## Title

An Evaluation of a Measure of the Proportion of the Treatment Effect Explained by a Surrogate Marker



## Authors

Paul W. Bycott
Jeremy M.G. Taylor

## Publication Date

2011-10-25

ELSEVIER

# An Evaluation of a Measure of the Proportion of the Treatment Effect Explained by a Surrogate Marker

**Paul W. Bycott, DrPH and Jeremy M.G. Taylor, PhD**
*Department of Biometrics, Parke-Davis Pharmaceutical Research, Ann Arbor, Michigan (P.W.B.); Department of Biostatistics, UCLA School of Public Health, Los Angeles, California (J.M.G.T.)*

**ABSTRACT:** Time-dependent markers, such as CD4 and viral load, are potential surrogate markers in AIDS clinical trials. A critical issue with surrogate markers is whether changes in these markers explain the beneficial effect of treatment on the real end point of the clinical trial. A statistic to measure the proportion of the treatment effect explained by the surrogate is $P^{(FGS)} = 1 - \gamma/\alpha$, where $\alpha$ is the treatment effect coefficient in a Cox model and $\gamma$ is the treatment effect coefficient from a time-dependent Cox model adjusted for the marker. In this article we evaluate the statistical properties of $P^{(FGS)}$. Using a Monte Carlo study we show that the statistic is not well calibrated, because it can fall outside the range zero to one, even in very large samples. In the simulation study we consider situations where the time-dependent marker is measured with error at a fixed number of times. We show that a method of fitting a time-dependent Cox model involving smoothing the marker reduces the bias in the estimate of $P^{(FGS)}$ compared with the standard method of using the current or last observed marker value. We also show that the estimate of $P^{(FGS)}$ has considerable variability and can have wide confidence intervals. We conclude that $P^{(FGS)}$ is only likely to be useful in large trials with a strong treatment effect. The methods are illustrated using CD4 counts from an AIDS clinical trial of zidovidine versus placebo. *Controlled Clin Trials* 1998;19:555–568 

KEY WORDS: *Clinical trials, longitudinal markers, time-dependent Cox model, ACTG protocol 019, CD4+ T-cell count*

## INTRODUCTION

There is considerable interest in conducting clinical trials expediently. The sooner the benefits of a new treatment can be determined, the more rapidly the treatment can be made available to the general population. A randomized phase III trial typically studies a clinical end point of primary interest. This end point, however, may occur only in a small fraction of trial participants and may take many years to develop, thus dictating a very large and long trial.

*Address reprint requests to: Dr. J.M.G. Taylor, Department of Biostatistics, UCLA School of Public Health, Los Angeles, California 90095-1772.*
*Received December 29, 1997; accepted June 24, 1998.*

Controlled Clinical Trials 19:555–568 (1998)

655 Avenue of the Americas, New York, NY 10010
0197-2456/98/$19.00
PII S0197-2456(98)00039-7

A surrogate marker represents a possible alternative end point that could allow more expedient trials, thus potentially reducing sample-size requirements and the cost of conducting the study. A marker that serves as a surrogate end point for a particular treatment must respond quickly to the treatment, must be prognostic for the true end point, and explain most of the effect of the treatment on the clinical end point [1]. Much recent discussion of surrogate markers in AIDS and other diseases has appeared in the literature with some cautionary words given by Machado et al. [2] and by Fleming and DeMets [3].

Prentice [4] formally defined a surrogate end point to be "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." This definition is rather restrictive and does not capture the concept that a marker may explain some but not all of the treatment effect.

Freedman et al. [5] proposed an alternative approach to investigating surrogate markers in the context of a binary end point and logistic regression. They estimate the proportion of information about the treatment's effect on the true end point of interest explained by the proposed surrogate marker. Using a marginal logistic model with only the treatment covariate and a joint model with both the marker and treatment entered as covariates, they examined the change in the estimated treatment parameter. The particular statistic suggested, which we denote by $P^{(FGS)}$, is the difference between treatment effect coefficients divided by the unadjusted treatment effect coefficient. Lin et al. [6] informally extended these ideas to the time-dependent Cox proportional hazards model and suggested comparing the treatment effect coefficient in a time-dependent Cox model with and without adjusting for the marker. Freedman et al. [5] suggested a procedure for using their statistic; in particular, they recommended placing less emphasis on the estimate of $P^{(FGS)}$ but rather determining whether the lower limit of a confidence interval for $P^{(FGS)}$ was greater than 0 or some specified value. Their evaluation suggested that only if the unadjusted treatment effect was more than four times its standard error would it be possible to validate that a surrogate was explaining some of the treatment effect. O'Brien et al. [7] used $P^{(FGS)}$ to evaluate the proportion of zidovidine's (AZT) effect on progression to AIDS explained by HIV viral RNA copy number and by CD4. The authors used estimates of $P^{(FGS)}$ to conclude that viral RNA explained a larger proportion of AZT's effect than CD4, but using them both as joint surrogates in a Cox model explained the largest proportion. This controversial conclusion [8] has prompted investigation into the statistical properties of the $P^{(FGS)}$ statistic [9]. Lin et al. [9] showed that $P^{(FGS)}$ can be quite variable. They also considered how to estimate standard errors for $P^{(FGS)}$. The current article is also concerned with an evaluation of the properties of $P^{(FGS)}$.

