

# ESTIMATING THE PROPORTION OF TREATMENT EFFECT EXPLAINED BY A SURROGATE MARKER

D. Y. LIN,<sup>1,\*</sup> T. R. FLEMING<sup>1</sup> AND V. DE GRUTTOLA<sup>2</sup>

<sup>1</sup>*Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, U.S.A.*

<sup>2</sup>*Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.*

## SUMMARY

In this paper, we measure the extent to which a biological marker is a surrogate endpoint for a clinical event by the proportional reduction in the regression coefficient for the treatment indicator due to the inclusion of the marker in the Cox regression model. We estimate this proportion by applying the partial likelihood function to two Cox models postulated on the same failure time variable. We show that the resultant estimator is asymptotically normal with a simple variance estimator. One can construct confidence intervals for the proportion by using the direct normal approximation to the point estimator or by using Fieller's theorem. Extensive simulation studies demonstrate that the proposed methods are appropriate for practical use. We provide applications to HIV/AIDS clinical trials. © 1997 by John Wiley & Sons, Ltd.

*Statist. Med.*, **16**, 1515–1527 (1997)

No. of Figures: 0    No. of Tables: 4    No. of References: 18

## 1. INTRODUCTION

The conventional approach to evaluate the efficacy of therapeutic agents is to conduct clinical trials with clinical endpoints that reflect tangible benefits to patients. Such endpoints include disease occurrence (for example, infection, cancer recurrence and heart attack) and death. Unfortunately, conventional clinical trials require hundreds of patients and take years to complete. Researchers and patients wish to assess the effectiveness of promising new agents as quickly as possible, which has led investigators to explore laboratory markers that may serve as 'surrogate' endpoints in clinical trials. Replacement of a rare or late-occurring clinical endpoint with a frequent or short-term outcome variable can lead to substantial reduction in sample size and trial duration. On the other hand, inappropriate choice of (surrogate) endpoints has led to misleading results and improper treatment for large groups of patients.

For the treatment comparison based on a surrogate response variable to have an unequivocal implication for the corresponding true endpoint treatment comparison, a test of the null hypothesis of no treatment difference in the surrogate endpoint should also be a valid test of the corresponding null hypothesis based on the true endpoint (Prentice<sup>1</sup>). This definition essentially requires the surrogate variable to capture any relationship between the treatment and the true

\*Correspondence to: D. Y. Lin, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, U.S.A.

Contract grant sponsor: National Institutes of Health  
Contract grant numbers: R01 AI-291968 and R01 AI-31789

endpoint. In practice, such a restrictive condition is unlikely to be satisfied completely. A more realistic expectation is that the surrogate variable accounts for a substantial portion of the treatment effect on the clinical endpoint. This latter criterion has received considerable recent attention in HIV/AIDS research (Choi *et al.*,<sup>2</sup> O'Brien *et al.*<sup>3</sup>).

In particular, Choi *et al.*<sup>2</sup> reported that CD4 count and net CD4 per cent at week 16 explained, respectively, 46 per cent and 74 per cent of the zidovudine's effect on subsequent progression to AIDS in persons with asymptomatic HIV infection. (These numbers pertain to the proportional reduction in the regression coefficient for the treatment assignment under the proportional hazards model after controlling for changes in the CD4 levels.) The usefulness of these findings depends on the accuracy of the estimates. In this paper, we study the variability associated with such estimators. We demonstrate that, in many practical settings, the estimators are highly variable and the 95 per cent confidence intervals are very broad.

Our paper expands upon an important work by Freedman *et al.*,<sup>4</sup> which provides Fieller's confidence interval for the proportion of treatment effect explained by an intermediate endpoint in the setting of a binary outcome variable. The focus here is the failure time endpoint. The inclusion of the time dimension into the problem increases the technical complexity, but is of practical importance as most phase III clinical trials involve failure time rather than binary outcomes. Although we focus our attention on the failure time endpoint, much of the new development in this paper, including the adjustment for model misspecification and the construction of variance estimators and confidence intervals using the  $\delta$ -method, applies to other endpoints.

The remainder of this paper is organized as follows. In the next section, we develop inference procedures for the proportion of treatment effect explained by a surrogate marker. In Section 3, we report the results of our simulation studies. In Section 4, we apply the proposed methods to several HIV/AIDS clinical trials. In Section 5, we discuss a number of related issues.

## 2. METHODS

Let  $R$  be the treatment indicator and  $W(t)$  a vector of possibly time-varying covariates that represent the history of the surrogate marker. We formulate the conditional hazard functions of the clinical event given  $R$  and  $\{R, W(\cdot)\}$  by the following two proportional hazards models.

$$\lambda(t|R) = \lambda_{10}(t)e^{\alpha R} \quad (1)$$

$$\lambda(t|R, W) = \lambda_{20}(t)e^{\beta R + \gamma W(t)} \quad (2)$$

where  $\lambda_{10}(\cdot)$  and  $\lambda_{20}(\cdot)$  are unspecified baseline hazard functions, and  $\alpha$ ,  $\beta$  and  $\gamma$  are unknown regression parameters. We define the proportion of treatment effect explained by the surrogate as

$$p = 1 - \frac{\beta}{\alpha}.$$

Note that  $p$  is a proportion in the mathematical sense only if  $0 \leq \beta/\alpha \leq 1$ .

A question naturally arises as to whether models (1) and (2) may hold simultaneously. To answer this question, suppose that model (2) holds with  $W$  being time-invariant. Then, we show in the Appendix that

$$\lambda(t|R) = \lambda_{20}(t)e^{\beta R} \frac{\int e^{\gamma \omega} \exp\{-\Lambda_{20}(t)e^{\beta R + \gamma \omega}\} dF(\omega|R)}{\int \exp\{-\Lambda_{20}(t)e^{\beta R + \gamma \omega}\} dF(\omega|R)} \quad (3)$$

where  $\Lambda_{20}(t) = \int_0^t \lambda_{20}(u) du$  and  $F(\omega|R)$  is the conditional distribution function of  $W$  given  $R$ . Because the ratio of the integrals on the right side of (3) is a function of  $t$  that depends on  $R$ , the hazard ratio  $\lambda(t|R = 1)/\lambda(t|R = 0)$  varies with  $t$ , violating the proportional hazards assumption. Nevertheless, model (1) provides a good approximation to (3) if  $\gamma$  or  $\Lambda_{20}(t)$  is small.

