BIOMETRICS 58, 803–812
December 2002

A Measure of the Proportion of Treatment Effect Explained by a Surrogate Marker

Yue Wang* **and Jeremy M. G. Taylor**

Department of Biostatistics, University of Michigan,
1420 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.
**email:* wangyue@umich.edu

SUMMARY. Randomized clinical trials with rare primary endpoints or long duration times are costly. Because of this, there has been increasing interest in replacing the true endpoint with an earlier measured marker. However, surrogate markers must be appropriately validated. A quantitative measure for the proportion of treatment effect explained by the marker in a specific trial is a useful concept. Freedman, Graubard, and Schatzkin (1992, *Statistics in Medicine* **11**, 167–178) suggested such a measure of surrogacy by the ratio of regression coefficients for the treatment indicator from two separate models with or without adjusting for the surrogate marker. However, it has been shown that this measure is very variable and there is no guarantee that the two models both fit. In this article, we propose alternative measures of the proportion explained that adapts an idea in Tsiatis, DeGruttola, and Wulfsohn (1995, *Journal of the American Statistical Association* **90**, 27–37). The new measures require fewer assumptions in estimation and allow more flexibility in modeling. The estimates of these different measures are compared using data from an ophthalmology clinical trial and a series of simulation studies. The results suggest that the new measures are less variable.

KEY WORDS: Surrogate marker; Validation.

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. Surrogate endpoints are usually intermediate biomarkers in disease development that can be observed and assessed earlier and are often easy to measure. They are generally proposed based on the biological process of a disease and their strong correlations with the primary endpoint (Ellenberg and Hamilton, 1989; Schatzkin et al., 1990; Fleming, 1992). However, correlation alone is not a good statistical criterion for surrogate validation. For example, in AIDS-related trials, it is known that the levels of CD4 counts and virus load are associated with AIDS and death. But data suggest that treatment-induced improvements in CD4 or viral RNA load do not reliably predict treatment-induced changes in the primary clinical outcomes (Lagakos and Hoth, 1992; Fleming and DeMets, 1996).

Prentice (1989) proposed a formal definition of surrogate endpoints and gave general operational criteria for validation of the surrogate endpoints. By his criteria, an appropriate surrogate endpoint is required to fully capture the treatment effect on the primary endpoint. This is rather too stringent a criterion and is unlikely to be satisfied completely. In practice, it is likely that a surrogate endpoint may explain part but not all the treatment effect. However, the more explained, the better we will perceive the marker as a surrogate endpoint. Thus, a quantitative measure of the proportion of the treatment effect that is explained by the surrogate marker would be a useful summary measure. For convenience, we refer to the proportion of the treatment effect explained by the biomarker as PE. Freedman, Graubard, and Schatzkin (1992) proposed such a quantitative measure for a single trial.

Daniels and Hughes (1997) proposed a meta-analysis method for the evaluation of a potential surrogate marker. A Bayesian approach was used to model the association between the treatment effects on the primary endpoints and the treatment effects on the surrogate endpoints from multiple clinical trials. A marker is considered as a valid surrogate if the association is significantly different from zero. Their approach enables one to obtain prediction intervals for the treatment effect on the primary outcome given the estimated treatment effect on the surrogate markers from a new trial.

Buyse et al. (2000) developed a new set of criteria for surrogate validation within the multiple clinical trial settings. They proposed a trial-level and an individual-level criteria. The trial-level criterion is the coefficient of determination (R^2) for prediction of the treatment effect on the primary endpoint conditioned on the treatment effect on the surrogate. The individual-level criterion is the correlation between the surrogate and the primary endpoint at the subject level.

In this article, we focus our attention on measures for PE in the context of a single randomized clinical trial. The following notation is adopted: T and S denote the random variables for the primary endpoint and surrogate markers, respectively. Z is the binary treatment indicator variable, with $Z = 1$ for treatment (or new treatment) and $Z = 0$ for placebo (or standard treatment). We will be assuming throughout the article that there is a treatment effect on the primary endpoint, i.e., $P(T \mid Z) \neq P(T)$, where $P(\cdot)$ denotes the probability distribution.

In a randomized clinical trial, a perfect surrogate occurs when S captures all the dependence of T on Z. In other words, $P(T \mid Z, S) = P(T \mid S)$. A useless surrogate can occur in cases where, conditional on the treatment assignments, the surrogate is independent of the primary endpoint, $P(T \mid Z, S) = P(T \mid Z)$. A useless surrogate can also occur when S is independent of the treatment indicator Z, $P(S \mid Z) = P(S)$. A partial surrogate occurs when the surrogate endpoint captures some but not all the dependence of T on Z. An ideal measure of PE will be one for a perfect surrogate, zero for a useless surrogate, and between zero and one for a partial surrogate.

In Section 2, we review the quantitative measure for PE proposed by Freedman et al. (1992). In Section 3, we propose new measures for PE and investigate some of their properties. In Section 4, we discuss approaches to estimation and inference for the new measures. In Section 5, data from a clinical trial in ophthalmology are used to estimate the new measures and Freedman's measure. In Section 6, simulation studies are carried out to compare the new measures with Freedman's measure in cases where T and S are binary.

2. Freedman's Measure for Proportion of Treatment Effect Explained

Freedman et al. (1992) proposed a measure for PE within the context of a binary primary endpoint. The measure, denoted P, is defined based on two logistic models,

$$\log \frac{p(T = 1 \mid S, Z)}{1 - p(T = 1 \mid S, Z)} = \mu_1 + \beta_S Z + \phi_Z S \qquad (1)$$

and

$$\log \frac{p(T = 1 \mid Z)}{1 - p(T = 1 \mid Z)} = \mu_2 + \beta Z, \qquad (2)$$

and $P = (\beta - \beta_S)/\beta = 1 - \beta_S/\beta$, the difference between the treatment effect with or without adjusting for the surrogate marker divided by the unadjusted treatment effect. Assuming treatment has a significant effect on the primary outcome, then $P = 1$ if $\beta_S = 0$ and $P = 0$ if $\beta_S = \beta$. The measure P generalizes in obvious ways for other nonbinary variables T with appropriate models, e.g., time-to-event outcome with proportional hazard (PH) models.

Freedman's P suffers a number of drawbacks (Daniels and Hughes, 1997; Lin, Fleming, and DeGruttola, 1997; Buyse and Molenberghs, 1998; Bycott and Taylor, 1998). First, models for $[T \mid S, Z]$ and $[T \mid Z]$ will be fitted simultaneously for estimation of P. However, in general, they will not both be true at the same time. Assuming the model for $[T \mid S, Z]$ is the correct one, integration with respect to $P(S \mid Z)$ will not usually result in the exact linear form as the model for $[T \mid Z]$ except for a few special cases. Thus, in general, at least one model fitted will be misspecified. Lin et al. (1997) studied the behavior of estimated coefficients from misspecified PH models for censored failure time data.