We were motivated by a particular AIDS clinical trial, ACTG019 part B [10], an early placebo-controlled randomized clinical trial of AZT in patients with CD4 counts less than 500 at enrollment in which the primary end points were survival and development of AIDS or advanced AIDS-related complex. The compliment part A of this trial in patients with CD4 counts of at least 500 is not discussed because that part of the trial closed at a much later date. We will consider the serially measured CD4 counts a potential surrogate for the primary end points. The Monte Carlo approach we use to evaluate the properties of

$P^{(FGS)}$ generates data that mimic the type of data collected in ACTG019. In particular, we consider a plausible stochastic process and measurement error for CD4 and realistic magnitude of the treatment effect on CD4 and on the clinical end point.

$Z(t)$ will denote a time-dependent marker, observed at times $t_1, \ldots, t_M$, and $X$ will denote the treatment group indicator. Two different Cox models will be considered, an unadjusted Cox model

$$\lambda(t;X) = \lambda_{1o}(t)e^{\alpha X}, \tag{1}$$

and an adjusted time-dependent Cox model,

$$\lambda(t,Z,X) = \lambda_{2o}(t)e^{\beta Z(t)+\gamma X}. \tag{2}$$

Then $P^{(FGS)}$, defined as $1 - \gamma/\alpha$, is estimated by fitting Eqs. (1) and (2) separately to the identical dataset. In theory, the more information about treatment that passes through $Z$, the closer $\gamma$ will be to zero and hence the closer $P^{(FGS)}$ will be to one. On the other hand, if $\gamma$ is close to $\alpha$, then $P^{(FGS)}$ will be close to zero, implying that the surrogate marker explains little or none of the information about the treatment.

It is unlikely that models (1) and (2) both fit the observed data. Imagine a true process underlying the generation of the observations consisting of a stochastic process for $Z(t)$, possibly influenced by treatment, and a model for the hazard of the event. Suppose Eq. (2) is the true hazard model, then the marginal model for the effect of $X$ on the hazard could be obtained, in principle, by integrating over $Z(t)$. It is difficult to imagine how this process could lead exactly to Eq. (1), except in trivial cases such as $\beta = 0$. Therefore, we need to give careful consideration to defining a measure based on a model that may not fit the data.

Next, we begin by describing ACTG019 trial. Then we consider calibration of $P^{(FGS)}$, in particular whether the population quantity being estimated is a reasonable measure. Using Monte Carlo simulations, we evaluate the statistical properties of $P^{(FGS)}$ in three scenarios in which the marker is a perfect, a partial, or not a surrogate, and we investigate the variability of $P^{(FGS)}$.

## ACTG019 CLINICAL TRIAL

### Description of ACTG019

The ACTG019 clinical trial part B was a randomized double-blind trial in asymptomatic HIV-infected adults who had CD4 counts of fewer than 500/$mm^3$ on entry into the study [10]. Subjects were randomly assigned to one of three treatment arms: placebo, AZT 500 mg/day, or AZT 1500 mg/day. The trial was conducted to determine the safety of AZT and its efficacy in prolonging survival and in delaying the onset of AIDS or advanced AIDS-related complex.

The protocol stipulated that patients' CD4 values were to be measured at baseline, 8, 16, 32, 48, 64, and 80 weeks. There were 428 individuals in the placebo arm, 453 in the low-dose arm, and 456 in the high-dose arm. All analyses here combine the low-dose and high-dose arms, because there is no significant difference in time to AIDS between the two arms. In the placebo arm, 33 subjects progressed to AIDS, whereas 25 progressed in the combined