To present results for models (1) and (2) in a compact form and to allow inclusion of other covariates (such as baseline prognostic factors) in the models, we rewrite models (1) and (2) in the following more general forms:

$$\lambda(t|Z_1) = \lambda_{10}(t)e^{\theta_1^T Z_1(t)} \tag{1}$$

$$\lambda(t|Z_2) = \lambda_{20}(t)e^{\theta_2^T Z_2(t)} \tag{2}$$

where  $Z_1$  and  $Z_2$  are  $d_1$ - and  $d_2$ -dimensional covariate vectors whose first components are the treatment indicator  $R$ , and  $\theta_1$  and  $\theta_2$  are the corresponding parameter vectors whose first components are  $\alpha$  and  $\beta$ , respectively. Note that  $Z_1$  is normally a subset of  $Z_2$ .

The data consist of  $n$  independent replicates of  $(X, \delta, Z_1, Z_2)$ , where  $X$  and  $\delta$  are, respectively, the observation time and failure indicator. Based on the data  $(X_i, \delta_i, Z_{ki})$  ( $i = 1, \dots, n$ ), the partial likelihood score function and information matrix for  $\theta_k$  ( $k = 1$  or  $2$ ) are

$$U_k(\theta) = \sum_{i=1}^n \delta_i \left\{ Z_{ki}(X_i) - \frac{S_k^{(1)}(\theta, X_i)}{S_k^{(0)}(\theta, X_i)} \right\}$$

and

$$J_k(\theta) = \sum_{i=1}^n \delta_i \left\{ \frac{S_k^{(2)}(\theta, X_i)}{S_k^{(0)}(\theta, X_i)} - \frac{S_k^{(1)}(\theta, X_i)^{\otimes 2}}{S_k^{(0)}(\theta, X_i)^2} \right\}$$

where  $S_k^{(r)}(\theta, t) = n^{-1} \sum_{i=1}^n I(X_i \geq t) e^{\theta^T Z_{ki}(t)} Z_{ki}^{\otimes r}(t)$  ( $r = 0, 1, 2$ ) with  $a^{\otimes 0} = 1$ ,  $a^{\otimes 1} = a$  and  $a^{\otimes 2} = aa'$ , and  $I(\cdot)$  being the indicator function.

The maximum partial likelihood estimator  $\hat{\theta}_k$  for  $\theta_k$  is the solution to the system of equations  $\{U_k(\theta) = 0\}$ . Correspondingly, we estimate  $p$  by

$$\hat{p} \equiv 1 - \frac{\hat{\beta}}{\hat{\alpha}}$$

where  $\hat{\alpha}$  and  $\hat{\beta}$  are the first components of  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , respectively.

Since in general models (1') and (2') cannot hold (exactly) at the same time, it is desirable to study the behaviour of  $\hat{\theta}_k$  ( $k = 1, 2$ ) under misspecified models. Define  $s_k^{(r)}(\theta, t) = E\{S_k^{(r)}(\theta, t)\}$  and  $s_k^{(r)}(t) = E\{n^{-1} \sum_{i=1}^n I(X_i \geq t) \lambda(t|Z_{ki}) Z_{ki}^{\otimes r}(t)\}$  ( $k = 1, 2; r = 0, 1, 2$ ), where  $E$  denotes expectation. Also, let  $\tau$  denote the maximum length of follow-up on a patient. Then  $\hat{\theta}_k$  converges in probability to  $\theta_k^*$ , which is the unique solution to the system of equations

$$\int_0^\tau \left\{ \frac{s_k^{(1)}(t)}{s_k^{(0)}(t)} - \frac{s_k^{(1)}(\theta, t)}{s_k^{(0)}(\theta, t)} \right\} s_k^{(0)}(t) dt = 0 \tag{4}$$

provided that

$$A_k \equiv \int_0^\tau \left\{ \frac{s_k^{(2)}(\theta, t)}{s_k^{(0)}(\theta, t)} - \frac{s_k^{(1)}(\theta, t)^{\otimes 2}}{s_k^{(0)}(\theta, t)^2} \right\} s_k^{(0)}(t) dt$$

is positive definite (Struthers and Kalbfleisch<sup>5</sup>, Lin and Wei<sup>6</sup>). Consequently,  $\hat{p}$  converges to

$$p^* \equiv 1 - \frac{\beta^*}{\alpha^*}$$

where  $\alpha^*$  and  $\beta^*$  are the first components of  $\theta_1^*$  and  $\theta_2^*$ , respectively.

Note that  $s_k^{(1)}(t)/s_k^{(0)}(t)$  and  $s_k^{(1)}(\theta, t)/s_k^{(0)}(\theta, t)$  are the weighted means of  $Z_k(t)$  among the patients under observation at time  $t$  with weights proportional to the true and assumed conditional hazard functions, respectively. The left side of (4) is an integrated difference between these two weighted means. Under misspecified models, we may call  $\theta_k^*$  the least false parameter value in that it minimizes a generalized Kullback–Leibler information criterion characterizing the distance between the true and assumed conditional hazard functions (Hjort<sup>7</sup>). Specifically,  $\alpha^*$  and  $\beta^*$  represent, respectively, some averages over  $[0, \tau]$  of the (true) log hazard ratios between the two treatments with and without adjusting for the surrogate marker, and  $p^*$  is still a useful measure for the proportion of treatment effect explained by the surrogate. If models (1) and (2) hold approximately, then  $\alpha^*$  and  $\beta^*$  are close to the hypothetical  $\alpha$  and  $\beta$ .

