Biometrics 61, 1102–1111
December 2005
DOI: 10.1111/j.1541-0420.2005.00380.x

Counterfactual Links to the Proportion of Treatment Effect Explained by a Surrogate Marker

Jeremy M. G. Taylor,[1,*] Yue Wang,[2] and Rodolphe Thiébaut[3]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Research Laboratory, Merck & Co., West Point, Pennsylvania 19486, U.S.A.
[3]INSERM E0338 Biostatistics, ISPED, Bordeaux 2 University, Bordeaux 33076, France
[*]*email:* jmgt@umich.edu

Summary. In a randomized clinical trial, a statistic that measures the proportion of treatment effect on the primary clinical outcome that is explained by the treatment effect on a surrogate outcome is a useful concept. We investigate whether a statistic proposed to estimate this proportion can be given a causal interpretation as defined by models of counterfactual variables. For the situation of binary surrogate and outcome variables, two counterfactual models are considered, both of which include the concept of the proportion of the treatment effect, which acts through the surrogate. In general, the statistic does not equal either of the two proportions from the counterfactual models, and can be substantially different. Conditions are given for which the statistic does equal the counterfactual model proportions. A randomized clinical trial with potential surrogate endpoints is undertaken in a scientific context; this context will naturally place constraints on the parameters of the counterfactual model. We conducted a simulation experiment to investigate what impact these constraints had on the relationship between the proportion explained (PE) statistic and the counterfactual model proportions. We found that observable constraints had very little impact on the agreement between the statistic and the counterfactual model proportions, whereas unobservable constraints could lead to more agreement.

Key words: Causal effects; Clinical trial; Counterfactual model; Direct effect; Indirect effect; Surrogate marker.

1. Introduction

Clinical trials with rare primary endpoints or long duration times often require large sample sizes and extensive periods of follow-up. Because of this, there has been increasing interest in using surrogate endpoints in lieu of the primary endpoints in these situations. A number of statistical articles concerned with evaluating surrogate markers have been written (Prentice, 1989; Freedman, Graubard, and Schatzkin, 1992; Daniels and Hughes, 1997; Buyse et al., 2000; Li, Meredith, and Hoseyni, 2001; Wang and Taylor, 2002; Ditlevsen et al., 2005). Surrogate endpoints are usually intermediate biomarkers in disease development, which can be assessed earlier and more easily. They are generally proposed based on the biological process of a disease and their strong associations with the primary endpoint.

Prentice (1989) proposed a formal definition of surrogate endpoints and gave general operational criteria for validation of surrogate endpoints. Prentice's criteria leads to consideration of a model for the treatment effect on the primary endpoint adjusting for the surrogate marker and statistical tests for $\tau_{\text{adj.}} = 0$, where $\tau_{\text{adj.}}$ is the adjusted treatment effect in the model. Prentice's criteria, which requires a surrogate endpoint to fully capture the treatment effect on the primary endpoint, is rather too stringent. In practice, it is more likely that a surrogate endpoint may explain part but not all the treatment effect. Thus, a quantitative measure of the proportion of the treatment effect that is explained by the surrogate marker was proposed by Freedman et al. (1992). This measure was given by $P = (\tau_{\text{unadj.}} - \tau_{\text{adj.}})/(\tau_{\text{unadj.}})$, where $\tau_{\text{unadj.}}$ is the treatment effect on the primary outcome without adjusting for the marker. The properties of this statistic are reviewed in Wang and Taylor (2002) and two alternative statistics F and F' for assessing the proportion of the treatment effect explained were proposed.

We use the following notation: T and S denote the primary endpoint and surrogate marker, respectively. They are assumed to be binary. Z is the treatment variable, with $Z = 1$ for treatment (or new treatment) and $Z = 0$ for placebo (or standard treatment). We assume a positive effect of the treatment, with $T = 1$ and $S = 1$ representing better outcomes. In a randomized clinical trial, a perfect surrogate occurs when S captures all the dependence of T on Z, that is, $P(T \mid Z, S) = P(T \mid S)$. A useless surrogate can occur when, conditional on the treatment, the surrogate is independent of the primary endpoint, that is, $P(T \mid Z, S) = P(T \mid Z)$, or when S is independent of the treatment group, that is, $P(S \mid Z) = P(S)$.

An alternative approach to the consideration of surrogate markers is through models of counterfactual variables. Such models are frequently used in the statistical literature on causal inference. The general idea of a counterfactual model

is to postulate all the values for both the surrogate S and the primary outcome T for a subject under the different possible interventions Z. Thus, for example, one postulates what the two values of S would be if either $Z = 0$ or $Z = 1$, even though only one of these could be observed for each subject. These two values are denoted by the pair (S_0, S_1). In the counterfactual framework the population is partitioned into subgroups, such that within each subgroup everyone would have the same both observed and counterfactual outcomes. The proportions of the population in each of these subgroups are parameters in the model. Two counterfactual frameworks have been suggested that are relevant to surrogate markers. The work of Robins and Greenland (1992) considered the problem of separating the direct effects of an exposure or treatment from the indirect effects relayed through an intermediate or surrogate variable. The ratio of the indirect effect to the total effect can be interpreted as the proportion of the treatment effect on the primary outcome explained by the surrogate. Using ideas from counterfactual models, Frangakis and Rubin (2002) proposed the concept of principal surrogacy and that the effects of treatment on the primary outcome can be considered as either associated with or disassociated from the effect of the treatment on the surrogate. In this approach, the ratio of the associative effect to the total treatment effect could be viewed as a measure of the proportion of treatment effect explained by the surrogate. The aim of this article is to investigate the link between the two previously proposed statistics F and F' and the proportions from these two counterfactual models. Because results and conclusions are similar for F and F', we present here only the results for F. We have changed the notation for this article: what was previously (Wang and Taylor, 2002) denoted by F is denoted by PE (proportion explained) in the current article.