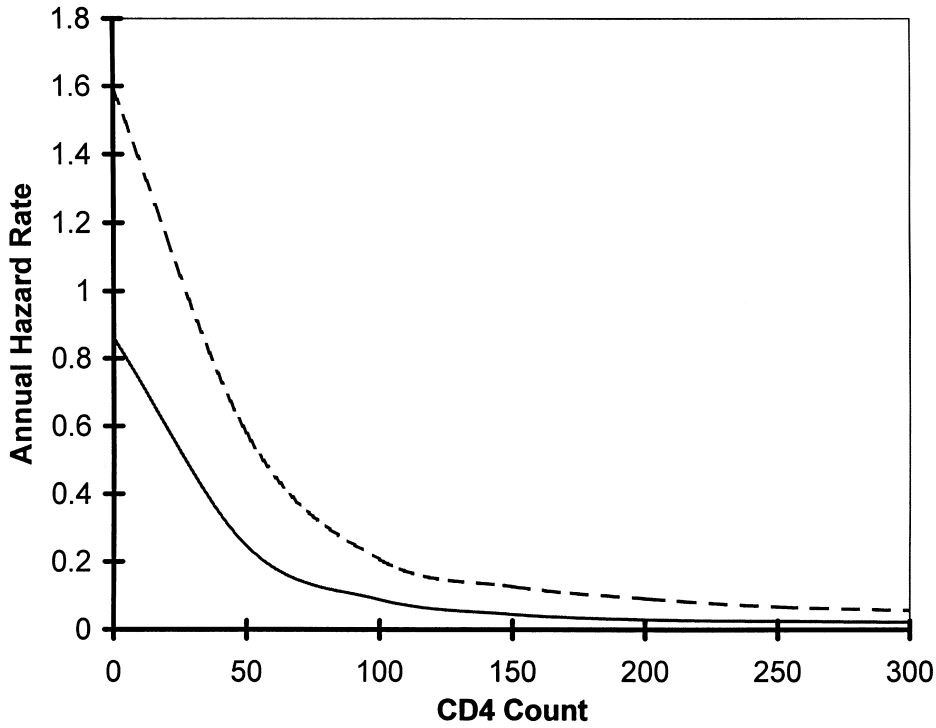


**Figure 1** Dependence of the hazard of AIDS on the current CD4 count by treatment arm. – – –, Placebo; ——, AZT.

treatment arm, showing a significant impact of AZT in reducing AIDS. The blinded part of the trial was stopped before many of the later enrollers had their 64- and 80-week follow-up measurements. Average follow-up was 50 weeks with a maximum of about 108 weeks. The median number of CD4 measurements per subject was 3.7, indicating many missing observations and considerable loss to follow-up.

### Surrogate Properties of CD4

Figure 1 depicts the estimated annual hazard of AIDS as a function of CD4 in the two treatment groups using the method proposed by Breslow [11] smoothed using cubic splines. To calculate the hazard, we first transform CD4 counts by taking fourth roots [12]. For a specific observed $CD4^{1/4}$ value $Y$ we let

$O$ = the observed number of failures with AIDS for $CD4^{1/4}$ values in $[Y - 1, Y + 1]$,
$E$ = the total exposure time with $CD4^{1/4}$ values in $[Y - 1, Y + 1]$.

Then, the estimate of the hazard for a specific value of $CD4^{1/4}$ is

$$\hat{\lambda}(Y) = \frac{O}{E}.$$

This graph shows that the hazard increases substantially in the placebo arm

when the CD4 count falls below 100 and in the treatment group when it falls below 75. The curves differ substantially, indicating that CD4 count does not capture all information about the effect of AZT. This suggests that CD4 is not a perfect surrogate marker for AZT. The graph, however, does not indicate whether CD4 is a partial surrogate marker. The estimate of $P^{(FGS)}$ from fitting Eqs. (1) and (2) is $-0.02$, obtained using the OBS method for fitting a time-dependent Cox model, as described below. This value of $P^{(FGS)}$ is near 0. Other estimates of $P^{(FGS)}$ using slightly different methods (not shown) are also near 0, suggesting that CD4 captures none of the effect of AZT. This article examines whether we can meaningfully interpret estimated values of $P^{(FGS)}$.

## CALIBRATION OF $P^{(FGS)}$

Because $P^{(FGS)}$ is supposed to represent the proportion of the treatment effect explained by the marker, we would expect $P^{(FGS)}$ to fall between 0 and 1; however, an estimate of $P^{(FGS)}$ derived from a set of observations can be negative or larger than 1. Consider a joint distribution for the stochastic process of $Z(t)$ and the event time. Let $\hat{P}^{(FGS)} = 1 - \hat{\gamma}/\hat{\alpha}$. Then $\hat{P}^{(FGS)}$ estimates a population quantity $P^*$, or functional, of this joint distribution. In this section we examine the value of $P^*$ under a variety of joint distributions of $Z(t)$ and the event time. We look at the value of $P^*$ under several combinations of $\beta$ and $\gamma$ in Eq. (2) and under differing effects of treatment on the longitudinal mean structure of $Z$.

We estimate the value of $P^*$ using the following procedure. First, the number of subjects is set at 6000 and then values of $Z(t)$ are simulated for each subject over a fine grid of time points for 27 months. The stochastic process for $Z(t)$ may differ between the treatment and the placebo arms. Then, event times are simulated for each subject assuming model (2), and fit both models (1) and (2) to obtain the estimate of $\alpha$ and $\gamma$ and hence $P^*$.

The stochastic process for $Z(t)$, based on our experience with analyzing CD4 counts [13, 14] is a model with an underlying mean structure, a random intercept, and Brownian motion. In previous work we [13, 14] and others [15] have shown that this model with the addition of measurement error describes longitudinal fourth root CD4 counts well. The mean structure for the placebo arm is a constant decline, whereas the mean structure for the treatment arm is a linear increase for 8 weeks followed by a linear decrease. The vector of $Z(t)$ values for subject $i$ is given by

$$\mathbf{Z_i} = \mathbf{X_i}\eta + \mathbf{V_i}\mathbf{b_i} + \mathbf{BM_i},$$

where $\mathbf{X_i}$ is a known matrix consisting of a column of 1's and a column of time points where $Z_i$ values are recorded for the placebo arm, and an additional column equal to $(t_{ij} - 0.15332)^+$ for the treatment arm. The time $t_{ij}$ denotes observation time $j$ for subject $i$ measured in units of years, and $(t_{ij} - 0.15332)^+$ is zero if this quantity is negative. The fixed effects vector is set as $\eta = (4.25, -0.20)^T$ for the placebo arm and as $\eta = (4.25, 1.10, -1.30)^T$ for the treatment arm. These values correspond approximately to those observed in ACTG019. $\mathbf{V_i}$ is a known matrix consisting of a column of 1's and $\mathbf{b_i}$ is the individual intercept that is assumed to have a normal distribution with mean 0 and variance 0.10 for both treatment arms. $\mathbf{BM_i}$, the Brownian motion term, is normally distributed with mean 0 and variance parameter $c^2 = 0.15$ for both