Lin and Wei<sup>6</sup> established the asymptotic normality of the maximum partial likelihood estimator under misspecified models. In our setting, it is necessary to ascertain the joint distribution between  $\hat{\theta}_1$  and  $\hat{\theta}_2$ . According to the Appendix of Lin and Wei,<sup>6</sup> the random vector  $n^{1/2}(\hat{\theta}_k - \theta_k^*)$  is asymptotically equivalent to  $A_k^{-1}n^{-1/2}\sum_{i=1}^n \zeta_{ki}$ , where

$$\zeta_{ki} = \delta_i \left\{ Z_{ki}(X_i) - \frac{s_k^{(1)}(\theta_k^*, X_i)}{s_k^{(0)}(\theta_k^*, X_i)} \right\} - \int_0^\tau \frac{I(X_i \geq t) e^{\theta_k^* Z_{ki}(t)}}{s_k^{(0)}(\theta_k^*, t)} \left\{ Z_{ki}(t) - \frac{s_k^{(1)}(\theta_k^*, t)}{s_k^{(0)}(\theta_k^*, t)} \right\} d\Pr(X \leq t, \delta = 1).$$

For each  $k$ , the random vectors  $\zeta_{ki}$  ( $i = 1, \dots, n$ ) are independent and identically distributed with zero means, though  $\zeta_{1i}$  and  $\zeta_{2i}$  (for the same  $i$ ) are correlated. By the multivariate central limit theorem and the Cramer–Wold device, the  $(d_1 + d_2)$ -dimensional random vector  $n^{1/2}[(\hat{\theta}_1 - \theta_1^*)', (\hat{\theta}_2 - \theta_2^*)']'$  is asymptotically multivariate normal with mean zero and with covariance matrix

$$V = \begin{bmatrix} A_1^{-1}B_{11}A_1^{-1} & A_1^{-1}B_{12}A_2^{-1} \\ A_2^{-1}B_{21}A_1^{-1} & A_2^{-1}B_{22}A_2^{-1} \end{bmatrix}$$

where  $B_{kl} = E(\zeta_{k1}\zeta'_{l1})$  ( $k, l = 1, 2$ ).

It is natural to estimate  $B_{kl}$  by  $\hat{B}_{kl} = n^{-1}\sum_{i=1}^n \hat{\zeta}_{ki}\hat{\zeta}'_{li}$ , where we obtain the  $\hat{\zeta}_{ki}$  from the  $\zeta_{ki}$  by replacing the unknown parameters in the latter with their sample estimators. Specifically,

$$\hat{\zeta}_{ki} = \delta_i \left\{ Z_{ki}(X_i) - \frac{S_k^{(1)}(\hat{\theta}_k, X_i)}{S_k^{(0)}(\hat{\theta}_k, X_i)} \right\} - \sum_{j=1}^n \frac{\delta_j I(X_i \geq X_j) e^{\hat{\theta}_k Z_{ki}(X_j)}}{n S_k^{(0)}(\hat{\theta}_k, X_j)} \left\{ Z_{ki}(X_j) - \frac{S_k^{(1)}(\hat{\theta}_k, X_j)}{S_k^{(0)}(\hat{\theta}_k, X_j)} \right\}.$$

The consistency of  $\hat{B}_{kl}$  for  $B_{kl}$  follows from the arguments given in the Appendix of Lin and Wei.<sup>6</sup> Furthermore, a consistent estimator of  $A_k$  is  $\hat{A}_k = \{n^{-1}\mathcal{J}_k(\hat{\theta}_k)\}^{-1}$ . Hence, a consistent estimator of the limiting covariance matrix  $V$  is

$$\hat{V} = \begin{bmatrix} \hat{A}_1^{-1}\hat{B}_{11}\hat{A}_1^{-1} & \hat{A}_1^{-1}\hat{B}_{12}\hat{A}_2^{-1} \\ \hat{A}_2^{-1}\hat{B}_{21}\hat{A}_1^{-1} & \hat{A}_2^{-1}\hat{B}_{22}\hat{A}_2^{-1} \end{bmatrix}.$$

It is interesting to note that  $\hat{V}$  takes the same form as the covariance matrix estimator for the marginal hazard modelling of multiple events data (Wei *et al.*<sup>8</sup>). The latter has been implemented

in the FORTRAN programs MULCOX (Lin<sup>9</sup>) and MULCOX2 (Lin<sup>10</sup>) as well as in the recent releases of SAS and S-plus. To obtain  $\hat{V}$  from MULCOX2, we construct the artificial bivariate survival data  $\{\tilde{X}_{ki}, \tilde{\delta}_{ki}, \tilde{Z}_{ki}(\cdot)\}$  ( $k = 1, 2; i = 1, \dots, n$ ), where  $\tilde{X}_{ki} = X_i, \tilde{\delta}_{ki} = \delta_i,$

$$\tilde{Z}_{1i} = \begin{bmatrix} Z_{1i} \\ 0 \end{bmatrix}, \quad \tilde{Z}_{2i} = \begin{bmatrix} 0 \\ Z_{2i} \end{bmatrix}$$

and fit the following model

$$\lambda_k(t|\tilde{Z}_{ki}) = \lambda_{k0}(t)e^{\eta\tilde{Z}_{ki}(t)}, \quad k = 1, 2.$$

Note that  $\eta_1 = \alpha$  and  $\eta_{d_1+1} = \beta$ . MULCOX2 will produce the estimate of  $\eta = (\theta'_1, \theta'_2)'$  and its covariance matrix estimate  $n^{-1}\hat{V}$ . (It is necessary to use non-zero  $\eta$  in the initial step of the Newton–Raphson algorithm; otherwise, the Hessian matrix is singular).

