_{(1)}(s_1, s_0) \equiv \Pr(Y(1)=1|V(1)=1, V(0)=1, S(1)=s_1, S(0)=s_0)$$

and

$$\mathrm{risk}_{(0)}(s_1, s_0) \equiv \Pr(Y(0)=1|V(1)=1, V(0)=1, S(1)=s_1, S(0)=s_0)$$

These risks condition on $V(1)=V(0)=1$ because the potential immune responses $S(1)$ and $S(0)$ are both defined only in this subpopulation. A contrast in $\mathrm{risk}_{(1)}(s_1, s_0)$ and $\mathrm{risk}_{(0)}(s_1, s_0)$ measures

DOI: 10.1002/sim

a population level or average causal effect on Y for subjects with $\{V_i(1)=V_i(0)=1, S_i(1)=s_1, S_i(0)=s_0\}$. In particular, we consider the causal estimand

$$\mathrm{VE}(s_1,s_0)\equiv 1-\frac{\mathrm{risk}_{(1)}(s_1,s_0)}{\mathrm{risk}_{(0)}(s_1,s_0)} \tag{1}$$

If the immune response S is measured near baseline, then VE is related to $\mathrm{VE}(s_1,s_0)$ as follows:

$$\begin{aligned}\mathrm{VE} &= 1-\frac{\Pr(Y(1)=1)}{\Pr(Y(0)=1)}\\ &\approx 1-\frac{\Pr(Y(1)=1|V(1)=V(0)=1)}{\Pr(Y(0)=1|V(1)=V(0)=1)}\\ &= 1-\frac{E[\mathrm{risk}_{(1)}(S(1),S(0))]}{E[\mathrm{risk}_{(0)}(S(1),S(0))]}\end{aligned}$$

where the expectations are with respect to the joint distribution of $(S(1),S(0))$. Henceforth, all probabilities involving $(S(1),S(0))$ are implicitly assumed to condition on $V(1)=V(0)=1$.

Gilbert and Hudgens [27] defined a principal surrogate (i.e. a *principal* SoP) as a biomarker satisfying the following two conditions:

Average causal necessity: $\mathrm{VE}(s_1,s_0)=0$ if $s_1=s_0$.

Average causal sufficiency: $\mathrm{VE}(s_1,s_0)>0$ for all $s_1>$ a constant $C\geqslant s_0$.

Average causal necessity states that a positive vaccine effect on the immune response is necessary for protection while average causal sufficiency states that a large enough vaccine effect on the immune response is sufficient for protection. A reviewer helpfully pointed out that the choice of constants 0 and $C\geqslant s_0$ for defining necessity and sufficiency depends on a subtle assumption about the meaning of $S(1)$ and $S(0)$. To explain this, note that $S(1)$ and $S(0)$ are measured using the identical assay and target antigen, but $S(0)$ measures the pre-existing/natural immune response whereas $S(1)$ measures this immune response inseparably combined with any immune response generated by the vaccine. Thus, for example, $S(1)$ may measure 'fresh' antibodies newly generated by the vaccine, whereas $S(0)$ measures older pre-existing antibodies. Consequently, even if the immune response is fully mechanistically causative of protection, $\mathrm{VE}(s_1,s_1)$ may exceed zero (and hence average causal necessity fails) because 'fresher' vaccine-induced memory B cells proliferate better than older memory B cells. In this case, the average causal necessity criterion may be misleading, and the surrogate evaluation could instead be based on the whole surface $\mathrm{VE}(s_1,s_0)$. Indeed, in the presence of the subtle interpretation of $(S(1),S(0))$, the extent to which $\mathrm{VE}(s_1,s_0)$ increases with s_1-s_0 should generally provide meaningful quantification of surrogate value.

While we do not require it in the definition, often a good surrogate will satisfy the condition that vaccine protection does not get worse with greater vaccine effects on the immune response:

Monotone VE: $\mathrm{VE}(s_1,s_0)$ is monotone nondecreasing in s_1-s_0.

While it may be useful to check the three properties given above, we stress that it is most useful to examine the whole surface $\mathrm{VE}(s_1,s_0)$ to provide full information on the predictive value of a biomarker as a surrogate endpoint. To reinforce the importance of this, consider that average causal necessity and sufficiency may hold yet $\mathrm{VE}(s_1,s_0)<0$ for $s_0<s_1\leqslant C$. This scenario is plausible for various pathogens due to the theoretical possibility that low levels of immune response enhance disease while high levels protect; respiratory syncytial virus and dengue illustrate this potential

DOI: 10.1002/sim

phenomenon [28, 29]. Estimating the whole surface $\mathrm{VE}(s_1, s_0)$ would provide a way to discover this phenomenon, whereas only checking necessity and sufficiency could not identify it.

For the case that S does not vary in the placebo arm (such that $S_i(0)=c$ for all i, with $c=0$ without loss of generality), the causal estimand is a curve $\mathrm{VE}(s_1, 0)$. This curve at s_1 has interpretation as the per cent reduction in risk for a vaccinated subpopulation with immune response s_1 compared with if it had not been vaccinated. The more the curve $\mathrm{VE}(s_1, 0)$ increases in s_1 the greater the capacity of the biomarker to predict vaccine efficacy; thus, even if average causal necessity and sufficiency fail, a CoR may still be useful as an SoP if the curve substantially increases in s_1.

For the general case that S may vary in the placebo arm, a marginal causal estimand of interest is

$$\mathrm{mVE}(s_1) \equiv 1 - \frac{\mathrm{risk}_{(1)}(s_1)}{\mathrm{risk}_{(0)}(s_1)} \tag{2}$$

where $\mathrm{risk}_{(Z)}(s_1) \equiv \Pr(Y(Z)=1 | V(1)=1, S(1)=s_1) = E[\mathrm{risk}_{(Z)}(s_1, S(0)) | V(1)=1, S(1)=s_1]$, for $Z=0, 1$. In this estimand the risks under each vaccination assignment average over the conditional cdf of $S(0)$ given $S(1)=s_1$. The estimand $\mathrm{mVE}(s_1)$ is causal because the two risks involved condition on the union of basic principal strata, $\bigcup_{s_0}\{S(1)=s_1, S(0)=s_0\}$. Value of the biomarker to predict vaccine efficacy is indicated by $\mathrm{mVE}(s_1)$ increasing in s_1. In the case that $S_i(0)$ is constant at zero, the marginal estimand (obviously) collapses to the joint estimand: $\mathrm{mVE}(s_1)=\mathrm{VE}(s_1, 0)$.