**Table 1** Calibration Results for $P^{(FGS)}$ Varying Treatment's Effect on $Z$ Longitudinally and Under Different Hazard Functions

| Longitudinal Effect | $\beta$ | $\gamma$ | $\hat{\alpha}$ | $\hat{\gamma}$ | $\hat{P}^{(FGS)}$ | Proportion of Events |
|---|---|---|---|---|---|---|
| Full | −2 | −0.5 | −0.80 | −0.46 | 0.43 | 0.08 |
| Full | −2 | −1 | −1.28 | −0.95 | 0.26 | 0.07 |
| Full | −2 | −2 | −2.20 | −1.90 | 0.14 | 0.06 |
| Half | −2 | −0.5 | −0.65 | −0.50 | 0.23 | 0.09 |
| Half | −2 | −1 | −1.05 | −0.91 | 0.13 | 0.08 |
| Half | −2 | −2 | −2.04 | −1.93 | 0.05 | 0.07 |
| Quarter | −2 | −0.5 | −0.54 | −0.48 | 0.11 | 0.09 |
| Quarter | −2 | −1 | −1.01 | −0.97 | 0.04 | 0.08 |
| Quarter | −2 | −2 | −1.94 | −1.92 | 0.01 | 0.07 |
| None | −2 | −0.5 | −0.43 | −0.47 | −0.09 | 0.10 |
| None | −2 | −1 | −0.88 | −0.94 | −0.07 | 0.08 |
| None | −2 | −2 | −1.86 | −1.93 | −0.04 | 0.07 |
| Full | 0 | −0.5 | −0.42 | −0.42 | −0.002 | 0.17 |
| Full | 0 | −1 | −0.96 | −0.96 | −0.001 | 0.15 |
| Full | 0 | −2 | −2.02 | −2.03 | −0.005 | 0.12 |

treatment arms. For Brownian motion, the covariance between values at $t_{ij}$ and $t_{ik}$ is $c^2\min(t_{ij},t_{ik})$. Failure times with AIDS occur under a proportional hazards model [Eq. (2)] with a constant baseline hazard ($\lambda_{20}(t) = 2.5$). There is no censoring other than at the end of the study. As well as the above "full" model for the treatment effect on $Z$, we also consider "half," "quarter," and "none" models. The difference between "full," "half," "quarter," and "none" for the treatment effect on the mean structure of $Z$ is obtained by changing the magnitude of the nonnegative slope for $Z$ from randomization until 8 weeks in the treatment arm and adjusting the slope of decline after 8 weeks to be parallel to the placebo arm. When the effect is "none," the placebo and treatment arms have the same slope of decline starting from baseline [$\eta = (4.25, -.20, 0)$]. For a "half" effect, the positive slope from baseline to 8 weeks is half way in magnitude between the "full" effect and "none" models [$\eta = (4.25, 0.45, -0.65)$]. For a "quarter" effect, $\eta = (4.25, 0.125, -0.325)$. We also consider a variety of values for $\beta$ and $\gamma$. The values $\beta = -2$ and $\gamma = -1$ correspond approximately to what we found in our analysis of ACTG019. When $\beta = 0$, we set the baseline hazard at 0.0025. Results are presented in Table 1. "Proportion of Events" is the proportion of individuals progressing to AIDS. These proportions are intentionally set low to reflect the fact that in the actual ACTG019 trial there was approximately 95% censoring. We note that the estimates of $\gamma$ are close to the true values of $\gamma$; similarly we expect the values for $\hat{P}^{(FGS)}$ to be a good approximation of $P^*$.

When $\beta = -2$, and treatment has a longitudinal effect ("full," "half," or "quarter"), as the magnitude of $\gamma$ increases in absolute value the estimate of $P^{(FGS)}$ decreases. This is because the proportion of the treatment effect on the hazard that goes unexplained by its effect on the mean structure of $Z$ increases. Similarly, when $\beta = -2$ and treatment has a "full" longitudinal effect on $Z$, the estimate of $P^{(FGS)}$ at all values of $\gamma$ is higher than when treatment has only

a "half" or "quarter" effect. The reason for this is that more of the overall effect of treatment is passing through $Z$, and therefore $Z$ explains more of the effect of treatment on progression to the end point.

When $\beta = -2$ and treatment has no effect longitudinally on $Z$, $P^{(FGS)}$ is consistently estimated to be negative, indicating that the statistic is poorly calibrated when the amount of information about treatment's effect on the true end point explained by the marker is small. Because the estimate of $P^{(FGS)}$ is always negative and never positive, we would not make the mistake in this scenario of claiming that the marker explains some positive percentage of the treatment effect when, in fact, it explains none of it. Finally, when $\beta = 0$, the estimate of $P^{(FGS)}$ is close to 0 for all levels of $\gamma$, implying that the effect of treatment on $Z$ longitudinally is irrelevant because $Z$ plays no role in the joint hazard function with treatment. In other simulations with different values of the longitudinal parameters (not shown), we have found situations where the estimate of $P^{(FGS)}$ is substantially more negative than for the "none" rows in Table 1 and situations where the estimate of $P^{(FGS)}$ is negative even in the presence of a true positive longitudinal effect of treatment on $Z$.