The joint distribution of  $(\hat{\alpha}, \hat{\beta})$  provides the basis for making inference about  $p^*$ . By the  $\delta$ -method, the random variable  $n^{1/2}(\hat{p} - p^*)$  is asymptotically zero-mean normal with variance

$$\sigma^2 = \frac{V_\beta}{(\alpha^*)^2} + \frac{(\beta^*)^2 V_\alpha}{(\alpha^*)^4} - 2\frac{\beta^* V_{\alpha\beta}}{(\alpha^*)^3} \tag{5}$$

for which a consistent estimator is

$$\hat{\sigma}^2 = \frac{\hat{V}_\beta}{\hat{\alpha}^2} + \frac{\hat{\beta}^2 \hat{V}_\alpha}{\hat{\alpha}^4} - 2\frac{\hat{\beta} \hat{V}_{\alpha\beta}}{\hat{\alpha}^3}.$$

Here,  $V_\alpha, V_\beta$  and  $V_{\alpha\beta}$  denote the variances and covariance of  $n^{1/2}\hat{\alpha}$  and  $n^{1/2}\hat{\beta}$ , and  $\hat{V}_\alpha, \hat{V}_\beta$  and  $\hat{V}_{\alpha\beta}$  denote their estimators, which are elements of the covariance matrix estimator  $\hat{V}$ .

We may rewrite (5) as

$$\sigma^2 = \frac{V_\alpha}{(\alpha^*)^2} \left\{ \frac{V_\beta}{V_\alpha} + (1 - p^*)^2 - 2(1 - p^*) \frac{V_{\alpha\beta}}{V_\alpha} \right\}$$

which shows that the standard error of  $\hat{p}$  depends on the coefficient of variation for  $\hat{\alpha}$  (that is, the inverse of the ratio of the unadjusted treatment effect over the standard error of its estimator), on the value of  $p^*$  itself as well as on the values of  $V_\beta$  and  $V_{\alpha\beta}$  relative to  $V_\alpha$ . If  $\gamma$  is small and the correlation between  $W$  and  $R$  is low, then  $V_\alpha \approx V_\beta \approx V_{\alpha\beta}$ , in which case

$$\frac{\sigma}{|p^*|} \approx \frac{V_\alpha^{1/2}}{|\alpha^*|} \tag{6}$$

that is, the coefficient of variation for  $\hat{p}$  is roughly that of  $\hat{\alpha}$ . Thus, if  $|\alpha^*|$  is small relative to the standard error of  $\hat{\alpha}$ , one must expect poor precision in estimating  $p^*$ . In most applications, formula (6) would underestimate the true variability of  $\hat{p}$  because  $V_\beta$  tends to be larger than  $V_\alpha$  due to a high correlation between  $W$  and  $R$ .

Based on the direct normal approximation for  $\hat{p}$ , the  $(1 - \psi)100$  per cent confidence interval of  $p^*$  is

$$\hat{p} \pm z_{1-\psi/2}(\hat{\sigma}^2/n)^{1/2} \tag{7}$$

where  $z_{1-\psi/2}$  is the  $(1 - \psi/2)100$ th upper percentile of the standard normal distribution. One may also construct confidence intervals for  $p^*$  using Fieller’s theorem.<sup>11</sup> The corresponding

Table I. Probability that  $p_L > f$  under  $p^* = 1$

	$\alpha^*/(V_\beta/n)^{1/2}$				
	2	4	6	8	10
$f = 0.5$	0.169	0.516	0.851	0.979	0.999
$f = 0.75$	0.072	0.169	0.323	0.516	0.705

$(1 - \psi)100$  per cent confidence limits of  $p^*$  are

$$1 - (1 - g)^{-1} \left[ \hat{q} - g \frac{\hat{V}_{\alpha\beta}}{\hat{V}_\alpha} \pm \frac{n^{-1/2} z_{1-\psi/2}}{|\hat{\alpha}|} \left\{ \hat{V}_\beta - 2\hat{q}\hat{V}_{\alpha\beta} + \hat{q}^2\hat{V}_\alpha - g \left( \hat{V}_\beta - \frac{\hat{V}_{\alpha\beta}^2}{\hat{V}_\alpha} \right) \right\}^{1/2} \right] \tag{8}$$

where  $\hat{q} = \hat{\beta}/\hat{\alpha}$  and  $g = z_{1-\psi/2}^2 n^{-1} \hat{V}_\alpha / \hat{\alpha}^2$ . We shall refer to (7) and (8) as the  $\delta$ -method and Fieller's method intervals, respectively.

Fieller's method requires that  $g < 1$ , that is

$$\frac{|\hat{\alpha}|}{\{n^{-1} \hat{V}_\alpha\}^{1/2}} > z_{1-\psi/2}. \tag{9}$$

Thus, it is unfeasible to construct a proper  $(1 - \psi)100$  per cent confidence interval for  $p^*$  using Fieller's theorem unless the unadjusted treatment effect is significant at the  $\psi$  level. The  $\delta$ -method does not require condition (9). Nevertheless, since  $\hat{\sigma}$  is inversely proportional to  $|\hat{\alpha}|/\hat{V}_\alpha^{1/2}$ , the interval (7) is likely to be wide if condition (9) is not met. Hence, the data are rather uninformative about  $p^*$  if the unadjusted treatment effect is less than twice its standard error.

The Prentice definition<sup>1</sup> requires the surrogate marker to capture all the net treatment effect on the clinical outcome, which corresponds to  $p^* = 1$ . Thus, the confidence interval for  $p^*$  excluding the value 1 would constitute a case where the Prentice criterion could be said to be not met. One may consider a surrogate marker important if the lower limit of the 95 per cent confidence interval of  $p^*$  is sufficiently large, say greater than 0.5 or 0.75, and may have particular interest in determining the power of obtaining such a result under the condition of  $p^* = 1$ . By the asymptotic normality of  $\hat{p}$ , the probability that the lower limit of the  $\delta$ -method  $(1 - \psi)100$  per cent confidence interval for  $p^*$  exceeds a given fraction  $f$  is

$$\Pr \left( \hat{p} - z_{1-\psi/2} n^{-1/2} \hat{\sigma} > f \right) \approx \Phi \left( \frac{p^* - f}{n^{-1/2} \sigma} - z_{1-\psi/2} \right) \tag{10}$$

where  $\Phi$  is the standard-normal distribution function.

