Published in final edited form as:

Stat Biopharm Res. 2010 May 1; 2(2): 229–238. doi:10.1198/sbr.2009.0070.

Evaluating the Proportion of Treatment Effect Explained by a Continuous Surrogate Marker in Logistic or Probit Regression Models

Jie Huang [Associate Director] and

Novartis Pharmaceuticals, Oncology Business Unit, East Hanover, NJ 07936
(jie.huang@novartis.com).

Bin Huang [Associate Professor]

Center for Epidemiology and Biostatistics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229 (bin.huang@cchmc.org).

Abstract

Using surrogate endpoints in clinical trials is desirable for drug development because the trials can be shortened and therefore more cost-effective. Validating a surrogate for the clinical endpoint is critical in this context. One of the key steps in statistical validation of a surrogate for a single trial is to estimate the proportion of treatment effect explained (PTE or PE) by a surrogate. Often the measure for PTE is estimated from the difference in coefficients of treatment from two models with or without adjusting for the surrogate for clinical endpoint. Inherent problems with the method are: the two models may not be valid simultaneously; and the estimate can often lie outside the interval $[0, 1]$. In this article, we provide alternative measures for evaluating the proportion of treatment effect explained by a surrogate in logistic or probit regression models. Our measures can be estimated easily with any statistical programs capable of binary linear regression modeling, and the interpretation of the measures can be illustrated using Ordinal Dominance (OD) curves. The concept can be visually understood by any practical user. Simulation shows our alternative measures yield more accurate estimates which are less biased, less variable, and with narrower confidence intervals. A clinical trial example is provided.

Keywords

Biomarker; Logistic; Mediator; Probit model; Ordinal dominance curve; Surrogate validation

1. Introduction

In clinical trials, study endpoints such as cancer survival or cardiovascular event require a prolonged follow-up time. The trials can be costly; it may be difficult to enroll patients, and even more difficult to follow and monitor patients. However, a surrogate endpoint can usually be observed early during a trial, and can thus be used as an attractive substitute for the clinical endpoint in studying treatment effect. Using surrogate endpoints, trials may be shortened, and it may be possible to avoid fatal clinical endpoints before drug approval. It makes efficacious treatment available to patients sooner, saves patients' lives, and reduces medical expenditure. However, a critical question must be answered beforehand: What may serve as a valid surrogate endpoint? Both the U.S. regulatory agency Food and Drug Administration (FDA) (Temple 1995) and the European Medicines Agency (EMA) (EMA/CHMP report 2007) have recognized the importance of developing and validating surrogate endpoints. The validation of biomarkers as surrogate endpoints is part of FDA's "critical path initiative." Workshops and

meetings were organized among the regulatory agencies, industry representatives, and the academics to discuss and set out position statements on surrogate endpoints (An NIDAID Workshop 1989; Biomarkers Definitions Working Group 2001; DeGruttola et al. 2001). In recent years, statistical validation of surrogate endpoints has attracted increased attention.

In his landmark article, Prentice (1989) defined a surrogate as “a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.” He defined three criteria for admitting a valid surrogate. One of the criteria requires that “the full effect of treatment on the true endpoint is captured by the surrogate, that is, $f(T|S, Z) = f(T|S)$.” This condition is rather stringent and unlikely to be true in practice. To validate this equation, Freedman, Graubard, and Schatzkin (1992) proposed a statistic to measure the proportion of the treatment effect explained by the surrogate (hereafter $PE^{(FGS)}$). This statistic is defined as the percentage change of treatment effects estimated from two models with or without adjusting for the surrogate marker. It is well known that $PE^{(FGS)}$ suffers from several serious drawbacks (DeGruttola et al. 1997; Bycott and Taylor 1998; Wang and Taylor 2002). In particular, either the point estimate or its confidence interval estimate can be out of the range $[0, 1]$. As a result, it often fails to provide a meaningful assessment of “the proportion of treatment effect explained by a surrogate marker.” Subsequently, Wang and Taylor (2002) provided the more desirable alternative measures $F(F')$ for PTE, which are less variable and have its point estimate within $[0, 1]$ under certain conditions. When surrogate markers and outcome endpoint are either all continuous or binary, $F(F')$ can be calculated easily. For the continuous surrogate marker and binary outcome, Wang and Taylor (2002) suggested two approaches to derive the measures, but one is computationally difficult, and the other is not easily interpretable. Evaluating a continuous surrogate for a binary endpoint is often encountered in clinical studies. For examples, blood pressure level has been accepted by clinical and regulatory as a surrogate for the incidence of stroke and congestive heart failure (Biomarkers Definitions Working Group 2001; DeGruttola et al. 2001). In an oncology imaging study, the use of post-therapy FDG-PET as a metabolic surrogate marker of tumor response in cervical cancer was prospectively validated (Schwarz et al. 2007). In this article, we provide alternative measures to evaluate the proportion of treatment effect explained by a surrogate in the logistic and probit regression.

The remainder of this article is organized as follows. Section 2 introduces the definition of PTE in general. Section 3 provides our simple alternative PTE measures for the case of continuous surrogate marker and binary outcome. Our measures can be obtained easily with any commonly used statistical computation package capable of binary linear regression modeling. Simulation results are shown in Section 4 for comparing our measures to $PE^{(FGS)}$. Results of our measures and $PE^{(FGS)}$ are compared to each other; both are compared to the estimated true PTE from the continuous latent outcome of the binary outcome. Interpretations of our measures are provided using Ordinal Dominance (OD) curves (Bamber 1975); a concept that can be easily understood by practical users. In Section 5, a clinical example is illustrated for the method. Discussions and conclusions are provided in Section 6.

