Biometrics 67, 1434–1441
December 2011

DOI: 10.1111/j.1541-0420.2011.01608.x

A Shrinkage Approach for Estimating a Treatment Effect Using Intermediate Biomarker Data in Clinical Trials

Yun Li,[*] Jeremy M. G. Taylor, and Roderick J. A. Little

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
[*]*email:* yunlisph@umich.edu

Summary. In clinical trials, a biomarker (S) that is measured after randomization and is strongly associated with the true endpoint (T) can often provide information about T and hence the effect of a treatment (Z) on T. A useful biomarker can be measured earlier than T and cost less than T. In this article, we consider the use of S as an auxiliary variable and examine the information recovery from using S for estimating the treatment effect on T, when S is completely observed and T is partially observed. In an ideal but often unrealistic setting, when S satisfies Prentice's definition for perfect surrogacy, there is the potential for substantial gain in precision by using data from S to estimate the treatment effect on T. When S is not close to a perfect surrogate, it can provide substantial information only under particular circumstances. We propose to use a targeted shrinkage regression approach that data-adaptively takes advantage of the potential efficiency gain yet avoids the need to make a strong surrogacy assumption. Simulations show that this approach strikes a balance between bias and efficiency gain. Compared with competing methods, it has better mean squared error properties and can achieve substantial efficiency gain, particularly in a common practical setting when S captures much but not all of the treatment effect and the sample size is relatively small. We apply the proposed method to a glaucoma data example.

Key words: Auxiliary variable; Biomarker; Missing data; Randomized trials; Ridge regression.

1. Introduction

An intermediate biomarker (S) in a clinical trial that is measured after randomization and is strongly associated with the true endpoint (T) can often provide information about T and hence the effect of the treatment (Z) on T. It is often an intermediate physical or laboratory indicator in a disease progression process and can be measured earlier and is easier to collect than T. Examples of these types of biomarkers include CD4 counts in AIDS, blood pressure in cardiovascular disease, and prostate-specific antigen in prostate cancer studies. In general, S will be a different entity than T, but early measurements are also used as biomarkers for the later measurements, such as the interim height for adult height in girls with Turner Syndrome by Venkatraman and Begg (1999). Different investigators use different terminology for the role of biomarkers. In this article, we call S a surrogate endpoint when the potential use of S is to completely replace T to evaluate whether the treatment is effective (Buyse and Molenberghs, 1998). Alternatively, when S is used to help provide information or enhance the efficiency of the estimator of the treatment effect on T, we call S an auxiliary variable (Fleming et al., 1994). In this article, we focus on the latter role. Intuitively, since S and T are often closely associated, incorporating the information from S in estimating the actual effect of Z on T (denoted by Q) should lead to more efficient estimates, narrower confidence intervals (CIs), and more powerful tests.

A number of authors have explored the role of intermediate biomarkers as auxiliary variables (Murray and Tsiatis, 1996; Faucett, Schenker, and Taylor, 2002). However, the opinions on their value have been mixed, as noted by Cook and Lawless (2001). Correlation has often been the focus of investigations into the extent of efficiency gain from using S to help estimate the treatment effect in a new trial (Buyse et al., 2000; Li and Taylor, 2010). In general, the information recovered from S appears to be very small unless S and T are very highly correlated (Venkatraman and Begg, 1999). In this article, we focus on the relationship between the extent of efficiency gain and the structural relationship among S, T, and Z, defined by the coefficients of a regression of T on S and Z. Even with fixed correlation between S and T given Z, if there is a strong structural relationship among S, T, and Z, a significant efficiency gain from using S is possible.

Here, we focus on a single trial setting where T is partially observed, and S and Z are measured on everyone. Both S and T are continuous and Z is binary. We assume a parametric model for the joint distribution of S and T given Z, and because of the time sequence in which S and T are typically measured, we factor this model as $f(T|S, Z)$ and $f(S|Z)$. We assume linear models, with the full model for $T|S, Z$ given by $T = \beta_0 + \beta_1 S + \beta_2 Z + \beta_3 SZ + \epsilon$, where ϵ is a normally distributed error term. Our goal is to examine the extent of efficiency gain through the use of S as an auxiliary variable rather than as a surrogate variable, but we borrow the terminology of surrogacy to describe the different structural relationships between S and T. In a landmark paper concerning surrogacy, Prentice (1989) called S a perfect surrogate endpoint (PES) when S fully captures the effect of Z on T. For our linear model, this condition becomes $\beta_1 \neq 0$ and $\beta_2 = \beta_3 = 0$. When $\beta_1 \neq 0$ and either $\beta_2 \neq 0$ or $\beta_3 \neq 0$, S explains some, but not all, of the association between T and Z, and S is called a partial surrogate (Wang and Taylor, 2002). More specifically, when $\beta_1 \neq 0$, $\beta_2 \neq 0$, and $\beta_3 = 0$, we call S an additive partial surrogate (APAS); when $\beta_1 \neq 0$, $\beta_2 \neq 0$, and $\beta_3 \neq 0$, we call S an interactive partial surrogate (IPAS). We are interested in estimating the effect of Z on T under these three structural relationships that describe the distribution of T given S and Z.

Our numerical studies suggest that the gain in efficiency from using S as an auxiliary variable depends strongly on whether or not the structural relationship satisfies the PES, APAS, or IPAS assumption. As we will show, if the PES structure is correctly assumed, there is the potential for substantial gains in efficiency. On the other hand, when PES is incorrectly assumed, substantial bias can occur in the estimated treatment effect. Since in practice the validity of PES is uncertain, there is the potential for an adaptive method that realizes this efficiency gain if PES is true or approximately true, but also limits the bias if PES is clearly not true. One such strategy is to apply model selection methods, using p-values to judge whether β_2 and/or β_3 equal to 0 and then fitting the selected model. However, this common practice ignores the model uncertainty and can lead to high type I errors (Albert et al., 2001) and substantial prediction error. From a biological point of view, there are often multiple pathways through which the treatment can affect T, and a marker seldom captures all the effects on T. On the other hand, partial surrogates that capture much but not all of the treatment effect are very plausible. For example, a biomarker can be a good partial surrogate if it is in one of the few important mechanistic pathways between Z and T or it can explain a large amount of the treatment effect on T. In these settings, we propose an adaptive approach using a targeted ridge regression method that shrinks β_2 and β_3 toward zero by an amount that is supported by the data. This method is a compromise between the perfect surrogacy and partial surrogacy models and provides better mean squared error properties by striking a data-driven balance between bias and variance.