In practice, data are collected in a scientific context from which there might be a considerable amount of a priori knowledge. For example, it will almost certainly be known that the primary and surrogate outcomes are correlated. Furthermore, we would not be contemplating evaluating whether S is a good surrogate unless there was a treatment effect on both S and T. These types of restrictions will imply constraints on the parameters in the counterfactual model. An aim of this article is to investigate the relationship between the PE statistic and the analogous measures from the counterfactual model, and how this relationship is impacted by subject-matter constraints. Because PE is easy to estimate, an important question is whether it can be given a causal interpretation as defined from the counterfactual models or whether it approximates the analogous proportion from the counterfactual models under certain conditions.

Throughout this article we will use a simple HIV clinical trial evaluating the efficacy of a new antiretroviral treatment to illustrate the models and statistics. The trial has two arms, a standard treatment ($Z = 0$) and a new treatment ($Z = 1$). The patients have CD4 counts measured at baseline and at 16 weeks in the trial. The change in CD4 count from baseline is a possible surrogate endpoint, for use in future trials, to replace the primary clinical endpoint. For this illustration, the change in CD4 count is considered as binary, with a decrease in CD4 considered as bad ($S = 0$) and an increase as good ($S = 1$). The primary clinical endpoint is whether the patient is alive ($T = 1$) or dead ($T = 0$) at 2 years. Assume the results of the trial were in the expected direction, that is, the new treatment gave a higher percentage with increasing CD4 and higher percentage alive at 2 years. Although we present this hypothetical trial for the purpose of illustrating the ideas, it is quite similar to some large randomized trials for which the role of early CD4 counts as a surrogate has been investigated (Delta, 1996; Aboulker et al., 1999; Hughes, 2000), and the present evaluation of antiretroviral therapy in HIV clinical trials is most often based on CD4 count and plasma HIV RNA (Lazzarin et al., 2003).

In Section 2, we describe the statistic PE, as proposed in Wang and Taylor (2002). In Section 3, we describe the counterfactual model and investigate algebraic relationships between PE and the counterfactual proportions. In Section 4, we discuss possible contextual restrictions. In Section 5, we present the results of a simulation experiment to compare PE with the counterfactual analogues.

2. A Measure for the Proportion of Effect Explained

Motivated by Tsiatis, De Gruttola, and Wulfsohn (1995), Wang and Taylor (2002) proposed a measure PE for the proportion of treatment effect explained, defined by

$$\mathrm{PE} = (M_{01} - M_0)/(M_1 - M_0)$$

where

$$\begin{aligned} M_0 &= P(T = 1 \mid Z = 0) \\ &= \textstyle\sum_s P(T = 1 \mid S = s, Z = 0) P(S = s \mid Z = 0), \\ M_1 &= P(T = 1 \mid Z = 1) \\ &= \textstyle\sum_s P(T = 1 \mid S = s, Z = 1) P(S = s \mid Z = 1), \\ M_{01} &= \textstyle\sum_s P(T = 1 \mid S = s, Z = 0) P(S = s \mid Z = 1). \end{aligned}$$

Here, M_{01} measures what the probability of being alive in the standard treatment group would be if the values of the surrogate are distributed as those in the new treatment group. Hence, $M_{01} - M_0$ can be interpreted as the change in the probability of being alive that is due to the new treatment induced effect on the surrogate marker in the standard treatment group. Thus, PE can be interpreted as a measure for the proportion of treatment effect on the survival at 2 years being explained by the change in CD4 values at 16 weeks.

Wang and Taylor (2002) gave sufficient conditions for PE to be between 0 and 1. These conditions reduce to

R1: $P(S = 1 \mid Z = 1) \geq P(S = 1 \mid Z = 0)$
R2: $P(T = 1 \mid S = 1,\ Z = z) \geq P(T = 1 \mid S = 0, Z = z)$ for all z
R3: $P(T = 1 \mid S = s,\ Z = 1) \geq P(T = 1 \mid S = s,\ Z = 0)$ for all s.

Conditions R1 and R2 are very natural and one would expect to be always true; while R3, although quite plausible, is less likely to be universally true; it says that the direction of the association between T and Z is not altered by knowing S.

For the HIV clinical trial R1 says more people in the new treatment group tend to have increasing CD4 values than in the standard treatment group. R2 says that within each treatment group, the percentage alive at 2 years is higher if the CD4 increased compared to if it decreased. R3 says among those with increasing CD4 the new treatment group still gives a higher percentage alive than the standard treatment group, and the same applies for those with decreasing CD4. The

results in Tsiatis et al. (1995) suggest that R3 is likely to be satisfied.

It is easy to show that PE can be reexpressed as:

$$\text{PE} = \delta\gamma_0/\tau \tag{1}$$

where,

$\delta = \Pr(S = 1 \mid Z = 1) - \Pr(S = 1 \mid Z = 0)$,
$\tau = \Pr(T = 1 \mid Z = 1) - \Pr(T = 1 \mid Z = 0)$, and
$\gamma_0 = \Pr(T = 1 \mid Z = 0, S = 1) - \Pr(T = 1 \mid Z = 0, S = 0)$.

Note that PE incorporates three aspects, the treatment effect (δ) on S, the treatment effect (τ) on T, and the association (γ_0) between S and T.

3. Causal Inference Models

3.1 *Counterfactual Characterization of Direct and Indirect Effects*

The concept of proportion of treatment effect explained implies that the overall causal effect of treatment consists of two parts—direct and indirect effects. Indirect effect refers to the part that is mediated through the surrogate marker. Direct effect refers to the part that does not involve the surrogate marker. The ratio of the indirect effect to the overall effect gives the proportion of treatment effect explained by S. If it is possible to block the effect of S on T using some kind of intervention, then randomization of such intervention within each level of Z will allow estimation of these two effects. However, S is usually affected by the treatment and not manipulatable. In this case, causal models using the counterfactual concept provide a way to state clearly what the direct and indirect causal effects are.

Robins and Greenland (1992) developed such a counterfactual model. They assumed 12 types of subjects and their associated proportions in the population as shown in Table 1.

In the table the numbers in parentheses represent those that are never observable unless one can manipulate S. For example, for a type 1 person, S would have the same value 1 whatever Z, and the value of T would be 1 whatever the value of Z; however, if you could manipulate the value of S to be 0 then T would change to 0 for both treatment groups. In terms of the HIV/AIDS example, for a type 1 person, both standard and new treatments result in an increase in CD4, and under both treatments the person is alive at 2 years. However, if you were able to block the increase of CD4 the person would be dead at 2 years in either treatment arm.