2.2.1. More on the interpretation of the two estimands $\mathrm{VE}(s_1, s_0)$ *and* $\mathrm{mVE}(s_1)$. The estimand $\mathrm{VE}(s_1, s_0)$ has a clear interpretation for measuring surrogate value, capturing the association between causal vaccine effects on the immune response and causal vaccine effects on the disease endpoint. However, this estimand is complicated by the fact that $(S(1), S(0))$ has an unobservable bivariate distribution, and $\mathrm{VE}(s_1, s_0)$ is only meaningful/defined for points (s_1, s_0) in the support of $(S(1), S(0))$. Therefore, knowledge and/or assumptions about the joint distribution of $(S(1), S(0))$ are needed for determining the region over which $\mathrm{VE}(s_1, s_0)$ is estimated. With $[l_1, u_1]$ the range of observed S's for $Z=1$ subjects and $[l_0, u_0]$ the range of observed S's for $Z=0$ subjects, for some applications it will be reasonable to restrict consideration to the region

$$\{(s_1, s_0) : s_1 \in [l_1, u_1], s_0 \in [l_0, u_0], s_1 \geqslant s_0\}$$

In contrast, the causal estimand $\mathrm{mVE}(s_1)$ is simpler, only having one argument, but because it does not condition on both $S(1)$ and $S(0)$, in general it does not reflect the relationship between causal biomarker effects and causal clinical effects. Therefore, in general, $\mathrm{mVE}(s_1)$ does not measure principal surrogate value. Under A1 and A2, $\mathrm{mVE}(s_1)$ has interpretation as the per cent reduction in risk for a vaccinated subpopulation with immune response s_1 compared with if it had not been vaccinated. As such this marginal estimand may be used for predicting the level of causal vaccine efficacy for vaccinated groups with different immunogenicity levels, and the comparison of $\mathrm{mVE}(s_1)$ and $\mathrm{mVE}(s_1')$ for $s_1 \neq s_1'$ quantifies the difference in protection that is expected given different immune response levels. Furthermore, for placebo-controlled trials for which $S_i(0)$ has much less variability than $S_i(1)$, $s_1 > s_1'$ for $S(1)$ approximately corresponds to a $s_1 - s_1'$ greater causal vaccine effect on the immune response. In this case $\mathrm{mVE}(s_1)$ approximately measures principal surrogate value, with interpretation similar to $\mathrm{VE}(s_1, s_0)$.

DOI: 10.1002/sim

2.3. Evaluation of a statistical SoP

There are major challenges for assessing level 1 SoPs within either the statistical or principal surrogate frameworks, as we now address in turn.

It is well known that statistical SoPs are difficult to validate [16–19, 30, 31], with a reasonably precise direct validation requiring a very large efficacy trial, usually much larger than typical Phase 3 trials conducted in practice. Furthermore, a larger sample size is needed to quantitate the surrogate value of a partially predictive level 1 SoP than for identifying a nearly 'perfect' level 1 surrogate. An approach to evaluating a level 1 statistical SoP can be broken down into two requirements: First, that the biomarker is a 'strong' CoR in each of the vaccine and placebo arms, and second, after controlling for the biomarker in a regression model, vaccination status does not predict the study endpoint. These criteria can only be evaluated directly if the immune responses vary in the placebo arm, because otherwise it is not possible to evaluate the biomarker as a CoR in the placebo arm, and it is conceptually challenging to check the second 'full mediation' condition.

As pointed out by Frangakis and Rubin [22], a drawback of the statistical surrogate framework is that checking the full mediation condition entails checking equality of the observed risks of $Y=1$ for vaccine *versus* placebo recipients with the same observed biomarker value $S=s$. Because the immune response is measured after randomization, this comparison is susceptible to post-randomization selection bias. Such comparator groups may differ in their host genetics or other factors, so that observed differences cannot be attributed to vaccine assignment. For example, in a controlled experiment someone who has $S=3$ without vaccination may have more exposures and/or a weaker immune system than someone with no prior exposure who achieves $S=3$ following vaccination. The implication is that a perfectly validated statistical SoP could potentially provide inaccurate predictions of vaccine efficacy, and *vice versa*, a biomarker with departures from the statistical SoP criteria could be a good predictor. To address this limitation, Frangakis and Rubin [22] proposed an alternative principal surrogate framework for evaluating surrogates, for which the estimand is a causal effect (e.g. the $\mathrm{VE}(s_1, s_0)$ estimand defined at (1), as addressed further below).

2.3.1. Correcting for selection bias in the statistical SoP definition through baseline covariates. Based on the above discussion, an estimand that could be used for assessing a validated CoR as a statistical surrogate is

$$\mathrm{VE}^S(s) = 1 - \frac{\Pr(Y=1|S=s, Z=1)}{\Pr(Y=1|S=s, Z=0)}$$

where $\mathrm{VE}^S(s)=0$ for all fixed s indicates a statistical surrogate. Incorporating sufficient baseline covariates can make this net effect estimand equal to the causal effect estimand $\mathrm{VE}(s_1, s_0)$. With $\Pr(Y(1)=1|S(1)=s, x)$ the probability of $Y(1)=1$ conditional on $S(1)=s$ and baseline covariates x, the needed assumption (plus A1–A2) is as follows.

Assumption B1

For all fixed s, the risk ratio $\mathrm{risk}_{(1)}(s,s,x)/\mathrm{risk}_{(0)}(s,s,x) \equiv \Pr(Y(1)=1|S(1)=s, S(0)=s, x)/\Pr(Y(0)=1|S(1)=s, S(0)=s, x)$ equals $\Pr(Y(1)=1|S(1)=s, x)/\Pr(Y(0)=1|S(0)=s, x)$.

Under A1 and A2, assumption B1 is equivalent to $\mathrm{VE}^S(s,x) = \mathrm{VE}(s,s,x) \equiv 1 - \mathrm{risk}_{(1)}(s,s,x)/\mathrm{risk}_{(0)}(s,s,x)$, such that after controlling for X the vaccine and placebo groups with observed $S=s$ have the same distribution of risk factors for Y, and differences in risk between $\{S=s, Z=1, X=x\}$

DOI: 10.1002/sim

and $\{S=s, Z=0, X=x\}$ are attributable to assignment to vaccine. Although B1 is untestable, it may be plausible in trials for which there is an excellent understanding of the biology of the pathogen and tested vaccine, and the risk factor covariates X are judiciously selected and collected.

2.4. Evaluation of a principal SoP

Before describing approaches to evaluating a level 1 principal SoP, we compare the interpretations of principal and statistical surrogates with a simple example. Consider a placebo-controlled vaccine trial where Y is infection and S is 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 II presents a perfect principal surrogate, wherein subjects in the 'not take' principal stratum have a 30 per cent chance of

Table II. Example illustrating a principal *versus* statistical surrogate, S binary with $S_i(0)=0$ for all i.

		Unknowable truth	
*Perfect principal surrogate but not a statistical surrogate**			
Principal stratum (PS)	$(S(1), S(0))$	Fraction in PS	$\Pr(Y(1)=1\mid S(1), S(0))$	$\Pr(Y(0)=1\mid S(1), S(0))$
Vaccine not take	$(0,0)$	$\frac{1}{3}$	0.3	0.3
Vaccine take	$(1,0)$	$\frac{2}{3}$	0.0	0.15

Observable data: *infection rates* (*proportion of volunteers*)

	Vaccine status	
S	$Z=1$	$Z=0$
0	0.3 $(\frac{1}{3})$	0.2 (1)
1	0.0 $(\frac{2}{3})$	— (0)

*$\mathrm{VE}=1-[(\frac{1}{3})\times 0.3+(\frac{2}{3})\times 0.0]/[(\frac{1}{3})\times 0.3+(\frac{2}{3})\times 0.15]=0.5$; $\Pr(Y=1|S=0,Z=1)=(1)\times 0.3=0.3$; $\Pr(Y=1|S=0,Z=0)=(\frac{1}{3})\times 0.3+(\frac{2}{3})\times 0.15=0.2$; $\mathrm{VE}(0,0)=1-0.3/0.3=0.0$; $\mathrm{VE}(1,0)=1-0.0/0.15=1.0$.