If  $p^* = 1$ , then  $\sigma = V_\beta^{1/2}/\alpha^*$ . Note that  $\alpha^*/(V_\beta/n)^{1/2}$  is the ratio of the unadjusted treatment effect to the standard error of the adjusted treatment effect estimator. It follows from (10) that

$$\Pr(p_L > f) \approx \Phi \left\{ (1 - f) \frac{\alpha^*}{(V_\beta/n)^{1/2}} - 1.96 \right\} \tag{11}$$

where  $p_L$  denotes the lower bound of the 95 per cent confidence interval. In Table I we tabulate the above probability for  $f = 0.5$  and 0.75 and selected values of  $\alpha^*/(V_\beta/n)^{1/2}$ . The results show that the probability that  $p_L > 0.5$  is 85 per cent or higher when  $\alpha^*/(V_\beta/n)^{1/2} \geq 6$ . On the other hand, the probability that  $p_L > 0.75$  remains low even for large values of  $\alpha^*/(V_\beta/n)^{1/2}$ .

Due to the complicated nature of expression (8), it is difficult to evaluate the power  $\Pr(p_L > f)$  for Fieller's method. Freedman *et al.*<sup>4</sup> provided a formula for such evaluation under the assumption of  $V_\alpha = V_\beta$ , that is, the equality of the variances for the estimators of the unadjusted and adjusted treatment effects.

In some applications, it is of interest to compare the proportions of treatment effect explained by two sets of surrogate markers. To this end, we introduce the following model for a second set of markers  $W_2$ :

$$\lambda(t|R, W_2) = \lambda_{20}^\dagger(t)e^{\beta_2 R + \gamma_2 W_2(t)}$$

which is analogous to model (2). Let  $p_2 = 1 - \beta_2/\alpha$ , estimated by  $\hat{p}_2 = 1 - \hat{\beta}_2/\hat{\alpha}$ , where  $\hat{\beta}_2$  is the maximum partial likelihood estimator of  $\beta_2$ . Also, let  $p_2^* = 1 - \beta_2^*/\alpha^*$ , where  $\beta_2^*$  is the probability limit of  $\hat{\beta}_2$ . Then,  $n^{1/2}\{(\hat{p}_2 - p_2) - (p_2^* - p^*)\}$  is asymptotically zero-mean normal with variance

$$\frac{V_\beta + V_{\beta_2} - 2V_{\beta\beta_2}}{(\alpha^*)^2} + \frac{(\beta^* - \beta_2^*)^2 V_\alpha}{(\alpha^*)^4} - 2 \frac{(\beta^* - \beta_2^*) (V_{\alpha\beta} - V_{\alpha\beta_2})}{(\alpha^*)^3}$$

where  $V_{\beta_2}$  is the variance of  $n^{1/2}\hat{\beta}_2$ , and  $V_{\beta\beta_2}$  (or  $V_{\alpha\beta_2}$ ) is the covariance between  $n^{1/2}\hat{\beta}_2$  and  $n^{1/2}\hat{\beta}$  (or  $n^{1/2}\hat{\alpha}$ ), which we can estimate in the same way as  $V_\alpha$ ,  $V_\beta$  and  $V_{\alpha\beta}$ .

### 3. SIMULATION STUDIES

We conducted a series of simulation studies to assess the performance of the methods developed in the previous section. We generated failure times from model (2) with equal numbers of patients on the two treatment arms and with  $W$  as a normal random variable with mean  $\mu_R$  and unit variance. By inserting the normal density function into (3) and by some simple algebraic manipulation, we see that

$$\lambda(t|R) = \lambda_{20}(t)e^{\beta R + \gamma\mu_R + 0.5\gamma^2}h(t; R)$$

where

$$h(t; R) = \frac{\int_{-\infty}^{\infty} \exp\{-0.5(\omega - \mu_R)^2 - \Lambda_{20}(t)e^{\beta R + \gamma\omega + \gamma^2}\}d\omega}{\int_{-\infty}^{\infty} \exp\{-0.5(\omega - \mu_R)^2 - \Lambda_{20}(t)e^{\beta R + \gamma\omega}\}d\omega}.$$

If  $\gamma$  or  $\Lambda_{20}(t)$  is small, then  $h(t; R) \approx 1$ , which implies that  $\lambda(t|R = 1)/\lambda(t|R = 0) \approx e^\alpha$ , where  $\alpha = \beta + \gamma(\mu_1 - \mu_0)$ . In our simulations, we set  $\lambda_{20}(t) = 1$  and let censoring times be uniformly distributed over  $[0, \tau]$ .

Table II displays the results for the set-up of  $\mu_0 = 0, \mu_1 = 2, \beta = 1$  and  $\gamma = \{0.25, 0.5, 1\}$ . We set the terminal time point  $\tau$  as the lower 25th percentile of the failure times, which yields 86.6, 86.3 and 85.1 per cent censorship for  $\gamma = 0.25, 0.5$  and  $1$ , respectively. Note that  $\alpha$  would be 1.5, 2 and 3, and  $p$  would be 1/3, 1/2 and 2/3 if  $h$  were identically equal to 1. The exact values of  $\alpha^*$  and  $p^*$  are difficult to determine analytically. We approximated  $p^*$  by the sampling mean of  $\hat{p}$  for  $n = 5000$  with 1000 simulation samples. The approximate values were 0.33, 0.49 and 0.64 for  $\gamma = 0.25, 0.5$  and  $1$ , respectively, which are almost identical to 1/3, 1/2 and 2/3.

As shown in Table II, the standard error of  $\hat{\beta}$  is greater than that of  $\hat{\alpha}$ , and the correlation between  $\hat{\alpha}$  and  $\hat{\beta}$  is high. The bias of  $\hat{p}$  and that of its standard error estimator are both negligible. The proposed confidence intervals have proper coverage probabilities except for the combination