Robins and Greenland characterize the direct and indirect effects. For subjects of types 3, 5, and 7, Z would be a direct cause of positive T, because $T = 1$ occurs only when $Z = 1$, and any manipulation of S given Z would not change the result of T. For type 4, Z would be an indirect cause of positive T, because for these subjects $T = 1$ only when $Z = 1$, and modification of S results in a modification of T. For types 2, 4, 5, and 9, Z would be a cause of positive S, as seen from the first two columns in the table. Hence, the excess of the good outcome for T due to direct effect of Z is $p_3 + p_5 + p_7$ and the excess of the good outcome for T due to indirect effect through S is p_4, thus the total overall effect is $(p_3 + p_5 + p_7 + p_4)$, and the proportion of treatment effect explained by S is given by $p_4/p_3 + p_5 + p_7 + p_4$. We denote this by proportion indirect (PI). This and other important quantities are summarized in Table 2.

There are 11 free parameters in this model, but they cannot be individually estimated. However, certain functions of the parameters are estimable. The four conditional probabilities $R_{zs} = P(T = 1 \mid S = s, Z = z)$ and the two conditional probabilities $U_z = P(S = 1 \mid Z = z)$, or other quantities derived from these six, are estimable. Which quantities are estimable and their definition in terms of the 12 probabilities are given in Table 2.

3.2. *Exchangeability Assumptions*

Under this model, Robins and Greenland show that estimation of the direct effect is not possible without further assumptions. They propose "exchangeability assumptions" E1 and E2:

(E1) Probability (R_{10}) of $T = 1$ for subjects with $Z = 1$ and $S_1 = 0$ equals the probability (R_{11B}) of $T = 1$ for subjects with $Z = 1$ and $S_1 = 1$ would have had the treatment effect through S been prevented.

Table 1
Counterfactual model; division of population into subtypes based on potential values of S and T. The numbers in parentheses are the potential values of T when S is changed.

	$Z=1$ $Z=0$		$Z=1$		$Z=0$		
			$S_1=1$ $S_1=0$		$S_0=1$ $S_0=0$		
Type	Values of S		Values of T				Expected proportion
0	1	1	1	(1)	1	(1)	p_0
1	1	1	1	(0)	1	(0)	p_1
2	1	0	1	(1)	(1)	1	p_2
3	1	1	1	(1)	0	(0)	p_3
4	1	0	1	(0)	(1)	0	p_4
5	1	0	1	(1)	(0)	0	p_5
6	0	0	(1)	1	(1)	1	p_6
7	0	0	(1)	1	(0)	0	p_7
8	1	1	0	(0)	0	(0)	p_8
9	1	0	0	(0)	(0)	0	p_9
10	0	0	(1)	0	(1)	0	p_{10}
11	0	0	(0)	0	(0)	0	p_{11}

Table 2
Definitions of important quantities

Quantity	Definitions	Algebraic expression	Estimable
Treatment effect (τ)	$P(T=1\,\vert\,Z=1) - P(T=1\,\vert\,Z=0)$	$p_3 + p_4 + p_5 + p_7$	Y
Surrogate effect (δ)	$P(S=1\,\vert\,Z=1) - P(S=1\,\vert\,Z=0)$	$p_2 + p_4 + p_5 + p_9$	Y
Association (γ_0)	$R_{01} - R_{00}$		Y
Association (γ_1)	$R_{11} - R_{10}$		Y
U_0	$P(S=1\,\vert\,Z=0)$	$p_0 + p_1 + p_3 + p_8$	Y
U_1	$P(S=1\,\vert\,Z=1)$	$p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_8 + p_9$	Y
R_{00}	$P(T=1\,\vert\,Z=0,\,S=0)$	$\frac{p_2+p_6}{p_2+p_4+p_5+p_6+p_7+p_9+p_{10}+p_{11}}$	Y
R_{01}	$P(T=1\,\vert\,Z=0,\,S=1)$	$\frac{p_0+p_1}{p_0+p_1+p_3+p_8}$	Y
R_{10}	$P(T=1\,\vert\,Z=1,\,S=0)$	$\frac{p_6+p_7}{p_6+p_7+p_{10}+p_{11}}$	Y
R_{11}	$P(T=1\,\vert\,Z=1,\,S=1)$	$\frac{p_0+p_1+p_2+p_3+p_4+p_5}{p_0+p_1+p_2+p_3+p_4+p_5+p_8+p_9}$	Y
Indirect effect		p_4	N
Direct effect		$p_3 + p_5 + p_7$	N
Associative effect		$p_4 + p_5$	N
Dissociative effect		$p_3 + p_7$	N
PE	$\delta\,\gamma_0/\tau$		Y
PI	Indirect/treatment	$p_4/(p_3 + p_4 + p_5 + p_7)$	N
PA	Associative/treatment	$(p_4 + p_5)/(p_3 + p_4 + p_5 + p_7)$	N
R_{00A}	$P(T=1\,\vert\,Z=0,\,S=(0\to 1))$	$\frac{p_2+p_6+p_4+p_{10}}{p_2+p_4+p_5+p_6+p_7+p_9+p_{10}+p_{11}}$	N
R_{10A}	$P(T=1\,\vert\,Z=1,\,S=(0\to 1))$	$\frac{p_6+p_7+p_{10}}{p_6+p_7+p_{10}+p_{11}}$	N
R_{01B}	$P(T=1\,\vert\,Z=0,\,S=(1\to 0))$	$\frac{p_0}{p_0+p_1+p_3+p_8}$	N
R_{11B}	$P(T=1\,\vert\,Z=1,\,S=(1\to 0))$	$\frac{p_0+p_2+p_3+p_5}{p_0+p_1+p_2+p_3+p_4+p_5+p_8+p_9}$	N

(E2) Probability (R_{00}) of $T=1$ for subjects with $Z=0$ and $S_0=0$ equals the probability (R_{01B}) of $T=1$ for subjects with $Z=0$ and $S_0=1$ would have had the treatment effect through S been prevented, and state that unless these are satisfied, direct and indirect effects are confounded and not separately identifiable when only the treatment is randomized (Robins and Greenland, 1992). The estimated treatment effect obtained using the conventional adjustment, as proposed in Freedman et al. (1992), is a biased estimate of the direct effect.