		Unknowable truth	
No value as a principal surrogate but a statistical surrogate†			
Principal stratum (PS)	$(S(1), S(0))$	Fraction in PS	$\Pr(Y(1)=1\mid S(1), S(0))$	$\Pr(Y(0)=1\mid S(1), S(0))$
Vaccine not take	$(0,0)$	$\frac{1}{3}$	0.2	0.4
Vaccine take	$(1,0)$	$\frac{2}{3}$	0.05	0.1

Observable data: *infection rates* (*proportion of volunteers*)

	Vaccine status	
S	$Z=1$	$Z=0$
0	0.2 $(\frac{1}{3})$	0.2 (1)
1	0.05 $(\frac{2}{3})$	— (0)

†$\mathrm{VE}=1-[(\frac{1}{3})\times 0.2+(\frac{2}{3})\times 0.05]/[(\frac{1}{3})\times 0.4+(\frac{2}{3})\times 0.1]=0.5$; $\Pr(Y=1|S=0,Z=1)=(1)\times 0.2=0.2$; $\Pr(Y=1|S=0,Z=0)=(\frac{1}{3})\times 0.4+(\frac{2}{3})\times 0.1=0.2$; $\mathrm{VE}(0,0)=1-0.2/0.4=0.5$; $\mathrm{VE}(1,0)=1-0.05/0.1=0.5$.

DOI: 10.1002/sim

becoming infected under either assignment vaccine or placebo (0 per cent protection), and subjects in the 'take' stratum have a 0 per cent chance of becoming infected under vaccine assignment and a 15 per cent chance under placebo assignment (100 per cent 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 $S_i=0$, the probabilities of infection $\Pr(Y=1|S=0, Z=z)$ 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 II presents a scenario where the vaccine efficacy is the same irrespective of the vaccine effect on the immune response yet is a statistical surrogate (a 'false positive').

The statistical surrogate definition fails in these cases because of the causal vaccine effect on S, with 67 per cent *versus* 0 per cent responders in the vaccine *versus* 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, placing them at relatively high risk for infection.

While the principal SoP definition advantageously is based on causal effects, the relevant estimand $\mathrm{VE}(s_1, s_0)$ is not identified under the standard assumptions A1 and A2, because we see either $(S_i(1), Y_i(1))$ or $(S_i(0), Y_i(0))$ but not both. The identifiability problem is partially ameliorated in the case that S does not vary in the placebo group, because then $S_i(0)$ is known for all subjects, and only the $S_i(1)$'s for placebo subjects must be predicted to achieve identifiability. Similarly, prediction of the $S_i(1)$'s for placebo recipients will suffice to identify the marginal estimand $\mathrm{mVE}(s_1)$ in the general case that $S_i(0)$ has arbitrary variability. We consider identifiability of $\mathrm{VE}(s_1, 0)$ $(=\mathrm{mVE}(s_1))$.

2.4.1. Identifiability of $\mathrm{VE}(s_1, 0)$ *(Equivalently of* $\mathrm{mVE}(s_1)$*).* To identify $\mathrm{VE}(s_1, 0)$, assumptions beyond A1 and A2 are needed. Gilbert and Hudgens [27] considered the following assumption.

A3: $V_i(1)=1$ if and only if $V_i(0)=1$, which states that any individual who did not experience the event $Y=1$ by t_0 would also not have experienced it by t_0 had they received the opposite randomization assignment. Together A1–A3 identify $\mathrm{risk}_{(1)}(s_1, 0)$ as $\Pr(Y=1|S=s_1, Z=1)$, which can be directly estimated in the CoR evaluation. A1–A3 do not identify the remaining piece of $\mathrm{VE}(s_1, 0)$, $\mathrm{risk}_{(0)}(s_1, 0)$, but do allow simplifying it to

$$\mathrm{risk}_{(0)}(s_1, 0)=\Pr(Y=1|S(1)=s_1, S(0)=0, Z=0) \tag{3}$$

We consider different assumptions, innovative study designs, and data collection techniques that can be used to identify $\mathrm{risk}_{(0)}(s_1, 0)$ and hence $\mathrm{VE}(s_1, 0)$.

2.4.2. Identifying $\mathrm{VE}(s_1, 0)$ *through baseline covariates.* In observational studies, typically all known risk factors are included in the analysis, and causal effects of interest are identified under a no unmeasured confounders assumption, together with a correctly specified model of outcome or treatment conditional on observed covariates (see for example [32]). In an analogous way, incorporating comprehensive baseline prognostic factors for the study endpoint can identify $\mathrm{risk}_{(0)}(s_1, 0, x) \equiv \Pr(Y(0)=1|S(1)=s_1, S(0)=0, x)$. In particular, assumption B2 defined as follows (in addition to A1–A3) implies that $\mathrm{risk}_{(0)}(s_1, 0, x)$ is identified by $\Pr(Y=1|Z=0, x)$.

Assumption B2: $\Pr(Y(0)=1|S(1)=s_1, S(0)=0, x)=\Pr(Y(0)=1|S(0)=0, x)$.

DOI: 10.1002/sim

B2 states that within levels of X, $S(1)$ does not predict $Y(0)$. In other words, conditional on baseline covariates, knowledge of the immune response to vaccine would not help predict the clinical endpoint for placebo recipients.

A1–A3 plus B2 imply that a CoR will automatically have some value as a level 1 principal SoP (i.e. $\mathrm{VE}(s_1, 0)$ increases with s_1). This follows by noting that $\Pr(Y=1|S=s_1, Z=1, x)/\Pr(Y=1|S=s_1', Z=1, x)$, which measures S as a CoR in the vaccine arm, equals $[\mathrm{risk}_{(1)}(s_1, 0, x)/\mathrm{risk}_{(0)}(s_1, 0, x)]/[\mathrm{risk}_{(1)}(s_1', 0, x)/\mathrm{risk}_{(0)}(s_1', 0, x)]$, which measures S as a principal SoP. This relationship can be expressed as

$$\frac{\Pr(Y=1|S=s_1, Z=1, x)}{\Pr(Y=1|S=s_1', Z=1, x)} = \frac{1-\mathrm{VE}(s_1, 0, x)}{1-\mathrm{VE}(s_1', 0, x)}$$

so that the relative risk of infection in the vaccine arm per 1-unit difference in $S(1)$ equals the ratio of one minus causal VE's for a 1-unit difference in $S(1)$. Therefore, demonstrating a CoR will demonstrate a biomarker's value as a principal SoP if sufficient risk factors are collected to justify B2. However, B2 is a (very) strong untestable assumption, and we are not aware of any examples where it is thought to hold. The requirement of such a bold presumption to make the first tier CoR assessment provide a direct inference about surrogacy reinforces the point that a correlate does not a surrogate make [17], and explicit SoP evaluations at levels 1 and 2 are necessary.

2.4.3. Identifying $\mathrm{VE}(s_1, 0)$ *through baseline predictors and/or close-out placebo vaccination.* Given an unwillingness to assume B2, one approach to identifying $\mathrm{VE}(s_1, 0)$ is to collect additional data that can be used to predict the $S(1)$'s of placebo recipients. Follmann [33] introduced two approaches to predicting $S(1)$. The *baseline irrelevant predictor* (BIP) approach incorporates a baseline variable that is measured in both the vaccine and placebo groups that correlates with $S(1)$ and does not predict clinical risk (i.e. is 'irrelevant') after accounting for $S(1)$ and baseline covariates; and the *closeout placebo vaccination* approach vaccinates uninfected placebo recipients at the end of the trial, and measures their immune response $S(1)$ to vaccine. Statistical methods have been developed for making inferences on $\mathrm{VE}(s_1, 0)$ that use either or both of these approaches, or variant approaches that drop the irrelevancy condition, and simulation studies have demonstrated their use [27, 33, 34]. The 'irrelevant' condition is a strong assumption, as it implies there are no unmeasured 'common causes' of S and Y in the sense of [35]. This condition will be more plausible if baseline covariates known to predict S and/or Y are included in the surrogate evaluation.