Table II. Summary statistics for the simulation studies

	$\gamma = 0.25$			$\gamma = 0.5$			$\gamma = 1$		
	$n = 250$	500	1000	$n = 250$	500	1000	$n = 250$	500	1000
Mean( $\hat{\alpha}$ )	1.56	1.52	1.50	2.15	2.01	1.97	3.47	2.91	2.78
S.E.( $\hat{\alpha}$ )	0.47	0.31	0.22	1.10	0.37	0.26	2.47	1.03	0.35
Mean( $\hat{\beta}$ )	1.07	1.03	1.01	1.18	1.05	1.02	1.69	1.16	1.04
S.E.( $\hat{\beta}$ )	0.58	0.38	0.28	1.17	0.43	0.31	2.48	1.05	0.39
Corr( $\hat{\alpha}$ , $\hat{\beta}$ )	0.80	0.79	0.81	0.96	0.84	0.85	0.99	0.98	0.91
Mean( $\hat{p}$ )	0.35	0.33	0.33	0.50	0.49	0.49	0.62	0.63	0.64
S.E.( $\hat{p}$ )	0.31	0.17	0.13	0.24	0.15	0.11	0.22	0.14	0.10
Mean( $n^{-1/2}\hat{\sigma}$ )	0.28	0.18	0.13	0.22	0.15	0.11	0.19	0.13	0.09
Mean width of 95 per cent CI									
$\delta$ -method	1.11	0.72	0.50	0.87	0.60	0.42	0.73	0.52	0.37
Fieller's method	1.69	0.84	0.52	1.30	0.66	0.44	0.90	0.57	0.38
Coverage of 95 per cent CI									
$\delta$ -method	0.96	0.96	0.96	0.94	0.95	0.95	0.90	0.94	0.94
Fieller's method	0.94	0.96	0.96	0.94	0.96	0.95	0.91	0.96	0.95

Each entry was based on 1000 simulation samples. Condition (9) was not met in 18 samples and 1 sample for  $\{\gamma = 0.25, n = 250\}$  and  $\{\gamma = 0.5, n = 250\}$ , respectively. Those cases were excluded from the calculation of the summary statistics for Fieller's method

of  $\gamma = 1$  and  $n = 250$ , in which case an average of 37.5 events seems too small to produce stable  $\hat{\alpha}$  and  $\hat{\beta}$  estimates. In general, the estimator  $\hat{p}$  is quite variable and the confidence intervals are broad. As mentioned previously, one cannot construct the 95 per cent confidence interval for  $p^*$  using Fieller's theorem unless  $|\hat{\alpha}|/s.e.(\hat{\alpha}) > 1.96$ . This condition failed in a few simulation samples when  $n = 250$ . Fieller's method tends to produce wider intervals than the  $\delta$ -method, though the two methods have similar coverage probabilities. Further inspection of the simulation results (not shown in Table II) reveals that, relative to the  $\delta$ -method interval, Fieller's interval tends to be shifted to the right and its width is more variable.

#### 4. APPLICATIONS TO AIDS

The findings by Choi *et al.*<sup>2</sup> mentioned in Section 1 were based on data from the ACTG (AIDS Clinical Trials Group) Protocol 019, which is a placebo-controlled, double-blind, randomized trial on the efficacy of zidovudine in the treatment of asymptomatic HIV-infected persons (Volberding *et al.*<sup>12</sup>). A total of 1075 patients were enrolled: 350 were given placebo and 725 were given one of two doses of zidovudine. After a maximum follow-up period of 90 weeks (median 55 weeks), 44 patients had progressed to AIDS – 24 in the placebo group and 20 in the zidovudine groups. The two-sided  $p$ -value for the logrank test is 0.04.

Choi *et al.* assessed the extent to which different CD4 measures at week 16 were surrogate markers for the subsequent development of AIDS in patients who had not progressed by week 16. Excluded from this analysis were 5 placebo patients and 1 zidovudine patient who had developed AIDS by week 16 as well as 23 placebo patients and 84 zidovudine patients who had been censored. The CD4 measurements at week 16 were the most recently measured preceding values. The authors found that CD4 count and net CD4 per cent accounted for about 46 per cent and 74

Table III. Analysis of the ACTG 019 study

	Progression after week 16		Progression after randomization	
	CD4 count	Net CD4%	CD4 count	Net CD4%
$\hat{\alpha}$	-0.53	-0.53	-0.62	-0.62
S.E.( $\hat{\alpha}$ )	0.33	0.33	0.31	0.31
$\hat{\alpha}/s.e.(\hat{\alpha})$	-1.62	-1.62	-1.99	-1.99
$\hat{\beta}$	-0.28	-0.14	-0.50	-0.38
S.E.( $\hat{\beta}$ )	0.32	0.33	0.30	0.31
$\hat{\beta}/s.e.(\hat{\beta})$	-0.88	-0.43	-1.66	-1.21
Corr( $\hat{\alpha}$ , $\hat{\beta}$ )	0.95	0.95	0.95	0.97
$\hat{p}$	0.46	0.74	0.19	0.38
S.E.( $\hat{p}$ )	0.31	0.47	0.16	0.22
$\hat{p}/s.e.(\hat{p})$	1.51	1.54	1.14	1.74
95% CI for $p^*$				
$\delta$ -method	(-0.14, 1.08)	(-0.20, 1.65)	(-0.13, 0.51)	(-0.05, 0.81)
Fieller's method	—	—	(-0.27, 7.75)	(0.12, 24.49)

per cent of the zidovudine effect. As shown in the left panel of Table III, the (estimated) standard errors are quite large for these two estimates. The  $\delta$ -method confidence intervals cover the entire [0, 1] interval. It is not possible to construct Fieller's confidence intervals because the unadjusted treatment effect is not significant at the 5 per cent (or even 10 per cent) level.

In the above analysis, the patients on different treatment arms were not comparable due to the exclusion of those who had progressed to AIDS by week 16. It seems more meaningful to determine what proportions of the entire zidovudine effect on the progression to AIDS from the time of randomization are explained by the CD4 measures at week 16. This type of analysis uses the date of randomization as the time origin and includes all the events that occur after randomization, which is consistent with the original treatment comparison. The corresponding results appear in the right panel of Table III. The estimates for the proportions of zidovudine's effect explained by CD4 count and net CD4 per cent at week 16 are much smaller than those of the previous analysis. The upper limits of the  $\delta$ -method confidence intervals are less than 1 while the lower limits are below 0. The Fieller intervals are quite unstable in this case because the unadjusted treatment effect is barely significant at the 5 per cent level. Using either type of analysis, one cannot reject, at the 5 per cent significance level, the null hypothesis that CD4 count (or net CD4 per cent) at week 16 explains none of the zidovudine effect on progression to AIDS.