3.3 *Conditions for PE to Estimate the PI*

The measures PE is derived from the quantities M_0, M_1, and M_{01}. It can be shown that

$$\begin{aligned} M_0 &= p_0 + p_1 + p_2 + p_6, \\ M_1 &= p_2 + p_4 + p_5 + p_0 + p_1 + p_3 + p_6 + p_7, \\ M_{01} &= \frac{(p_0+p_1)(1-p_6-p_7-p_{10}-p_{11})}{p_0+p_1+p_3+p_8} \\ &\quad + \frac{(p_2+p_6)(p_6+p_7+p_{10}+p_{11})}{(1-p_0-p_1-p_3-p_8)}. \end{aligned}$$

The sum of direct ($p_3 + p_5 + p_7$) and indirect effects (p_4) equals the overall treatment effect ($M_1 - M_0$), and it is hoped that $M_{01} - M_0$ would estimate the indirect effect, while $M_1 - M_{01}$ would estimate the direct effect. From

$$\begin{aligned} M_1 - M_{01} = {} & (\underline{E[T \mid S_0=1, Z=1] - E[T \mid S_1=1, Z=0]}) \\ & \times P(S_1=1 \mid Z=1) \\ & + (\underline{E[T \mid S_0=0, Z=1] - E[T \mid S_1=0, Z=0]}) \\ & \times P(S_1=0 \mid Z=1), \end{aligned}$$

we can see that $M_1 - M_{01}$ is a weighted sum of the *adjusted effect* (the underlined quantities) within each level of observed S.

Robins and Greenland (1992) state that, if E1 and E2 are satisfied, then

$$\begin{aligned} & E[T \mid S_1=1, Z=1] - E[T \mid S_0=1, Z=0] \\ & \quad = E[T \mid S_1=0, Z=1] - E[T \mid S_0=0, Z=0] \\ & \quad = p_3 + p_5 + p_7 = \text{direct effect}. \end{aligned}$$

However, we think two additional complementary exchangeability conditions should be satisfied for this to be true. These extra conditions are E3 and E4:

(E3) Probability(R_{11}) of $T=1$ for subjects with $Z=1$ and $S_1=1$ equals the probability(R_{01A}) of $T=1$ for subjects with $Z=1$ and $S_1=0$ would have had the treatment effect through S been added.

(E4) Probability(R_{01}) of $T=1$ for subjects with $Z=0$ and $S_0=1$ equals the probability(R_{00A}) of $T=1$ subjects with $Z=0$ and $S_0=0$ would have had the treatment effect through S been added.

For the HIV trial illustration, E3 says that the probability of being alive at 2 years is equal for two groups of people. The first group is those who receive the new antiretroviral treatment and show an increase in CD4. The second group is those who receive the new antiretroviral treatment and had a decrease in CD4, but who would have had an increase in CD4 if they were given some other immunostimulating agent that is targeted at CD4 and caused it to increase.

The complete set of exchangeability conditions are:

(E1) $R_{11B} = R_{10}$,
(E2) $R_{01B} = R_{00}$,
(E3) $R_{10A} = R_{11}$,
(E4) $R_{00A} = R_{01}$,

where the subscripts A and B refer to adding and blocking of the surrogate effects, and the expression in terms of the 12 proportions are given in Table 2.

R_{11B} is the incidence of $T = 1$ that subjects with $Z = 1$, $S = 1$ would have had if the effect mediated through S is blocked, requiring this to equal R_{10}, which is the incidence of $T = 1$ for subjects with $Z = 1$, $S = 0$, is not unreasonable. A similar logic is used to justify $R_{01B} = R_{00}$, $R_{10A} = R_{11}$, and $R_{00A} = R_{01}$.

It can be shown that if E1 − E4 are satisfied, $\text{PE} = p_4/(p_3 + p_4 + p_5 + p_7)$.

3.4 *Principal Stratification Approach*

Frangakis and Rubin (2002) proposed a different counterfactual framework for assessing surrogacy. This framework differed from that of Robins and Greenland (1992) because it did not include manipulation of S. They proposed a concept of principal stratification based on which principal causal effect is defined. The principal strata are constructed based on the pair (S_{i0}, S_{i1}), the values of which would not change even though the treatment may have an effect on S such that $S_{i0} \neq S_{i1}$. For the counterfactual model in Table 1, there are three principal strata, those for which $(S_0, S_1) = (0, 0)$ consisting of types 6, 7, 10, and 11, those for which $(S_0, S_1) = (0, 1)$ consisting of types 2, 4, 5, and 9, and those for which $(S_0, S_1) = (1, 1)$ consisting of types 0, 1, 3, and 8. The three principal strata are the three rows in Table 3, labeled PS = (0, 0), PS = (0, 1), and PS = (1, 1).

With this stratification, the principal causal effect always has a causal interpretation because it compares potential outcomes for a common set of subjects. Frangakis and Rubin (2002) proposed to evaluate surrogacy of S through the effects of treatment on T that are associative and dissociative with effects on S. The effect on T from the comparison between $\{T_{i1}\colon S_{i1} = S_{i0}\}$ and $\{T_{i0}\colon S_{i1} = S_{i0}\}$ is defined as dissociative with respect to the effect on S, because any difference between T_1 and T_0 is not accompanied by a difference between S_1 and S_0. The effect on T from the comparison between $\{T_{i1}\colon S_{i1} \neq S_{i0}\}$ and $\{T_{i0}\colon S_{i1} \neq S_{i0}\}$ is defined to be associative with respect to the effect on S, that is, the difference between T_1 and T_0 is associative when $(S_0, S_1) = (0, 1)$. In other words, information about the direct effect is obtained by comparing subjects with $S_0 = S_1$. Considering the scenario illustrated in Table 3, the treatment difference corresponding to the dissociative effect are $p_7 + p_3$ and the treatment difference corresponding to the associative effect are $p_4 + p_5$. We define $PA = (p_4 + p_5)/(p_3 + p_4 + p_5 + p_7)$, denoting the proportion associative.