The article is organized as follows. In Sections 2 and 3, we conduct analytic and numerical studies to explore the efficiency gain from S under the various structural assumptions. In Section 4, we introduce the generalized ridge regression method. In Section 5, we describe simulations comparing this shrinkage approach with competing methods including model selection and inverse probability weighting (IPW). In Section 6, we apply the proposed method to a glaucoma data set. In Section 7, we summarize and discuss our findings.

2. Treatment Effect Estimation and Surrogacy Assumptions

Suppose that the number of study participants is $n = n_0 + n_1$ with n_0, n_1 in the $Z = 0$, 1 groups, respectively. The biomarker, S, is measured on all n patients; T is available for a subset of r_j patients in the $Z = j$ group ($j = 0$, 1) and $r = r_0 + r_1$. The fraction of the subjects for whom T is not observed is $p = 1 - r/n$.

When S is an IPAS, we assume that the joint distribution $f(T_i, S_i \,|Z_i)$ for participant i is given by two models:

$$\begin{aligned} T_i &= \beta_0 + \beta_1 S_i + \beta_2 Z_i + \beta_3 S_i Z_i + \epsilon_{ti} \\ S_i &= \alpha_0 + \alpha_1 Z_i + \epsilon_{si}, \end{aligned} \tag{1}$$

where $\epsilon_{ti} \sim N(0, \sigma^2_{t\,|\,s})$ and $\epsilon_{si} \sim N(0, \sigma^2_{ss})$. For this model, the marginal average treatment effect is

$$\begin{aligned} Q_{IPAS} &= E(T \,|\, Z = 1) - E(T \,|\, Z = 0) \\ &= EE(T \,|\, S, Z = 1) - EE(T \,|\, S, Z = 0) \\ &= \beta_1\alpha_1 + \beta_2 + \beta_3\alpha_0 + \beta_3\alpha_1. \end{aligned}$$

We assume the missing data on T are missing at random (MAR; Little and Rubin, 2002) for which the probability of missingness depends only on observed data measures. In our setting, this implies that we consider the missingness depends on S and Z only. Under MAR, the likelihood of $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \alpha_0, \alpha_1, \sigma^2_{t\,|\,s}, \sigma^2_{ss})$ based on the observed data is given by: $L(\theta \,|\, S, T, Z) = \prod_{i=1}^{r} f(T_i \,|\, S_i, Z_i, \theta) \prod_{i=1}^{n} f(S_i \,|\, Z_i, \theta)$. The estimate of Q_{IPAS}, $\hat{Q}_{IPAS}$, can be obtained by substituting maximum likelihood estimates (MLEs) for the unknown parameters. The large sample covariance matrix of $\hat{\theta}$ can be calculated as the inverse of the observed information matrix $I^*_{IPAS}(\theta)$. Let $D_{IPAS}(Q) = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \beta_2}, \frac{\partial Q}{\partial \beta_3}, \frac{\partial Q}{\partial \alpha_0}, \frac{\partial Q}{\partial \alpha_1}) = (0, \alpha_1, 1, \alpha_0 + \alpha_1, \beta_3, \beta_1 + \beta_3)$. The asymptotic variance of $\hat{Q}_{IPAS}$ can be calculated using the delta method as

$$V(\hat{Q}_{IPAS}) = D_{IPAS}(Q)^T I^*_{IPAS}(\theta)^{-1} D_{IPAS}(Q).$$

Its estimate $\hat{V}(\hat{Q}_{IPAS})$ can be obtained by replacing θ with the MLE $\hat{\theta}$. Under the missing completely at random (MCAR) assumption, for which the probability of missingness does not depend on observed or unobserved data measures, Little and Rubin (2002) noted that $V(\hat{Q}_{IPAS})$ can be approximated by

$$\frac{\sigma^2_{tt0}}{r_0}\left(1 - \rho_0^2 \frac{n_0 - r_0}{n_0}\right) + \frac{\sigma^2_{tt1}}{r_1}\left(1 - \rho_1^2 \frac{n_1 - r_1}{n_1}\right), \tag{2}$$

where ρ_0 and ρ_1 denote the correlation between S and T in the $Z = 0$, 1 group, respectively; σ^2_{tt0} and σ^2_{tt1} refer to $V(T|Z = 0)$ and $V(T|Z = 1)$, respectively. Calculations in the Web Appendix show that $\rho_0^2 = \frac{\beta_1^2 \sigma^2_{ss}}{\sigma^2_{t\,|\,s} + \beta_1^2 \sigma^2_{ss}}$, $\rho_1^2 = \frac{(\beta_1 + \beta_3)^2 \sigma^2_{ss}}{\sigma^2_{t\,|\,s} + (\beta_1 + \beta_3)^2 \sigma^2_{ss}}$, $\sigma^2_{tt0} = \sigma^2_{t\,|\,s}/(1 - \rho_0^2)$, and $\sigma^2_{tt1} = \sigma^2_{t\,|\,s}/(1 - \rho_1^2)$. The approximation (2) shows that the correlations, the fractions of missingness, σ^2_{tt0}, and σ^2_{tt1} are important factors that impact the variance of $\hat{Q}_{IPAS}$.