For further illustration, we consider the BW (Burroughs Wellcome) Protocol 02 study which first demonstrated the clinical benefit of zidovudine in adults with symptomatic HIV infection (Fischl *et al.*<sup>13</sup>). The study enrolled 281 patients, among whom 160 had AIDS and 121 had advanced AIDS-related complex. There were 144 patients assigned to the zidovudine group and 137 to the placebo group. The patients were followed for a minimum of 8 weeks. By the end of the trial, opportunistic infections had developed in 51 patients who received placebo, as compared with 25 who received zidovudine. To determine the extent to which CD4 count at week 8 was a surrogate endpoint for the time to opportunistic infection, we fit the two Cox models shown in Table IV. The observed value of  $\hat{p}$  is 0.283 with estimated standard error of 0.115. The 95 per cent confidence intervals are (0.057, 0.509) and (0.100, 0.666) based on the  $\delta$ -method and Fieller's

Table IV. Analysis of the BW 02 study

Covariate	Estimate	Model 1	Model 2
Treatment	Coefficient	-0.92	-0.66
	Standard error	0.24	0.25
	Coefficient/s.e.	-3.76	-2.59
Status	Coefficient	0.82	0.40
	Standard error	0.26	0.26
	Coefficient/s.e.	3.17	1.52
Week 8 CD4 count	Coefficient	—	-0.0060
	Standard error	—	0.0016
	Coefficient/s.e.	—	-3.76

The failure time is the time from randomization to the occurrence of opportunistic infection. Treatment takes the value 1 if the patient was on zidovudine and 0 otherwise. Status takes the value 1 if the patient had AIDS at the time of randomization and 0 otherwise

theorem, respectively. We conclude that CD4 count at week 8 explained a small fraction of the zidovudine effect. Incidentally, several researchers (De Gruttola *et al.*,<sup>14</sup> Lin *et al.*,<sup>15</sup> Tsiatis *et al.*<sup>16</sup>) have shown that the Prentice criterion for a surrogate endpoint was not met in the BW 02 study when the entire time-course of CD4 measurements was considered.

We have also applied the proposed methods to other HIV/AIDS clinical trials, including a placebo-controlled trial of zidovudine on mildly symptomatic HIV patients (ACTG Protocol 016) and a trial comparing zidovudine and ddI (ACTG Protocols 116b/117), and again found no evidence that CD4 measures explain a substantial fraction of the treatment effect (De Gruttola *et al.*<sup>17</sup>). In fact, the estimate  $\hat{p}$  is very close to zero for the 016 study.

## 5. DISCUSSION

With great interest in determining the proper role of surrogate endpoints in clinical trials that evaluate interventions in diseases such as HIV/AIDS, cancer and cardiovascular disorders, there has been much attention directed to the computation of  $p^*$ , the proportion of the net treatment effect explained by the surrogate marker. This paper provides a rigorous methodology for such assessments. In particular, it enables the construction of confidence intervals for  $p^*$  in the very common setting in which the primary clinical endpoint is a failure time variable subject to censorship. Simulations and applications to HIV/AIDS data reveal the high variability of estimates of  $p^*$ , while formulae presented here for their standard errors provide insights into the conditions required for precise estimation of  $p^*$ .

The theoretical results of Section 2 hold for other types of endpoints as well, though the forms of  $V_\alpha$ ,  $V_\beta$  and  $V_{\alpha\beta}$  depend on the particular types of models being employed. Freedman *et al.*<sup>4</sup> provided the formulae for  $V_\alpha$ ,  $V_\beta$  and  $V_{\alpha\beta}$  under logistic regression models. One could modify their formulae to account for model misspecification using the techniques of White<sup>18</sup> and those given in Section 2 of this paper.

Recently, O'Brien *et al.*<sup>3</sup> analysed data from the Veterans Affairs Cooperative Study that compared immediate and deferred zidovudine therapy. They estimated the proportions of the treatment difference in progression to AIDS accounted for by the changes in plasma HIV-1 RNA

and CD4 counts, and used a bootstrap approach to calculate the 95 per cent confidence intervals. They did not explain how they performed the bootstrap, nor did they justify its validity. The asymptotic theory of  $\hat{p}$  developed in this paper is essential in establishing the validity of a bootstrap procedure. We concentrated on the analytic approach in this paper because there exists a simple expression for the variance of  $\hat{p}$ , which provides useful insights into the sources of variation in  $\hat{p}$  and the conditions required for precise estimation of  $p^*$ , and which permits direct evaluation of several important quantities.

The variance formula is particularly useful in designing studies intended to investigate the degree to which a biological marker is surrogate for a clinical event. As our analytical and numerical studies have indicated, reasonably precise estimation of  $p^*$  requires that a treatment effect on the development of the clinical event is much larger (relative to its standard error) than that needed to show a significant treatment effect on the clinical event, which implies the requirement of larger or/and longer trials or the use of meta-analyses.

The methods for estimating  $p^*$  described in this report can provide valuable insights into mechanisms of disease and drug action. Interpretation of this quantity is not straightforward, however, even when it can be estimated with precision. Because disease processes are complex and because drugs have many mechanisms of action, validation of a surrogate marker must rely on an understanding of the underlying biology, not simply estimation of  $p^*$ . A value of  $p^*$  near 1 is not sufficient for inferring that the marker is a good surrogate for the clinical endpoint, since a variety of factors such as drug toxicity, non-compliance with study medications, and incomplete marker information can artificially raise this value to 1 (or greater) even for poor surrogates (De Gruttola *et al.*<sup>17</sup>).