For the HIV trial illustration the people who have an associative effect are those who would be alive at 2 years if given the new treatment but dead if given the standard treatment, and would have an increase in CD4 if given the new treatment and a decrease if given the standard treatment. The dissociative effect comes from those who would be alive at 2 years if given the new treatment but dead if given the standard treatment, and would have an increase in CD4 under both treatment arms or a decrease in CD4 under both treatment arms.

3.5 *Conditions for PE to Estimate the Associative Proportion*

The following two conditions will result in PE = PA.

A: $(p_2 + p_4 + p_5)/(p_2 + p_4 + p_5 + p_9) = (p_0 + p_1)/(p_0 + p_1 + p_3 + p_8)$
B: $(p_6)/(p_6 + p_7 + p_{10} + p_{11}) = (p_2)/(p_2 + p_4 + p_5 + p_9)$.

Each of these ratios is a conditional probability of $T = 1$ given a specific treatment within one of the three principal strata in Table 3. With these equalities calculation of PE within each principal strata gives appropriately PE = 0 in PS = (0, 0) and PS = (1, 1). Calculation of PE within PS = (0, 1) requires defining $P(T = 1 \mid S = 1, Z = 0)$ which is nonexistent for this stratum. Substituting in this probability from PS = (1, 1) leads to the desired result that PE = 1 in PS = (0, 1). Other than this we do not see an obvious natural interpretation of conditions A and B.

4. Contextual Restrictions

In the general counterfactual model given by Table 1, the only restriction on the parameters is that the 12 values of p add to 1. However, in a clinical trial there will very likely be a considerable amount of scientific subject-matter knowledge that makes some set of parameter values much less likely than others. We consider two types of restrictions, those that are empirically observable and those that are not observable. Only the first type could be verified from data. The second type would need to rely on knowledge of the underlying biological mechanism or intuition of subject-matter specialists.

4.1. *Empirical Restrictions*

4.1.1 Association restrictions. In any clinical trial a marker S would not even be considered for a surrogate unless it had been previously observed to be associated with T and most probably to also have a plausible biological mechanism as to why it is in the pathway for the occurrence of T. Thus, we might expect the values of γ_0 and γ_1, which measure the association between S and T given $Z = 0$ and $Z = 1$, respectively, to be large but not excessively large. Also we might also expect the association not to be too different between

Table 3
Classification of population into subtypes by possible values of (S_0, S_1) *and* (T_0, T_1)

Principal strata	(S_0, S_1)	$(T_0, T_1) = (0, 0)$	$(T_0, T_1) = (0, 1)$	$(T_0, T_1) = (1, 1)$
PS = (0,0)	(0,0)	type = 10,11	type = 7	type = 6
PS = (0,1)	(0,1)	type = 9	type = 4,5	type = 2
PS = (1,1)	(1,1)	type = 8	type = 3	type = 0,1

the two groups. In terms of p_0 to p_{11}, the parameters γ_0 and γ_1 are defined in Table 2.

4.1.2. Effect size on primary outcome. We would not be considering S as a possible surrogate in the trial unless there was a real treatment effect on T, but not too large that one would not be interested in a surrogate. Thus we may restrict $P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0) = p_3 + p_4 + p_5 + p_7$ to a given range of values. Also we expect $P(T = 1 \mid Z = 1)$ to be bounded away from 1, because otherwise there would not be a need to think about surrogate endpoints. Another possible restriction may bound $P(T = 1 \mid Z = 0)$ away from 0, because randomized trials do not tend to be performed when standard treatment is totally ineffective.

4.1.3. Effect size on surrogate. If S is going to be useful as a surrogate it has to be affected by treatment. We would probably expect the effect on the surrogate to be quite strong, because it is frequently chosen to be something that is known to be altered by the treatment, thus we restrict $P(S = 1 \mid Z = 1) - P(S = 1 \mid Z = 0) = p_2 + p_4 + p_5 + p_9$ to a range appropriate to the context.

4.1.4. F in [0,1]. The conditions $R1$, $R2$, and $R3$ given for PE to lie between 0 and 1 can be regarded as empirical restrictions. Condition $R2$ is satisfied if $R_{00} \leq R_{01}$ and $R_{10} \leq R_{11}$ and $R3$ is true if $R_{00} \leq R_{10}$ and $R_{01} \leq R_{11}$.

Condition $R1$ is always true for any set of $(p_0, \ldots, p_{11})$. $R2$ is something we would require to be true, so this represents a restriction on the set $(p_0, \ldots, p_{11})$. $R3$ is an example of a restriction for which there is often, but not always, a reasonable rationale.

4.2 *Unobservable Restrictions*

4.2.1 Exchangeability restrictions. The conditions for exchangeability described in Section 3 are an example of an unobservable restriction. They cannot be verified from data, but they do have plausibility associated with them at least as an approximation. Thus, we might restrict the difference between each pair, $R_{11B} - R_{10}$, $R_{01B} - R_{00}$, $R_{10A} - R_{11}$, $R_{00A} - R_{01}$ to be not excessively far from 0.

4.2.2 Frailty restrictions. In clinical research the concept of frailty is quite plausible, that is, there are some people who will always respond to anything and others who will never respond to anything. For the outcome S this means that $P(S_1 = s,\ S_0 = s) \geq P(S_1 = s) * P(S_0 = s)$, with a similar restriction for T. By the construction of the 12 subgroups in Table 1 these are always satisfied.