If T is fully observed, without any distributional assumption, the estimated treatment effect would be $\hat{Q}_{ALL} = \sum_{i=1}^{n_1} T_i/n_1 - \sum_{i=1}^{n_0} T_i/n_0$ with variance $V(\hat{Q}_{ALL}) = \sigma^2_{tt0}/n_0 + \sigma^2_{tt1}/n_1$. When T is partially observed, the treatment effect estimated solely based on the observed T is $\hat{Q}_{CC} = \sum_{i=1}^{r_1} T_i/r_1 - \sum_{i=1}^{r_0} T_i/r_0$ and its variance is $V(\hat{Q}_{cc}) = \sigma^2_{tt0}/r_0 + \sigma^2_{tt1}/r_1$.

When S is an APAS, the treatment effect on T is $Q_{APAS} = \beta_2 + \beta_1\alpha_1$. Under the MAR assumption, the asymptotic variance of $\hat{Q}_{APAS}$ can be calculated in the same way as that of $\hat{Q}_{IPAS}$, but noting that I^*_{APAS} is a 5 × 5 information matrix. Under the MCAR assumption, the large sample variance $V(\hat{Q}_{APAS})$ can also be approximated by

$$\frac{\sigma^2_{tt}}{r_0}\left(1 - \rho^2 \frac{n_0 - r_0}{n_0}\right) + \frac{\sigma^2_{tt}}{r_1}\left(1 - \rho^2 \frac{n_1 - r_1}{n_1}\right), \tag{3}$$

where $\rho^2 = \rho_0^2 = \rho_1^2 = \frac{\beta_1^2\sigma_{ss}^2}{\sigma_{t\,|\,s}^2 + \beta_1^2\sigma_{ss}^2}$ and $\sigma_{tt}^2 = \sigma_{t\,|\,s}^2/(1-\rho^2)$. When the percent of missingness and σ_{tt} are fixed, ρ^2 is the single most important factor that determines the extent of efficiency gain from S.

When S is a perfect surrogate, the marginal treatment effect on T is $Q_{PES} = \beta_1\alpha_1$. Under the MAR assumption, the calculation of the asymptotic variance $V(\hat{Q}_{PES})$ follows closely those for $V(\hat{Q}_{IPAS})$ and $V(\hat{Q}_{APAS})$ with I^*_{PES} being a 4 × 4 information matrix. Under the MCAR assumption, as shown in the Web Appendix, the asymptotic variance can be approximated by

$$\frac{\alpha_1^2\sigma_{tt}^2(1-\rho^2)}{r\sigma_{ss}^2 + \left(r_1 - \dfrac{r_1^2}{r}\right)\alpha_1^2} + \frac{\beta_1^2\sigma_{ss}^2}{n_1 - \dfrac{n_1^2}{n}}. \qquad (4)$$

Under the PES assumption, the factors that impact the efficiency gain include not only the correlation and the factors associated with the correlation, but also α_1.

3. Information Recovery and Surrogacy Assumptions

We conduct numerical studies based on the asymptotic variances to examine the impact of different factors and different surrogacy assumptions on the efficiency gain from S. We assume that $n_0 = n_1 = 500$ and the missingness mechanism is MCAR. The true model is PES, APAS, or IPAS. We choose different combinations of θ, p, $\sigma_{t\,|\,s}^2$, and σ_{ss}^2. The variances of the estimated treatment effect on T are calculated for the five different estimates as $V(\hat{Q}_{ALL})$, $V(\hat{Q}_{CC})$, $V(\hat{Q}_{IPAS})$ in (2), $V(\hat{Q}_{APAS})$ in (3), and $V(\hat{Q}_{PES})$ in (4). We compute the relative efficiency (RE) defined by the ratios of the variance of $V(\hat{Q}_{ALL})$ to other variance estimates.

Numerical studies show that generally there is some improvement in the precision of $\hat{Q}$ by incorporating S (see Web Appendix). We plot the RE against ρ^2 and α_1 in Figure 1 when the true model is PES. When the fitted model assumes IPAS or APAS, the higher the correlation between S and T, the higher the extent of efficiency gain from S. When we fit PES, the amount of information recovery from S depends on the correlation and α_1. When everything else holds constant, the smaller the value of α_1, the higher the amount of information recovery from S. When the correlation increases, the extent of efficiency gain also increases, and reaches a maximum RE (larger than 1) compared with ALL when ρ^2 is approximately 0.8 in this setting of true parameter values.

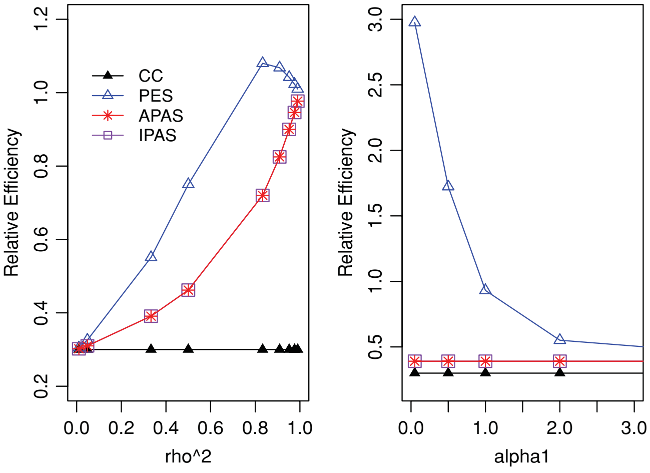

Figure 1. Asymptotic relative efficiency (RE) compared with that obtained from original data (ALL). Left: $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma_{t\,|\,s}^2 = 1$, $p = 0.7$, and ρ^2 varies. Right: $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\sigma_{t\,|\,s}^2 = 1$, $p = 0.7$, $\sigma_{ss}^2 = 0.5$, $\rho^2 = 0.333$, and α_1 varies ($n = 1000$). This figure appears in color in the electronic version of this article.