3. EXAMPLES OF THE EVALUATION OF A CoR AND A LEVEL 1 SoP

3.1. 1998–2003 HIV vaccine efficacy trial

The first placebo-controlled HIV vaccine efficacy trial showed that the tested monomeric recombinant glycoprotein 120 vaccine did not protect against HIV infection [36]. However, levels of certain *in vitro* antibody measurements significantly inversely correlated with the hazard rate of HIV infection in the vaccine arm, i.e. were identified as CoRs of the primary study endpoint. This result was detected using a Cox model and a case-cohort sampling design, wherein antibody responses were measured for all infected vaccine recipients and for a 5 per cent simple random sample of uninfected vaccine recipients [37–39].

DOI: 10.1002/sim

Based on this result, questions were raised about whether vaccine recipients with higher antibody levels were more likely to be protected than vaccine recipients with lower antibody levels, or, in our parlance, whether the antibody levels have value as a level 1 SoP. Based on the data available in the trial alone, it is not possible to empirically evaluate an SoP. This is the case for a statistical SoP because almost all placebo recipients had no antibody response and for a principal SoP because there are insufficient knowledge and covariate data to warrant making assumption B2, and no information is available for predicting the immune response to vaccine $S(1)$ for placebo recipients. This illustrates the common situation, apparently under-recognized in the vaccine field, that a standard efficacy trial design only permits evaluation of the 'mere correlation' CoR level of an immunological correlate of protection.

However, biological knowledge and a follow-up study generated information supporting that the CoR had no value as an SoP. Specifically, 28 HIV pseudo-viruses were created from blood samples of 28 participants who acquired HIV infection during the vaccine trial (14 each from the vaccine and placebo groups), and the pre-infection sera of 85 randomly sampled vaccine recipients were evaluated for their ability to neutralize each of the 28 HIV pseudo-viruses. The vaccinee sera did not generate antibodies that significantly neutralized any of the HIV strains. Since vaccinologists know that induction of antibodies that neutralize the exposing virus are most likely necessary for protection, this follow-up study supports that the neutralizing antibody levels are not able to predict vaccine efficacy and therefore are not a level 1 SoP. This example illustrates the role of biological knowledge for evaluating an SoP; in this case it could not be evaluated from the available trial data alone.

3.2. 1942–1943 influenza vaccine efficacy trial

For our second example, a level 1 SoP can be evaluated because the immune response of interest has substantial variability in the placebo arm and there is a means for predicting missing immune responses $S(1)$ of placebo recipients. The published data are from a 1942–1943 influenza vaccine trial in which 1776 men were arranged alphabetically and inoculated alternately with placebo or a vaccine containing the three flu strains Weiss type A, PR8 type A, and Lee type B [40]. The primary endpoint was hospitalization with strain-specific influenza isolated in throat culture. We first evaluate the antibody titers for Weiss strain A and for PR8 strain A as potential CoRs for hospitalization with strain-specific influenza infection. Figure 1 shows distributions of the $\log_2$ strain-specific antibody titers and Figure 2 shows the rates of strain-specific infection by antibody titer. Figure 2 suggests that both Weiss Strain A antibody titers and PR8 Strain A antibody titers are CoRs for infection in both study groups, with Weiss Strain A titers being a stronger CoR. Results from logistic regression models support these results (Table III).

Next we evaluate the strain-specific antibody titers as potential level 1 statistical SoPs. Based on incidence rates the estimated vaccine efficacy against hospitalization with Weiss strain A is $1-0.225/0.845=0.73$ (95 per cent CI 0.57–0.84) and against hospitalization with PR8 strain A is $1-0.225/0.822=0.73$ (95 per cent CI 0.55–0.83). To evaluate potential statistical SoPs, logistic regression models were fit with independent variables vaccination status and strain-specific antibody titer, and the observed and predicted strain-specific case incidences by antibody titer were plotted (Figure 2). For Weiss Strain A, vaccination status has a coefficient estimate near zero in the model after controlling for antibody titer (Table III), suggesting that the antibody titer mediates much of the vaccine effect on incidence. Figure 2 shows nearly identical incidence curves as a function of antibody titer in the vaccine and placebo groups, supporting that Weiss Strain A

DOI: 10.1002/sim

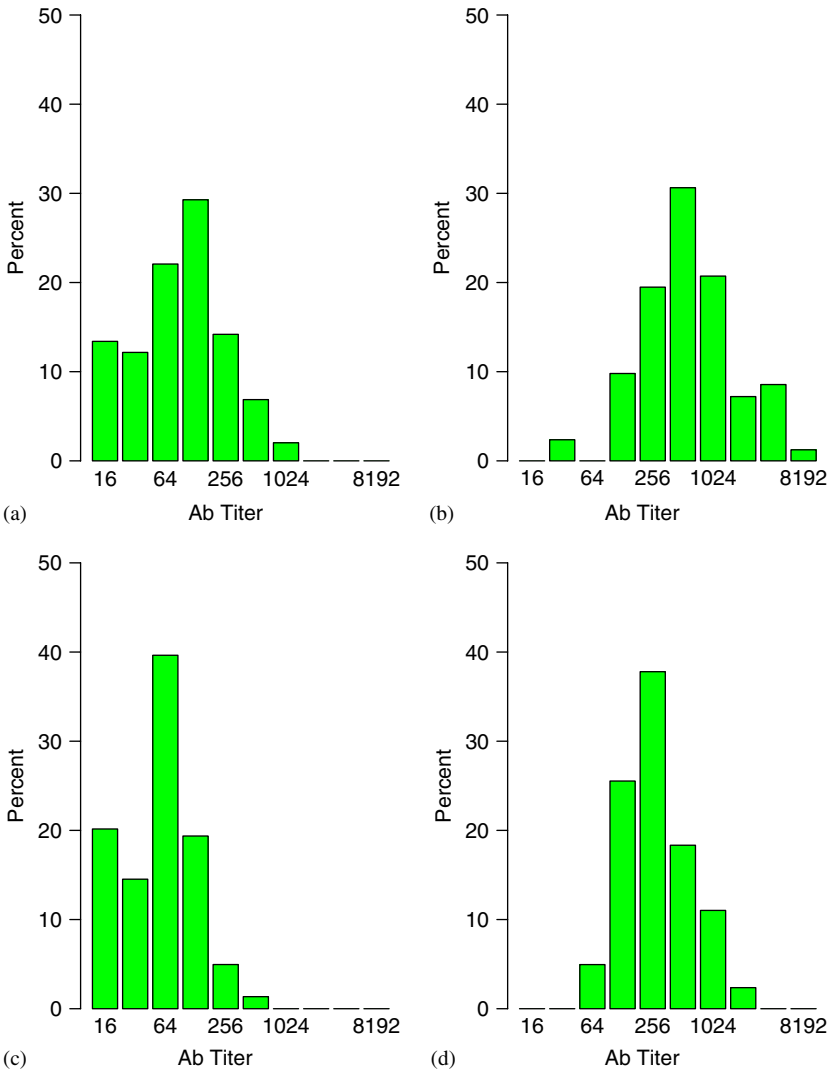

Figure 1. Frequency distributions of anti-Weiss Strain A ((a) placebo and (b) vaccine) and anti-PR8 Strain A ((c) placebo and (d) vaccine) antibody levels for the placebo and vaccine study groups of the influenza vaccine field trial.

antibody levels (nearly) completely explain the observed protection. Based on the model relating case incidence to Weiss strain A titers in the placebo group and the observed titer distribution in vaccinees, the predicted vaccine efficacy based on Weiss strain A titers is 0.82, close to the vaccine efficacy estimate computed ignoring the biomarker, 0.73. It is notable that this is one of the first examples of a biomarker that has been empirically validated to satisfy the Prentice criteria as a 'perfect' surrogate endpoint.

