Biometrics 64, 149–156
March 2008

DOI: 10.1111/j.1541-0420.2007.00834.x

Semiparametric Inference for Surrogate Endpoints with Bivariate Censored Data

Debashis Ghosh

Department of Biostatistics, University of Michigan, 1420 Washington Heights,
Ann Arbor, Michigan 48105, U.S.A.
email: ghoshd@umich.edu

Summary. Considerable attention has been recently paid to the use of surrogate endpoints in clinical research. We deal with the situation where the two endpoints are both right censored. While proportional hazards analyses are typically used for this setting, their use leads to several complications. In this article, we propose the use of the accelerated failure time model for analysis of surrogate endpoints. Based on the model, we then describe estimation and inference procedures for several measures of surrogacy. A complication is that potentially both the independent and dependent variable are subject to censoring. We adapt the Theil–Sen estimator to this problem, develop the associated asymptotic results, and propose a novel resampling-based technique for calculating the variances of the proposed estimators. The finite-sample properties of the estimation methodology are assessed using simulation studies, and the proposed procedures are applied to data from an acute myelogenous leukemia clinical trial.

Key words: Clayton–Oakes model; Copula; Linear regression; Prentice criterion; Stochastic perturbation; *U*-statistic.

1. Introduction

Recently, surrogate endpoints have become of great interest in clinical research (Biomarkers Working Group, 2001). These are endpoints that can be collected in a shorter time period and/or using fewer subjects than those normally considered in a classical clinical trial (e.g., survival). Surrogate endpoints are proposed based on biological considerations within a progression model of disease. One example is CD4 count levels in AIDS; the CD4 count can potentially serve as a surrogate endpoint for death. Another example from cancer studies is using tumor shrinkage as a surrogate endpoint for survival or disease-free survival. Formulation of an appropriate surrogate endpoint to use depends on the mechanism upon which the treatment acts. For example, a new treatment to prevent stroke may directly influence blood pressure; therefore, measuring the effect of treatment against blood pressure rather than the incidence of stroke would be appropriate. Here and in the sequel, we assume that an appropriate choice of endpoint, based upon pathway considerations, has been made.

Modeling surrogate endpoints has been the focus of much recent statistical research (Burzykowski, Molenberghs, and Buyse, 2005). A seminal paper in this area is that of Prentice (1989), who gave several conditions for inference for a surrogate endpoint to be the same as that for the true clinical endpoint. However, the criteria proposed by Prentice are very strict. In addition, they have come under criticism by several authors (Begg and Leung, 2000; Berger, 2004). Instead of using the testing-based criteria of Prentice (1989), we focus on estimation of measures of surrogacy, several of which have been proposed in the literature. We describe these quantities in Section 2.2.

Our interest is in the situation where the true and surrogate endpoints are right-censored failure timed. The major work in this area has been that of Lin, Fleming, and DeGruttola (1997) for proportion of treatment effect explained (PTE) and Burzykowski et al. (2001) for adjusted association and relative effects (RE; Buyse and Molenberghs, 1998). The regression model used by these authors was the proportional hazards model (Cox, 1972). While its use is widespread in the analysis of time-to-event data, there are several limitations in the surrogate endpoint setting. First, as noted by Lin et al. (1997), when fitting failure time models for the true endpoint and treatment in the absence and in the presence of the surrogate endpoint, proportional hazards cannot hold for both models. Second, the failure time is not directly modeled with the Cox model; instead, it is the hazard function, which is a nonintuitive quantity for clinicians. In this article, we consider inference for surrogate endpoints based on the accelerated failure time (AFT) model (Cox and Oakes, 1984, Section 5.2). It has a very similar form to that of a linear regression model and is much simpler to interpret than the proportional hazards model. We consider four measures of surrogacy in the article. For some of these measures, a major complication is that we will have to consider AFT models with right-censored covariates. To handle this problem, we extend the famous Theil–Sen estimator (Theil, 1950; Sen, 1968); we also develop a novel resampling-based scheme for variance estimation. The structure of the article is as follows. In Section 2, we define the data structures and describe four measures of surrogacy that have been proposed in the literature. In Section 3, we develop the new estimation procedure that extends the Theil–Sen estimator. There, the asymptotic properties are derived as well. Novel resampling techniques for variance estimation are provided. We then describe estimation and variability assessment of the measures of surrogacy. The finite-sample behavior of the proposed methods are assessed by simulation studies and applications to data from an acute myelogenous leukemia clinical trial in Section 4. Finally, we conclude with some discussion in Section 5.

2. Preliminaries and Definitions

2.1 *Data Structures*

Let $a \wedge b$ denote the minimum of two numbers a and b. Define $I(A)$ to be the indicator function for the event A. Let T be the failure time of the clinical endpoint, S the failure time for the surrogate endpoint, and C time to independent censoring. Let Z denote a treatment indicator (0 for control, 1 for treatment). We assume that (T, S) is independent of C given Z. We observe the data $(X_i, \delta_i^X, Y_i, \delta_i^Y, Z_i)$, $i = 1, \ldots, n$, n independent and identically distributed observations from $(X, \delta^X, Y, \delta^Y, Z)$, where $X = \log S \wedge \log C$, $\delta^X = I(\log S \leq \log C)$, $Y = \log T \wedge \log C$ and $\delta^Y = I(\log T \leq \log C)$.