The extent of efficiency gain also highly depends on which model we fit. When ρ^2 and α_1 hold constant, fitting either an IPAS or APAS model can result in a similarly modest amount of information recovery except when the correlation is unusually high. For large sample sizes, even though IPAS has one additional parameter compared to APAS, they have similar efficiencies. When the sample size is smaller (e.g., $n_1 = n_2 = 60$ in Figure 2), APAS gives more efficient estimates than IPAS. By fitting the PES model, however, we can uniformly improve the efficiency gain to a much greater extent. On the other hand, if we make an incorrect PES assumption, the estimates of the marginal treatment effect can be substantially biased (Figure 2 and Web Tables 3–8). Thus, the surrogacy assumption plays a central role in both the bias and the extent of efficiency gain. In the next section, we propose a shrinkage approach, a generalized ridge regression method (denoted by Ridge), that avoids the need to make the surrogacy assumptions, sacrifices some bias to gain efficiency in a data adaptive way, and gives better mean squared error properties.

4. Generalized Ridge Regression

We first consider the situation when $\beta_3 = 0$. As explained in the Introduction, a biomarker is rarely a perfect surrogate in practice, but it is more common for S to be a strong partial surrogate and capture a large portion of the treatment effect on T. In these settings, a reasonable assumption is that β_2 is close to but not exactly 0. We impose a prior distribution on β_2 such that $\beta_2 \sim N(0, \sigma_{b_2}^2)$, where $\sigma_{b_2}^2$ is used to capture the uncertainty about the departure from the perfect surrogacy assumption. By assuming this prior distribution, the generalized ridge regression model induces a shrinkage effect on $\hat{\beta}_2$, which will data-adaptively shrink $\hat{\beta}_2$ toward 0 with the amount of shrinkage determined by how much S is close to being a perfect surrogate. Note that the frequentist counterpart of the ridge regression is an L2 penalized regression. Here, we describe two estimates, the first is a full Bayes version, where we treat $\sigma_{b_2}^2$ as a hyperparameter with its own prior distribution; the second is an empirical Bayes version, where $\sigma_{b_2}^2$ is estimated directly from the data.

4.1 *Full Bayes Estimator*

When S is APAS, the joint distribution $f(T_i,\ S_i\,|\,Z_i)$ is expressed by two models in (1) with $\beta_3 = 0$. We assume $\beta_2 \sim N(0, \sigma_{b_2}^2)$. We specify a proper but diffuse prior of $N(0, a = 100^2)$ for $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ and Gamma$(c,\ d)$ for $(\sigma_{t\,|\,s}^{-2}, \sigma_{b_2}^{-2}, \sigma_{ss}^{-2})$, where the mean and variance of Gamma$(c,\ d)$ are cd and cd^2, and $c = 0.001$, $d = 1000$. We use Gibbs sampling to draw from the conditional posterior distributions (see the Web Appendix) and obtain the joint posterior distributions of the parameters. We can then easily obtain the posterior distribution of the treatment effect estimate $(\beta_2 + \beta_1\alpha_1)$ and use the posterior mean as the estimate for Q, $\hat{Q}_{Ridge-FB}$ and the variance of the posterior distribution as the variance estimate, $\hat{V}(\hat{Q}_{Ridge-FB})$.

4.2 *Empirical Bayes Estimator*

The advantage of the full Bayes estimation is that it accounts for all the uncertainty associated with estimating every parameter. However, it is computationally intensive, particularly for large sample sizes, so we consider an alternative empirical Bayes estimator that is faster to compute.

We first consider the situation when $\beta_3 = 0$. The model $T|S$, Z is given by $T_i = \beta_0 + \beta_1 S_i + \beta_2 Z_i + \epsilon_{ti}$, where $\epsilon_{ti} \sim N(0, \sigma^2_{t\,|\,s})$. We specify the prior for β_2 as $N(0, \sigma^2_{b_2})$. Let $\beta^T = (\beta_0, \beta_1, \beta_2)$, $X_t = (1,\ S,\ Z)$, and $K = \text{diag}(0,\ 0,\ k_2)$ where $k_2 = \sigma^2_{t\,|\,s}/\sigma^2_{b_2}$. Suppose $\sigma^2_{b_2}$ and $\sigma^2_{t\,|\,s}$ are known and noninformative prior distributions are assumed for β_0 and β_1. The posterior distribution of β follows a normal distribution with mean and variance:

$$\begin{aligned} \mathrm{E}(\hat{\beta}\,|\,X_t, T) &= \left(X_t^T X_t + K\right)^{-1} X_t^T T, \\ \mathrm{V}(\hat{\beta}\,|\,X_t, T) &= \left(X_t^T X_t + K\right)^{-1} \sigma^2_{t\,|\,s}. \end{aligned} \tag{5}$$

In practice, $\sigma^2_{b_2}$ and $\sigma^2_{t\,|\,s}$ are unknown and are estimated directly from the data. Given β_2, $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\beta_2})$, we obtain the joint distribution of $(\hat{\beta}_2, \beta_2)$ by multiplying the densities of $\hat{\beta}_2\,|\,\beta_2$ and β_2 together, yielding the marginal density of $\hat{\beta}_2$ as $N(0, \sigma^2_{\beta_2} + \sigma^2_{b_2})$. The quantity $\sigma^2_{\beta_2}$ can be estimated from the maximum likelihood fit to $T_i = \beta_0 + \beta_1 S_i + \beta_2 Z_i + \epsilon_{ti}$. Since $E(\hat{\beta}_2) = 0$, $E\{(\hat{\beta}_2)^2\} = \sigma^2_{\beta_2} + \sigma^2_{b_2}$, and an estimate of $\sigma^2_{b_2}$ is $\max\{0, (\hat{\beta}_2)^2 - \hat{\sigma}^2_{\beta_2}\}$, alternatively $\hat{\beta}_2^2$ can be considered as a computationally easier and more conservative estimate of $\sigma^2_{b_2}$. The two estimates of $\sigma^2_{b_2}$ give similar results in our simulations, thus we present results using $\hat{\beta}_2^2$. We then fit the model $T|S$, Z to get an MLE of $\sigma^2_{t\,|\,s}$. Then, we obtain the empirical Bayes estimate of β and its variance by replacing $\sigma^2_{t\,|\,s}$ and $\sigma^2_{b_2}$ in (5) with their estimates.