2. Measure of the Proportion of Treatment Effect Explained by the Surrogate

Freedman, Graubard, and Schatzkin (1992) defined the proportional of treatment effect explained by the surrogate ($PE^{(FGS)}$) as the percentage change of the coefficient of the treatment effect from two models, that is, with or without adjusting for a surrogate in the model. Their article focused on the binary outcome in the logistic model. The method provides an intuitive concept, but the two logistic models may not be valid at the same time. Estimation of $PE^{(FGS)}$ is shown to be quite variable; the point estimate and its confidence interval can be outside of interval $[0, 1]$ (Freedman 2001). When the endpoint is the time to event, Lin, Fleming, and DeGruttola (1997) defined PTE based on the two Cox proportional hazard models with or

without adjusting for the surrogate in the model. The two models can have different baseline hazards. Although both models may be approximated by valid Cox PH models, the confidence interval is again very wide.

Considering estimating PTE in the generalized linear model setting, the outcome observation (t_i) is assumed to be from a natural exponential family with density function:

$$f(t_i; \theta_i, \phi) = \exp \left[\frac{t_i \theta_i - b(\theta_i)}{a(\phi)} + c(t_i, \phi) \right].$$

The outcome (t_i) relates to treatment (z_i), and the surrogate (s_i) by a generalized model $[T|Z, S]$,

$$E(T_i | z_i, s_i) = h^{-1}(\beta_0 + \beta_1 s_i + \beta_2 z_i).$$

$h(\cdot)$ is the link function for the generalized linear model. If this full model holds, the marginal model of t_i , conditional on treatment only, would be obtained by integrating over the distribution function of $f(s|z_i)$. The marginal model $[T|Z]$ is

$$f(t_i | z_i) = \int f(t_i | z_i, s) f(s | z_i) ds.$$

If both $f(t_i | z_i, s_i)$ and $f(s_i | z_i)$ are normal density functions, the marginal model $[T|Z]$ can be obtained easily and follows normal distribution as of $[T|Z, S]$. In general, the marginal model may not be analytically integrable; $f(t_i | z_i, s_i)$ and $f(s_i | z_i)$ are different types of density functions. For example, if $h(\cdot)$ is a logit link function, numerical approximation is necessary to derive the marginal model $[T|Z]$.

In practice, a reduced model is fitted for outcome (t_i) and treatment (z_i) by using the same link function of the full model by omitting the surrogate variable as

$$E(T_i | z_i, s_i) = h^{-1}(\gamma_0 + \gamma_1 z_i).$$

The $PE^{(FGS)}$ is estimated by

$$PE^{(FGS)} = 1 - \hat{\beta}_2 / \hat{\gamma}_1.$$

This definition of $PE^{(FGS)}$ is intuitive and easy to calculate. There are inherited problems; first, the full model and the marginal model may not be valid simultaneously except in special cases; second, the $PE^{(FGS)}$ estimate can be extremely variable, and its point estimate and confidence interval can lie outside the interval $[0,1]$. For these reasons, seeking more reliable statistical measures of PTE would be necessary.

3. Alternative Measure of the Proportion of Treatment Effect Explained by the Surrogate: A Better Approach

3.1 Definition of $F(F')$ as Alternative Measures of PTE

Motivated by the idea of Tsiatis, DeGruttola, and Wulfsohn (1995), Wang and Taylor (2002) defined alternative measures (F and F') for the proportion of treatment effect explained by a surrogate. The measures are defined by two ratios, that is,

$$F = \frac{AA - AB}{AA - BB}$$

and

$$F' = \frac{BA - BB}{AA - BB}.$$

Terms AA , BB , AB , and BA are defined as

$$\begin{aligned} AA &= \tilde{h} \left(\int g_A(s) dP_A(s) \right), \\ AB &= \tilde{h} \left(\int g_A(s) dP_B(s) \right), \\ BA &= \tilde{h} \left(\int g_B(s) dP_A(s) \right), \end{aligned}$$