Nonetheless, the estimation of  $p^*$  can be useful in providing evidence to evaluate hypotheses about mechanisms of drug action. For example, early in the AIDS epidemic, there was widespread belief that the CD4 count reflected both the principle mechanisms of the destruction of the immune system and of anti-HIV drug benefits. Although the  $p^*$  associated with CD4 counts in antiviral drug trials could be near 1 even if this were not true,  $p^*$  must be near 1 if it were true. The fact that a number of analyses demonstrated that the  $p^*$  was nearer to 0 than 1 provided evidence that the CD4 count was not capturing all important drug effects. Although interpretation of values of  $p^*$  near 1 is less straightforward than interpretation of  $p^*$  near 0, values of  $p^*$  near 1 for a given marker across many studies and across drugs with different mechanisms of action may help to support hypotheses about marker biological mechanisms.

#### APPENDIX: DERIVATION OF FORMULA (3)

Let  $f$  and  $F$  denote the density and distribution function. Then

$$\begin{aligned} f(t|R) &= \int f(t|R, \omega) dF(\omega|R) \\ &= \int \lambda_{20}(t) e^{\beta R + \gamma \omega} \exp\{-\Lambda_{20}(t) e^{\beta R + \gamma \omega}\} dF(\omega|R). \end{aligned}$$

In addition,

$$\begin{aligned} F(t|R) &= \int_0^t f(u|R) du \\ &= \int \left[ \int_0^t \lambda_{20}(u) e^{\beta R + \gamma \omega} \exp\{-\Lambda_{20}(u) e^{\beta R + \gamma \omega}\} du \right] dF(\omega|R) \end{aligned}$$

$$\begin{aligned}
&= \int \left[ \int_0^t \exp\{-\Lambda_{20}(u)e^{\beta R + \gamma \omega}\} d\Lambda_{20}(u)e^{\beta R + \gamma \omega} \right] dF(\omega|R) \\
&= 1 - \int \exp\{-\Lambda_{20}(t)e^{\beta R + \gamma \omega}\} dF(\omega|R).
\end{aligned}$$

Thus,

$$\lambda(t|R) = \lambda_{20}(t)e^{\beta R} \frac{\int e^{\gamma \omega} \exp\{-\Lambda_{20}(t)e^{\beta R + \gamma \omega}\} dF(\omega|R)}{\int \exp\{-\Lambda_{20}(t)e^{\beta R + \gamma \omega}\} dF(\omega|R)}.$$

#### ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health grants R01 AI-291968 (for Lin and Fleming) and R01 AI-31789 (for De Gruttola). The authors are grateful to Dr. Larry Freedman for his helpful discussions and to two referees for their useful comments. Part of this work was conducted while the first author was visiting Limburgs Universitair Centrum in Belgium.

#### REFERENCES

1. Prentice, R. L. 'Surrogate endpoints in clinical trials: definition and operational criteria', *Statistics in Medicine*, **8**, 431–440 (1989).
2. Choi, S., Lagakos, S. W., Schooley, R. T. and Volberding, P. A. 'CD4<sup>+</sup> lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine', *Annals of Internal Medicine*, **118**, 674–680 (1993).
3. O'Brien, W. A., Hartigan, P. M., Martin, D., Esinhart, J., Hill, A., Benoit, S., Rubin, M., Simberkoff, M. S., Hamilton, J.D., and the Veterans Affairs Cooperative Study Group on AIDS. 'Changes in plasma HIV-1 RNA and CD4<sup>+</sup> lymphocyte counts and the risk of progression to AIDS', *New England Journal of Medicine*, **334**, 426–431 (1996).
4. Freedman, L. S., Graubard, B. I. and Schatzkin, A. 'Statistical validation of intermediate endpoints for chronic diseases', *Statistics in Medicine*, **11**, 167–178 (1992).
5. Struthers, C. A. and Kalbfleisch, J. D. 'Misspecified proportional hazard models', *Biometrika*, **73**, 363–369 (1986).
6. Lin, D. Y. and Wei, L. J. 'The robust inference for the Cox proportional hazards model', *Journal of the American Statistical Association*, **84**, 1074–1078 (1989).
7. Hjort, N. L. 'On inference in parametric survival data models', *International Statistical Review*, **60**, 355–387 (1992).
8. Wei, L. J., Lin, D. Y. and Weissfeld, L. 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association*, **84**, 1065–1073 (1989).
9. Lin, D. Y. 'MULCOX: a computer program for the Cox regression analysis of multiple failure time variables', *Computer Methods and Programs in Biomedicine*, **32**, 125–135 (1990).
10. Lin, D. Y. 'MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data', *Computer Methods and Programs in Biomedicine*, **40**, 279–293 (1993).
11. Fieller, E. C. 'The biological standardization of insulin', *Journal of the Royal Statistical Society*, **7**, Supplement 1–15 (1940).
12. Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour, H. H., Reichman, R. C., Bartlett, J. A., Hirsch, M. S., Murphy, R. L., Hardy, W. D., Soeiro, R., Fischl, M. A., Bartlett, J. G., Merigan, T. C., Hyslop, N. E., Richman, D. D., Valentine, F. T., Corey, L. and the AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases. 'Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter', *New England Journal of Medicine*, **322**, 941–949 (1990).

13. Fischl, M. A., Richman, D. D., Grieco, M. H., Gottlieb, M. S., Volberding, P. A., Laskin, O. L., Leedom, J. M., Groopman, J. E., Mildvan, D., Schooley, R. T., Jackson, G. G., Durack, D. T., King, D. and the AZT Collaborative Working Group. 'The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex', *New England Journal of Medicine*, **317**, 185–191 (1987).
14. De Gruttola, V., Wulfsohn, M., Fischl, M. A. and Tsiatis, A. A. 'Modeling the relationship between survival and CD4-lymphocytes in patients with AIDS and AIDS-related complex', *Journal of AIDS*, **6**, 359–365 (1993).
15. Lin, D. Y., Fischl, M. A. and Schoenfeld, D. A. 'Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials', *Statistics in Medicine*, **12**, 835–842 (1993).
16. Tsiatis, A. A., De Gruttola, V. and Wulfsohn, M. 'Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS', *Journal of the American Statistical Association*, **90**, 27–37 (1995).
17. De Gruttola, V., Fleming, T. R., Lin, D. Y. and Coombs, R. 'Validating surrogate markers – are we being naive?', *Journal of Infectious Diseases*, **175**, 237–246 (1997).
18. White, H. 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25 (1982).