Let $\alpha^T = (\alpha_0, \alpha_1)$ and $X_s = (1,\ Z)$, then the estimate $\hat{\alpha}$ follows a normal distribution with mean and variance:

$$\begin{aligned} \mathrm{E}(\hat{\alpha}\,|\,X_s, S) &= \left(X_s^T X_s\right)^{-1} X_s^T S, \\ \mathrm{V}(\hat{\alpha}\,|\,X_s, S) &= \left(X_s^T X_s\right)^{-1} \sigma^2_{ss}. \end{aligned}$$

We obtain the variance of $\hat{\alpha}$ by replacing σ^2_{ss} in $\mathrm{V}(\hat{\alpha})$ with its estimate.

Let $D_{Ridge-EB}(Q) = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \beta_2}, \frac{\partial Q}{\partial \alpha_0}, \frac{\partial Q}{\partial \alpha_1}) = (0, \alpha_1, 1, 0, \beta_1)$. The treatment effect estimate $\hat{Q}_{Ridge-EB}$ follows a normal distribution with mean and variance estimated by:

$$\begin{aligned} \hat{E}(\hat{Q}_{Ridge-EB}) &= \hat{\beta}_1 \hat{\alpha}_1 + \hat{\beta}_2, \\ \hat{\mathrm{V}}(\hat{Q}_{Ridge-EB}) &= D(\hat{Q}_{Ridge-EB})^T \begin{bmatrix} \hat{\mathrm{V}}(\hat{\beta}) & 0 \\ 0 & \hat{\mathrm{V}}(\hat{\alpha}) \end{bmatrix} D(\hat{Q}_{Ridge-EB}), \end{aligned}$$

where the parameter estimate of β is the empirical Bayes estimate.

For both full Bayes and empirical Bayes versions of the generalized ridge regression, we can easily extend the method to the situation when $\beta_3 \neq 0$ by assuming an additional prior distribution of $N(0, \sigma^2_{b_3})$ for β_3 and following analogous procedure to those described above.

5. Simulation Studies

5.1 *The Setup*

We conduct extensive simulations to examine the proposed methods and compare them with competing methods. We generate 400 data sets using the models in (1) with the following true parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, and $\sigma^2_{t\,|\,s} = 1$. We first set $\beta_3 = 0$. We vary β_2 to reflect different degrees of departure from the perfect surrogacy assumption. Each data set contains the observations from either 60, 120, or 480 subjects per treatment group. We observe all of S, but only 20% of T ($p = 0.8$). The missing data mechanism is MCAR. For each method and each data set, we obtain the point estimate of Q and the corresponding estimated standard error (SE), and an indicator variable for the coverage for whether or not the 95% CI contains the true value. We measure each method's performance by the average empirical bias (Bias), the average SE, the empirical standard deviation (ESD), the empirical mean squared error (MSE = ESD^2 + Bias^2) and the coverage rate (CR). For the Ridge-FB method, the SE is given by the standard deviation of the posterior distribution.

Many additional simulations are also performed. We vary the values of α_1, β_1, and $\sigma^2_{t\,|\,s}$, and we also conduct all the simulations under $\beta_3 \neq 0$. We examine the properties of these methods when there is no missingness. Even though MCAR is often the primary missing mechanism in clinical trials, we repeat the simulations under the more general MAR assumption for missingness by allowing the probability of missingness to depend on S and Z, specifically, $\text{logit}(p) = \gamma_0 + \gamma_1 Z + \gamma_2 S$ with $\gamma_0 = 0.5$, $\gamma_1 = 0.2$, and $\gamma_2 = 0.18$. In addition, to examine how the methods perform under different degrees of correlations, we repeat the simulations listed above with $\beta_3 = 0$ under various σ^2_{ss} ranging from 0.1, 0.5, 1, to 5 that correspond to $\rho^2 = 0.1$, 0.33, 0.5, 0.83, respectively.

5.2 *Simulation Results*

Figure 2 shows the MSE and Bias of $\hat{Q}_{Ridge-FB}$ relative to those of $\hat{Q}_{PES}$, $\hat{Q}_{APAS}$, and $\hat{Q}_{IPAS}$ and illustrates the data-adaptive property of Ridge. For completeness, we also include the MSE and Bias of $\hat{Q}_{CC}$. Since the simulations are conducted under MCAR, $\hat{Q}_{CC}$ and $\hat{Q}_{ALL}$ only differ by a multiplicative factor of r/n. The estimated variances $\hat{V}(\hat{Q}_{IPAS})$, $\hat{V}(\hat{Q}_{APAS})$, and $\hat{V}(\hat{Q}_{PES})$ are calculated based on the observed information matrix. When $\beta_2 = 0$, fitting an APAS or IPAS model can result in much larger MSEs and smaller efficiency gains relative to fitting a PES model. All methods give unbiased estimates. When β_2 departs further from 0, fitting a PES model leads to increasingly larger Bias and MSE compared to fitting APAS and IPAS models. When β_2 is 0 or close to 0, Ridge-EB retains a lot of the efficiency gain achieved by fitting a PES model without introducing appreciable bias. When β_2 is much different from 0, Ridge-EB gives estimates with MSEs similar to those obtained by fitting an APAS or IPAS model without the substantial bias resulted from fitting an incorrect PES model. Hence, Ridge-EB appears to strike a good balance between efficiency gain and bias, depending on the true nature of the relationship between S and T. This illustrates the data-adaptive capacity of Ridge-EB. These properties are more pronounced in small samples than in large samples; for example, β_2 can be relatively larger for Ridge-EB to retain a