and

$$BB = \tilde{h} \left(\int g_B(s) dP_B(s) \right), \quad (1)$$

where $P_A(s)$ and $P_B(s)$ denote the probability distribution function of surrogate (S) in treatment A group and treatment B group (treatment indicator $Z = 0$ for A or $Z = 1$ for B), respectively; $g_A(s)$ and $g_B(s)$ are functions of the conditional distributions of outcome $[T|s]$ in two treatment groups, and $\tilde{h}(\cdot)$ is a monotone link function. Citation $\{AA, AB, BA, BB\}$ are four types of potential outcomes corresponding to the two treatment groups and surrogate distribution in two treatment groups. In other words, if one could manipulate the experiment, so that the distribution of the surrogate marker can be randomly assigned to different treatment conditions, we should observe four types of outcomes, labeled AA , AB , BA , BB . However, since such experimental manipulation is not possible in reality, we could only observe AA and BB corresponding to the expected outcome under treatment A and B. In order to evaluate the effect of the surrogate marker, the potential outcome of AB and BA are introduced. $(AA - BB)$ measures the overall treatment difference. While $(AA - BB) = (AA - AB) + (AB - BB)$, $(AA - AB)$ measures the change of the probability of outcome ($T = 1$) for patients in treatment A group, had their surrogate distribution changed to $P_B(S)$ from $P_A(S)$. $(AB - BB)$ measures the change in the probability of outcome ($T = 1$) due to the change of treatment assuming surrogate follows the distribution in treatment B. Thus F measures the proportion of the overall treatment difference that is attributable to the change of distribution of surrogate or “the overall treatment difference explained by a surrogate.” F' measure is similarly defined, as a mathematically symmetric measure to F . Similar interpretation holds for F' and $(BA - BB)$ while $(AA - BB) = (AA - BA) + (BA - BB)$. Wang and Taylor (2002) conducted research and simulation for cases when outcome and surrogate are either continuous or binary variables. When both T and S are normally distributed continuous variables, $F = F' = \text{PE}^{(\text{FGS})}$. Although this may not be the case in the nonnormal setting, a simulation study from Wang and Taylor’s work and our work in this article shows F and F' are very close in value. The average of the F and F' has thus been suggested. In the case of binary outcome and continuous surrogate marker, Wang and Taylor suggested two approaches. One is to define the treatment effect on the probability scale, that is, assume $\text{logit}(g_Z(S))$ is a linear model for Z and S , which leads to a messy expression for $F(F')$; another is to model $g_Z(S)$ by a linear function, which makes the interpretation of $F(F')$

difficult. Neither approach provides good solutions for the practical user. However, if we use the probit model for $[T|S, Z]$ to define treatment effect and probit function for $P_Z(s)$, we are able to derive the analytical solutions for F and F^* . When the logistic model is desired for $[T|S, Z]$, F and F^* can be derived by approximating logistic function with probit function. For the remainder of the article, we examine the measures F (F^*) and provide an easy solution for the setting of logistic and probit regression models of binary clinical outcome. Interpretations of our measures are provided by graphics.

3.2 Estimation of $F(F^*)$ for Binary Outcome in Randomized Clinical Study

In the setting of a randomized clinical trial, T is a binary outcome, Z is a binary indicator of treatment, and S is a continuous surrogate marker. Without loss of generality, we assume that an increased surrogate will more likely result in an event ($T=1$ for event and $T=0$ for nonevent). Let us assume $P_A(S)$ and $P_B(S)$ follow the normal distribution functions (Φ denotes for normal distribution function)

$$\begin{aligned} P_A(s) &= \Phi\left(\frac{s-\mu_A}{\sigma_A}\right), \\ P_B(s) &= \Phi\left(\frac{s-\mu_B}{\sigma_B}\right). \end{aligned} \quad (2)$$

Let $g_A(S)$ and $g_B(S)$ be the mean of the conditional distribution of $[T|S, Z=0]$ and $[T|S, Z=1]$, that is, $g_A(S) = \Pr(T=1|S, Z=0)$ and $g_B(S) = \Pr(T=1|S, Z=1)$. More specifically, let the conditional probability of the outcome (T) given S take the probit functional forms

$$\begin{aligned} g_A(s) &= \Pr(T=1|s, Z=0) = \Phi(\beta_0 + \beta_1 s) \\ g_B(s) &= \Pr(T=1|s, Z=1) = \Phi(\beta_0 + \beta_1 s - \omega). \end{aligned} \quad (3)$$

The $\omega (= -\beta_2)$ represents the effect of treatment on the outcome. Setting ω as nonnegative will ensure reduced odds of event ($T=1$) for treatment versus placebo given $S=s$. In the expression for $F(F^*)$, $(AA - AB)$ measures the change in the probability $[T=1|Z=A]$ for patients in treatment A group, had their surrogate distribution followed $P_B(S)$ instead of $P_A(S)$. Similar interpretation holds for $(BA - BB)$. Both F and F^* measure the treatment effect that is due to the change in the distributions of surrogate S . From the definition of (2) and (3), it can be shown that

$$\begin{aligned} AA &= \Phi\left(\frac{\beta_0/\beta_1 + \mu_A}{\sqrt{1/\beta_1^2 + \sigma_A^2}}\right), \\ BB &= \Phi\left(\frac{(\beta_0 - \omega)/\beta_1 + \mu_B}{\sqrt{1/\beta_1^2 + \sigma_B^2}}\right), \\ AB &= \Phi\left(\frac{\beta_0/\beta_1 + \mu_B}{\sqrt{1/\beta_1^2 + \sigma_B^2}}\right), \\ BA &= \Phi\left(\frac{(\beta_0 - \omega)/\beta_1 + \mu_A}{\sqrt{1/\beta_1^2 + \sigma_A^2}}\right). \end{aligned} \quad (4)$$

Each of the parameters in (4) can be estimated from fitted models (2) and (3). $\hat{F}$ ($\hat{F}^*$) can thus be calculated easily. The choice of $\tilde{g}(\cdot)$ and $\tilde{h}(\cdot)$ may vary. In the setting of interest, we use identity function for $\tilde{h}(\cdot)$ and mean function for $\tilde{g}(\cdot)$. When the logit function is desired for the conditional probability $[T|S]$, the analytical solution to the integrations for each of the terms in (4) is not available. However, one may use the good approximation of

$$\frac{\exp(y)}{1+\exp(y)} \approx \Phi\left(\frac{\sqrt{3}}{\pi}y\right)$$

to obtain the estimates of AA , BB , AB , and BA .