Another drawback is that calculation of P requires there to be no significant interaction term between the surrogate S and the treatment Z in model (1). When the data suggest an interaction, P is not well defined. A third drawback for P to be a useful measure is that it has a large variability. To get a reasonably precise estimate for P requires a highly significant unadjusted treatment effect or a large sample size. Otherwise, the point estimate $\hat{P}$ can lie outside $[0, 1]$ and the confidence interval for P frequently covers the whole $[0, 1]$ interval and is too wide to be useful (Freedman et al., 1992; DeGruttola et al., 1996; Lin et al., 1997; Bycott and Taylor, 1998).

3. A New Measure for Proportion Explained

3.1 *Definition of F*

Tsiatis, DeGruttola, and Wulfsohn (1995) studied the relationship between survival and longitudinal CD4 counts. A model for the CD4 trajectory in each treatment group is developed. Let $S(t, \mathrm{Tr})$ and $S(t, \mathrm{Pl})$ denote the predicted survival curve in the treatment group and the placebo group, respectively. To see how much survival benefit could have been predicted by just the increase in CD4 counts, using the hazard function for the placebo group, the predicted survival curve $S(t, \mathrm{mix})$ for an average CD4 trajectory from the treatment group was computed. If CD4 counts serve as a useful surrogate endpoint, $S(t, mix)$ would to be close to $S(t, \mathrm{Tr})$, whereas if changes in CD4 explain very little of the treatment effect, $S(t, \mathrm{mix})$ would be close to $S(t, \mathrm{Pl})$. The relative position of $S(t, \mathrm{mix})$ between $S(t, \mathrm{Tr})$ and $S(t, \mathrm{Pl})$ was suggested as a way to assess the proportion of treatment effect explained.

Motivated by their work, we propose a new measure F for PE. We refer to the placebo group ($Z = 0$) as group A and to the treatment group ($Z = 1$) as group B. The new measure is $F = (AA - AB)/(AA - BB)$, where $AA = h(\int g_A(s) dP_A(s))$, $BB = h(\int g_B(s) dP_B(s))$, and $AB = h(\int g_A(s) dP_B(s))$. Here $P_A(S)$ and $P_B(S)$ denote the distribution of S in groups A and B, respectively. $g_A(S)$ and $g_B(S)$ are functions of the conditional distribution of T given S in the placebo and treatment groups, respectively. For example, $g_A(S)$ and $g_B(S)$ can be the mean of this distribution. $h(\cdot)$ is a monotonic function. In general, $h(\cdot)$, $g_A(S)$, and $g_B(S)$ are chosen such that $(AA - BB)$ will be the desired measure of treatment effect on the primary endpoint T.

To explain the idea, let T be a binary variable. Choose $h(u) = u$, $g_A(S) = \Pr(T = 1 \mid S, Z = 0)$, and $g_B(S) = \Pr(T = 1 \mid S, Z = 1)$. Then population quantities $AA = \Pr(T = 1 \mid Z = 0)$ and $BB = \Pr(T = 1 \mid Z = 1)$. The treatment effect $(AA - BB)$ is the difference on the probability scale between the two groups. AA is the weighted mean of the conditional probability $g_A(S)$ with weight given by the density of S in the placebo group. Similarly, $AB = \int g_A(s) dP_B(s)$ is the weighted mean of $g_A(S)$ with the density of S in the treatment group as weight. Hence, AB measures what $\Pr(T = 1)$ in the placebo group would be if the values of the surrogate are distributed as those in the treatment group. If $T = 1$ represents disease occurrence, then $AA - AB$ can be interpreted as the change in the risk that is due to the change in distribution of S induced by the treatment. A complementary form of F is $F' = (BA - BB)/(AA - BB)$ with $BA = h(\int g_B(s) dP_A(s))$. Ideally, the values of F and F' will be close to each other.

The choices of $h(\cdot)$, $g_A(S)$, and $g_B(S)$ define the treatment effect. For binary T, we could express the treatment effect on the probability scale using $h(u) = u$ as shown above. Alternatively, we could let $h(u) = \log u/(1-u)$ and $g(S) = P(T = 1 \mid S, Z)$ if we believe the logit is the more natural scale on which to assess probabilities. In this case, the treatment effect is expressed by log(odds ratio) with $AA = \text{logit}[\Pr(T = 1 \mid Z = 0)]$ and $BB = \text{logit}[\Pr(T = 1 \mid Z = 1)]$. For a time-to-event variable T, we may choose $h(u) = u$, $g_A(S) = \Pr(T > c \mid S, Z = 0)$ and $g_B(S) = \Pr(T > c \mid S, Z = 1)$, where c is a prespecified time, say 5 years. Then the treatment effect $(AA - BB)$ is the difference in the probability of surviving at least 5 years between two groups.

3.2 *Conditions for Measure to Be Bounded Within* [0, 1]

As a measure of proportion, we would prefer F and F' to have values in the range from 0 to 1 with 0 indicating a useless surrogate and 1 a perfect one. However, this is not guaranteed. It is possible for F and F' to be less than 0 or greater than 1. In this section, we present and discuss the conditions for the population quantities F and F' to be bounded within $[0, 1]$. We will also discuss situations where these population quantities will be outside $[0, 1]$. By population quantities, we mean F and F' as functions of the joint distribution of T and S, which can be interpreted as the values that would arise from an infinitely large sample.

For simplicity, assume $h(u) = u$. For a perfect surrogate, the conditional distribution $[T \mid S, Z]$ is the same as $[T \mid S]$. Thus, $g_A(S) = g_B(S)$ and $F = F' = 1$. For a useless surrogate, either $P_B(S) = P_A(S)$, or $g_A(S)$ and $g_B(S)$ are constant with respect to S. Both result in $F = F' = 0$. Hence, for the perfect or useless scenario, the population quantities F and F' are equal and take the desired values.

For a partial surrogate, in special cases, e.g., when the treatment effect on T is not a function of S, in other words, $g_A(S) = g_B(S) + c$ where c is a constant, F will be equal to F'. In general, F' is likely to have different values from F. As a measure of proportion, ideally, F and F' will fall between zero and one in partial surrogate scenarios. However, this does not hold in all situations. Certain conditions need to be satisfied for F and F' to be bounded within $(0, 1)$.