4.2.3 Frailty association restrictions. We can consider the three pairs $(S_1 = 0, S_0 = 0)$, $(S_1 = 1, S_0 = 0)$, $(S_1 = 1, S_0 = 1)$ as an ordered classification of the people from least responsive to most responsive with respect to S. The three pairs $(T_1 = 0, T_0 = 0)$, $(T_1 = 1, T_0 = 0)$, and $(T_1 = 1, T_0 = 1)$ have an analogous ordering. Together these give a 3×3 contingency table with ordered categories as shown in Table 3. Association between S and T would imply positive association in 2×2 subtables of the 3×3 table. In terms of the 12 parameters this could be expressed as a set of inequalities

$p_6/(p_6 + p_7 + p_{10} + p_{11}) \leq p_2/(p_2 + p_4 + p_5 + p_9) \leq (p_0 + p_1)/(p_0 + p_1 + p_3 + p_8)$

$(p_{10} + p_{11})/(p_6 + p_7 + p_{10} + p_{11}) \geq p_9/(p_2 + p_4 + p_5 + p_9) \geq p_8/(p_0 + p_1 + p_3 + p_8)$

$p_6/(p_0 + p_1 + p_2 + p_6) \leq p_7/(p_3 + p_4 + p_5 + p_7) \leq (p_{10} + p_{11})/(p_8 + p_9 + p_{10} + p_{11})$

$(p_0 + p_1)/(p_0 + p_1 + p_2 + p_6) \geq p_3/(p_3 + p_4 + p_5 + p_7) \geq p_8/(p_8 + p_9 + p_{10} + p_{11})$.

One interpretation of the association in the 3×3 table is that it is due to unmeasured confounders, for example, a genetic factor that influences both S and T. It is certainly plausible that there are such unmeasured or unknown factors; it is also unlikely that such a factor would have an excessively large impact on the associations between S and T. Thus, we could place further restrictions on the 12 proportions listed above.

5. Simulation Experiment

The relationship between PE, the indirect proportion PI, and the associative proportion PA was investigated in a simulation study. The set of parameters p_0 to p_{11} were simulated 1,000,000 times and for each set the values of PE, PI, and PA were calculated. Then as measures of the link between PE and PI and between PE and PA we calculated the Spearman's correlation coefficient and the proportion of times they were different by less than 0.2. Values of $p_0, \ldots, p_{11}$ were generated from an exponential distribution, then rescaled to make them add to one, to give uniform coverage over the high dimensional space of possible values of p.

To assess the impact of the observable constraints we divided the simulated sets into categories based on the magnitude of the treatment effect on T, the treatment effect on S, and the association between S and T. We restricted the set of p's to the 537,028 sets where PE was between 0 and 1. Figure 1 shows graphs of the correlation between PE and PI and between PE and PA as the treatment effect (τ) on T, the treatment effect (δ) on S, and the association (γ_0) of S and T increase. We see that the average correlation between PE and PI is quite low, around 0.23, whereas the average correlation between PE and PA is higher, around 0.45. As the treatment effect increases there is a modest increase in the correlation between PE and PI, and a more pronounced increase in the correlation between PE and PA. As the effect of Z on S increases there is a clear decrease in the correlation between PE and the other indices. As the association between S and T increases there is not much impact on the correlation between PE and PI or between PE and PA. Overall, these results suggest that observable restrictions, suggested by the context of a randomized trial, are likely to lead to less close links between the PE statistic and the proportions from the counterfactual model.

The joint impact of the observable restrictions was assessed. Table 4 shows a summary of the values of PE, PI, and PA before and after restrictions were applied. The restrictions that were applied were based on a broad range of observable quantities. Comparing the second row with the first row shows that in combination the likely observable restrictions lead to a weaker link between PE and the other two measures.

The impact of the closeness to exchangeability on the correlations in the empirically restricted sample is illustrated in Figure 2a and 2b. The horizontal axis represents the difference