DOI: 10.1002/sim

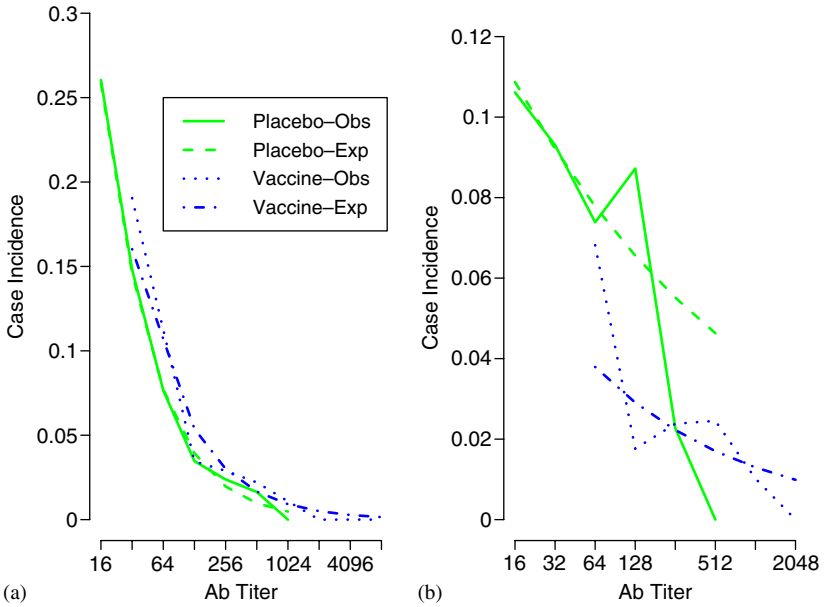

Figure 2. Nonparametric and parametric (logistic regression model based) estimates of the incidence of hospitalization with Weiss Strain A (a) or PR8 Strain A (b) influenza infection for the placebo and vaccine study groups.

Table III. Logistic regression models of strain-specific $\log_2$-neutralizing antibody titers as a predictor of hospitalization with strain-specific influenza infection.

	Weiss Strain A		PR8 Strain A	
	Coef. est. (s.e.)	p-Value	Coef. est. (s.e.)	p-Value
Intercept	1.62 (0.45)	0.0003	−1.27 (0.53)	0.017
Log_2 antibody titer	−0.98 (0.12)	<0.0001	−0.29 (0.13)	0.031
Vaccination status	−0.33 (0.32)	0.31	−0.89 (0.34)	0.0085

In contrast, the analysis suggests that PR8 strain A titers only partially mediate the vaccine efficacy, as evidenced by the facts that vaccination status is significant in the logistic regression model that includes PR8 Strain A-specific antibody titer, and the case incidence curves are visibly different for the vaccine and placebo groups (Figure 2(b)). Furthermore, the predicted vaccine efficacy based on PR8 strain A titers is 0.33 compared with the estimate 0.73 computed ignoring the titers. Apparently, the vaccine protects against PR8 strain A through mechanisms that are not fully captured in the PR8 strain A neutralization assay.

To evaluate the strain-specific antibody titers as potential level 1 principal SoPs, we consider the data suggesting that pre-vaccination anti-flu antibody titers in adults are inversely correlated with post-vaccination titers [41–44]. For example, Gorse *et al*. [45] measured pre-vaccination and post-vaccination serum hemagglutination inhibition (HAI) antibody titers to influenza A virus from 400 adults and found a strong inverse correlation of the pre- and post-measurements. Unfortunately

DOI: 10.1002/sim

these data are not available to us; if they were, a fitted regression model would be used to impute the missing titers $S(1)$ of placebo recipients.

Instead, we use the observed titers $S_i(0)$ of placebo recipients to impute the $S_i(1)$ values under an 'anti-equipercentile' or 'inverse rank preserving' assumption, wherein placebo subjects with lowest rank of $S_i(0)$ are assumed to have the highest rank of $S_i(1)$, placebo subjects with the second lowest rank of $S_i(0)$ are assumed to have the second highest rank of $S_i(1)$, and so on. The following table shows the distinct observed $S(1)$ and $S(0)$ values for the vaccine and placebo groups as well as the imputed $S(1)$ for each distinct observed $S(0)$.

Weiss Strain A										
Observed $S_i(1)$	—	32	—	128	256	512	1024	2048	4096	8192
Observed $S_i(0)$	16	32	64	128	256	512	1024	—	—	—
Imputed $S_i(1)$	8192	4096	2048	1024	512	256	128/32*	—	—	—
PR8 Strain A										
Observed $S_i(1)$	—	—	64	128	256	512	1024	2048	—	—
Observed $S_i(0)$	16	32	64	128	256	512	—	—	—	—
Imputed $S_i(1)$	2048	1024	512	256	128	64	—	—	—	—

*Placebo recipients with $S_i(0)=1024$ were randomly assigned $S_i(1)=128$ or 32 with chance one-half.

Based on the imputed data sets, for each influenza strain we estimated the marginal vaccine efficacy curve at each distinct observed $S_i(1)$ value s_1 by

$$\widehat{\mathrm{mVE}}(s_1)=1-\frac{\widehat{\Pr}(Y=1|S=s_1,Z=1)}{\widehat{\Pr}(Y=1|\text{imputed } S(1)=s_1,Z=0)}$$

where the probabilities in the numerator and denominator are estimated either nonparametrically (by empirical fractions) or by logistic regression. Figure 3 (top panel) displays the estimated curves $\mathrm{mVE}(s_1)$, showing that for Weiss strain A titers the curve increases from 0 to 1 as titers rise, supporting the high value of the titers as a level 1 principal SoP. In contrast, the estimated $\mathrm{mVE}(s_1)$ curve for PR8 strain A increases as s_1 increases but less steeply, suggesting that PR8 strain titers have partial value as a level 1 principal SoP. Because these results depend on the imputation model, a sensitivity analysis was performed in which an extreme opposite imputation model was used. Specifically, an equipercentile (i.e. rank preserving) assumption was made, which supposes that the ranks of the $S_i(0)$ values in placebo recipients are the same as the ranks of the $S_i(1)$ values. Under the equipercentile assumption the estimated $\mathrm{mVE}(s_1)$ curve still increases considerably with s_1 for Weiss Strain A (Figure 3, bottom panel), supporting some robustness of the surrogate endpoint result.

We stress that the main value of this analysis is to illustrate the evaluation of a level 1 principal SoP, and the substantive result should be interpreted with caution. Both the anti-equipercentile and equipercentile assumptions are strong and unverifiable. Had the data on pre-vaccination titers been available, it would have been possible to estimate the bivariate distribution of $(S(1), S(0))$ in the vaccine arm, which would provide a more credible technique for evaluating $\mathrm{mVE}(s_1)$. In future vaccine trials in populations with prior exposure to the pathogen under examination, it may be fruitful to incorporate such pre-vaccination titers into the level 1 SoP evaluation.