2.2 *Measures of Surrogacy*

Before describing the measures of surrogacy, we first point out that to calculate the appropriate measures, we will need methods of estimation for the following regression models:

$$\log T = \beta Z + \epsilon_1, \tag{1}$$

$$\log S = \alpha Z + \epsilon_2, \tag{2}$$

and

$$\log T = \eta \log S + \gamma Z + \epsilon_3, \tag{3}$$

where ϵ_1, ϵ_2, and ϵ_3 are error terms and β, α, and (η, γ) are unknown regression coefficients to be estimated. If we use the AFT formulation for the three regression models, then fitting models (1) and (2) is straightforward, while fitting (3) is not. In Section 3.1, we propose a new method for its estimation.

The first measure of surrogacy that appears to have been proposed in the literature is the proportion of treatment effect explained (PTE), proposed by Freedman, Graubard, and Schatzkin (1992). The idea behind PTE is analogous to the attributable fraction in epidemiology. We wish to seek how much of the effect of treatment on the true endpoint is mediated through the surrogate endpoint. If the surrogate endpoint is valid in the sense of Prentice (1989), then the population PTE should equal one. At the other extreme, a population PTE of zero implies that the surrogate has no mediation effect between treatment with the true endpoint. The PTE is given by the following formula:

$$\mathrm{PTE} = \frac{\beta - \gamma}{\beta}, \tag{4}$$

where β and γ are the regression coefficients for Z in models (1) and (3).

The PTE measure has been criticized by Molenberghs et al. (2002) and Wang and Taylor (2002) for a variety of reasons. One criticism is related to that of Lin et al. (1997), namely the potential incompatabilility of models for T given Z and T, given S and Z, which are required for the calculation of PTE. Another problem is that PTE cannot handle interactions between S and Z in a natural way; in fact, PTE is not even well defined for this situation.

Buyse and Molenberghs (1998) have argued that for a surrogate endpoint to be useful, the treatment effect on the surrogate endpoint should be able to predict the treatment effect on the true endpoint. This motivates two new measures of surrogacy. The first is RE, defined by Buyse and Molenberghs (1998) as the ratio of the treatment effects on the true and surrogate endpoints. If the treatment effects on the two endpoints are the same, then RE would equal one. Buyse and Molenberghs (1998) argue that it might also be of interest to study the correlation between T and S, adjusting for the effects of Z. They term this the adjusted association measure. When T and S have a multivariate normal distribution, the adjusted association, on a standardized scale, reduces to the partial correlation of T and S given Z. Note, however, that its interpretation depends upon standardizing based on the variance of T and S. The framework of Buyse and Molenberghs is much different from that of Prentice in that the former only considers bivariate distributions and their associations, while the latter is based on a causal framework in which for a surrogate to be valid, it must lie on the causal pathway between Z and T and must capture the entire effect of Z on T. The Prentice framework cannot be verified in any strict statistical way, while the Buyse and Molenberghs framework can. Because these paradigms have such different starting points, conceptually it is hard to resolve analyses based on these two approaches.

Inspired by a technique utilized by Tsiatis et al. (1995), Wang and Taylor (2002) proposed an alternative measure to PTE, which they call the F measure. Their approach is to compare the distribution of the surrogate endpoints between the two treatment groups in which the distributions are weighted by the conditional distribution of the true endpoint, given treatment and the surrogate endpoint. With a time-to-event endpoint, Wang and Taylor (2002) suggest to make the endpoint binary based on dichotomization at a single time point. The F measure of Wang and Taylor (2002), adapted to our setting, is defined as the following:

$$F \equiv \frac{\displaystyle\int \Pr(T > c \mid Z = 0, s) f(s \mid Z = 0)\,\mathrm{ds} - \int \Pr(T > c \mid Z = 0, s) f(s \mid Z = 1)\,\mathrm{ds}}{\displaystyle\int \Pr(T > c \mid Z = 0, s) f(s \mid Z = 0)\,\mathrm{ds} - \int \Pr(T > c \mid Z = 1, s) f(s \mid Z = 1)\,\mathrm{ds}}, \tag{5}$$

where $c > 0$ is a prespecified time point. Note that we have suppressed dependence of F on c. Also, note that the surrogacy measure (5) requires specification of (2) and (3). Wang and Taylor (2002) argue consequently that because of this property, the F measure is more flexible than the PTE measure because it does not require specification of (1). Ideally, an F value of zero corresponds to the surrogate endpoint being useless, while that of one implies it is a perfect surrogate. However, there are issues with the F measure as well. First, it is not guaranteed to be between zero and one as a population quantity; Wang and Taylor (2002) provide a set of necessary and sufficient conditions for these to hold. Second, because of the ratio nature of the quantity, the confidence limits for the F measure can go outside the limits of zero and one. This criticism also applies to the PTE and RE measures.

Some further intuition about the F measure can be gleaned in the situation with uncensored data. Suppose we have models (1) and (2) holding jointly with (ϵ_1, ϵ_2) having a zero-mean bivariate normal distribution. Then for this situation, Wang and Taylor (2002) show that their F measure reduces to the PTE. We use a slightly different definition of the F measure with censored data because that of Wang and Taylor (2002) would use the mean of $\log T$ given $\log S$ and Z, which is harder to estimate with censored data. However, because we lose the feature of linearity of the mean as a function of the regression parameters, in our definition of (5), it is possible for the F measure of a perfect surrogate to be less than one.