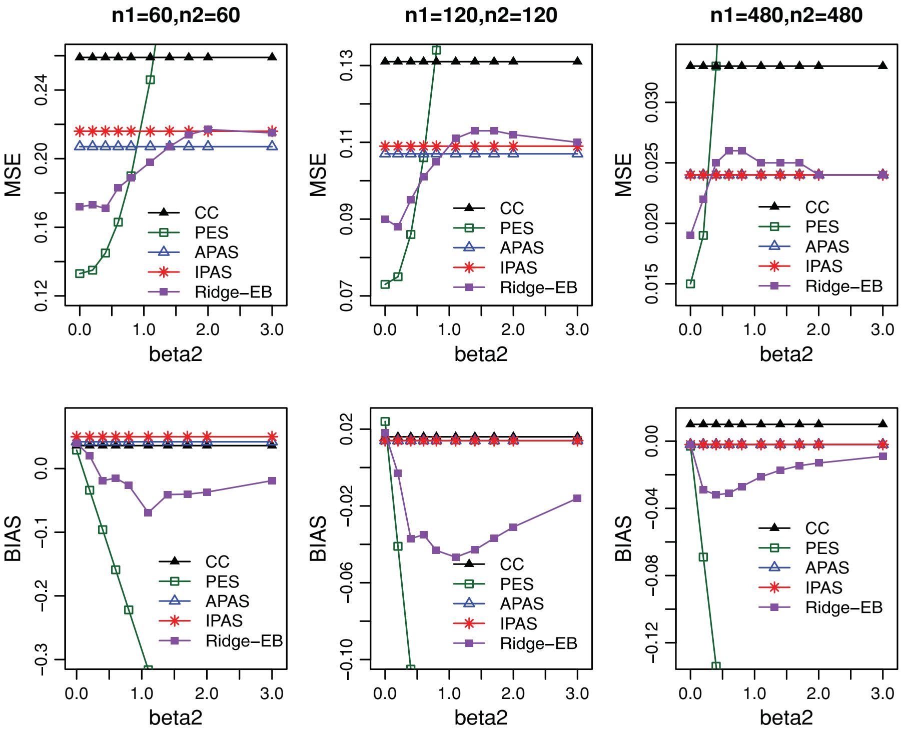

Figure 2. Comparison of Ridge-EB, IPAS, APAS, PES, and CC in terms of MSE and bias by sample size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_{t\,|\,s} = 1$, $\rho^2 = 0.333$, and $p = 0.8$. This figure appears in color in the electronic version of this article.

large amount of efficiency gain in smaller samples ($n_1 = n_2 = 60$) than that in larger samples ($n_1 = n_2 = 480$). Note that as β_2 deviates further from 0, the shape of MSE line for Ridge-EB generally increases before decreasing, a property that is largely driven by the bias.

We then compare Ridge-EB, Ridge-FB with alternative methods, including a two-stage model selection method (MdlSel) and an IPW method (Horvitz and Thompson, 1952). These are all methods that might be used in practice. The MdlSel method first tests that model among APAS, IPAS, and PES is not contradicted by the data. Specifically, we used the backward elimination approach (selection criterion: p-value <0.05) to select the model. The selected model is then used as the correct model to obtain the estimate of Q ($\hat{Q}_{MdlSel}$) and its variance $\hat{V}(\hat{Q}_{MdlSel})$. The IPW method is mostly used to reduce bias but can also be applied to utilize the information from auxiliary variables when T is partially observed. Let Δ_i be the indicator for whether T_i is observed or not (1 for being observed and 0 otherwise). Denote $\pi_i = Pr(\Delta_i = 1)$. We obtain the estimated π_i ($\hat{\pi}_i$) by fitting the saturated model: $\text{logit}\{Pr(\Delta_i = 1)\} = \delta_0 + \delta_1 S_i + \delta_2 Z_i + \delta_3 S_i Z_i$. The treatment effect can be estimated by: $\hat{Q}_{IPW} = \{\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} T_i I(Z_i = 1)/\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} I(Z_i = 1)\} - \{\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} T_i I(Z_i = 0)/\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} I(Z_i = 0)\}$.

The comparisons of the MSE and CR properties of $\hat{Q}_{Ridge-EB}$, $\hat{Q}_{Ridge-FB}$, $\hat{Q}_{MdlSel}$, $\hat{Q}_{IPW}$, and $\hat{Q}_{CC}$ are illustrated in Figure 3. On average, CC gives the highest MSEs. Both Ridge-FB and Ridge-EB are data adaptive. When the sample size is large, Ridge-FB and Ridge-EB have very similar performances. However, there are subtle differences, particularly in small samples where Ridge-EB gives below-nominal-level CRs and Ridge-FB offers uniformly higher and closer-to-nominal CRs than any other method. Unlike its competitors, Ridge-FB accounts for all the uncertainty associated with estimating the variance parameters. Generally, there is more shrinkage toward 0 using Ridge-FB than using Ridge-EB and the MSEs from Ridge-FB are often smaller than Ridge-EB when β_2 is 0 or close to 0; however, Ridge-EB is more robust and less biased when there is a large departure in β_2 from 0 and often leads to smaller MSEs than Ridge-FB in these situations. MdlSel is also data adaptive, but, unlike Ridge, its performance depends on the available power to choose the correct model. When the power is small (e.g., when β_2 and β_3 are moderate in size, or when the sample size is small), Ridge can achieve smaller MSEs than MdlSel. On the other hand, when the power is sufficient (e.g., when the size is 120 or 480 per group and when β_2 and β_3 are either ≈ 0 or very large), MdlSel and Ridge have similar performances. In general, MdlSel underestimates the variance, more so in smaller samples which results in lower-than-nominal-level CRs. The IPW method does not have the data-adaptive property and cannot take advantage of the various plausible surrogacy assumptions. Regardless of the magnitude of β_2, the amount of efficiency gain from utilizing S to estimate Q_{IPW} stays the same. When β_2 is close to 0, Ridge has a clear advantage over IPW and gives considerably smaller MSEs particularly for small sample sizes. The biases of these estimates can be found in Web Tables 5 and 6. The Ridge methods often result in estimates with larger biases than CC; they also give larger biases than IPW except for in very small samples. For