In the special case $\omega = 0$, we have $AB - BB = 0$ and $AA - BA = 0$, thus $F = F^* = 1$. If $\mu_A = \mu_B$ and $\sigma_A = \sigma_B$, that is, the distributions for S are the same in both treatment groups, then $AA = AB \Rightarrow F = 0$ and $BA = BB \Rightarrow F^* = 0$. From the definitions of AA , BB , AB , and BA , it is easy to see that each term represents a probability value confined within $[0, 1]$. Without loss of generality, it is assumed that $AA \geq BB$. In order to ensure that F and F^* value are bounded within $[0, 1]$, the sufficient and necessary conditions are: $AA \geq AB \geq BB$ and $AA \geq BA \geq BB$. Only the sufficient conditions are of practical interest, they are (Wang and Taylor 2002):

R1: $P_A(\cdot)$ is stochastically higher than $P_B(\cdot)$;

R2: $g_A(s)$ and $g_B(s)$ are nondecreasing function of s ;

R3: $g_A(s) \geq g_B(s)$ for all s .

Such conditions are generally met if S is a surrogate endpoint. In particular, for the setting of interest, if the variances of S are equal for both treatment groups, then all three conditions are met.

Unlike $PE^{(FGS)}$, definition of $F(F^*)$ measures is not based on the model assumptions. It can be applied for any model as long as the calculation of each term is feasible.

3.3 Graphical Presentation of $F(F^*)$ for the Continuous Surrogate

By definition each term of (AA, BB, AB, BA) corresponds to a probability term which is defined by two distribution functions $F(\cdot)$ and $G(\cdot)$; that is, $\int F(x)dG(y)$. Thus, each term could be presented by an area under an ordinal dominance (OD) curve (Bamber 1975) connecting $(0, 0)$ and $(1, 1)$ in a two-dimensional probability space. Figure 1 illustrates a clinical example that will be presented in detail later in Section 4. OD curves are plotted for AA (solid line), BB (long dashed), AB (dotdashed), and BA (dotted dash). The areas under each curve are the values for AA , BB , AB , and BA . Thus the area bounded by AA and BB curves corresponds to the $(AA-BB)$ value; that is, the overall treatment effect. Further, the area of $(AA-BB)$ can be divided into two mutually exclusive areas of $(AA-AB)$ and $(AB-BB)$. Therefore the F statistic could be represented by the ratio between two areas, the area defined by the solid line and short dashed lines $(AA-AB)$ versus the area defined by the solid and long dashed lines $(AA-BB)$. Similarly, F^* is represented by the ratio between the areas defined by $(BA-BB)$ and $(AA-BB)$. This graphic presentation is easy to understand, and directly corresponds to the definitions of F and F^* statistics. When the curves of AA and BB overlap, it suggests no treatment effect; the larger the $(AA-BB)$ area is, the stronger the treatment effect is. On the other hand, the closer together the AB and BB curves (or BA and AA curves) are, the higher the percentage of the effect of treatment is explained by the surrogate. When AA overlaps AB or BA overlaps BB , $F(F^*) = 0$, suggesting the surrogate is useless. When AA overlaps BA or AB overlaps BB , $F(F^*) = 1$, suggesting the surrogate is perfect.

4. Simulation Study

A simulation study is conducted for various scenarios (Table 1) where PTE varies from useless to perfect (0%, 16.67%, 33.33%, 66.67%, 80%, and 100%). The purpose of the simulation is to compare the performance of $F(F^*)$ measures to $PE^{(FGS)}$ measure in the probit model. For the specific setting of the problem (i.e., binary outcome and treatment with continuous

surrogate), $PE^{(FGS)}$ and $F(F')$ by definition are not the same. However, if the outcome endpoint is a continuous one (T_c), and the relationships between treatment (Z), surrogate (S), and outcome (T_c) are linearly related by $S = \alpha_0 + \alpha_1 Z + e$ and $T_c = \beta_0 + \beta_1 S + \beta_2 Z + \varepsilon$ (e and ε are iid normal). Measures of $PE^{(FGS)}$ and $F(F')$ are all the same under the assumption that both

T_c and S are normally distributed, that is, $PE^{(FGS)} = F = F' = \frac{\alpha_1 \beta_1}{\alpha_1 \beta_1 + \beta_2}$ (Wang and Taylor 2002). We refer to this value as the true latent PTE. A binary random variable T is derived from the latent continuous measure T_c , that is, T is a random variable from a Bernoulli distribution with probability $\Phi(t_c)$. Simulation results are compared to this true latent PTE to assess the performances of $PE^{(FGS)}$ and $F(F')$. Unless the surrogate is perfect or useless, F and F' usually differ slightly; thus we also include the estimate $FM (= (F + F')/2)$. Theoretical expected values of F , F' , FM , and latent PTE are calculated and presented in Table 1.