Without the loss of generality, we assume $AA > BB$. Under this assumption, the conditions for F and F' to be bounded within $(0, 1)$ are

$$\text{C1}: F > 0 \quad \text{iff} \quad \int g_A(s)d[P_A(s) - P_B(s)] > 0$$

$$\text{C2}: F' > 0 \quad \text{iff} \quad \int g_B(s)d[P_A(S) - P_B(S)] > 0$$

$$\text{C3}: F < 1 \quad \text{iff} \quad \int [g_A(s) - g_B(s)]dP_B(S) > 0$$

$$\text{C4}: F' < 1 \quad \text{iff} \quad \int [g_A(s) - g_B(s)]dP_A(S) > 0.$$

These conditions are best understood in terms of a plot of $g_A(S)$ and $g_B(S)$ versus S, as shown in Figure 1. $[S \mid A]$ and $[S \mid B]$ are the distributions of the surrogate marker in groups A and B, respectively. $g(S \mid Z)$ is a function of the distribution of $[T \mid S, Z]$ for each value of Z. It could be the probability of disease or survivorship up to a certain time point or hazard at a certain time point. The position AA indicates approximately the value of $\text{E}_A[g_A(S)]$; similarly, AB approximates $\text{E}_B[g_A(S)]$, BA approximates $\text{E}_A[g_B(S)]$, and BB approximates $\text{E}_B[g_B(S)]$. For the perfect surrogate scenario, $g_A(s) = g_B(s)$, and it is easy to see $F = F' = 1$ from the plot. For the useless surrogate scenario, either $[S \mid Z = 1]$ is identical to $[S \mid Z = 0]$ or $g_A(S)$ and $g_B(S)$ become horizontal lines (constant with respect to S). Both give $F = F' = 0$. For the partial surrogate scenario, graphically what is required for both F and F' to lie between zero and one is that both AB and BA lie between AA and BB.

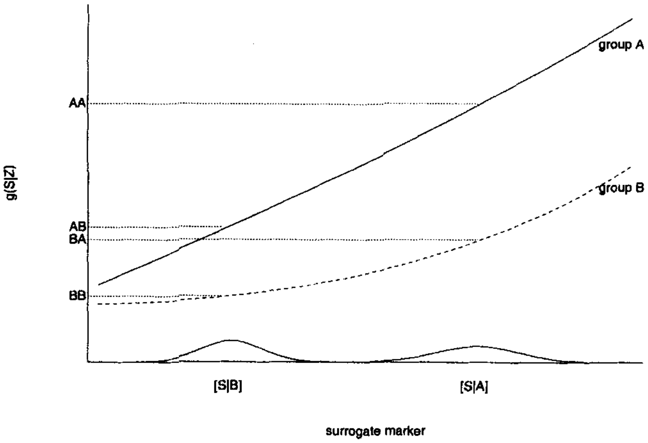

Figure 1. Illustration plot of $g_A(S)$, $g_B(S)$ versus S. $g_A(S)$ and $g_B(S)$ are functions of the conditional distributions of $[T \mid S, Z]$ for group A (the placebo group) and group B (the treatment group). The solid line in the plot is $g_A(S)$ and the dashed line is $g_B(S)$. AA indicates the position that is approximately the value of $\text{E}_A[g_A(S)]$; similarly, AB approximates $\text{E}_B[g_A(S)]$, BA approximates $\text{E}_A[g_B(S)]$, and BB approximates $\text{E}_B[g_B(S)]$.

The four necessary and sufficient conditions C1–C4 can be further simplified depending on the shape of $g_A(S)$ and $g_B(S)$ and the distributions of S in the two groups. Plots similar to Figure 1 enable us to visualize relatively easily the necessary and sufficient conditions. Consider the cases where the distributions $P_A(S)$ and $P_B(S)$ have a stochastic ordering. Without loss of generality, assume $P_A(S)$ is stochastically higher than $P_B(S)$, i.e., $\Pr(S \leq s \mid \text{group B}) \geq \Pr(S \leq s \mid \text{group A})\ \forall\ s$. It is easy to see that

R1: $P_A(S)$ stochastically higher than $P_B(S)$

R2: $g_A(S)$ and $g_B(S)$ are nondecreasing functions of S

R3: $g_A(S) - g_B(S) \geq 0$ for all S

are sufficient but not necessary conditions for C1–C4 to be satisfied and hence for F and F' to be bounded within $[0, 1]$. What is shown in Figure 1 is a typical case of such situations. Note $g_A(S)$ and $g_B(S)$ need not be parallel. In other words, the treatment effect can be a function of S. R1–R3 can be further weakened. For example, $g_A(S) - g_B(S) \geq 0$ can be relaxed by allowing $g_A(S)$ and $g_B(S)$ to cross each other at certain points. Nondecreasing $g_A(S)$ and $g_B(S)$ can also be relaxed to allow certain regions of $g(S \mid Z)$ to be decreasing.

R1–R3 are simpler than C1–C4 and are easier to understand and check. Condition R1 is an assumption on how treatment affects the surrogate marker, and it is reasonable to think that it is satisfied for plausible markers being seriously considered in a randomized trial. Condition R2 is likely to be satisfied for any marker that is strongly associated with the primary endpoint. Condition R3 is the one that may not be satisfied in every trial. For example, if $g_A(S)$ and $g_B(S)$ represent the risk in the control and treatment groups, respectively, R3 requires that, conditioning on each value for the surrogate marker, the risk for the treated group should be consistently lower than the risk in the control group. R3 may not be true when there are unexpected aspects of the pathway between treatment, surrogate marker, and primary endpoint. For example, unintended adverse effects can occur such that the risk is higher in the treated group than in the control group. In these cases, $g_A(S)$ is lower than $g_B(S)$ instead of higher, as shown in Figure 1. If the magnitude of the adverse effect is strong, $AA-BB$ can be negative while $AA-AB$ and $BA-BB$ are positive due to $R1$ and $R2$. As a result, F and F' have negative values. If the magnitude of the adverse effect is weak, $AA-BB$ will be positive but less than $AA-AB$ or $BA-BB$ and results in values greater than one for F and F'.

In cases where conditions C1–C4 or R1–R3 are satisfied, the scale from zero to one corresponds to a meaningful transition from useless surrogacy to perfect surrogacy. However, as described above, the population quantities F (F') can have values greater than one or less than zero, which most likely indicates the existence of unintended effects.

3.3 *Interpretation of F (F') in Special Cases*

F is constructed based on the concept of trying to measure what the treatment effect would be if the surrogate marker in the nontreated group has the treatment-induced distribution and vice versa for F'. However, AB and BA are two hypothetical quantities that are not observable. In this section, we investigate F (F') in several special situations where they can be represented by familiar quantities.

3.3.1 *Normally distributed T and S.* In cases where both T and S are normally distributed with linear mean structure,

$$S_i = \alpha_0 + \alpha_1 Z_i + \epsilon_{Si}$$
$$T_i = \gamma_0 + \gamma_1 Z_i + \epsilon_{Ti},$$

where

$$\begin{pmatrix}\epsilon_{Si}\\ \epsilon_{Ti}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_S^2 & \sigma_{ST}\\ & \sigma_T^2\end{pmatrix}\right),$$

the model for T given (S, Z) is a linear model without an interaction term,

$$T = \beta_0 + \beta_1 S + \beta_2 Z + \epsilon^*,$$

where

$$\beta_0 = \gamma_0 - \frac{\sigma_T}{\sigma_S}\rho\alpha_0,$$
$$\beta_1 = \frac{\sigma_T}{\sigma_S}\rho,$$
$$\beta_2 = \gamma_1 - \frac{\sigma_T}{\sigma_S}\rho\alpha_1,$$
$$\rho = \frac{\sigma_{ST}}{\sigma_S\sigma_T},$$