3. Proposed Methodology

3.1 *Semiparametric Regression Model and Estimation*

We now describe the main new methodological development in this article. To construct the PTE and F measures, we need to be able to fit model (3), where $\xi \equiv (\eta, \gamma)$ are regression coefficients and ϵ_3 is an error term with an unspecified distribution. Note that a complicating feature of this model is that both the covariate and the response variable are potentially censored. For $\gamma = 0$, model (3) was previously considered by Akritas, Murphy, and LaValley (1995). They propose to estimate η using the Theil–Sen estimator (Theil, 1950; Sen, 1968). Note that in (3), η has a simple interpretation as a dependence parameter between $\log T$ and $\log S$, adjusting for Z. Equivalently, η measures the change in the average of $\log T$ associated with a one-unit change in $\log S$, holding treatment constant. Because of censoring, it is difficult to estimate an intercept term in the linear regression model (3).

Another advantage of (3) is that it is possible for both (3) and (1) to hold simultaneously. From results on ordinary linear models, the two models can potentially hold simultaneously because of linearity. By contrast, the proportional hazards versions of these models can never be true at the same time (Lin et al., 1997). This led Lin et al. (1997) to consider inference for the estimated regression coefficients based on model misspecification theory. The model and attendant theory presented here is conceptually simpler.

We now consider estimation of ξ in (3). We can estimate the regression coefficients mimicking the approach of Akritas et al. (1995). In particular, we define the following class of estimating functions:

$$\begin{aligned} U(\xi) &\equiv U(\eta;\gamma) \\ &= \sum_{i<j} \delta_i^X \delta_j^X \{I(X_i < X_j) - I(X_j < X_i)\} \\ &\quad \times \big\{\delta_i^Y I(Y_i - \eta X_i - \gamma Z_i < Y_j - \eta X_j - \gamma Z_j) \\ &\qquad - \delta_j^Y I(Y_j - \eta X_j - \gamma Z_j < Y_i - \eta X_i - \gamma Z_i)\big\}. \quad (6) \end{aligned}$$

Then $\hat{\xi} \equiv (\hat{\eta}, \hat{\gamma})$ is defined to be the solution to $U(\xi) = 0$. Note that for fixed γ, U is a monotone function of η and is hence guaranteed to have a zero-crossing. We can solve (6) to get $\hat{\eta}$ using iterated bisection root search. By symmetry, for a fixed value of η, the estimating function (6) is monotone in γ, so bisection root search can be used there as well. Thus, our algorithm is a componentwise iterated bisection root search. Let η_0 and γ_0 denote the true values of η and γ. We show in the Web Appendix that $E\{U(\xi_0)\} = 0$.

Assume that $\Pr(T > x, S > y \mid Z)\Pr(C > x \mid Z)$ is bounded away from zero for sufficiently large x and y. We know that in a neighborhood of (η_0, γ_0), $n^{-2} U(\eta; \gamma)$ converges uniformly to $E[U(\eta_0; \gamma_0)] \equiv 0$. By the strong law of large numbers for U-statistics (Van der Vaart, 2000, p. 192) and the continuous mapping theorem, $(\hat{\eta}, \hat{\gamma})$ is consistent for (η_0, γ_0). In the Web Appendix, we prove the following theorem.

THEOREM 1. *(a) Under the usual regularity conditions, $n^{-3/2} U(\xi_0)$ converges in distribution to a normal random vector with mean zero and variance matrix*

$$\mathbf{J} \equiv \mathbf{J}(\xi) = 4\rho,$$

where $\rho = \mathrm{Var}[E\{g(X_1, Y_1, Z_1, X_2, Y_2, Z_2) \mid X_1, Y_1, Z_1\}]$ *and*

$$\begin{aligned} & g(X_1, Y_1, Z_1, X_2, Y_2, Z_2) \\ &\quad = \delta_1^X \delta_2^X \{I(X_1 < X_2) - I(X_2 < X_1)\} \\ &\qquad \times \big\{\delta_1^Y I(Y_1 - \eta X_1 - \gamma Z_1 < Y_2 - \eta X_2 - \gamma Z_2) \\ &\qquad\quad - \delta_2^Y I(Y_2 - \eta X_2 - \gamma Z_2 < Y_1 - \eta X_1 - \gamma Z_1)\big\}. \end{aligned}$$

(b) $n^{1/2}(\hat{\xi} - \xi_0)$ converges in distribution to a normal random vector with mean zero and variance

$$\mathbf{A}^{-1}\mathbf{J}\mathbf{A}^{-1}, \quad (7)$$

where

$$\mathbf{A} = \lim_{n\to\infty} n^{-2} \frac{dU(\xi_0)}{d\xi}.$$

Note that if we want to estimate the asymptotic variance of $\hat{\xi}$ consistently based on (7), then this will require consistent estimation of a density function. Smoothing is needed and may be sensitive to the choice of bandwidth. We will use a resampling method for estimating the variance of the test statistics. Note that (6) has the following form:

$$U(\xi) = \sum_{i<j} T_{ij}(\xi), \quad (8)$$

where

$$\begin{aligned} T_{ij}(\xi) &= \delta_i^X \delta_j^X \{I(X_i < X_j) - I(X_j < X_i)\} \\ &\quad \times \big\{\delta_i^Y I(Y_i - \eta X_i - \gamma Z_i < Y_j - \eta X_j - \gamma Z_j) \\ &\qquad - \delta_j^Y I(Y_j - \eta X_j - \gamma Z_j < Y_i - \eta X_i - \gamma Z_i)\big\}. \end{aligned}$$