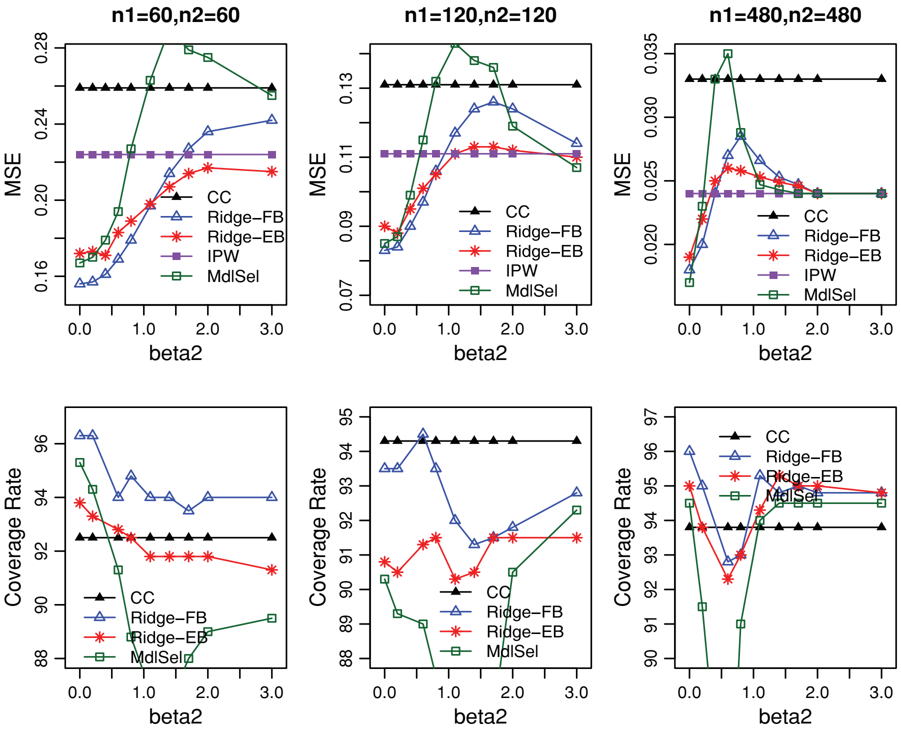

Figure 3. Comparison of Ridge-EB, Ridge-FB, MdlSel, IPW, and CC in terms of MSE and coverage rate by sample size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_{t\,|\,s} = 1$, $\rho^2 = 0.333$, and $p = 0.8$. This figure appears in color in the electronic version of this article.

MdlSel, the extent of bias relative to Ridge also depends on the available power to choose the correct model.

Additional simulation results can be found in the Web Appendix. With different values of α_1, β_1, and $\sigma^2_{t\,|\,s}$, the findings are similar as above. When $\beta_3 \neq 0$, the results show similar patterns for all methods considered. With different correlations, the findings across simulations are also very similar; in addition, we find that the greater the magnitude of the efficiency gain can achieve from PES compared with APAS and IPAS under $\beta_2 = \beta_3 = 0$, the greater the amount of efficiency gain Ridge can retain. When there is no missingness, the findings regarding PES, APAS, IPAS, MdlSel, and Ridge are similar to those given above but IPW is not applicable. When the missingness depends on S and Z under MAR, the estimates from CC and PES are prone to large biases; however, the properties of APAS, IPAS, IPW, MdlSel, and Ridge are similar to those under MCAR. As a reviewer points out, missingness may be explained by covariates other than S and Z and if so, we need to incorporate these other covariates in our models to obtain valid estimates.

6. Application to a Glaucoma Study

We apply these methods to data from the Collaborative Initial Glaucoma Treatment Study (CIGTS) (Musch et al., 2009). Glaucoma is a group of diseases that cause vision loss and is a leading cause for blindness. Elevated pressure in the eyes (i.e., intraocular pressure, IOP), is a major risk factor of glaucoma. The Advanced Glaucoma Intervention Study (AGIS) demonstrated that when IOP reduction from baseline is substantial, progression of visual field loss can be prevented (Musch et al., 2009). The CIGTS is a randomized trial to compare the effects of two initial treatment strategies, immediate filtration surgery ($Z = 1$) and medications ($Z = 0$), on reducing IOP for newly diagnosed open-angle glaucoma patients. Patients were enrolled between 1993 and 1997. The IOP level (in mmHg) has been measured at different time points following randomization. We define the IOP measurements at the 102nd month as T and IOP at the 12th month as S. Due to dropout, there are many fewer patients at the later periods than at the earlier periods. A total of 160 patients have IOP measured at both months 12 and 102, and 413 patients have IOP measured only at month 12. We fit a logistic regression for the probability of missingness that is found not to be significantly associated with either S or Z. The correlation between S and T is 0.456. Summary statistics are presented in Table 1.

Table 1
Summary statistics from CIGTS data. IOP at the 102nd month is the true endpoint and IOP at the 12th month is the biomarker.

	Medicine	Surgery
IOP observed at 12th and 102nd month		
Number of patients	86	74
IOP at 12th month: mean (SE)	17.9 (3.29)	14.1(4.96)
IOP at 102nd month: mean (SE)	17.5 (4.67)	15.1 (4.61)
IOP missing at 102nd month		
Number of patients	206	207
IOP at 12th month: mean (SE)	18.2 (3.80)	14.3 (5.19)

Table 2

Quantity of interest: difference in the IOP reduction at the 102nd month between medicine and surgery treatments. Estimates from eight methods are presented here. Note that the CI from IPW is obtained using bootstrapping and the p-value is calculated as the probability of an observation from a standard normal distribution that is less or equal to the ratio of the estimate over its standard error (*or the posterior standard deviation when Ridge-FB is considered*)*. IOP at the 102nd month as true endpoint and IOP at the* 12*th month as the biomarker.*