and

$$\epsilon^* \sim \mathrm{N}\left(0, \sigma_T^2\left(1-\rho^2\right)\right).$$

If we choose the treatment effect to be the difference in mean, then

$$\begin{aligned} g_A(S) &= \mathrm{E}[T \mid S, Z=0] = \beta_0 + \beta_1 S,\\ g_B(S) &= \mathrm{E}[T \mid S, Z=1] = \beta_0 + \beta_1 S + \beta_2,\\ AA &= \mathrm{E}[T \mid Z=0] = \gamma_0,\\ BB &= \mathrm{E}[T \mid Z=1] = \gamma_0 + \gamma_1,\\ AB &= \mathrm{E}_{S\mid Z=1}\mathrm{E}[T \mid S, Z=0] = \beta_0 + \beta_1(\alpha_0+\alpha_1),\\ BA &= \mathrm{E}_{S\mid Z=0}\mathrm{E}[T \mid S, Z=1] = \beta_0 + \beta_2 + \beta_1\alpha_0, \end{aligned}$$

and

$$F = F' = \frac{\beta_1\alpha_1}{\beta_2 + \beta_1\alpha_1}.$$

Thus, F (F') depends on the strength of the association between T and S (measured by β_1), the effect of Z on S (α_1), and the adjusted treatment effect (β_2). In this example, the linear form of the mean structure for $[T \mid Z]$ can be obtained exactly by integrating $\mathrm{E}[T \mid S, Z]$ with respect to $P(S \mid Z)$. We note that

$$P = 1 - \frac{\beta_2}{\gamma_1} = \frac{\beta_1\alpha_1}{\beta_2 + \beta_1\alpha_1},$$

which is equal to F (F'). Also note that, in this case, condition R1 is satisfied if $\alpha_1 < 0$ and R2 is satisfied if $\beta_1 \geq 0$ and condition R3 is satisfied if $\beta_2 \leq 0$.

3.3.2 *Binary primary endpoint and binary surrogate marker.* Many diagnostic markers, with appropriate cutoff values, are used in medical applications to predict disease. A good test would have high positive predictive value (PPV) and high negative predictive value (NPV). Let p^+ denote Pr(disease) and $p^- = 1 - p^+$. If a binary surrogate marker S is defined such that

$$S = \begin{cases} 1 & \text{if positive diagnosis} \\ 0 & \text{otherwise} \end{cases}$$

and the treatment effect is the difference on the probability scale, then $F = \delta\gamma_A/\tau$ and $F' = \delta\gamma_B/\tau$, where

$$\delta = \Pr(S=1 \mid Z=0) - \Pr(S=1 \mid Z=1),$$

which is the treatment effect on S,

$$\tau = \Pr(T=1 \mid Z=0) - \Pr(T=1 \mid Z=1),$$

which is the treatment effect on T,

$$\gamma_A = \Pr(T=1 \mid Z=0, S=1) - \Pr(T=1 \mid Z=0, S=0),$$

and

$$\gamma_B = \Pr(T=1 \mid Z=1, S=1) - \Pr(T=1 \mid Z=1, S=0).$$

γ_A and γ_B are equal to $[(PPV - p^+) + (NPV - p^-)]$ in the placebo group and the treatment group, respectively; thus, γ reflects the strength of the association between S and T, with larger γ indicating higher association. F and F' are influenced by two components: the ratio of the treatment effect on S and T and the accuracy of the prediction of T based on S. For a given treatment effect on the primary outcome (τ), the larger the effect on the surrogate marker (δ) or the stronger the association between S and T, the higher F and F' will be. The two components of F (F') are very similar to the measures Buyse and Molenberghs (1998) proposed for surrogate validation, relative effect (RE) and adjusted association.

Condition R1 is satisfied if $\delta > 0$ and R2 is satisfied if $\gamma_A \geq 0$ and $\gamma_B \geq 0$. R3 is satisfied if $NPV_A \leq NPV_B$ and $PPV_A \geq PPV_B$, where (NPV_A, PPV_A) and (NPV_B, PPV_B) are the NPV and PPV in the placebo and treatment groups, respectively. Note in this case $F = F'$ if $(PPV + NPV)$ are the same in the treatment and placebo groups.

3.3.3 *Binary primary endpoint T.* Now consider cases where T is binary and the effect of S on T in group B is related to that for group A by a linear shift on the logistic scale. In other words,

$$\operatorname{logit}(\Pr(T = 1 \mid S, Z)) = \beta_0 + \beta_1 S - \omega Z.$$

Assume ω to be a nonnegative constant so that treatment will reduce the odds of $T = 1$ given $S = s$. If we define the treatment effect $(AA - BB)$ on the probability scale, then

$$g_A(S) = \frac{\exp(\beta_0 + \beta_1 S)}{1 + \exp(\beta_0 + \beta_1 S)},$$
$$g_B(S) = \frac{\exp(\beta_0 + \beta_1 S - \omega)}{1 + \exp(\beta_0 + \beta_1 S - \omega)},$$

which leads to messy expressions for F and F'. If instead we define the treatment effect on the logit probability scale and choose $g_A(S)$ and $g_B(S)$ to be $g_A(S) = \beta_0 + \beta_1 S$ and $g_B(S) = \beta_0 + \beta_1 S - \omega$, then

$$F = F' = \frac{\int \beta_1 S(dP_A(S) - dP_B(S))}{\int \beta_1 S dP_A(S) - \int (\beta_1 S - \omega) dP_B(S)},$$

which simplifies to

$$\frac{\beta_1(\mu_{S,A} - \mu_{S,B})}{\beta_1(\mu_{S,A} - \mu_{S,B}) + \omega} = 1 - \frac{\omega}{\beta_1(\mu_{S,A} - \mu_{S,B}) + \omega},$$

where $\mu_{S,A}$ and $\mu_{S,B}$ are the expectation of S in the two groups.

F' equals F because $g_A(s)$ is a linear shift of $g_B(s)$. In this form, F (F') combines three components: the magnitude of the effect of the treatment on the marker $(\mu_{S,A} - \mu_{S,B})$, the strength of the association between the marker and the endpoint (β_1), and the adjusted effect of the treatment on the endpoint (ω). This form of F (F'), which we can think of as a transformation of the original one, also gives some insight into the meaning of F (F'). However, the original F on the probability scale may have a more informative interpretation in terms of the proportion of treatment effect being explained. In this case, where treatment effect is defined on the logit probability scale, condition R3 is satisfied if ω is nonnegative and R2 is satisfied if $\beta_1 \geq 0$. The condition R1 is equivalent to $\mu_{S,A} > \mu_{S,B}$.

4. Estimation and Inference

4.1 *Estimation*

Estimation of Freedman's P requires joint modeling of $[T \mid S, Z]$ and $[T \mid Z]$, which leads to the possibility that one of the models may be misspecified. In addition, P is only defined when there is no significant interaction between S and Z for model $[T \mid S, Z]$. In contrast, F (F') requires the specification of $[T \mid S, Z]$ and $[S \mid Z]$. Unlike P, estimation of F is not necessarily tied to the linear models and can be much more flexible. For example, we can fit a generalized linear model for $[T \mid S, Z]$ to obtain estimates $g_A(s)$ and $g_B(s)$. Such models could include interactions between S and Z if needed. Or we may choose estimation approaches that impose less parametric structures on $g_A(S)$, $g_B(S)$, and their relationship. This flexibility in choice of estimation approaches enables F (F') to be used in more general design settings than is P, especially if the linear no-interaction model is not preferred.