We generate n independent and identically distributed normal (0, 1) random variables $(G_1, \ldots, G_n)$ and calculate perturbations of (8):

$$U^*(\xi) = \sum_{i<j} T_{ij}(\hat{\xi})(G_i \times G_j). \quad (9)$$

Notice that in (9), the only stochastic components are $(G_1, \ldots, G_n)$. The only requirement is having mean and variance one. In the Web Appendix, we sketch a proof of the following theorem:

THEOREM 2. *The conditional distribution of* $n^{-3/2} U^*(\xi_0)$ *and the unconditional distribution of* $n^{-3/2} U(\xi_0)$ *have the same limiting distribution.*

This leads to the following algorithm for calculating the variance of $\hat{\xi}$:

(1) Generate $(G_1, \ldots, G_n)$.
(2) Set (9) equal to zero and solve for $(\hat{\eta}^*, \hat{\gamma}^*)$.
(3) Repeat steps 1 and 2 M times.

This resampling procedure is quite fast. In practice, we usually take $M = 1000$. Using Theorem 2 and Taylor series arguments, it can be shown that the conditional distribution of $n^{1/2}(\hat{\xi}^* - \hat{\xi})$ approaches that of the unconditional distribution of $n^{1/2}(\hat{\xi} - \xi_0)$. We can construct confidence intervals in two ways. The first is to calculate an estimator of standard error based on the empirical distribution of $\hat{\xi}^*$. The second is based on the $\alpha/2$th and $(1 - \alpha/2)$th percentiles of the empirical distribution of $\hat{\xi}^*$. The algorithm described here can be viewed as being related to the commonly used bootstrap algorithm.

3.2 *Estimation of Surrogacy Measures*

We now return to the four measures of surrogacy described in Section 2.2 and describe their estimation and inference procedures. We first start with PTE. We estimate β in (1) using any existing estimation procedure for the AFT model with univariate censored data (e.g., Jin et al., 2003); we estimate ξ in (3) using the estimation procedure described in Section 3.1. Based on these estimates, we estimate PTE by $\widehat{\mathrm{PTE}} = (\hat{\beta} - \hat{\gamma})/\hat{\beta}$. To calculate the variance of PTE, we use the nonparametric bootstrap (Efron and Tibshirani, 1986). Note that we are fitting models (1) and (3) simultaneously so that the estimation algorithms are applied to the same bootstrapped data sets.

Next, we consider the RE measure. To calculate it, we estimate β and α from models (1) and (2) using rank-based estimating equations for the AFT model with censored data (Jin et al., 2003), obtain estimators $\hat{\beta}$ and $\hat{\alpha}$ and estimate RE as $\widehat{\mathrm{RE}} = \hat{\beta}/\hat{\alpha}$. We can again use the nonparametric bootstrap (Efron and Tibshirani, 1986) to obtain confidence intervals for RE in a manner similar to that described in the previous paragraph.

To calculate the adjusted association measure, we formulate dependence between $\log T$ and $\log S$, adjusted for Z, using a Clayton–Oakes model (Clayton, 1978; Oakes, 1986), where the marginal failure time distributions are modeled using (1) and (2). For the Clayton–Oakes model, the crossratio function is constant; specifically,

$$\frac{\lambda(\tilde{t} \mid \tilde{S} = s)}{\lambda(\tilde{t} \mid \tilde{S} \geq s)} = \theta, \qquad (10)$$

where $\lambda(\tilde{t} \mid \tilde{S} = s)$ and $\lambda(\tilde{t} \mid \tilde{S} \geq s)$ are the hazard functions of $\tilde{T} \equiv \log T$ conditional on $\tilde{S} \equiv \log S = y$ and $\tilde{S} \geq y$, respectively. This has been the dependence model used by Burzykowski et al. (2001); however, the marginal models used there were based on the proportional hazards model.

For $i = 1, \ldots, n$, define e_i^T and e_i^S to be the residual for the ith individual from the estimates in model (1) and (2). A key result of Fine and Jiang (2000) is that the population residuals corresponding to (1) and (2) do not depend on the error distribution. Using the Clayton–Oakes model, one can estimate the dependence parameter θ in (10) using the following formula:

$$\hat{\theta} = \frac{\sum_{i<j} R_{ij}\psi_{ij}}{\sum_{i<j} R_{ij}(1 - \psi_{ij})},$$

where $R_{ij} = I(X_i \wedge X_j \leq C_i \wedge C_j,\ Y_i \wedge Y_j \leq C_i \wedge C_j)$ and $\psi_{ij} = I\{(e_i^T - e_j^T)(e_i^S - e_j^S) > 0\}$, $i, j = 1, \ldots, n$. Note that we are using the estimator from Fine and Jiang (2000, Section 2) with the weight function being equal to one. More general weight functions could be considered. Alternatively, one could take the residuals and estimate dependence using the pseudolikelihood method of Shih and Louis (1995). In the example in Section 5, we will use the procedure of Fine and Jiang (2000). Note that the residuals come from the estimates from the two models using the method of Jin et al. (2003).