## MONTE CARLO STUDY: EVALUATION OF $P^{(FGS)}$

### Measurement Error

This section describes properties of the Freedman et al. approach to assessing a surrogate marker in the presence of substantial measurement error. We assume that the observed value ($Z_i$) equals the "true" value ($Z_i^*$) plus independent measurement error.

### Fitting a Time-Dependent Cox Model

Estimation of $P^{(FGS)}$ requires fitting a time-dependent Cox model, which requires knowing the value of the time-dependent variable at every event time. We adopt a two-stage approach in which we first impute marker values, denoted by $\hat{Z}_i^*$, at the event times and then uses these to fit the Cox model. In the simulation study we consider four different approaches for imputing values of $Z_i^*$: using the true (but unobserved) values of $Z^*$ (denoted by TRUE), using the last observed value of $Z$ (denoted by OBS), using the Tsiatis et al. [16] method with a random intercept and Brownian motion model (denoted by BM), and using the method of fitting a straight line separately to each subject's set of $Z$ values and using the interpolation from that [17] (denoted by IRL). Bycott and Taylor [14] give a more complete description and comparison of these methods for fitting a time-dependent Cox model.

### Design of the Simulation Study

We simulate data when CD4 is a perfect surrogate, a partial surrogate, and not a surrogate. In each of these scenarios, we simulate 200 data sets, each with 300 subjects randomized with equal probability to either the treatment or placebo arm and followed for up to 27 months. Observed $Z$ values are recorded

**Table 2** Parameter Values for Perfect, Partial, and Nonsurrogate Cases

| Group | β | γ | Treatment Affects Z | η | $\sigma_b^2$ |
|---|---|---|---|---|---|
| | | Perfect Surrogate | | | |
| Placebo | −2 | 0 | Yes | (4.2520, −0.2812) | 0.1347 |
| Treatment | −2 | 0 | Yes | (4.2143, 1.1476, −1.2097) | 0.1224 |
| | | Partial Surrogate | | | |
| Placebo | −2 | −1 | Yes | (4.2520, −0.2812) | 0.1347 |
| Treatment | −2 | −1 | Yes | (4.2143, 1.1476, −1.2097) | 0.1224 |
| | | Nonsurrogate | | | |
| Placebo | −2 | −1 | No | (4.2520, −0.2812) | 0.1347 |
| Treatment | −2 | −1 | No | (4.2143, −.2812,0.0) | 0.1224 |

at 3-month intervals starting at $t = 0$ according to a random intercept, plus Brownian motion, plus measurement error model of the form

$$Z_i = Z_i^* + \epsilon_i,$$

where $Z_i^*$ is the true $CD4^{1/4}$ value defined by the mixed effects model above. In this model, the Brownian motion term has mean 0 and variance $c^2 = 0.15$ and $\epsilon_i \sim N(0, 0.10)$ for both treatment arms. We simulate failure times with AIDS to occur under a proportional hazards model $\lambda_{2o}(t)e^{\beta Z^*(t)+\gamma X}$, with $\gamma_{2o}(t) = 2.5$. The parameter values for the three scenarios are described in Table 2. The values in scenarios 1 and 2 are designed to mimic that observed in analysis of data from ACTG019. The percentage of censored observations under the three scenarios are 85%, 88%, and 86%, respectively. To each dataset we fit the unadjusted Cox model [Eq. (1)] and the Cox model adjusted by $\hat{Z}^*$ (the predicted true value), using the four methods described above.

For $P^{(FGS)}$ we consider bias, variability, and coverage rate of $100(1 - \alpha)\%$ confidence intervals. The asymptotic $100(1 - \alpha)\%$ confidence interval for $P^{(FGS)}$ is given by $\hat{P}^{(FGS)} \pm Z_{1-\alpha/2}(\text{var}(\hat{P}^{(FGS)}))^{1/2}$. The variance of $P^{(FGS)}$ for each simulated dataset is calculated using an asymptomatic approximation

$$\text{var}(\hat{P}^{(FGS)}) = \frac{\hat{\gamma}^2}{\hat{\alpha}^2}\left[\frac{\text{var}(\hat{\gamma})}{\hat{\gamma}^2} + \frac{\text{var}(\hat{\alpha})}{\hat{\alpha}^2} - \frac{2\text{cov}(\hat{\alpha},\hat{\gamma})}{\hat{\alpha}\hat{\gamma}}\right]. \quad (3)$$

The parameters $\hat{\gamma}$ and $\hat{\alpha}$ are the maximum partial log-likelihood estimates from the two Cox models and $\text{var}(\hat{\gamma})$ and $\text{var}(\hat{\alpha})$ are obtained from the inverse information matrix. Because the term $\text{cov}(\hat{\alpha},\hat{\gamma})$ is not trivial to calculate, we use a "bootstrap-like" approximation in which we first estimate $\text{corr}(\hat{\alpha},\hat{\gamma})$ by the correlation coefficient across the 200 simulations. We then obtain $\text{cov}(\hat{\alpha},\hat{\gamma})$ from the expression $\text{cov}(\hat{\alpha},\hat{\gamma}) = \text{corr}(\hat{\alpha},\hat{\gamma})(\text{var}(\hat{\alpha}))^{1/2}(\text{var}(\hat{\gamma}))^{1/2}$, using the same correlation for all of the 200 datasets.