To estimate F and F', we need estimates for the distribution $[S \mid Z]$ and estimates for $g_A(S)$ and $g_B(S)$, denoted by $\hat{g}_A(S)$ and $\hat{g}_B(S)$, respectively. $[S \mid Z]$ can be estimated simply by the empirical distribution, which results in $\hat{AA} = (1/N_A) \times \Sigma_{i=1}^{N_A} \hat{g}_A(S_i)$, $\hat{BB} = (1/N_B) \Sigma_{i=1}^{N_B} \hat{g}_B(S_i)$, and $\hat{AB} = (1 \div N_B) \Sigma_{i=1}^{N_B}, \hat{g}_A(S_i)$, where N_A and N_B are the number of subjects in the placebo and treatment groups, respectively. If parametric forms are used for $[S \mid Z]$, Monte Carlo methods can be used to evaluate the integrals based on estimates of $g_A(S)$, $g_B(S)$, $P_A(S)$, and $P_B(S)$ from the data.

4.2 *Confidence Intervals*

F (F') is a ratio of parameters. For simplicity, denote F (F') by $r = \theta_1/\theta_2$. Let σ_{11}, σ_{22}, and σ_{12} be the variance of the estimates $\hat{\theta}_1$, $\hat{\theta}_2$, and their covariance, respectively. Three different approaches can be used to construct the $(1 - \alpha)\%$ confidence interval for F (F'). The first one is based on Fieller's Theorem. It assumes that $(\hat{\theta}_1, \hat{\theta}_2)$ follows a bivariate normal distribution. The asymptotic $(1 - \alpha)\%$ confidence set for r solves $H(r)^2 \leq Z_{1-\alpha/2}^2$, where $H(r) = (\hat{\theta}_1 - r\hat{\theta}_2)/(\hat{\sigma}_{11} - 2r\hat{\sigma}_{12} + r^2\hat{\sigma}_{22})^{1/2}$, which could be a finite interval, a disjoint interval, or the real line.

When the bivariate normal assumption is not appropriate or in cases where σ_{11}, σ_{22}, and σ_{12} are not easy to compute, the bootstrap technique (Efron and Tibshirani, 1986) can be used to obtain the confidence interval. The bias-corrected (BC) percentile method is implemented here. Let $\hat{G}(s)$ be the c.d.f. of the bootstrapped statistics $\hat{r}^*$, $\Phi(\cdot)$ (the standard normal c.d.f.) and $\hat{r}$ (the sample estimate). $z_0 = \Phi^{-1}\{\hat{G}(\hat{r})\}$ and $r_{BC}[\alpha] = \hat{G}^{-1}(\Phi\{2z_0 + \Phi^{-1}(\alpha)\})$. Then the $(1 - \alpha)\%$ BC confidence interval for r will be $(r_{BC}[\alpha], r_{BC}[1 - \alpha])$.

The third approach is based on Hwang (1995), who suggested that, in order to have good coverage probabilities, one should bootstrap the pivot quantity, $H(\hat{r})$. Then the confidence set for the parameter of interest solves $l \leq H(r) \leq u$, where l and u are the lower and upper limit of the $(1-\alpha)\%$ confidence interval for $H(r)$ obtained from the BC percentile method. When the variances and covariance are not easy to calculate, a nested bootstrap computation may be needed to estimate σ_{11}^*, σ_{22}^*, and σ_{12}^*.

The delta method could also be used to derive the confidence intervals for F and F'. Freedman (2001) compared, through computer simulations, the properties of confidence intervals for P based on Fieller's Theorem with the delta method and found that Fieller's method is superior to the delta method. In this article, we do not use the delta method to construct confidence intervals.

5. Application to Ophthalmology Data

5.1 *Ophthalmology Data*

This dataset (Buyse and Molenberghs, 1998) is from a randomized clinical trial in ophthalmology studying the effects of interferon-α in patients with age-related muscular degeneration (ARMD). Patients in the treatment group received interferon-α and those in the control group received placebo. A patient's visual acuity was assessed through the ability to read lines on a vision chart. The primary endpoint of the trial was the proportion of patients who lost at least three lines of vision at 1 year. Buyse et al. used the loss of at least two lines of vision at 6 months as the surrogate.

In summary, S and T are defined as

$$S = \begin{cases} 0 & \text{if patient had lost less than} \\ & \text{two lines of vision at 6 months,} \\ 1 & \text{otherwise,} \end{cases}$$

$$T = \begin{cases} 0 & \text{if patient had lost less than} \\ & \text{three lines of vision at 1 year,} \\ 1 & \text{if otherwise,} \end{cases}$$

and the data are

Z	S	No. of $T = 0$	No. of $T = 1$	Total
0	0	56	9	65
0	1	8	30	38
1	0	31	9	40
1	1	9	38	47

5.2 *Estimation*

For this dataset, we choose $h(u) = u$ and $g_A(s)$, $g_B(s)$ to be the conditional probability $\Pr(T = 1 \mid S, Z)$. $P_A(s)$, $P_B(s)$, $g_A(s)$, and $g_B(s)$ are empirically estimated from the sample proportions. The data suggest that the use of interferon expedites the degradation process of vision for patients. The overall effect of Z on T and the effect of Z on S are in the same direction. $\hat{F}$ and $\hat{F}'$ estimate how much the negative effect of interferon on vision can be explained by S. $\widehat{AA} = 39/103 = 0.379$, $\widehat{BB} = 47/87 = 0.540$, $\widehat{AB} = (30/38)(47/87) + (9/65)(40/87) = 0.490$, and $\widehat{BA} = (38/47) \times (38/103) + (9/40)(65/103) = 0.440$, leading to $\hat{F} = 0.690$ and $\hat{F}' = 0.619$. To estimate P, model (1) and model (2) are both fitted to the data. The regression coefficient (and standard error) for the unadjusted treatment effect ($\hat{\beta}$) is 0.657 (±0.296) and for the adjusted treatment effect ($\hat{\beta}_S$) is 0.364 (±0.377), giving $\hat{P} = 0.445$. The point estimates for F, F', and P are all between zero and one, suggesting a partial surrogate.

Condition R1 will be satisfied if $P(S = 1 \mid Z = 1) > P(S = 1 \mid Z = 0)$. R2 will be satisfied if $P(T = 1 \mid S = 1, Z) \geq P(T = 1 \mid S = 0, Z), Z = 0, 1$. And R3 will be satisfied if $P(T = 1 \mid S, Z = 1) \geq P(T = 1 \mid S, Z = 0), S = 0, 1$. The data in the table suggest that R1, R2, and R3 are true.