The sample sizes are set to be 500 for each treatment group. For each scenario, simulation results are obtained from 1000 replicates, and are summarized in Table 2. The bootstrap estimates are obtained based on 1000 bootstrap samples. Since the distributional form for the F and F' statistics are unknown, following Wang and Taylor's (2002) suggestion, bias-corrected (BC) Bootstrap (Efron and Tibshirani 1986) method is used for constructing the confidence intervals of F and F' . The two following models are used for the simulation.

$$S = \alpha_0 + \alpha_1 Z + e; \quad \text{Model 1}$$

$$\Phi^{-1}(P(T=1)) = \beta_0 + \beta_1 S + \beta_2 Z. \quad \text{Model 2}$$

To calculate PTE, the marginal model below is estimated,

$$\Phi^{-1}(P(T=1)) = \gamma_0 + \gamma_1 Z. \quad \text{Model 3}$$

Estimates of $PE^{(FGS)}$ or FM are summarized in Table 2. The distributions of these estimates are plotted in histograms as shown in Figure 2. The results show that, in any of the considered cases, $F(F')$ and FM has little bias and smaller standard deviations (SD) compared to $PE^{(FGS)}$. In all cases, the estimated $PE^{(FGS)}$ is biased toward 0 and seriously underestimates the true PTE except for the design case C6, in which $\beta_2 = 0$ and the surrogate is perfect. The variability of estimated $PE^{(FGS)}$ can be more than twice that of FM except in design C3. The variability is particularly large when the marker is perfect or useless (cases C1 and C6). In design C3, the SD of $PE^{(FGS)}$ is smaller due to the fact that the mean estimate is around zero (compared to the true effect of 1/3). In fact, when the true PTE is less than 1/3, the estimated $PE^{(FGS)}$ is negative. The bias of the estimated $PE^{(FGS)}$ improves as the true surrogate effect approaches one; however, the standard deviation remains larger than estimates of $F(F')$. As presented in Figure 2, the distributions of $PE^{(FGS)}$ are more skewed to the left than $F(F')$, particularly for design C1 (true effect = 0). Also in Table 2, the average FM is close to the true PTE. The SDs of F and F' are similar. The problems of $PE^{(FGS)}$ identified in our simulation are consistent with the simulation findings from other studies (Bycott and Taylor 1998; Wang and Taylor 2002).

Table 3 presents the 95% CI (confidence interval) and its coverage rate for F , F' , FM , and $PE^{(FGS)}$. $PE^{(FGS)}$ has the poorest coverage rate, especially when the true PTE is small (only 4.8% when true PTE = 66.7%). On the other hand, F , F' , and FM have nominal coverage rates (>93%) over all considerations. Table 4 shows the numbers of times (out of 1000 simulations)

that 95% CIs lie within $[0, 1]$, lower bound ≥ 0 , and upper bound ≤ 1 . As discussed earlier, $PE^{(FGS)}$ underestimates the true value by 30% or more. Therefore its 95% CI is shifted toward the negative direction. The lower CI is often below 0 and upper CI is below 1. This can be seen from the results of cases C1–C4. PTE is worse than FM in terms of lower bound ≥ 0 or CI within $[0, 1]$. They are comparable in terms of upper bound ≤ 1 . In the cases of C5 and C6, although the results for $PE^{(FGS)}$ and FM are comparable, both are often within $[0, 1]$, the point estimate of $PE^{(FGS)}$ and its CI are biased toward zero.

With probit model, it could be derived that Model 3 takes the form of

$$\Phi^{-1}(P(T=1|Z)) = \frac{(\beta_0 + \alpha_0 \beta_1)}{\sqrt{1 + \beta_1^2 \sigma_s^2}} + \frac{(\beta_2 + \alpha_1 \beta_1)}{\sqrt{1 + \beta_1^2 \sigma_s^2}} Z.$$

Thus, $PE^{(FGS)}$ by definition can be derived, that is,

$$PE^{FGS} = 1 - \frac{\beta_2 \sqrt{1 + \beta_1^2 \sigma_s^2}}{\beta_2 + \alpha_1 \beta_1}.$$

Compare $PE^{(FGS)}$ to the latent PTE ($= \frac{\alpha_1 \beta_1}{\alpha_1 \beta_1 + \beta_2}$ from Section 3.2), the bias is

$$\frac{\beta_2 \left(\sqrt{1 + \beta_1^2 \sigma_s^2} - 1 \right)}{\beta_2 + \alpha_1 \beta_1},$$

which is always ≥ 0 , suggesting $PE^{(FGS)}$ almost always underestimate the true effect. It is unbiased only when $\beta_2 = 0$ or $\beta_1 = 0$. When the logit model is desired for Model 2, an approximation of relationship

$$\frac{\exp(y)}{1 + \exp(y)} \approx \Phi\left(\frac{\sqrt{3}}{\pi} y\right)$$

can be used to estimate PTE, and same method is applicable.

Our simulation shows FM is a more accurate and less variable measure than $PE^{(FGS)}$. The point estimate is bounded within $[0, 1]$. It is unavoidable that the confidence interval of F (F^*) can be out of interval $[0, 1]$ when the true $PE^{(FGS)}$ is close to 0 or 1. Otherwise, its 95% CI is mostly within $[0, 1]$ when the true PTE is not near 0 or 1. The precision, variability, and coverage probability are much improved versus the $PE^{(FGS)}$ estimate.