Fine and Jiang (2000) show that $n^{1/2}(\hat{\theta} - \theta)$ has an asymptotically normal distribution with mean zero and variance that can be consistently estimated by observed data quantities. Here, we propose an alternative method of variance estimation. The procedure proceeds as follows:

(1) Generate n independent and identically distributed normal (0, 1) random variables with mean one, $(V_1, \ldots, V_n)$.
(2) Calculate for this bth data set,

$$\hat{\theta}_b^* = \frac{\sum_{i<j} R_{ij}\psi_{ij} V_i V_j}{\sum_{i<j} R_{ij}(1 - \psi_{ij}) V_i V_j}.$$

(3) Repeat steps 1 and 2 M times.

This procedure is computationally quite fast. In practice, we take $M = 1000$. Using Taylor series arguments and modifying the proof of Theorem 2 (see Web Appendix) to the estimating function for θ, it can be shown that the asymptotic distribution of $n^{1/2}(\hat{\theta}^* - \hat{\theta})$, conditional on data, is equivalent to that of $n^{1/2}(\hat{\theta} - \theta)$. This theoretically justifies the use of the perturbation algorithm described above.

The last measure of surrogacy we consider is the F measure of Wang and Taylor (2002). We can construct a model-based estimate of (5) by using estimates from model (2) to estimate $F(s \mid z)$ and those from model (3) to estimate $\Pr(T > u \mid z, s)$ for $z = 0$, 1. Note that differentiating $F(s \mid z)$ gives $f(s \mid z)$. For these computations, the baseline distributions for the models (2) and (3) are required. We estimate $F(s \mid z)$ from (2) by $\hat{F}_0\{s \exp(-\hat{\alpha} z)\}$, where $\hat{\alpha}$ is the estimate from Jin et al. (2003), and $\hat{F}_0$ is one minus the Kaplan–Meier estimator of the transformed survival times $(X_i e^{-\hat{\alpha} Z_i}, \delta_i^X), i = 1, \ldots, n$. Similarly, we estimate $\Pr(T > t \mid z, s)$ as $\hat{S}_0\{t \exp(-\eta \log s - \gamma z)\}$, where $\hat{\eta}$ and $\hat{\gamma}$ are the estimates using the methodology in Section 3.1, and $\hat{S}_0$ is the Kaplan–Meier estimator applied to the times $(Y_i e^{-\hat{\eta} \log X_i - \hat{\gamma} Z_i}, \delta_i^Y), i = 1, \ldots, n$. Again, the nonparametric bootstrap can be used for constructing confidence intervals, following the recommendations of Wang and Taylor (2002).

4. Numerical Examples

4.1 *Simulation Studies*

We conducted simulation studies to assess the finite-sample properties of the proposed estimators in the article. First, we assess the finite-sample properties of the estimation methodology in Section 3.1. Let Z denote a binary indicator. Data are generated using a bivariate normal distribution for $(\log T, \log S)$ conditional on Z with mean vector $(\mu_{\mathrm{TZ}}, \mu_{\mathrm{SZ}})$ and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Note that we assume a common covariance matrix for both populations. Since we are mimicking a clinical trial situation, we assume that $P(Z = 1) = 0.5$ for the simulation. We set $\mu_{T1} - \mu_{T0} = 1.2$ and $\mu_{S0} = \mu_{S1} = 0$. The conditional distribution of $\log T$ given $\log S$ and Z satisfies model (3) with $\eta = \rho$ and $\gamma = \mu_{T1} - \mu_{T0}$. We considered $n = 50$, 100, and 150 and $\rho = 0$, 0.2, and 0.8. Independent censoring (C) was generated using a Uniform[0,5] random variable distributed independently of (S, T). This yielded on average 30% censoring for S and 45% censoring for T. For each simulation setting, 1000 simulation samples were generated and 1000 resamplings were generated for each simulation sample. The results are presented in Table 1. We find that the proposed estimators perform satisfactorily in finite samples.

Next, we conducted simulation studies of the PTE and F measure to assess their finite-sample performance. We focus on these two measures because their estimation requires that from Section 3.1. We use the same simulation setup as above. Following the arguments of Section 3.3 of Wang and Taylor (2002), we find that

$$\mathrm{PTE} = \frac{\rho(\mu_{S1} - \mu_{S0})}{(\mu_{T1} - \mu_{T0})}. \tag{11}$$

As mentioned earlier, the population F value from (5) will be different from (11) because of the presence of the conditional survival functions in lieu of the mean, as done by Wang and Taylor (2002). We set $c = 2$. Several settings were used. In the first, we set $(\mu_{T0}, \mu_{T1}) = (1, 2)$, $(\mu_{S0}, \mu_{S1}) = (1, 2)$ and $\rho = 1$. This yields a population PTE measure of one, corresponding to a perfect surrogate. The F measure, calculated using (5), is 0.97; it is less than one because of the use of the survival probability in (5). Second, we set $(\mu_{T0}, \mu_{T1}) = (1, 2)$, $(\mu_{S0}, \mu_{S1}) = (0.5, 0.5)$ and $\rho = 0$, which yields a population value in (11) of zero and corresponds to the surrogate being useless; here the F measure will also be zero. Finally, we set $(\mu_{T0}, \mu_{T1}) = (1, 2)$, $(\mu_{S0}, \mu_{S1}) = (1, 2)$ and $\rho = 0.5$, which yields a (11) of 0.5 and represents a practical value of the surrogate. The F measure will be 0.48. The censoring percentages for S and T were similar to those in the previous paragraph. Again, sample sizes $n = 50$, 100, 200 were considered. We again generated 1000 simulated data sets with 1000 perturbed data sets within each simulation. The results are given in Table 2. Based on the table, we find that the F measure and the PTE measure estimators perform well, with the F measure being slightly more efficient. There is some slight negative bias in the estimators for smaller sample sizes, but it diminishes in larger samples. A small fraction of the confidence intervals fall outside (0, 1), but this tends to be a fairly small percentage.