Three methods, as described earlier, are used to construct the confidence intervals for F, F', and P. The estimated variances of $\hat{\beta}$ and $\hat{\beta}_S$ from fitted models (1) and (2) are used as $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$, and $\hat{\sigma}_{12} = \widehat{\text{corr}}(\hat{\beta}, \hat{\beta}_s)(\hat{\sigma}_{11}\hat{\sigma}_{22})^{1/2}$, where $\widehat{\text{corr}}(\hat{\beta}, \hat{\beta}_s)$ is obtained from bootstrap samples. In this example, the numerator and denominator of F and F' are functions of the multinomial probabilities, which are asymptotically normal. By the delta method, the variances and covariance for the numerator and denominator of $\hat{F}$ ($\hat{F}'$) can be computed. For the bootstrap methods, parametric bootstrapping is used in which 1000 samples are generated from multinomial distributions.

The resulting 95% confidence intervals for F, F', and P estimated from the ophthalmology data are shown in Table 1. Based on the result, the confidence intervals for F and F' are smaller than the corresponding intervals for P. In the first and second methods, there is at least a 30% reduction in the interval width of F (F') compared with P. Although the upper bounds of all the confidence intervals are above one, the lower bounds of the confidence intervals for F or F' are larger than zero, while the lower bounds for P are less than zero. The intervals obtained by the three different methods are fairly close to each other, which suggests the normal assumption for the numerator and denominator is reasonable. We also observe that there are some differences between the confidence intervals for F and F' in terms of region covered, with the intervals for F' shifted to smaller values.

For a small fraction of the bootstrap samples, the estimates in the denominator of F, F', or P are close to zero. Of the bootstrapped $\hat{F}$ ($\hat{F}'$), 1.2% (1%) are outside the range $[-4, 4]$ and 2.8% of the bootstrapped $\hat{P}$ are outside the range $[-4, 4]$. $\hat{F}$, $\hat{F}'$, and P that lie within $[-4, 4]$ are plotted against the unadjusted treatment effect β (as in model 2) (Figure 2). The plot also suggests that $\hat{P}$ has a larger variation and that $\hat{F}$ and $\hat{F}'$ are more likely than $\hat{P}$ to be concentrated within the range $[0, 1]$.

6. Simulation Study

Simulation studies were carried out to compare the behavior of F and F' to P in cases where both T and S are binary. Different sets of data, $\{Z = (0, 1), S = (0, 1), \text{and } T = (0, 1)\}$, are generated according to the following two logistic models:

$$\log \frac{p(T = 1 \mid S, Z)}{1 - p(T = 1 \mid S, Z)} = \beta_0 + \beta_1 S + \beta_2 Z + \beta_3 ZS \quad (3)$$

Table 1
95% Confidence intervals of F, F', and P for ophthalmology data

	Method 1, Fieller Theorem	Method 2, bootstrap directly	Method 3, bootstrap pivot
95% confidence interval for F	(0.17, 3.12)	(0.21, 3.11)	(0.19, 3.33)
95% confidence interval for F'	(0.02, 2.57)	(0.18, 2.87)	(0.07, 2.56)
95% confidence interval for P	(−0.30, 4.35)	(−0.31, 4.25)	(−0.30, 3.04)

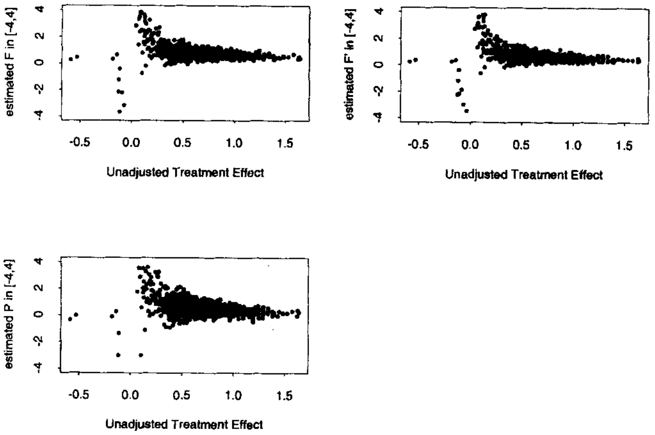

Figure 2. Plots of $\hat{F}$, $\hat{F}'$, and $\hat{P}$ from bootstrap samples for ophthalmology data against the unadjusted treatment effect. Only the estimated values that are between $[-4, 4]$ are plotted.

$$\log \frac{p(S = 1 \mid Z)}{1 - p(S = 1 \mid Z)} = \alpha_0 + \alpha_1 Z. \qquad (4)$$

Datasets are created for three scenarios: perfect, useless, and partial surrogates. In each scenario, we simulate 500 datasets and each dataset has a total of 200 subjects randomized to the treatment group or the placebo group with equal probabilities. The parameter values for each scenario are listed in Table 2. In all simulated cases, conditions R1–R3 are all satisfied.

We study the bias, variability, and coverage rate of 95% confidence intervals. The true value of F (F') can be calculated algebraically based on (3) and (4). For P, a large dataset ($n = 2000$) is simulated and then logistic models (1) and (2) are fitted on this large dataset to get the estimate for the true value of P. Ninety-five percent confidence intervals for F, F', and P are calculated for each simulated dataset based on Fieller's Theorem as described earlier.

We show in Table 3 for each scenario the true value of F, F', and P, the lower and upper 2.5% percentiles and median of the estimates from 500 datasets, and coverage rate for 95% confidence intervals. The number of times (out of 500) that the 95% confidence intervals for F, F', and P lie within $[0, 1]$, have lower bounds no smaller than zero, or have upper bounds no larger than one are shown in Table 4.

The results show that the bias for all three estimates are close to zero. The coverage rates of the 95% confidence intervals based on Fieller's Theorem are close to the nominal level. Overall, $\hat{F}$ and $\hat{F}'$ are less variable than $\hat{P}$. The ranges of 2.5% and 97.5% percentiles for $\hat{F}$ and $\hat{F}'$ are mostly smaller and are about 30–85% of the ranges for $\hat{P}$. In the partial surrogate scenarios, the 2.5% percentiles of $\hat{F}$ and $\hat{F}'$ are mostly greater than 0.2, while the 2.5% percentiles of $\hat{P}$ are mostly below or near zero. Table 4 shows that 95% confidence intervals for F and F' are, on average, 11 times more likely to fall between $[0, 1]$ than P. In the partial surrogate cases, the number of times that the lower bounds for F and F' are above zero are 1.5 to 5.5 times those for P. These results suggest that the variabilities of F and F' are smaller than P. In the partial scenarios, which are most likely to happen in practice, the lower bounds of the confidence intervals for F and F' are much more likely to be above zero than are the confidence intervals for P, which indicates that F and F' are more useful as measures for surrogacy.