### Results for $P^{(FGS)}$

Table 3 shows results for estimation of $P^{(FGS)}$. Occasional values of $\hat{P}^{(FGS)}$ across the 200 simulations are extreme outliers. To reduce the influence of these points, we present the 10% trimmed mean (the arithmetic mean after discarding the lowest 10% and highest 10% of the data) of $\hat{P}^{(FGS)}$ across the simulations for

**Table 3** Simulation Results: Estimate, Standard Deviation, Coverage Rate of Nominal 90% Confidence Intervals, and Relative Efficiency for $P^{(FGS)}$

| Model | 10% Trimmed Mean of $\hat{P}^{(FGS)}$ | Standard Deviation | Coverage Rate | Ratio of MSEs | IQR ($Q_{25}$, $Q_{75}$) | Range (Min, Max) |
|---|---|---|---|---|---|---|
| | | | Perfect Surrogate ($P^* = 1.0$) | | | |
| TRUE | 1.04 | 0.53 | 0.88 | 1.00 | (0.74, 1.33) | (−13.14, 20.88) |
| OBS | 0.79 | 0.46 | 0.73 | 1.03 | (0.55, 1.00) | (−9.15, 21.27) |
| IRL | 0.99 | 0.49 | 0.87 | 0.94 | (0.71, 1.27) | (−12.03, 21.69) |
| BM | 1.03 | 0.53 | 0.89 | 1.00 | (0.70, 1.30) | (−12.49, 25.12) |
| | | | Partial Surrogate ($P^* = 0.389$) | | | |
| TRUE | 0.38 | 0.15 | 0.95 | 1.00 | (0.29, 0.46) | (−0.07, 1.15) |
| OBS | 0.30 | 0.16 | 0.85 | 1.55 | (0.23, 0.36) | (0.01, 0.87) |
| IRL | 0.37 | 0.14 | 0.93 | 0.91 | (0.29, 0.44) | (−0.07, 1.06) |
| BM | 0.38 | 0.14 | 0.93 | 0.91 | (0.30, 0.47) | (−0.02, 1.19) |
| | | | Nonsurrogate ($P^* = -0.321$) | | | |
| TRUE | −0.33 | 0.42 | 0.91 | 1.00 | (−0.61, −0.11) | (−237.76, 0.37) |
| OBS | −0.25 | 0.32 | 0.81 | 0.64 | (−0.42, −0.08) | (−200.66, 0.20) |
| IRL | −0.31 | 0.42 | 0.91 | 1.02 | (−0.49, −0.09) | (−271.22, 0.34) |
| BM | −0.34 | 0.44 | 0.92 | 1.11 | (−0.59, −0.11) | (−268.57, 0.26) |

each method. The ratio of MSEs (where MSE is defined to be the variance of an estimate plus the squared bias of the estimate) is from this trimmed sample relative to TRUE. The "true" values of $P^{(FGS)}$, denoted by $P^*$, are 0.39 and $-0.32$ for the partial and nonsurrogate cases, respectively, calculated using the methods described above. We use these estimates as the "true" values for the purpose of constructing confidence intervals and MSE calculations. In the perfect surrogate case, the TRUE approach gives a trimmed mean of 1.04, the results for the BM and IRL approaches are also close to the true value of 1. The OBS approach has a trimmed mean value for $P^{(FGS)}$ of 0.79, well below the target value of 1 and substantially less than the smoothing approaches.

For the perfect and partial surrogate cases, because the OBS approach does not account for measurement error, the risk parameter estimate for $\beta$ is biased downward. This approach then explains less about the treatment effect in the joint model. $\hat{P}^{(FGS)}$ is, therefore, biased downward, leading to actual coverage of a nominal 90% confidence interval for $P^{(FGS)}$ well below the nominal level. Smoothing substantially reduces this bias. The two smoothing approaches give ratios of the mean squared errors of approximately 1 (Table 3, column 5). In these scenarios, smoothing gives more accurate estimates of $P^{(FGS)}$ than simply using the nearest preceding value of $Z$. Moreover, these smoothed estimates have good efficiency properties.

For the nonsurrogate case, the OBS approach has the largest bias in the estimate of $P^{(FGS)}$, but its considerably smaller variance estimate makes this approach more efficient than the TRUE approach. The TRUE approach and all the smoothing approaches in general have the actual coverage rate around 90%. Once again, because of the bias, the OBS approach gives actual coverage well below the nominal 90% level.

## VARIABILITY OF $P^{(FGS)}$

In examining the properties of $P^{(FGS)}$, we found it has tremendous variability, which is a glaring drawback of this measure for determining the amount of information explained by a potential surrogate end point. In columns 6 and 7 of Table 3, we present the interquartile range (IQR) and the range across the 200 simulation runs for each of the three scenarios. In scenarios one and two, the range of $\hat{P}^{(FGS)}$ for the various techniques essentially covers the whole parameter space for $P^{(FGS)}$ between 0 and 1, as well as values well outside of the acceptable range. For scenario three, $\hat{P}^{(FGS)}$ never gets close to 1 for any of the approaches but does on several occasions have values below, and often substantially below, 0.