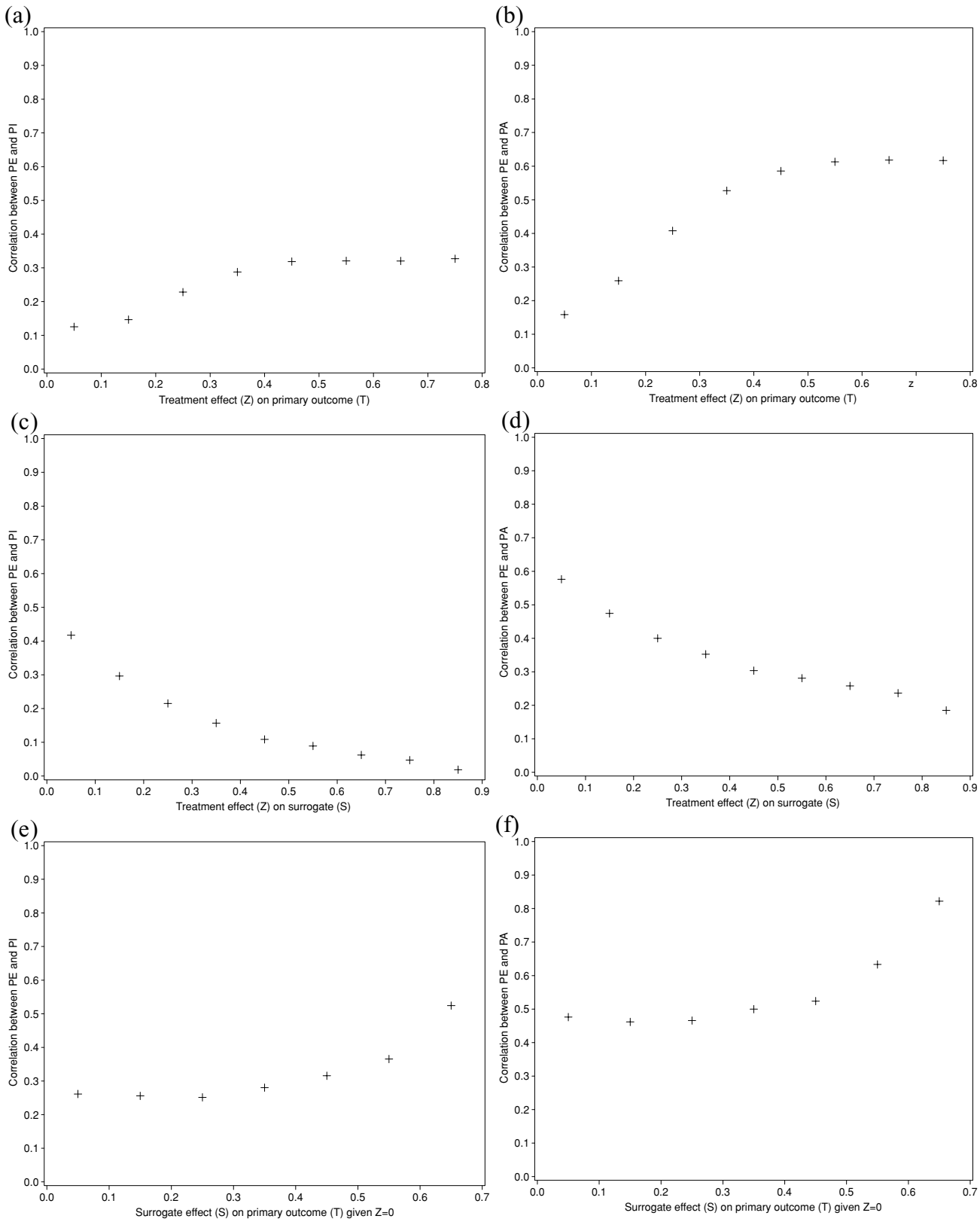

Figure 1. Impact of observable effects on correlation between PE measures. Association between PE and PI (a, c, e) and PE and PA (b, d, f) according to observable effects.

Table 4

Results of simulation experiment: impact of applying empirical and unobservable restrictions on the relationship between PE measures

		Median values			Correlations		% agree $\lvert\Delta\rvert < 0.2$	
Restrictions	n	PE	PI	PA	ρ(PE, PI)	ρ(PE, PA)	(PE, PI)	(PE, PA)
None, PE in [0,1]	537,028	.27	.22	.53	.23	.45	60	45
Empirical[1]	23,452	.42	.25	.58	.09	.21	45	53
Unobservable[2], empirical								
Exchangeability	4077	.26	.26	.58	.44	.28	69	38
Frailty association	3009	.56	.22	.46	.19	.49	28	75
All restrictions	313	.45	.27	.45	.45	.47	56	77

[1] $P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0)$ in (.1, .7), $P(S = 1 \mid Z = 1) - P(S = 1 \mid Z = 0)$ in (.2, .8), $P(T = 1 \mid Z = 1) < .9$, $P(T = 1 \mid Z = 0) > .1$, γ_0 and γ_1 in (.2, .8), $|\gamma_0 - \gamma_1| < .2$.

[2] $|R_{11B} - R_{10}|$, $|R_{01B} - R_{00}|$, $|R_{10A} - R_{11}|$, and $|R_{00A} - R_{01}|$ all $< .2$, frailty association ratios all > 1.0.

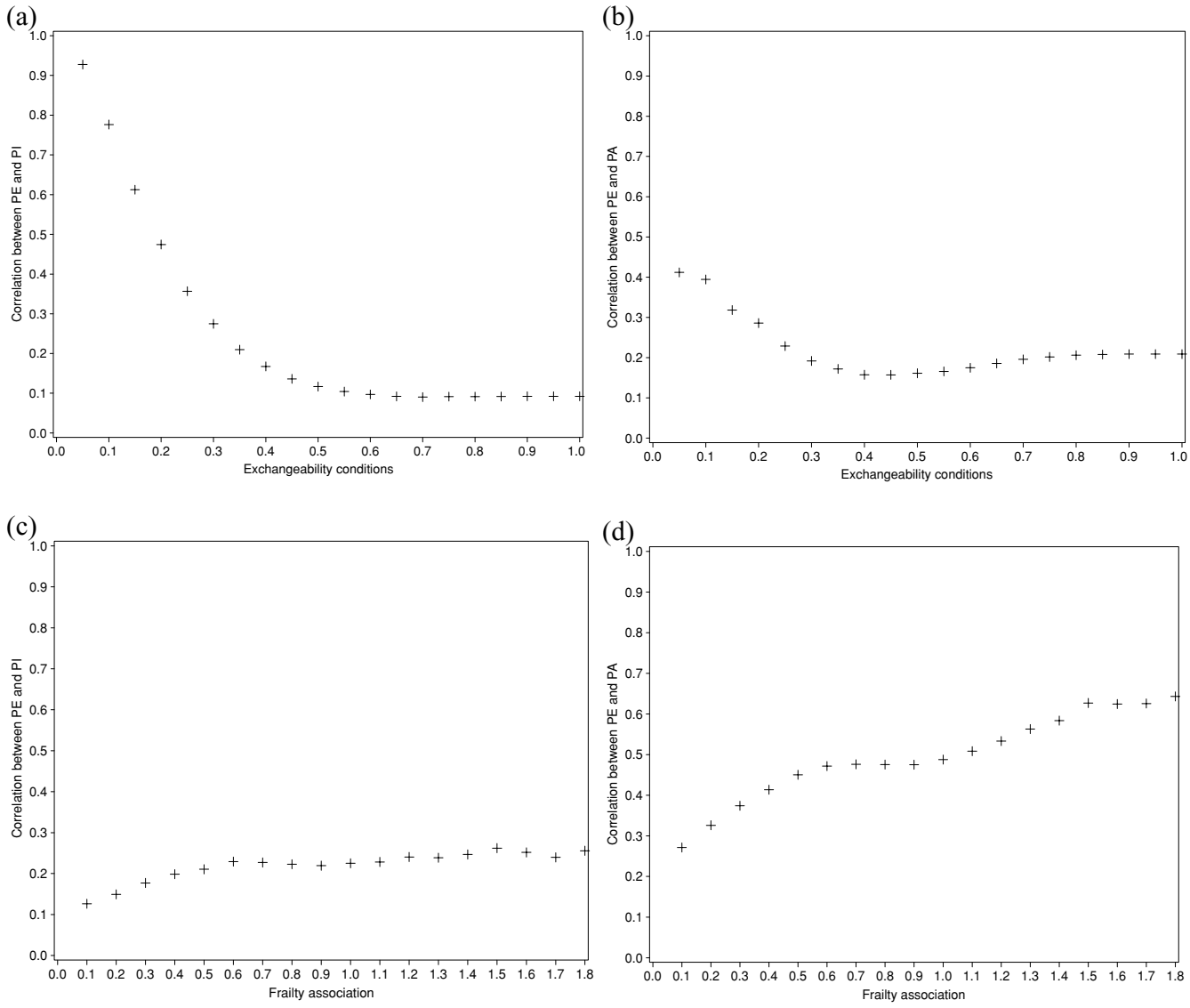

Figure 2. Impact of unobservable effects on correlation between PE measures. Association between PE and PI (a, c) and between PE and PA (b, d) according to the exchangeability conditions (a, b) and frailty association (c, d) in empirically restricted sample.

in proportion between the two ratios in each of the four conditions E1 to E4. Thus, for example, the point above 0.2 has all four of the differences in probability less than 0.2. We see that as the conditions of exchangeability become closer to being satisfied, the correlation between PE and PI becomes closer to 1, and the correlation between PE and PA also increases.