4.2 *Leukemia Data*

We now apply the proposed methodology to a Phase III clinical trial dealing with therapies for acute leukemia (Berman et al., 1991). Acute leukemia is a severe form of hematological cancer in which blood-forming cells undergo changes resulting in uncontrolled, malignant growth. This disease is characterized by excessive numbers of abnormal white blood cells that are limited in their ability to fight infections. Because the

Table 1
Simulation results for proposed estimator

		Estimate of γ					Estimate of η				
n	β	Bias	SE	SEE	CI1	CI2	Bias	SE	SEE	CI1	CI2
50	0	0.02	0.35	0.42	0.96	0.97	0.01	0.30	0.26	0.97	0.96
	0.2	0.01	0.37	0.40	0.95	0.96	0.01	0.28	0.26	0.97	0.96
	0.8	−0.01	0.39	0.41	0.96	0.96	0.03	0.30	0.26	0.97	0.96
100	0	0.02	0.25	0.24	0.95	0.95	0.02	0.21	0.20	0.96	0.95
	0.2	−0.01	0.24	0.26	0.96	0.96	0.01	0.21	0.20	0.96	0.95
	0.8	−0.02	0.27	0.28	0.95	0.95	−0.01	0.21	0.20	0.96	0.95
150	0	0.01	0.19	0.19	0.95	0.94	0.0	0.17	0.16	0.96	0.95
	0.2	0.02	0.21	0.20	0.94	0.94	0.02	0.17	0.17	0.96	0.95
	0.8	0.01	0.22	0.21	0.95	0.94	0.02	0.18	0.18	0.96	0.95

Note: SEE denotes average of standard errors based on empirical distribution of $\hat{\xi}^*$; CI1 denotes coverage probability of 95% CI using standard error calculated from empirical distribution of $\hat{\xi}^*$; CI2 denotes coverage probability of 95% CI using 2.5th and 97.5th percentiles calculated from empirical distribution of $\hat{\xi}^*$.

Table 2
Simulation results for F and PTE measures

		F measure					PTE				
n	ρ	Bias	SE	SEE	CI	NO	Bias	SE	SEE	CI	NO
50	0	−0.07	0.32	0.29	0.97	1	−0.08	0.41	0.37	0.96	1
	0.5	−0.06	0.34	0.31	0.96	2	−0.06	0.43	0.38	0.97	1
	1	−0.09	0.36	0.32	0.96	1	−0.07	0.42	0.37	0.97	2
100	0	−0.02	0.21	0.21	0.95	2	0.02	0.30	0.27	0.96	1
	0.5	−0.01	0.23	0.21	0.96	2	0.01	0.30	0.28	0.96	1
	1	−0.02	0.24	0.22	0.95	1	−0.01	0.28	0.27	0.95	3
200	0	−0.01	0.17	0.17	0.95	1	0.01	0.21	0.20	0.95	1
	0.5	0.00	0.18	0.17	0.94	1	0.02	0.20	0.20	0.95	2
	1	−0.01	0.17	0.17	0.94	1	−0.02	0.19	0.18	0.95	2

Note: See note to Table 1. CI denotes coverage probability of 95% CI using standard error calculated from empirical distribution of $\hat{\xi}^*$. NO (number of confidence intervals outside [0,1]) indicates number of 95% confidence intervals where the left endpoint is smaller than zero or the right endpoint is bigger than one.

cancerous cells replace the blood-forming cells in the bone marrow, leukemia patients will also have low numbers of platelets in the circulating blood. Here, the standard therapy was daunorubicin (DNR) and cytosine arabinoside (Ara-C), while the experimental therapy was idarubicin (IDR) and Ara-C, all of which are chemotherapeutic agents. Note that a placebo group with no treatment is not utilized in this study because the standard of care is the DNR/Ara-C combination therapy.

We explore the issue of the usefulness of the effect of IDR/Ara-C on survival, as mediated through the surrogate endpoint of complete remission. Remission refers to the endpoint of leukemia cells being killed by the chemotherapies. Thus, T represents time to death, while S represents time to complete remission, both of which are potentially right-censored random variables.

There are 130 patients in the study, 65 in each treatment arm. Kaplan–Meier plots of the survival distributions for T and S, by treatment group, are given in Figures 1 and 2. A log-rank test for time to remission between the two treatment arms yielded test statistic of 9.04, which corresponds to a p-value of 0.003. A log-rank test for time to death between the two treatment arms yielded a test statistic of 5.3, which corresponds to a p-value of 0.02. Based on these analyses, we find that IDR/Ara-C leads to earlier remission times and increased survival, suggesting its benefit relative to DNR/Ara-C.

Next, we fit marginal AFT models for time to complete remission and for time to death with the covariate being treatment. This was done using the estimation procedure proposed by Jin et al. (2003). Based on these models, the estimated effect of treatment on complete remission is −0.69 with an associated standard error of 0.34, while the corresponding value for survival is 0.51, with an associated standard error of 0.24. This yields a RE of 0.74 in magnitude. Interestingly, the same analysis using proportional hazards regression model yields a RE estimate of 0.79, which is very similar to the AFT analysis. A nonparametric bootstrap using 1000 replications yields a 95% CI of (0.40, 1.15) for the magnitude of the RE using the percentile method.