5. An Example From an Ophthalmology Study

We use an example from published literature (Buyse et al. 1998; Wang and Taylor 2002). This is a randomized clinical trial study investigating the effect of interferon- α in 190 patients with age-related macular degeneration (ARMD). The surrogate S and clinical end point T are defined as S = change of visual acuity at six months (lines loss of vision in six months)

$$T = \begin{cases} 0 & \text{if patient lost less than three lines of} \\ & \text{vision at one year compared to baseline} \\ 1 & \text{if patient lost three or more lines of} \\ & \text{vision at one year compared to baseline} \end{cases}.$$

A Box–Cox plot suggests that five extreme data points (≥ 8 or ≤ -6) could be outliers. The Surrogate variable presents a reasonable normal distribution within each treatment group. Two groups t -test on surrogate was not significant for all 190 patients, but became significant after excluding the five patients with extreme values. It might be arguable that the five patients are outliers. For the purpose of illustrating Prentice's validation criteria, these five patients' data are excluded from analysis. Three of the criteria are met, that is, treatment is significant for clinical endpoint T , surrogate is also significant for clinical endpoint T , and there is no interaction between treatment and surrogate. The criteria that treatment effect is captured by surrogate for the clinical endpoint is assessed by the PTE estimation.

Point estimates of PTE, F and F^* , bootstrap estimates and their confidence intervals (CIs) are reported in Table 5. The bootstrap is used for constructing CI with replicates of 1000. The latent PTE is calculated based on the continuous clinical endpoint (vision at one year), which is 78%. Our median estimate of $\hat{F}$ and $\hat{F}^*$ are similar (62% and 64.5%). They are similar to the estimates reported by Wang and Taylor (2002) (69% and 61.9%) where the surrogate is treated as binary data. The CIs of F , F^* , and FM are clearly narrower than that of $PE^{(FGS)}$; and they are about 30% narrower than that reported by Wang and Taylor (2002). The point estimates of F and F^* are all within (0, 1), and the lower bound of the 95% CI are above 0, suggesting change of visual acuity at six months is a potential surrogate for the clinical endpoint (loss of three or more lines of vision at one year).

Figure 1 presents the effect of treatment in the placebo group (AA curve, solid line), and in the interferon- α group (BB curve, long dashed line). The BB curve lies below the AA curve suggesting that interferon-treatment may be effective in preventing more than three lines of loss of vision in ARMD patients by the end of one year. The treatment effect is presented by the inner area between the AA and BB curves. To demonstrate how much of the treatment effect could be explained by the surrogate measure, AB (short dashed line) and BA (dotted dash line) curves were plotted together with AA and BB curves. The area between the AA and AB curves is about $2/3$ ($F = 0.628$) of the ($AA-BB$) area, suggesting that a substantial amount of treatment-effect could be explained by the surrogate measure. Similarly, the $F^* (= 0.647)$ measure is presented by the area of ($BA-BB$) out of the area of ($AA-BB$).

6. Conclusion and Discussion

This article provides a useful and more accurate estimate for the proportion of treatment effect explained by a surrogate when the clinical endpoint is modeled by logistic or probit regression. The estimate can be calculated easily with any existing statistical computation package that is capable of binary regression modeling, such as SAS, SPSS, or S-Plus. This measure can be presented graphically by the ratio of the areas under ordinal dominant curves.

The derived measures $F(F^*)$ have much smaller bias and smaller variability in contrast to commonly used $PE^{(FGS)}$. The expected value of $F(F^*)$ is often inside $[0, 1]$ under a few mild conditions, while $PE^{(FGS)}$ often lies outside of $[0, 1]$. Unlike $PE^{(FGS)}$, measures $F(F^*)$ have no requirement that the fitted marginal model $[T|Z]$ and the full model $[T|S, Z]$ be the same type. In the logistic model $[T|Z, S]$, the marginal model of $[T|Z]$ is no longer in the logistic model

setting, and it is not easily obtained. $F(F')$ can be easily estimated by proper transformation of logit to probit function, and it is shown to be more reliable and efficient. In the setting of the probit model of $[T|S, Z]$, the marginal model $[T|S]$ follows a valid probit model as well. $PE^{(FGS)}$ estimate is shown to be unreliable and seriously underestimates the true PTE by as much as 30% or more.

Although theoretically both F and F' measures are to quantify the proportion of the treatment effect explained by a surrogate marker, the two could be slightly different in a nonnormal setting. FM as an average of both is less bias and has smaller variability than F and F' , and is much better than $PE^{(FGS)}$. We recommend the use of FM . Many other alternative measures have been proposed to assess the proportion of the treatment effect that is explained by a surrogate marker or mediator. A closely related proposal is the relative indirect effect (RIND) measure proposed by Huang et al. (2004). The RIND measure was proposed by first estimating an imaginary quantity, that is, the expected potential outcome given a specific treatment had the effect of treatment not acted through the intermediate surrogate endpoint. When taking the $g(\cdot)$ as the expected mean of the conditional distribution of $[T|S, Z]$, this expected potential outcome is equivalent to the AB measure in our definition, and the RIND is equivalent to the F measure of Wang and Taylor's (2002). There are slight differences in the RIND and $F(F')$ measures. First, RIND is defined by setting $g(\cdot)$ as the expected conditional mean; second, unlike the $F(F')$ measure, the RIND measure does not require the treatment to be a binary measure; third, RIND was defined based on a system of two generalized regression equations, and thus generally requires linear assumptions between the treatment Z and the outcome T . However, such assumptions could be easily relaxed. Both $F(F')$ and RIND measure were proposed for single trial only. Although it is accepted to evaluate the effect of surrogacy from multiple trials (Buyse et al. 2000), validation based on a single trial can provide preliminary evidence in the early study of surrogate or biomarker. Surrogate validation using $F(F')$ can be valid in a large single trial. Recently, a new measure was proposed based on information theory which can be applied to a wide variety of settings (Alonso and Molenberghs 2007). The information-theoretic based measure inherits a difficulty in providing a hard cut-off value for discriminating good or bad surrogates.