Figure 2c and 2d show the impact of the magnitude of the frailty association in the empirically restricted sample. This is a cumulative plot, with increasing strength of association between S and T represented by large values on the horizontal axis. Thus the point above 1.2 represents all sets of p's that have all of the eight frailty association inequalities satisfied by a factor of at least 1.2. From the left-hand end of the plot we see that eliminating scenarios where the frailty association is substantially less than 1 has an effect of increasing the correlation between PE and PA, but less of an effect on the correlation between PE and PI.

Lastly, we considered the joint impact of all the observable and unobservable contextual restrictions. The unobservable restrictions are chosen to be fairly conservative, that is, only eliminating scenarios where the exchangeability conditions are far from being satisfied or frailty association is negative. Table 4 shows the impact of adding the unobservable restrictions. The results show a modest increase in the link between PE and both the other two measures when the exchangeability and frailty association restrictions are applied. Despite this stronger link there still are cases where PE is close to 1 and either PI or PA is close to 0, or vice versa. It is also worth noting that even though the restrictions were chosen to be fairly mild, they do substantially reduce the number of sets of p's to a minute fraction of all possible p's. This nicely illustrates how general the counterfactual model is, and one should not assume that any set of p's from a counterfactual model are plausible even before data are collected, and in fact the majority may not be. The model for simulating the p's together with the restrictions could be considered as a prior distribution if a Bayesian approach to analyzing a data set is being considered.

6. Discussion

Different approaches for assessing the surrogacy for a biomarker in clinical trials are proposed in the literature. Daniels and Hughes (1997) and Buyse et al. (2000) considered the setting of multiple trials. Ditlevsen et al. (2005) developed the concept of the mediation proportion, based on a latent multivariate normal distribution. Prentice (1989) suggests a strict validation criteria; Freedman et al. (1992), Wang and Taylor (2002), and Li et al. (2001) investigate the concept of the proportion explained. In previous work we suggested a statistic PE to represent the proportion explained and investigated its statistical properties. In the current article we investigate to what extent PE can be given a causal interpretation, as defined by a counterfactual model.

The quantity PE can be estimated from data, whereas PI and PA cannot be estimated without untestable assumptions. The most ideal result of this work would have been finding regions of the parameter space, defined by observable quantities, for which the link between PE and either PI or PA was strong. We were unable to find such regions. We did find restrictions, based on unobservable quantities, most notably exchangeability but also to a lesser extent frailty association, for which there was a stronger link between PE and PA or PI. Thus in the context of the trial under consideration, if these conditions seem plausible, it does give some assurance that PE can be used and will tend to have at least approximately a causal interpretation as defined by a counterfactual model.

Although the principles that motivate PI and PA are different, the algebraic formulas are quite similar and only differ by the extra term p_5 in the numerator of PA, thus they will be quite correlated with each other, but PA will always be larger than PI. As the name suggests, the proportion associated (PA) potentially incorporates the concept of association, particularly frailty association in the 3×3 table (Table 3). Thus it is not surprising that PA and PE are closer when the frailty association is high.

The exchangeability conditions essentially require the outcome to depend on the final value of the surrogate and not whether it was manipulated to arrive at this final value. Under such conditions it is expected that the PE statistics, which is motivated by considering what the outcome would be if the distribution of the surrogate in the placebo group was changed to that of the treatment group, would equal the PI. However, when exchangeability is not satisfied then the outcome will depend on whether the final surrogate value was arrived at by manipulation or not, thus the statistics PE, which can be estimated from data in which manipulation did not occur, will differ from PI. As we showed in Figure 2, the closeness of PE and PI is directly determined by the degree to which the exchangeability conditions are a good approximation.

The practical implications of the results in this article are that of the three measures, PE is the only one that can be estimated from the data, but may not have a causal interpretation as defined by a counterfactual model, the degree to which it has an approximate causal interpretation will depend on the scientific context, and the degree to which one is willing to make untestable assumptions.

We have considered only the situation of a binary endpoint and a binary surrogate. The calculation of PE in more complex situations such as continuous or longitudinal S and continuous or censored T is certainly possible. The extent to which PE can be given causal interpretations in such settings will need to be investigated.

References

Aboulker, J., Babiker, A., Flandre, P., Gazzard, B., Loveday, C., Nunn, A. (1999). An evaluation of HIV RNA and CD4 cell count as surrogates for clinical outcome. *AIDS* **13,** 565–573.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). Validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics* **1,** 49–68.

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

Delta. (1996). Delta: A randomised double-blind controlled trial comparing combinations of zidovudine plus

didanosine or zalcitabine with zidovudine alone in HIV-infected individuals. *Lancet* **348,** 283–291.

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M., and Keiding, N. (2005). The mediation proportion: A structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology* **16,** 114–120.

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 disease. *Statistics in Medicine* **11,** 167–178.

Hughes, M. E. A. (2000). Human immunodeficiency virus type 1 RNA level and CD4 count as prognostic markers and surrogate end points: A meta-analysis. HIV Surrogate Marker Collaborative Group. *AIDS Research and Human Retroviruses* **16,** 1123–1133.

Lazzarin, A., Clotet, B., Cooper, D., et al. (2003). Efficacy of enfuvirtide in patients infected with drug-resistant HIV-1 in Europe and Australia. *New England Journal of Medicine* **348,** 2186–2195.

Li, Z., Meredith, M., and Hoseyni, M. (2001). A method to assess the proportion of treatment effect explained by a surrogate endpoint. *Statistics in Medicine* **20,** 3175–3188.

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

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

Tsiatis, A., De Gruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90,** 27–37.

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

Received June 2004. *Revised January* 2005.
Accepted February 2005.