DOI: 10.1002/sim

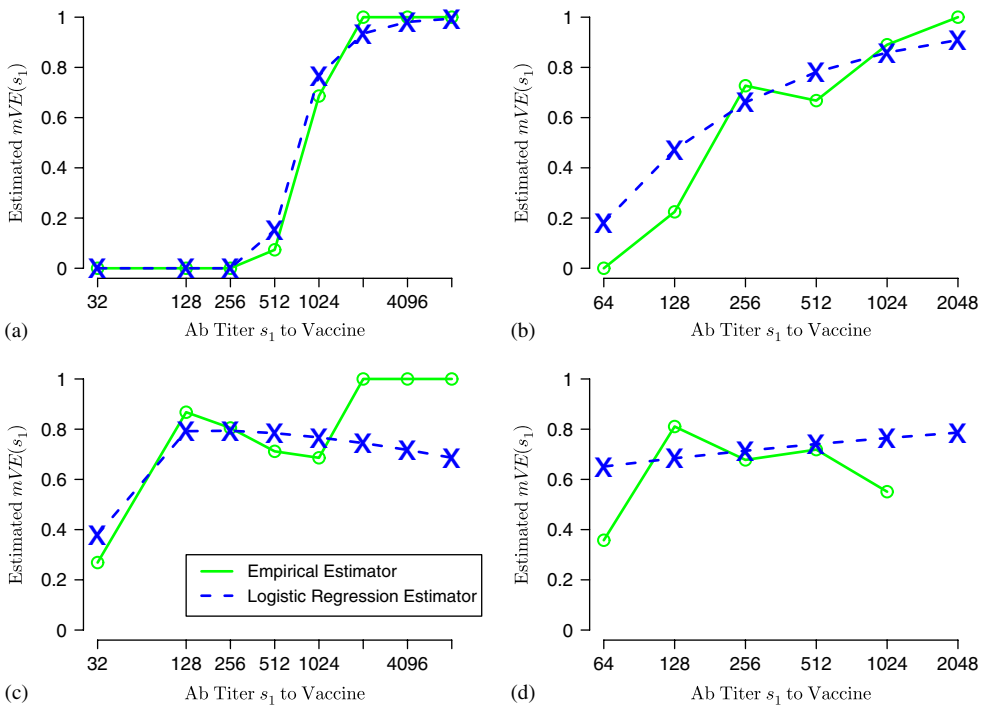

Figure 3. Nonparametric and parametric (logistic regression model based) estimates of mVE(s_1) for (a) Weiss Strain A and (b) PR8 Strain A under the anti-equipercentile assumption. As a sensitivity analysis, panels (c) and (d) show these estimates under the equipercentile assumption.

3.3. Design of the PAVE-100 HIV vaccine efficacy trial

The PAVE-100 HIV vaccine efficacy trial, sponsored by the U.S. National Institutes of Health, the U.S. Military HIV Research Program, the International AIDS Vaccine Initiative, and the Centers for Disease Control and Prevention, is currently being planned. In the current design 8500 HIV negative volunteers from the Americas, East Africa, and Southern Africa will be randomized to a prime-boost vaccine regimen (DNA prime:Adenovirus 5 vector boost) or placebo in a 1:1 allocation. The design is event driven with a planned total of 280 HIV infections. A secondary objective of the trial is to evaluate the magnitude of CD8+ T cell response levels, as measured by the ELISpot assay from blood samples drawn at the week 26 visit after randomization, as a CoR and as a level 1 principal SoP for HIV infection. In this section, we briefly consider how well the BIP, close-out placebo vaccination (CPV) and combined (BIP+CPV) augmented trial designs would be able to evaluate an SoP. We base this consideration on the discrete failure time method developed by Qin *et al*. [34] for evaluating a level 1 principal SoP. Simulations reported there verified that under A1–A3 and the BIP and/or CPV assumptions the method provided unbiased estimates of regression parameters measuring surrogate value, and Wald-confidence intervals about these parameters had correct coverage levels.

We conducted a small simulation study to match the PAVE-100 trial design. The total sample size is 4250 subjects in each of the vaccine and placebo arms, and we assume 50 per cent vaccine

DOI: 10.1002/sim

efficacy, such that at the time of analysis there are an expected 187 placebo recipients and 94 vaccine recipients HIV infected. We suppose the ELISpot T cell response is measured in all infected vaccine recipients and a 25 per cent simple random sample of uninfected vaccine recipients. For the augmented designs with CPV, we suppose that 25 per cent of uninfected placebo recipients receive the AIDS vaccine at study close-out. For the BIP and BIP+CPV designs, we suppose that the titer of neutralizing antibodies to the Adenovirus 5 serotype vector that carries the HIV genes is measured from all trial participants, as is planned for the trial. This measurement is chosen as the BIP because it has been shown to inversely correlate with ELISpot response levels [46] and plausibly does not independently predict the rate of HIV infection.

The BIP (Adenovirus 5 titers) and ELISpot response $S(1)$ were generated from a bivariate normal distribution with mean zero and variance 0.4 for each component (reflecting the variance

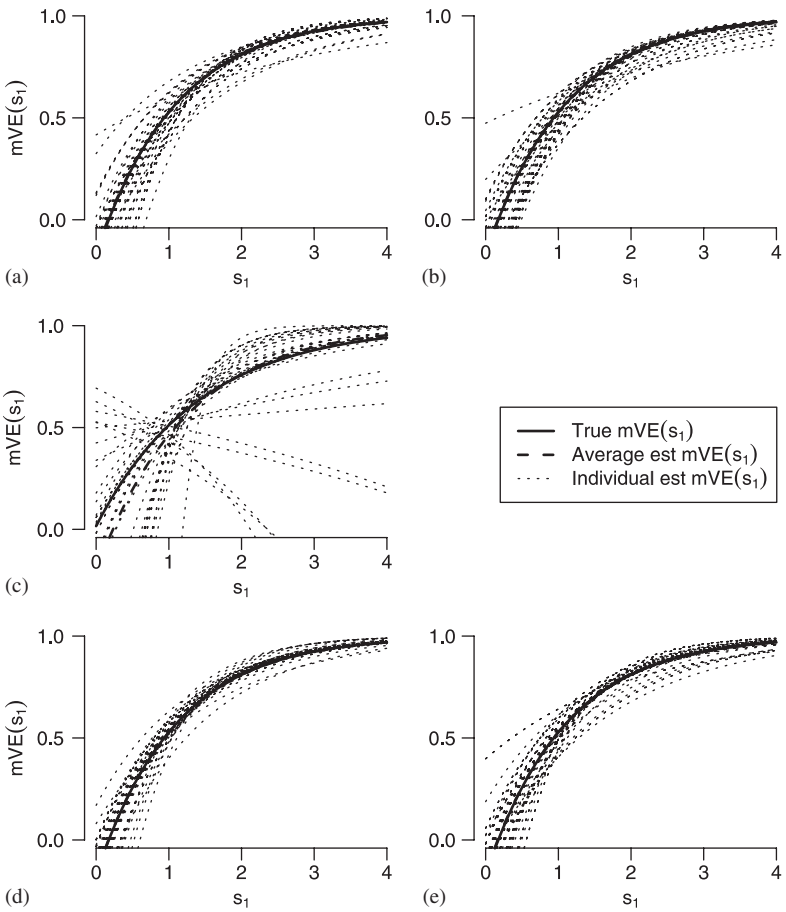

Figure 4. For data simulated to reflect the design of the PAVE-100 HIV vaccine efficacy trial, estimates of $\mathrm{VE}(s_1, 0) = \mathrm{mVE}(s_1)$ under (a, b) the baseline irrelevant predictor design; (c) the close-out placebo vaccination design; and (d, e) the combined design. The solid line is the true $\mathrm{mVE}(s_1)$ curve, the dashed line is the empirical average estimate of $\mathrm{mVE}(s_1)$ over the 500 simulated data sets, and the dotted lines are 50 estimates from 50 simulated data sets. $\rho = 0.5$ for (a), (d) and $\rho = 0.9$ for (b), (e).