Comparing the smoothing techniques (BM and IRL) across the three scenarios, we see that they do not reduce this tremendous variability compared with OBS, and they are not noticeably worse than TRUE. To decrease this tremendous variability and obtain more precise estimates of the treatment parameters $\gamma$ and $\alpha$, the sample size for each individual trial would have to be made substantially larger. An explanation for the cause of the wide variability in $P^{(FGS)}$ is that for a few datasets $\hat{\alpha}$ is very small and $\hat{\gamma}$ is moderate sized, causing $P^{(FGS)}$ to blow up, giving extreme outliers.

**Table 4** Median Value (Median of the Estimated Variances) of $P^{(FGS)}$ Categorized by Significance of Log-Rank Test

| $p$ Value | Perfect Surrogate | Partial Surrogate | Nonsurrogate |
|---|---|---|---|
| $\leq 0.001$ | 0.65(0.04) | 0.32(0.01) | −0.02(0.02) |
| (0.001, 0.01) | 0.78(0.09) | 0.39(0.02) | −0.17(0.04) |
| (0.01, 0.05) | 0.98(0.22) | 0.40(0.04) | −0.24(0.10) |
| (0.05, 0.10) | 1.30(0.47) | NA | −0.54(0.24) |
| $>0.10$ | 1.70(1.56) | 0.04(225.86) | −0.95(1.04) |

Table 4 shows the median of the estimates of $P^{(FGS)}$ and the median of the estimated variances of $P^{(FGS)}$ [using Eq. (3)] when the true $Z$ values are measured without error. For each of the three scenarios, we categorize the median estimated value of $P^{(FGS)}$ and its variance by the significance level of the log-rank test of $\alpha = 0$ in the marginal Cox model.

Our simulation studies show that to obtain a precise estimate of $P^{(FGS)}$, the parameter estimate of $\alpha$ must be at least three to four times larger than its standard error, especially when the true value of $P^{(FGS)}$ is near 0 or 1. This is far larger than is required to observe a significant impact of treatment on delaying the progression to an event and lends strong support to larger and/or longer clinical trials if this method of evaluating a possible surrogate marker is going to be used. Also, even under the perfect surrogate scenario and $\hat{\alpha} \geq$ four times its standard error, the median width of the 90% confidence intervals for $P^{(FGS)}$ is still 0.63, which covers everything from a moderately weak surrogate to a nearly perfect surrogate.

Table 4 also implies that even though we may gain precision in our estimate of $P^{(FGS)}$ when $\alpha$ is very significant, this estimate may be quite biased. The more significant the marginal treatment effect in Table 4, the further, in general, our median estimates of $P^{(FGS)}$ are from the "true" value. This observation holds particularly for the perfect and nonsurrogate cases and less so for the partial surrogate case.

Another way to think about $P^{(FGS)}$ is that it would be nice to say with some certainty that the marker explains at least 50% or 75% of the effect of treatment. Again, using the true $Z$ values measured without error under the perfect surrogate scenario, we can evaluate this by looking at the lower limit of a 90% confidence interval and seeing how often it is at least 0.5 or 0.75. The lower limit is at least 0.5 only 33% of the time and greater than 0.75 only 4.5% of the time.

In summary, the fact that $Z$ is measured with error seems to have a large effect on $P^{(FGS)}$. Under all these scenarios, $P^{(FGS)}$ is quite variable, especially when $Z$ is a perfect surrogate or a nonsurrogate. The various smoothing techniques do not reduce this large variability but do reduce bias in the estimate of $P^{(FGS)}$. Because of the tremendous variability of $P^{(FGS)}$, we recommend that it be used only as a measure of the amount of information explained by a potential surrogate end point if the parameter estimate of $\alpha$ is at least three times the size of its standard error. Only under this condition is reasonable precision achieved for $P^{(FGS)}$; however, in a trial with such a strong treatment effect, it seems unlikely that even a perfect surrogate end point can explain all or nearly

all of the information about the effect of treatment on the true end point of interest. Also, even though the estimate of $P^{(FGS)}$ may be reasonably precise, we saw in Table 4 that it may be considerably biased.

## DISCUSSION AND CONCLUSIONS

To expedite clinical trials and contain costs, an alternative end point to the true one is often considered to evaluate the performance of a new therapeutic agent. For a marker to be a surrogate for the true end point for a particular agent, it must be prognostic of the true end point, respond quickly to the treatment, and explain nearly all of the effect of the treatment on the true end point. Expecting a marker to explain nearly all of the effect of treatment on the true end point is restrictive. In the context of logistic regression, Freedman et al. [5] proposed a statistic, denoted by $P^{(FGS)}$, which aims to measure the proportion of information about treatment's effect on the true end point explained by the marker. Following Lin et al. [6], we considered this statistic in a time-dependent Cox model. Our studies show that $P^{(FGS)}$ is poorly calibrated. Estimates can and will frequently be outside the range 0 to 1 in small samples. Therefore, $P^{(FGS)}$ cannot be considered a proportion.

When a covariate is measured with error and used in the time-dependent Cox model, we found the estimate of $P^{(FGS)}$ was, on average, biased toward 0. In the scenarios we considered, the estimate of $P^{(FGS)}$ was highly variable, especially when the marker was a perfect surrogate or a nonsurrogate, corresponding to the extremes of the parameter space for $P^{(FGS)}$. Confidence intervals for $P^{(FGS)}$ were frequently quite wide, enclosing a substantial fraction of the 0 to 1 range.