Table 2
Parameter settings for simulations

Scenario	cases	α_1	β_1	β_2	β_3
Perfect	1	1.0	3.0	0	0
	2	1.5	3.0	0	0
	3	1.0	4.0	0	0
Useless	1	0	3.0	0.8	0
	2	0	1.5	0.8	0
	3	0.5	0	0.8	0
Partial	1	1.0	6.0	0.3	0
	2	1.0	3.0	0.45	0
	3	1.0	2.0	0.45	0
	4	0.5	2.0	0.68	0
	5	1.0	2.3	0.68	0.3
	6	1.0	2.3	0.68	−0.3
	7	1.0	2.3	0.68	0.68
	8	1.0	2.3	0.68	−0.68

Although the variability of F (F') is smaller than that of P, the results suggest that the confidence intervals for F (F') are still wide. The majority of the intervals for F and F' extend beyond zero and/or one, especially the latter, as shown in Table 4. As discussed earlier, F (F') are not true proportions with estimates bounded within $[0, 1]$; thus, it is plausible for the values of F and F' to be outside $[0, 1]$. Table 3 shows that the 2.5% and 97.5% percentiles for $\hat{F}$ and $\hat{F}'$ from the 500 simulated datasets can go outside $[0, 1]$, especially on the upper sides. It is not surprising that the confidence intervals can also include values less than zero and greater than one. Freedman (2001) shows that the lengths of the confidence intervals for P using Fieller's Theorem decrease as the unadjusted treatment effect becomes more significant. For adequate statistical power to validate surrogate endpoints using P, the unadjusted effect would need to be five or six times its standard error. It is likely that a stronger treatment effect will also result in shorter confidence intervals for F and F'.

In the partial surrogacy scenarios, as expected, increasing the treatment effect on S and decreasing the treatment effect on T adjusting for S results in larger values for F, F', and P. We observe that $\hat{P}$ has consistently lower values than $\hat{F}$ or $\hat{F}'$ inside the interval $(0, 1)$, which could be due to the different metrics used in estimating PE.

We also observe that, in the useless surrogate scenarios (cases 1 and 2), where treatment has no effect on the surrogate marker ($\alpha = 0$), the large-sample limit for P is negative while the large-sample limit for F and F' are zero. This suggests that P is not measuring PE appropriately in these situations.

7. Discussion

Appropriate statistical validation of the surrogacy for a biomarker is important. Prentice (1989) suggests a strict valida-

Table 3

Results from simulation studies for F, F', and P. Shown are the large-sample limit, median of the estimates, upper and lower 2.5% percentiles, and coverage rate of confidence interval based on Fieller's Theorem.

		True value			Median			$Q_{(2.5\%,97.5\%)}$			Coverage		
Scenario	Cases	F	F'	P	F	F'	P	F	F'	P	F	F'	P
Perfect	1	1	1	1	0.98	0.97	0.96	(0.45, 3.12)	(0.47, 3.14)	(0.15, 4.86)	0.94	0.95	0.96
	2	1	1	1	1.01	0.98	1.00	(0.58, 2.04)	(0.58, 1.99)	(0.44, 2.68)	0.95	0.95	0.96
	3	1	1	1	0.99	0.98	0.94	(0.58, 1.93)	(0.58, 2.03)	(0.13, 3.48)	0.95	0.95	0.96
Useless	1	0.00	0.00	−0.56	0.00	0.00	−0.54	(−1.84, 1.28)	(−1.61, 1.23)	(−3.42, 1.31)	0.95	0.98	0.96
	2	0.00	0.00	−0.14	0.00	0.00	−0.13	(−0.78, 0.33)	(−0.85, 0.35)	(−1.16, 0.20)	0.98	0.99	0.94
	3	0.00	0.00	0.00	0.00	0.00	−0.00	(−0.24, 0.24)	(−0.19, 0.27)	(−0.17, 0.18)	0.99	1.0	0.94
Partial	1	0.91	0.85	0.67	0.90	0.84	0.66	(0.64, 1.36)	(0.50, 1.49)	(−0.24, 2.33)	0.95	0.96	0.95
	2	0.71	0.68	0.50	0.69	0.68	0.47	(0.36, 1.32)	(0.34, 1.34)	(−0.03, 1.60)	0.94	0.96	0.95
	3	0.54	0.57	0.45	0.54	0.56	0.44	(0.20, 1.52)	(0.24, 1.63)	(0.06, 1.56)	0.95	0.96	0.95
	4	0.29	0.30	0.11	0.28	0.28	0.10	(−0.10, 0.78)	(−0.11, 0.86)	(−0.38, 0.79)	0.93	0.98	0.95
	5	0.45	0.49	0.27	0.45	0.48	0.25	(0.21, 0.82)	(0.23, 0.94)	(−0.08, 0.83)	0.95	0.97	0.96
	6	0.57	0.52	0.41	0.57	0.50	0.39	(0.26, 1.32)	(0.23, 1.25)	(0.06, 1.38)	0.95	0.97	0.96
	7	0.42	0.48	0.20	0.41	0.48	0.19	(0.20, 0.73)	(0.23, 0.87)	(−0.15, 0.64)	0.96	0.97	0.96
	8	0.73	0.55	0.55	0.71	0.55	0.54	(0.27, 2.61)	(0.21, 1.92)	(0.12, 3.05)	0.95	0.96	0.96

tion criterion, which requires $P(T \mid S, Z)$ to be equal to $P(T \mid S)$. Freedman et al. (1992) raise the concept of PE in surrogacy validation and propose a quantitative measure of the role a surrogate marker plays in the therapeutic pathways of a treatment. Both of these two approaches focus on the conditional distribution $[T \mid S, Z]$ in the assessment for surrogate endpoints. Daniels and Hughes (1997) and Buyse et al. (2000) develop a different concept. Let θ_T denote the treatment effect on the primary endpoint and θ_S the treatment effect on the surrogate marker. θ_T is based on $[T \mid Z]$ and is θ_S based on $[S \mid Z]$. In their approaches, the main concerns are the relationship between θ_T and θ_S, the prediction of θ_T based on θ_S, and the precision of the prediction. The concept of assessing surrogacy by precision of predicting θ_T based on the relationship between θ_T and θ_S is a useful one. However, in this approach, the therapeutic pathways through which treatment takes effect is not a major concern as long as the statistical association between θ_T and θ_S is strong. Also, this type of approach is appropriate only if there are multiple trials available on the same or similar treatments so that the relationship between θ_T and θ_S can be studied.

In this article, we focus on the situation of a single trial and the validation of biomarkers by estimating PE. The underlying motivation for PE comes from thinking of the therapeutic

Table 4

95% Confidence intervals based on Fieller's Theorem are constructed for each scenario. Shown are the numbers of times (out of 500) that the confidence intervals lie between $[0, 1]$, that lower bounds are greater than or equal to zero, and that upper bounds are less than or equal to one for F, F', and P.