Next, we study the dependence of the surrogate and true endpoints, adjusting for the treatment using the Clayton–Oakes copula model, which leads to computing the RE and adjusted association. The estimate of the dependence parameter using the Fine and Jiang (2000) estimator is 0.61; the 95% confidence interval based on the resampling method proposed in Section 2.2 with 1000 resamplings is (0.47, 0.83). Thus, after adjusting for treatment, there is statistically significant negative association between the surrogate endpoint (time to remission) and the true endpoint (time to death). This is consistent with the biology of the disease (Haferlach et al., 2005).

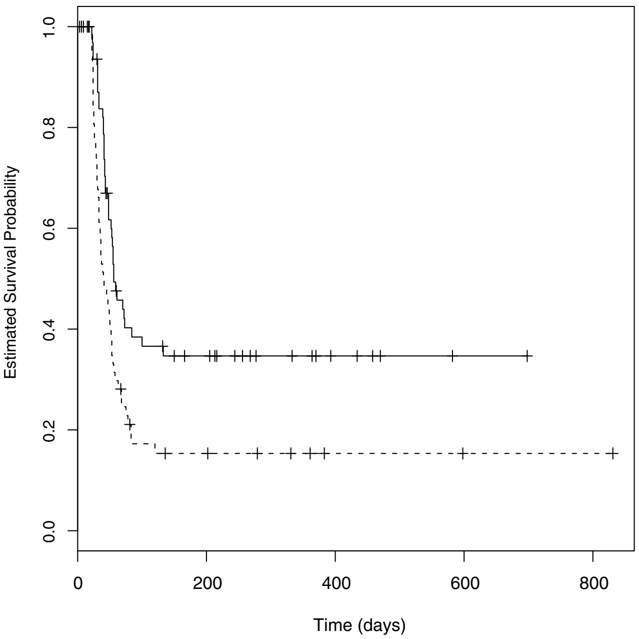

Figure 1. Kaplan–Meier plots for time to remission by treatment group: the dashed line represents the IDR/Ara-C arm, while the solid line represents the DNR/Ara-C arm.

Next, we construct the estimates of PTE and F using the methods in Section 3. First, we fit the linear regression model (3) to get an estimate of the treatment effect on death, adjusting for remission status. The parameter estimates are

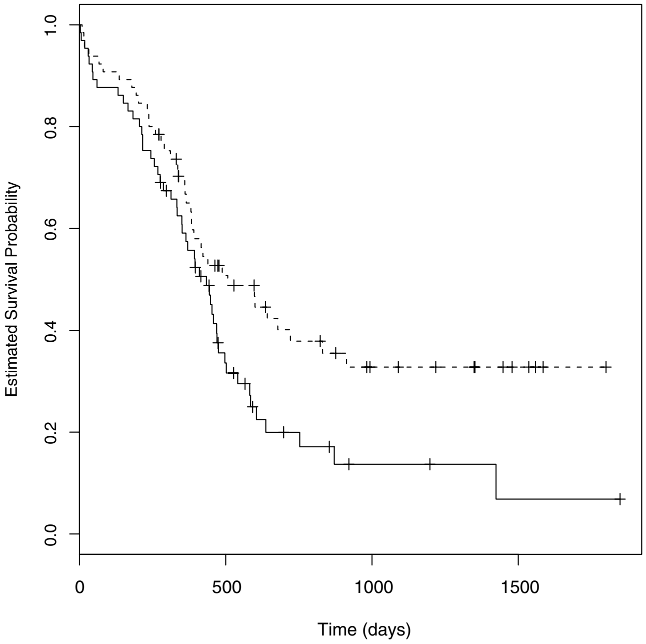

Figure 2. Kaplan–Meier plots for time to death by treatment group: the dashed line represents the IDR/Ara-C arm, while the solid line represents the DNR/Ara-C arm.

Table 3

Regression coefficient estimates and associated 95% *confidence intervals of* (3) *for AML clinical trial data*

Leukemia type	Estimate	95% CI (percentile)	95% CI (SE)
Time to remission	−0.23	(−0.01, −0.47)	(−0.02, −0.42)
Treatment	0.09	(−0.10, 0.34)	(−0.04, 0.20)

Note: Treatment is coded as 0 for DNR/Ara-C and 1 for IDR/ARA-C. Time to remission and survival time are on a logarithmic scale.

summarized in Table 3. Based on the results, we still find a significant effect of time to remission on death, adjusted for treatment. However, the treatment effect on death is no longer significant if we adjust for the surrogate.

Based on the results in Table 3, we are now ready to calculate the measures of surrogacy described in Section 3.2. The PTE by remission is $(0.51 - 0.09)/0.51 = 0.82$, with an associated 95% CI of (0.54, 1.05) using the nonparametric bootstrap. This suggests that remission mediates a large proportion of the effect of IDR/Ara-C on survival. The estimates for (5) for several values of c, along with associated 95% confidence intervals using the nonparametric bootstrap with 1000 replications, is given in Table 4. Note that most of the estimates of (5), along with the upper confidence limits are greater than one. This reflects the underlying variability associated with the use of the ratio-based measures of surrogacy. A practical recommendation is to conclude that an endpoint serves as a useful surrogate for the true endpoint if the lower limit for the 95% CI for the PTE exceeds 0.5 (Freedman et al., 1992; Petrylak et al., 2006). Based on this rule, we find that there is evidence that the treatment effect on survival is mediated through the remission endpoint. The same conclusion is obtained if we apply the same rule to the F measure. Given the biology of leukemias, this is consistent with the current medical opinion as to managing the disease (Haferlach et al., 2005).