DOI: 10.1002/sim

of the ELISpot assay) and correlation $\rho=0.5$ or $\rho=0.9$. Continuous failure times were generated from the Cox model $\lambda(t|Z,S(1))=\lambda_0(t)\exp\{\beta_1 Z+\beta_2 S(1)+\beta_3 ZS(1)\}$ and were binned into six equal-length time intervals to reflect a semi-annual schedule of testing for HIV infection. The true parameters were set at $\beta_2=-1.109$ and $\beta_3=-0.91$, reflecting a strongly predictive level 1 principal SoP with a 10-fold lower causal relative risk $\mathrm{RR}(S(1))=1-\mathrm{VE}(S(1))$ per 4 standard deviation higher immune response $S(1)$ (spanning the range of common immune responses). In addition, β_1 and the constant baseline hazard $\lambda_0(t)=\lambda_0$ were calibrated to give overall $\mathrm{VE}=0.5$ and 187 expected infections in the placebo arm.

Based on 500 simulated vaccine trials, the BIP, CPV, and BIP+CPV augmented designs had estimated power 0.988, 0.198, and 0.992 for rejecting the null hypothesis $\beta_3=0$ of no surrogate value in the case of $\rho=0.5$, and power 0.996, 0.198, and 0.998 in the case of $\rho=0.9$. This demonstrates that the designs including a BIP that is at least 50 per cent correlated with $S(1)$ will provide high power to detect an excellent level 1 SoP, whereas CPV alone confers low power. Each panel of Figure 4 shows the true $\mathrm{VE}(s_1,0)$ curve, the average estimated curve over the simulated data sets, and estimated curves for 50 randomly sampled individual data sets. Follmann [33] and Gilbert and Hudgens [27] provide simulation results for a similar method that treats HIV infection as a binary endpoint, which more fully describe how fast power increases with ρ. It is of interest to perform additional simulations, wherein A3, the BIP condition, and/or the CPV assumption do not hold, to assess the resulting bias in parameter estimation.

4. LEVEL 2 SoP

In practice, usually the main utility of a surrogate endpoint is to predict clinical efficacy of a treatment for a new setting not studied in an efficacy trial. We refer to an immunological SoP that provides reliable predictions of vaccine efficacy for a new setting as a level 2 SoP. To illustrate the nature of a level 2 SoP, suppose an HIV vaccine efficacy trial is conducted in South Africa in men and women exposed to subtype C HIVs through heterosexual sex. HVTN 503 is evaluating Merck's Adenovirus 5 vector vaccine in men and women exposed to subtype C HIVs through heterosexual sex. Suppose hypothetically the trial identifies beneficial $\mathrm{VE}>0$ and an excellent level 1 SoP. A question of interest would then be whether measurements of this immune response in intravenous drug users that are exposed to subtype B HIVs through needle sharing in the urban United States reliably predict vaccine efficacy. In this complex example, the predictive bridge of interest spans different viral genetics, host genetics, cultural characteristics, and routes of exposure, and as such may be quite difficult to validate. A much 'shorter bridge' to evaluate would consider whether the level 1 SoP, applied to the same population as studied in the efficacy trial, reliably predicts vaccine efficacy of a product identical to the tested product except it is produced by a more efficient manufacturing process.

One approach to evaluating a level 2 SoP is meta-analysis of multiple efficacy and/or proof-of-concept trials [31, 47, 48], possibly including post-licensure studies. The evaluation of a level 2 SoP is specific to the type of predictive bridge, so that the meta-analytic unit as well as the target of the prediction must be appropriately chosen. For example, to predict vaccine efficacy against the predominant influenza strain in next year's flu season, the appropriate meta-analytic unit would be the predominant circulating influenza strain across a set of years, and N strain-specific assessments of immune responses and vaccine efficacies across N annual flu seasons would be required. The observed relationship between the N estimated vaccine efficacies and summary contrasts of

DOI: 10.1002/sim

immune responses (vaccine *versus* placebo groups) could be used to predict vaccine efficacy for the new setting based on a sample of immune responses in that setting (e.g. HAI antibody titers to the influenza strain that is predicted to be predominant in the next flu season), and provide a way to estimate the error in the prediction.

The meta-analysis approach is data intensive and may not always be feasible. An indirect strategy for assessing a level 2 SoP requires a conceptual leap from an identified level 1 SoP to a level 2 SoP. Without a meta-analytic assessment, this leap can only be made based on indirect inferences and through incorporation of biological knowledge of mechanisms of vaccine protection. Moreover, predictions based on meta-analysis rely on the assumption that the vaccine effects on the immune response and the study endpoint for the new setting are sampled from the bivariate distribution of the vaccine effects for the N meta-analytic units, which is not fully verifiable. Therefore, even if large meta-analyses are conducted, the incorporation of biological information is critically important for quantifying the value of an immunological measurement as a level 2 SoP.

5. DISCUSSION

In this paper, a statistical companion to Qin *et al*. [8], we have described a framework for evaluating the utility of a biomarker measurement for predicting clinical treatment efficacy at three different levels, ordered by scientific importance and by the extent of data requirements for making the assessment. While this three-tier framework may be useful for randomized placebo-controlled Phase IIb/III trials in many disease areas, for illustration we have focused on the assessment of an immunological measurement as a surrogate endpoint for vaccine efficacy to prevent clinically significant infection. At the first tier, the assessment of a CoR is relatively straightforward and may be achievable with standard efficacy trial designs, although even at this level the assessment is challenged by potential measurement error of the putative CoR, time variations of the CoR, and the task of developing an efficient sampling design that optimally incorporates participant covariate information.

An inference in a trial that VE>0 plus validation of an immunological CoR does not imply that the immune response has any value as a surrogate endpoint. An immune response with no capacity for predicting causal vaccine efficacy may be a CoR because it mirrors innate immunity or some other factor such as risk behavior. In addition, for trials that conclude VE is zero, an immunological CoR cannot be a surrogate endpoint unless certain vaccine effects on the immune response (e.g. low-level effects) predict enhanced disease risk if vaccination is received. If VE$=0$ a level 1 SoP must satisfy $\mathrm{VE}(s_1, s_0)<0$ for some values of (s_1, s_0) and $\mathrm{VE}(s_1, s_0)>0$ for other (s_1, s_0).

At the second tier of an immune correlate, the evaluation of a level 1 SoP can be approached using methods for evaluating a surrogate endpoint based on a single large clinical trial. We have considered methods based on the statistical and principal surrogate frameworks and provided an example (the analysis of the 1942–1943 influenza vaccine trial) demonstrating that both frameworks can identify an excellent level 1 SoP. However, in general it is quite difficult to evaluate a level 1 SoP *via* either framework, with pros and cons for each. Incorporating baseline covariates, such as on host genetics or on innate immunity, can potentially overcome the challenges. In fact, for both approaches we have noted that under strong assumptions about not missing any risk factors, the assessment of a CoR is equivalent to the assessment of a level 1 SoP. The strong assumptions are unverifiable, however, so that sensitivity analyses would be needed, that ideally account for

DOI: 10.1002/sim

the available biological knowledge of mechanisms of protection. In addition, we have summarized novel study designs and data collection that can be used to evaluate a level 1 SoP under lighter assumptions. These assumptions are not fully verifiable, however, suggesting the value of sensitivity analysis and the need for further research.