In this article, we have considered only a single surrogate endpoint in the clinical trial. In many situations, there could be multiple biomarkers on the pathway to the clinical endpoint, where the combination of biomarkers may serve as a surrogate endpoint. One may follow the two-stage approach, by first finding a best linear combination of these markers with the available statistical methods (Pepe 2003), then applying the method presented in this article to validate the surrogate. If multiple surrogate markers assume a joint distribution, $F(F')$ can be derived under multiple integration.

Acknowledgments

The authors thank the Pharmacological Therapy for Macular Degeneration Study Group, F. Hoffman-LaRoche, and Dr. Marc Buyse and Dr. Geert Molenberghs for providing the authors with data and the permission to use the data. This work is partially supported by the NIDA R01DA019965-01A1 awarded to Dr. Bin Huang.

REFERENCES

- Alonso A, Molenberghs G. Surrogate Marker Evaluation from an Information Theory Perspective. *Biometrics* 2007;63:180–186. [PubMed: 17447943]
- Bamber D. The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph. *Journal of Mathematical Psychology* 1975;12:387–415.
- Biomarkers Definitions Working Group. Biomarkers and Surrogate Endpoints: Preferred Definitions and Conceptual Framework. *Clinical Pharmacology and Therapeutics* 2001;69(3):89–95. [PubMed: 11240971]

- Buyse M, Molenberghs G. Criteria for the Validation of Surrogate Endpoint in Randomized Experiments. *Biometrics* 1998;54:1014–1029. [PubMed: 9840970]
- Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The Validation of Surrogate Endpoints in Meta-analyses of Randomized Experiments. *Biostatistics* 2000;1:49–67. [PubMed: 12933525]
- Bycott PW, Taylor J. An Evaluation of a Measure of the Proportion of the Treatment Effect Explained by a Surrogate Marker. *Controlled Clinical Trials* 1998;19:555–568. [PubMed: 9875835]
- DeGruttola V, Fleming T, Lin DY, Coombs R. Perspective: Validating Surrogate Markers—Are We Being Naive? *Journal of Infectious Diseases* 1997;175:237–246. [PubMed: 9203643]
- DeGruttola V, Clax PC, Demets DL, Downing GJ, Ellenberg SS, Freedman L, Gail MH, Pretence R, Wittes J, Zeger SL. Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institutes of Health Workshop. *Controlled Clinical Trials* 2001;22:485–502. [PubMed: 11578783]
- Efron B, Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* 1986;1:54–77.
- EMA/CHMP report. Innovative Drug Development Approaches: Report From the EMA / CHMP-Think-Tank Group on Innovative Drug Development. [March 2007]. 2007
<http://www.EMA.EUROPA.EU/PDFS/HUMAN/ITF/12731807EN.PDF>
- Freedman LS. Confidence Intervals and Statistical Power of the ‘Validation’ Ratio for Surrogate or Intermediate Endpoint. *Journal of Statistical Planning and Inference* 2001;96:143–153.
- Freedman LS, Graubard BI, Schatzkin A. Statistical Validation of Intermediate Endpoints for Chronic Disease. *Statistics in Medicine* 1992;11:167–178. [PubMed: 1579756]
- Huang B, Sivaganesan S, Succop P, Goodman E. Statistical Assessment of Mediation Effects for Logistic Mediation Methods. *Statistics in Medicine* 2004;23:2713–2728. [PubMed: 15316954]
- Lin DY, Fleming TR, DeGruttola V. Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker. *Statistics in Medicine* 1997;16:1515–1527. [PubMed: 9249922]
- NIAID Workshop. Statistical Issues for HIV Surrogate Endpoints: Point/Counterpoint. *Statistics in Medicine* 1989;17:2435–2462.
- Pepe, MS. The Evaluation of Medical Tests for Classification and Predictions. Oxford University Press; 2003. Oxford Statistical Science Series
- Prentice R. Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria. *Statistics in Medicine* 1989;8:431–440. [PubMed: 2727467]
- Temple, RJ. A Regulatory Authority’s Opinion About Surrogate Endpoints. In: Nimmo, WS.; Tucker, GT., editors. *Clinical Measurement in Drug Evaluation*. Wiley; New York: 1995. p. 3–22.
- Tsiatis AA, DeGruttola V, Wulfsohn MS. 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* 1995;90:27–37.
- Wang Y, Taylor J. A Measure of the Proportion of Treatment Effect Explained by a Surrogate Marker. *Biometrics* 2002;58:803–812. [PubMed: 12495134]
- Schwarz JK, Siegel BA, Dehdashti F, Grigsby PW. Association of Posttherapy Positron Emission Tomography With Tumor Response and Survival in Cervical Carcinoma. *Journal of American Medical Association* 2007;298(19):2289–2295.