Table 4

Estimates of the F measure from (5) *and associated* 95% *confidence intervals for AML clinical trial data*

c	Estimate	95% CI (percentile)
1 years	1.10	(0.85, 1.16)
2 years	1.15	(0.75, 1.34)
3 years	1.10	(0.72, 1.46)
5 years	1.47	(0.61, 1.84)
7 years	1.06	(0.71, 2.03)

Note: c represents the cutoff time in (5). Estimate is the estimated value of F, defined in (5). 95% CI is based on the 2.5th and 97.5th percentiles of the empirical distribution based on 1000 bootstrap samples.

5. Discussion

As argued by many authors, the Prentice criterion is a notoriously difficult one to prove in order to demonstrate validness of a surrogate. We have taken the approach of quantifying surrogacy using several measures. We have focused on the use of a linear regression model for the analysis of surrogate endpoints. There are advantages of using such a model in terms of interpretability of regression coefficients and compatibility of the models being fit. While much of the surrogacy calculations are straightforward, what becomes more difficult to calculate are measures of surrogacy such as PTE (Freedman et al., 1992) or the F measure of Wang and Taylor (2002). This is because these approaches require fitting a model for T given S and Z, and such a model has right-censored covariates and response variables. We have extended the estimation procedure of Akritas et al. (1995) to this setting and developed novel variance estimation procedures along with the needed asymptotic theory. Our simulation study results suggest that the proposed estimation procedures perform well in small samples. In principle, our methodology can be extended to the general case in which there are arbitrary numbers of right-censored and uncensored covariates in the model. One issue is that potentially the maximum of the F measure (5) may be less than one. How to calibrate it is an open question and should be studied further.

For the example considered in the article, it was possible for subjects to die without having achieved remission. This raises the issue of whether one should treat death as a competing risk. The topic is controversial, although there has been work done on treating the data as so-called semicompeting risks data (Fine, Jiang, and Chappell, 2001). For the purposes of the article, we ignored that issue here.

We used both the perturbation algorithm and the nonparametric bootstrap to assess the variability of estimators in the algorithms used here. Efron (1981, p. 314) writes that the validity of the bootstrap with censored data requires that the censoring mechanism not "look into the future." While the simulation studies suggest that the use of the bootstrap is valid, further theoretical investigation is needed here.

We focus on both the PTE and the F measure in this article. One could imagine letting the value of c vary in (5), which yields a time-dependent extension of the F measure proposed here. This extension is currently under exploration.

6. Supplementary Materials

The Web Appendix referenced in Sections 3.1 and 3.2 is available under the Paper Information link at the *Biometrics* website `http://www.tibs.org/biometrics`.

Acknowledgements

The author would like to thank Dr Jeremy Taylor for useful discussions. This research is supported in part by the National Institutes of Health through the University of Michigan's Cancer Center Support Grant (5 P30 CA46592) and by the Early Detection Research Network. The author would also like to thank the associate editor and two referees, whose comments have greatly improved the manuscript.

References

Akritas, M. G., Murphy, S. A., and LaValley, M. P. (1995). The Theil–Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association* **90,** 170–177.

Begg, C. B. and Leung, D. H. Y. (2000). On the use of surrogate endpoints in randomized trials (with discussion). *Journal of the Royal Statistical Society, Series A* **163,** 15–28.

Berger, V. W. (2004). Does the Prentice criterion validate surrogate endpoints? *Statistics in Medicine* **23,** 1571–1578.

Berman, E., Heller, G., Santorsa, J., McKenzie, S., Gee, T., Kempin, S., Gulati, S., Andreeff, M., Kolitz, J., and Gabrilove, J. (1991). Results of a randomized trial comparing idarubicin and cytosine arabinoside with daunorubicin and cytosine arabinoside in adult patients with newly diagnosed acute myelogenous leukemia. *Blood* **77,** 1666–1674.

Biomarkers Working Group. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* **69,** 89–95.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics* **50,** 405–422.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints.* New York: Springer-Verlag.

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

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76,** 312–319.

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

Fine, J. P. and Jiang, H. (2000). On association in a copula with time transformations. *Biometrika* **87,** 559–571.

Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika* **88,** 907–919.

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

Haferlach, T., Kern, W., Schnittger, S., and Schoch, C. (2005). Modern diagnostics in acute leukemias. *Critical Reviews in Oncology and Hematology* **56,** 223–234.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90,** 341–353.

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

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* **23,** 607–625.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73,** 353–361.

Petrylak, D. P., Ankerst, D. P., Jiang, C. S., et al. (2006). Evaluation of prostate-specific antigen declines for surrogacy in patients treated on SWOG 99-16. *Journal of the National Cancer Institute* **98,** 516–521.

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

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* **63,** 1379–1389.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51,** 1384–1399.

Theil, H. (1950). A rank invariant method of linear and polynomial regression analysis. *Koninklijke Nederlandse Akademie var Wetenschappen Proceedings* **53,** 386–392.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90,** 27–37.

Van der Vaart, A. (2000). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

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

Received June 2006. *Revised February* 2007.
Accepted March 2007.