		CIs within $[0, 1]$			CIs with lower bound ≥ 0			CIs with upper bound ≤ 1		
Scenario	Cases	F	F'	P	F	F'	P	F	F'	P
Perfect	1	14	8	4	299	293	207	16	10	17
	2	12	11	12	453	451	439	12	11	15
	3	9	10	1	375	369	223	10	11	15
Useless	1	3	0	0	8	2	1	224	216	221
	2	0	0	0	0	0	0	328	328	326
	3	0	0	4	0	0	8	349	344	351
Partial	1	33	37	1	430	419	134	34	42	46
	2	85	86	20	428	402	195	96	102	119
	3	101	73	41	376	323	243	128	119	128
	4	74	35	2	134	78	25	275	265	281
	5	293	206	55	443	389	152	332	285	334
	6	114	103	67	409	318	244	135	181	185
	7	360	265	52	447	408	109	404	343	397
	8	42	43	24	330	226	233	49	104	73

pathway. In particular, it is an attempt to construct a scalar summary that captures the concept that the measure should be one if all of the effect of Z on T is through S, zero if either there is no effect of Z on S or S is not associated with T, and between zero and one if part of the effect of Z on T is through S. PE is a useful concept for surrogacy validation and F (F') are proposed as statistical measures for it.

The measure Freedman et al. (1992) proposed has been shown to be problematic. In this article, we propose alternative measures F and F' and have compared and contrasted them with P. The new measures F and F' are estimated based on the distributions $[S \mid Z]$ and $[T \mid Z, S]$, which we think is a logical approach because of the temporal relationship between the marker and the primary outcome. F is based on a factorization of the joint distribution $[T, S \mid Z]$ into $[S \mid Z]$ and $[T \mid S, Z]$, whereas the measure P is based on consideration of $[T \mid S, Z]$ and $[T \mid Z]$, which together do not necessarily specify the joint distribution $[T, S \mid Z]$. It also seems natural for a measure of surrogacy to depend directly on how treatment affects the marker. In addition, unlike P, estimation for F is not tied to linear models and can be estimated with more flexibility and fewer assumptions. We think these differences will give F and F' better properties and allow for generalizations. In the case of binary T, our results from the ophthalmology data and the simulation studies suggest that F is less variable than P. In the partial surrogate scenarios, which are most likely to happen in practice, the lower confidence bounds for F and F' are more likely to be greater than zero than the confidence bounds for P and hence suggests that the new measures are more useful for surrogacy validation.

In cases with repeated measurement of markers, it is often the case that there are dependent censoring or dropouts during the trial. Fitting model $[T \mid Z]$, as required for P, will yield a biased estimate for the treatment effect. However, for F (F'), by joint modeling of $[T \mid S, Z]$ and $[S \mid Z]$, we may reduce the bias in the estimates of treatment effect.

We have proposed two complementary forms of the measure of PE, i.e., F and F'. An alternative measure based on these two, $(F + F')/2$, can be used. Although population quantities F and F' do not equal each other in all situations, ideally, they would be close to each other. It would be interesting to further investigate the meaning of the difference between F and F' in different settings and the properties of their averages.

A drawback for both F (F') and P is in the interpretation of the estimated values. While values of zero and one have clear interpretations, it is not so easy to understand the exact meaning of an intermediate value. The relative size of F for two different biomarkers in a trial does give an indication of their relative usefulness as a surrogate. The graphical interpretation in Figure 1 is also a useful aid to the interpretation. We recommend the approach suggested by Freedman et al. (1992) and judge a potential surrogate by whether the lower bound of a confidence interval for F is above a certain value.

Although the variability of F (F') is smaller than P, the results suggest that the confidence intervals for F (F') are still wide. The relatively large variability of F (F'), like P, is most likely due to its inherent limitation as a ratio of estimates. If the denominator is not clearly different from zero, then calculation of the ratio is problematic. It is likely that, for F (F') to be useful in validating a surrogate endpoint, a strong treatment effect of Z on T would be needed.

In this article, we have focused mainly on the situations of a binary endpoint and a binary surrogate. For future research, extensions of F (F') to more complex situations will be required. For example, multiple biomarkers or repeated measurements of biomarkers can be considered as the surrogate endpoint. Meta-analysis that combine data across studies may also be used to increase power and also to examine the consistency of the underlying proportion F (F') across similar studies.

Acknowledgements

This work was partially supported by the Early Disease Research Network grant CA86400 from the National Cancer Institute. The authors thank the editor, the associate editor, and the referees for their helpful suggestions.

Résumé

Les essais randomisés impliquant des critères principaux dont l'occurrence est rare, ou dont la mesure est pénalisée par des temps de survenue importants, s'avèrent onéreux. C'est la raison de l'intérêt croissant porté aux méthodes consistant à remplacer le véritable critère clinique par un critère de substitution, disponible plus précocement. Cependant, ces critères de substitution doivent être correctement validés. A cet égard, une mesure quantitative—spécifique à chaque essai—de la proportion de l'effet traitement expliquée par le critère de substitution s'avère un concept utile. Freedman et al. (1992) ont ainsi suggéré de mesurer la qualité de la substitution en calculant le rapport des coefficients de régression associés à l'effet traitement dans deux modèles séparés, ajustés ou non par le critère de substitution. Cette mesure se révèle hélas extrêmement variable, sans compter qu'il n'y a aucune garantie que chacun des deux modèles soit approprié. En nous inspirant d'une idée formulée par Tsiatis et al. (1995), nous proposons, pour le calcul de cette proportion expliquée, des méthodes alternatives qui requièrent moins d'hypothèses sous-jacentes au niveau de l'estimation et permettent davantage de flexibilité au niveau de la modélisation. A partir de données issues d'un essai en ophtalmologie, et en utilisant également un certain nombre de simulations, nous comparons les estimations obtenues à l'aide de ces nouvelles mesures, lesquelles semblent de fait présenter moins de variabilité.

References

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

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

Bycott, P. W. and Taylor, J. M. G. (1998). An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. *Controlled Clinical Trials* **19,** 555–568.

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

DeGruttola, V., Fleming, T. R., Lin, D. Y., and Coombs, R. (1996). Validating surrogate markers: Are we being naive? *Journal of Infectious Disease* **175,** 237–246.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1,** 54–77.

Ellenberg, S. S. and Hamilton, J. M. (1989). Surrogate endpoints in clinical trials: Cancer. *Statistics in Medicine* **8,** 405–413.

Fleming, T. R. (1992). Evaluating therapeutic interventions: Some issues and experiences. *Statistical Science* **7,** 428–456.

Fleming, T. R. and DeMets, D. L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine* **125,** 605–613.

Freedman, L. S. (2001). Confidence intervals and statistical power of the "validation" ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference* **96,** 143–153.

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

Hwang, G. J. T. (1995). Fieller's problems and resampling techniques. *Statistica Sinica* **5,** 161–171.

Lagakos, S. W. and Hoth, D. F. (1992). Surrogate markers in AIDS: Where are we? Where are we going? *Annals of Internal Medicine* **116,** 599–601.

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

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

Schatzkin, A., Freedman, L. S., Schiffman, M. H., and Dawsey, S. M. (1990). Validation of intermediate end points in cancer research. *Journal of the National Cancer Institute* **82,** 1746–1752.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (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.

Received February 2001. *Revised April* 2002.
Accepted April 2002.