Estimation method	Estimate	95 % CI	CI width	p-value
CC	−2.391	(−3.844, −0.937)	2.907	0.00058
IPW	−2.387	(−3.726, −1.048)	2.678	0.00024
IPAS	−2.419	(−3.792, −1.046)	2.746	0.00028
APAS	−2.400	(−3.765, −1.034)	2.731	0.00029
PES	−1.833	(−2.490, −1.176)	1.315	2.30×10^{-9}
MdlSel	−1.833	(−2.490, −1.176)	1.315	2.30×10^{-9}
Ridge-EB	−2.094	(−3.138, −1.049)	2.089	0.000047
Ridge-FB	−2.019	(−3.033, −1.006)	2.027	0.000043

The $S|Z$ model is based on all 413 patients and the $T|S, Z$ models are based on 160. By assuming IPAS, we obtain the MLEs and their 95% CIs: $\hat{\beta}_1 = 0.61$ (0.012, 1.20), $\hat{\beta}_2 = 0.87$ (−4.99, 6.74), and $\hat{\beta}_3 = -0.094$ (−0.44, 0.25). Assuming APAS, we have: $\hat{\beta}_1 = 0.45$ (0.29, 0.61), $\hat{\beta}_2 = -0.69$ (−2.16, 0.78). Assuming PES, we have $\hat{\beta}_1 = 0.48$ (0.33, 0.63). While the model selection method supports the PES assumption, there is considerable uncertainty about the validity of that assumption because the number of complete cases is relatively small and power is limited. However, the preliminary analysis implies that S can capture most of the treatment effect on T. Table 2 shows the estimates of the treatment difference, their 95% CIs and p-values. The Ridge method assumes $\beta_3 = 0$. For Ridge-FB, we choose $c = 0.001$ and $d = 1000$ in the prior distributions. Although the treatment difference between two groups is statistically significant simply based on CC without using S, we can investigate the properties of different methods based on the CIs and p-values. Fitting either the IPAS or APAS model or applying the IPW method results in CIs with width slightly narrower than that from the CC method, suggesting limited efficiency gain from utilizing S. Fitting the PES model leads to substantial efficiency gain; however, the estimate is quite different from others, perhaps suggesting bias from failure of the PES assumption. Results from fitting Ridge-FB and Ridge-EB are comparable, giving estimates between those of IPAS and PES, with lower variances than IPAS. The results illustrate the data-adaptive and bias-variance trade-off feature of the Ridge methods.

7. Discussion

In this article, we propose the use of generalized ridge regression to incorporate the information from S to estimate the treatment effect on T when the underlying relationship between the biomarker and the true endpoint is not fully known. Without the need to make surrogacy assumptions, ridge regression can directly take advantage of the structural relationship between S and T, and increase the information recovery from S and, hence increase precision. When S captures much of the treatment effect, the generalized ridge regression method can retain most of the considerable efficiency gain achieved under the perfect surrogacy assumption. When S only captures a modest amount of the treatment effect, our method can achieve efficiency comparable to that under partial surrogacy assumptions, while limiting the bias resulting from an incorrect perfect surrogacy assumption. Our method achieves better mean squared error properties by data-adaptively making the bias and variance trade-off, particularly in a common setting when S captures most but not all of the treatment effect and the sample size is relatively small. Note that although generalized ridge regression provides a biased estimator of the treatment effect in finite samples, the estimator is consistent and the bias goes to zero when the sample sizes are infinitely large.

The ridge regression methods outperform the model selection procedure in terms of MSE, bias, and CR in situations where the power to detect the correct assumption is relatively small and the uncertainty of a model selection procedure is very large. Unlike the model selection method, the ridge regression method does not remove any variable, so it cannot achieve full efficiency when the true parameter β_2 is exactly equal to 0. However, this may not be a serious limitation as previous empirical studies have shown that it is unlikely for S to be a perfect surrogate (Fleming and DeMets, 1996).

Utilizing S in predicting a treatment effect when T is partially observed is essentially a missing data problem. Compared with the generalized ridge regression, the IPW method is robust; however, it requires us to model the probability of missingness, and it neither has the variance-bias trade-off feature of the ridge regression nor directly takes advantage of the nature of the relationship between S and T. Hence, when S is close to being perfect surrogate, our ridge method can give smaller MSEs and achieve more efficiency gain than IPW. A comparison with the alternative IPW methods (Scharfstein, Rotnitzky, and Robins, 1999) is also worthy of investigation.

Many extensions of the generalized ridge regression method can be made in the biomarker context. When multiple biomarkers are considered, there could be even stronger motivation for the use of a ridge regression method, since a greater percentage of the treatment effect may be captured by the biomarkers. The idea can also be extended to the cases when S and T are different data types, such as time-to-event data. In summary, generalized ridge regression is an area worthy of further study and implementation in the biomarker context.

8. Supplementary Materials

Web Appendices and Tables referenced in Sections 2, 3, 4.1, and 5.2 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors would like to thank Dr Bhramar Mukherjee for helpful discussion and Dr Brenda Gillespie for providing us with the CIGTS data. This research was partially supported by National Institutes of Health Grant CA129102.

References

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology* **154,** 687–693.

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). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1,** 49–67.

Cook, R. J. and Lawless, J. F. (2001). Some comments on efficiency gains from auxiliary information for right-censored data. *Journal of Statistical Planning and Inference* **96,** 191–202.

Faucett, C. L., Schenker, N., and Taylor, J. M. G. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* **58,** 37–47.

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

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials with potential applications in cancer and AIDS research. *Statistics in Medicine* **13,** 955–968.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

Murray, S. and Tsiatis, A. A. (1996). Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* **52,** 137–151.

Musch, D. C., Gillespie, B. W., Lichter, P. R., Niziol, L. M., and Janz, N. K. (2009). Visual field progression in the collaborative initial glaucoma treatment study: The impact of treatment and other baseline factors. *Ophthalmology* **116,** 200–207.

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

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association* **94,** 1096–1120.

Venkatraman, E. S. and Begg, C. B. (1999). Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics* **55,** 1171–1176.

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

Received May 2010. *Revised March* 2011.
Accepted March 2011.