Lastly, evaluation of a level 2 SoP can be approached *via* meta-analyses of multiple efficacy or proof-of-concept trials and possibly post-licensure trials. These methods are limited by the difficulty in defining the class of studies that form an appropriate basis for predicting vaccine efficacy for a new setting and by imprecision in this prediction [48], again underscoring the importance of drawing upon biological knowledge for helping to justify building a predictive bridge from a level 1 SoP to a level 2 SoP. A follow-up efficacy trial for the new setting may be required to credibly support that the immunological measurement can reliably predict protection in that setting.

ACKNOWLEDGEMENTS

The authors are grateful to Misrak Gezmu for organizing the workshop and for the referees for helpful comments that led to improvements. This work is supported by NIH grant 2 R01 AI54165-04.

REFERENCES

1. Fauci AS, Haynes BF, Pantaleo G. Toward an understanding of the correlates of protective immunity to HIV infection. *Science* 1996; **271**:324–328.
2. Clements-Mann ML. Lessons for AIDS vaccine development from non-AIDS vaccines. *AIDS Research and Human Retroviruses* 1998; **14**(Suppl. 3):S197–S203.
3. Burton DR, Desrosiers RC, Doms RW, Koff WC, Kwong PD, Moore JP, Nabel GJ, Sodroski J, Wilson IA, Wyatt RT. HIV vaccine design and the neutralizing antibody problem. *Nature Immunology* 2004; **5**:233–236.
4. Siber GR. Methods for estimating serological correlates of protection. *Developments in Biological Standardization* 1997; **89**:283–296.
5. Chan ISF, Shu L, Matthews H, Chan C, Vessey R, Sadoff J, Heyse J. Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine* 2002; **21**:3411–3430.
6. Dunning AJ. A model for immunological correlates of protection. *Statistics in Medicine* 2006; **25**:1485–1497.
7. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. Springer: New York, 2005.
8. Qin L, Gilbert PB, Corey L, McElrath J, Self SG. A framework for assessing an immunological correlate of protection in vaccine trials. *Journal of Infectious Diseases* 2007; **196**:1304–1312.
9. Szmuness W, Stevens CE, Zang EA, Harley EJ, Kellner A. A controlled clinical trial of the efficacy of the hepatitis B vaccine (Heptavax B): a final report. *Hepatology* 1981; **1**:377–385.
10. Lanata CF, Black RE, del Aguila R *et al*. Protection of Peruvian children against rotavirus diarrhea of specific serotypes of one, two, or three doses of the RIT 4237 attenuated bovine rotavirus vaccine. *Journal of Infectious Diseases* 1989; **159**:452–459.
11. Clemens JD, Sack DA, Harris JR *et al*. Field trial of oral cholera vaccines in Bangladesh: results from three-year follow-up. *The Lancet* 1990; **335**:270–273.
12. Villa LL, Costa RL, Petta CA *et al*. Prophylactic quadrivalent human papillomavirus (types 6, 11, 16, and 18) L1 virus-like particle vaccine in young women: a randomised double-blind placebo-controlled multicentre phase II efficacy trial. *The Lancet Oncology* 2005; **6**:271–278.
13. Mehrotra DV, Li X, Gilbert PB. A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial. *Biometrics* 2006; **62**:893–900.
14. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**:1014–1029.
15. Huang Y, Pepe M, Feng Z. Evaluating the predictiveness of a continuous marker. 2007, DOI: 10.1111/j.1541-0420.2007.00878.x.
16. Fleming TR. Surrogate markers in AIDS and cancer trials. *Statistics in Medicine* 1994; **13**:1423–1435.
17. DeMets DL, Fleming TR. Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* 1996; **125**:605–613.

DOI: 10.1002/sim

18. DeGruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials. *Controlled Clinical Trials* 2002; **22**:485–502.
19. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8**:431–440.
20. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–961.
21. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association* 2005; **100**:322–331.
22. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**:21–29.
23. Robins JM. An analytic method for randomized trials with informative censoring: Part I. *Lifetime Data Analysis* 1995; **1**:241–254.
24. Rubin DB. Statistics and causal inference: which ifs have causal answers. *Journal of the American Statistical Association* 1986; **81**:961–962.
25. Hudgens MG, Halloran ME. Causal vaccine effects on binary post-infection outcomes. *Journal of the American Statistical Association* 2006; **101**:51–64.
26. Gilbert PB, Bosch RJ, Hudgens MG. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* 2003; **59**:531–541.
27. Gilbert PB, Hudgens MG. Evaluating causal effect predictiveness of candidate surrogate endpoints. 2006, submitted.
28. Kim HW, Canchola JG, Brandt CD, Pyles G, Chanock RM, Jensen K, Parrott RH. Respiratory synctial virus disease in infants despite prior administration of antigenic inactivated vaccine. *American Journal of Epidemiology* 1969; **89**:405–421.
29. Kliks SC, Nisalak A, Brandt WE, Wahl L, Burke DS. Antibody-dependent enhancement of dengue virus growth in human monocytes as a risk factor for dengue hemorrhagic fever. *American Journal of Tropical Medicine and Hygiene* 1989; **40**:444–451.
30. Hughes MD. Evaluating surrogate endpoints. *Controlled Clinical Trials* 2002; **23**:703–707.
31. Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* 2002; **23**:607–625.
32. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
33. Follmann D. Augmented designs to assess immune response in vaccine trials. *Biometrics* 2006; **62**:1161–1169.
34. Qin L, Gilbert PB, Follmann D, Li D. Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. *Annals of Applied Statistics* 2006, submitted.
35. Pearl J. *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge University Press: Cambridge, 2000.
36. Flynn NM, Forthal DN, Harro CD, Mayer KH. The rgp120 HIV Vaccine Study Group. Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *Journal of Infectious Diseases* 2005; **191**:654–665.
37. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1–11.
38. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58.
39. Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, Sinangil F, Burke D, Berman PW. 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* 2005; **191**:666–677.
40. Salk JE, Menke Jr WJ, Francis Jr T. A clinical, epidemiological and immunological evaluation of vaccination against epidemic influenza. *American Journal of Hygiene* 1943; **42**:57–93.
41. Ennis FA, Yi-Hua Q, Schild GC. Antibody and cytotoxic T lymphocyte responses of humans to live and inactivated influenza vaccines. *Journal of General Virology* 1982; **58**:273–281.
42. Clements ML, Tierney EL, Murphy BR. Response of seronegative and seropositive adult volunteers to live attenuated cold-adapted reassortant influenza A virus vaccine. *Journal of Clinical Microbiology* 1985; **21**:997–999.
43. Gorse GJ, Belshe RB. Enhancement of anti-influenza A virus cytotoxicity following influenza A virus vaccination in older, chronically ill adults. *Journal of Clinical Microbiology* 1990; **28**:2539–2550.
44. Treanor JJ, Roth FK, Betts RF. Use of live cold-adapted influenza A H1N1 and H3N2 virus vaccines in seropositive adults. *Journal of Clinical Microbiology* 1990; **28**:596–599.

DOI: 10.1002/sim

45. Gorse GJ, O'Connor TZ, Newman FK, Mandava MD, Mendelman PM, Wittes J, Peduzzi PN. Immunity to influenza in older adults with chronic obstructive pulmonary disease. *Journal of Infectious Diseases* 2004; **190**:11–19.
46. Catanzaro AT, Koup RA, Roederer M *et al*. Safety and immunogenicity evaluation of a multiclade HIV-1 candidate vaccine delivered by a replication-defective recombinant adenovirus vector. *Journal of Infectious Diseases* 2006; **194**:1638–1649.
47. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**:1965–1982.
48. Gail MH, Pfeiffer R, Van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**:231–246.

DOI: 10.1002/sim