Smoothing techniques did not effectively reduce the estimated variability of $P^{(FGS)}$ but did effectively reduce the bias in the estimate of $P^{(FGS)}$. Because of the tremendous variability of this statistic, we recommend using it as a summary measure only if the marginal treatment effect was three to four times its standard error. In this situation, the variability of $P^{(FGS)}$ was reduced, but the estimates were biased. Thus, to use $P^{(FGS)}$ to validate a new marker, a trial of substantial size and/or length would need to be conducted to obtain a reasonably precise estimate of the amount of information explained by the intermediate end point. Longer and/or larger trials can be prohibitively expensive and time consuming if they need to be conducted to revalidate old markers or validate new markers for each new class of drugs that is developed.

We found the estimate of $P^{(FGS)}$ for the ACTG019 trial to be $-0.02$, calculated using the OBS method, with 90% confidence interval $(-0.32, 0.28)$. Use of the various smoothing techniques did not significantly alter this estimate. In view of the poor properties of $\hat{P}^{(FGS)}$, we believe it would be unwise to give any strong interpretation of this value other than demonstrating that CD4 is not a perfect surrogate. We believe that other methods of investigating surrogate markers are likely to be more fruitful, for example, a graphical approach [16] or by meta-analysis [18].

It is worth considering if there are circumstances when $P^{(FGS)}$ might be useful and how it might be used. If many events are expected in a short period of time, surrogate markers are unnecessary so $P^{(FGS)}$ is likely to be contemplated only when there is a high proportion of censored observations. Furthermore, implicit in the definition of $P^{(FGS)}$ is the assumption that the treatment arms will

differ in their effect on the real end point. We see no role for $P^{(FGS)}$ if treatment has either no effect or a minor effect on the real end point.

Even if one could establish a high value of $P^{(FGS)}$ for a particular marker, it does not explain how to use that marker as the end point in a trial. One could consider end points such as change in the marker from baseline, or time to cross a specified threshold value, or changes of a certain magnitude considered to be important.

$P^{(FGS)}$ itself does not constitute an end point in a trial; rather it summarizes some analyses of data from the trial. $P^{(FGS)}$ can be calculated after the trial is complete. If $P^{(FGS)}$ is found to be near 1, or just "high enough," one might be prepared to risk using this surrogate as the end point in the next trial of a similar type of drug.

A second use of $P^{(FGS)}$ might be to compare two potential markers. If one marker shows a higher value of $P^{(FGS)}$ than the other marker, then it might be the preferred surrogate end point for the next trial. When two markers are available, a combination of the two may be a better surrogate than either one alone.

A third possible use of $P^{(FGS)}$ is as a guide to stopping a trial (i.e., continually update $P^{(FGS)}$ as the trial evolves); if at some point before the intended conclusion of the trial $P^{(FGS)}$ can be confirmed as high, then switch the primary end point to the surrogate and stop the trial if there is a significant effect of the treatment on the surrogate. We doubt, however, that this procedure would save much time, because the amount of time required to confirm that $P^{(FGS)}$ is high is likely to be just as long as the time required to show that the treatment has an effect on the real end point.

This work was partially supported by National Institutes of Health grants AI07370 and AI29196.

## REFERENCES

1. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, eds. *Clinical Measurement in Drug Evaluation*. New York: Wiley; 1995.
2. Machado SG, Gail MH, Ellenberg SS. On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection. *J Acquir Immune Defic Syndr* 1990;3:1065–1073
3. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;125:605–613.
4. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8:431–440.
5. Freedman LS, Graubard BL, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992;11:167–178.
6. Lin DY, Fischl MA, Schoenfield DA. Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. *Stat Med* 1992;12:835–842.
7. O'Brien WA, Hartigan PM, Martin D, et al. Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. *N Engl J Med* 1996;334:426–431.
8. DeGruttola V, Fleming T, Coombs R. Viral load and response to treatment of HIV [letter]. *N Engl J Med* 1996;334:1671–1672.

9. Lin DY, Fleming TR, De Gruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat Med* 1997;16:1515–1527.
10. Volberding PA, Lagakos SW, Koch MA, et al. Zidovidine in asymptomatic human immunodeficiency virus infection. *N Engl J Med* 1990;322:941–949.
11. Breslow N. Covariance analysis of censored data. *Biometrics* 1974;30:89–99.
12. Taylor JMG, Tan S-J, Detels R, et al. Applications of a computer simulation model of the natural history of CD4 T-cell number in HIV-infected individuals. *AIDS* 1991;5:159–167.
13. Taylor JMG, Cumberland WG, Sy JP. A stochastic model for analysis of longitudinal AIDS data. *J Am Statist Assoc* 1994;89:727–736.
14. Bycott PW, Taylor JMG. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Stat Med* 1998;17:(in press).
15. LaValley MP, DeGruttola V. Models for empirical Bayes estimators of longitudinal CD4 counts. *Stat Med* 1996;15:2289–2305.
16. Tsiatis AA, DeGruttola V, Wolfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *J Am Statist Assoc* 1995;70:27–37.
17. Raboud J, Reid N, Coates RA, et al. Estimating risks of progressing to AIDS when covariates are measured with error. *J. R Statist Soc Ser A* 1993;156:393–406.
18. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med* 1997;10:1965–1982.