Figure 1.

Graphic presentation of $F(F')$ in terms of the Ordinal Dominance (OD) curves for a clinical case study. Areas under each curve are the values for AA , BB , AB , and BA . OD curves are plotted with different symbols for AA (— solid line), BB (— — long dashed), AB (- - - short dashed) and BA (- · - dotted-dash), where $F = (AA - AB)/(AA - BB)$ and $F' = (BA - BB)/(AA - BB)$. For the study, $F = 0.628$ and $F' = 0.647$, $\alpha_0 = -0.144$, $\alpha_1 = 0.232$, $\beta_0 = -0.747$, $\beta_1 = 0.478$, $\beta_2 = 0.124$.

Figure 2. Histogram of the distribution of estimate $PE^{(FGS)}$ and FM for different simulation settings. FM has smaller biases and standard deviations (SD) comparing to $PE^{(FGS)}$.

Table 1

Simulation design setting with true values

Design	Scenario	Model 1		Model 2		Latent		True		True	
		α_1	β_1	β_2	β_2	PTE	F	F'	F	F'	FM
C1	Useless	0	2	1	0	0	0	0	0	0	0
C2	Partial	0.2	2	2	1/6	0.164	0.152	0.152	0.152	0.152	0.158
C3	Partial	1	1	2	1/3	0.236	0.382	0.382	0.382	0.382	0.309
C4	Partial	1	2	1	2/3	0.709	0.646	0.646	0.646	0.646	0.677
C5	Partial	2	1	0.5	0.80	0.753	0.882	0.882	0.882	0.882	0.818
C6	Perfect	1	2	0	1	1	1	1	1	1	1

Note: Fix $\alpha_0 = 1$ and $\beta_0 = -3$ for all.

Table 2

Summary of asymptotic results from simulation study for F , F^* , FM , and $PE^{(FGS)}$

Design	F			F*			FM			PE ^(FGS)		
	Mean	Med.	STD	Mean	Med	STD	Mean	Med	STD	Mean	Med	STD
C1	-0.03	0.01	0.22	-0.03	0.01	0.23	-0.03	0.01	0.22	-1.37	-1.25	0.59
C2	0.16	0.17	0.07	0.15	0.15	0.06	0.16	0.16	0.06	-0.92	-0.89	0.22
C3	0.24	0.24	0.05	0.38	0.38	0.05	0.31	0.31	0.04	0.05	0.05	0.06
C4	0.71	0.71	0.05	0.65	0.65	0.06	0.68	0.68	0.05	0.24	0.24	0.11
C5	0.76	0.77	0.13	0.88	0.89	0.05	0.82	0.82	0.08	0.72	0.72	0.10
C6	1.01	1.00	0.09	1.01	1.00	0.09	1.01	1.00	0.08	1.02	1.01	0.21

Note: Med is shorten for median; STD for standard deviation.

Table 3

95% confidence intervals (CI) by BC pivotal percentile, and coverage

Design	95% CI by BC bootstrap				Coverage			
	F	F*	FM	PE ^(FGS)	F	F*	FM	PE ^(FGS)
C1	(-1.04, 0.48)	(-1.09, 0.50)	(-1.06, 0.48)	(-4.41, 0.08)	0.966	0.975	0.969	0.027
C2	(0.01, 0.28)	(0.01, 0.27)	(0.02, 0.27)	(-1.43, -0.56)	0.952	0.952	0.960	0.000
C3	(0.15, 0.35)	(0.30, 0.48)	(0.24, 0.40)	(-0.07, 0.14)	0.961	0.938	0.951	0.000
C4	(0.62, 0.80)	(0.54, 0.76)	(0.59, 0.78)	(0.01, 0.46)	0.937	0.938	0.939	0.048
C5	(0.50, 1.01)	(0.78, 0.98)	(0.66, 0.98)	(0.56, 0.94)	0.934	0.939	0.933	0.870
C6	(0.86, 1.20)	(0.86, 1.21)	(0.86, 1.20)	(0.68, 1.52)	0.938	0.938	0.938	0.935

Table 4

Shown are the numbers of times (out of 1000) that the 95% confidence intervals (CIs) 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 FM and $PE^{(FGS)}$

Design	CI within [0,1]		CI with lower bound ≥ 0		CI with upper bound ≤ 1	
	FM	$PE^{(FGS)}$	FM	$PE^{(FGS)}$	FM	$PE^{(FGS)}$
C1	22	0	22	0	986	977
C2	632	0	632	0	1000	1000
C3	1000	167	1000	167	1000	1000
C4	1000	534	1000	534	1000	1000
C5	624	701	1000	1000	624	701
C6	35	38	1000	1000	35	38

Table 5Estimates and 95% confidence interval of F , F' , and $PE^{(FGS)}$ for ARMD data

	Point est.	CI bootstrap directly	Median	CI bootstrap pivot
F	0.628	(0.08, 2.08)	0.620	(0.07, 1.91)
F'	0.647	(0.02, 1.93)	0.645	(0.10, 1.94)
FM	0.638	(0.06, 1.97)	0.635	(0.10, 1.82)
$PE^{(FGS)}$	0.467	(−0.38, 2.26)	0.464	(−0.36, 2.24)

Note: The latent PTE is 0.78.