Published in final edited form as:

Biometrics. 2012 March ; 68(1): 226–232. doi:10.1111/j.1541-0420.2011.01633.x.

Meta-analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling

Debashis Ghosh¹, Jeremy M. G. Taylor², and Daniel J. Sargent³

¹Departments of Statistics and Public Health Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A

²Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48103, U.S.A

³Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, U.S.A

Summary

There has been great recent interest in the medical and statistical literature in the assessment and validation of surrogate endpoints as proxies for clinical endpoints in medical studies. More recently, authors have focused on using metaanalytical methods for quantification of surrogacy. In this article, we extend existing procedures for analysis based on the accelerated failure time model to this setting. An advantage of this approach relative to proportional hazards model is that it allows for analysis in the semicompeting risks setting, where we model the region where the surrogate endpoint occurs before the true endpoint. Several estimation methods and attendant inferential procedures are presented. In addition, between-and within-trial methods for evaluating surrogacy are developed; a novel principal components procedure is developed for quantifying trial-level surrogacy. The methods are illustrated by application to data from several studies in colorectal cancer.

Keywords

Copula; Dependent censoring; Latent factor; Linear regression; Multivariate failure time data; Singular value decomposition

1. Introduction

Biomedical researchers, and particularly those in the pharmaceutical industry, have great interest in using surrogate markers, if they can be shown to be valid. When the true endpoints are rare, occur later or are very expensive, the use of valid surrogate markers can substantially reduce clinical trial duration and size, lower the trial's expense, and lead to earlier decision making.

Correspondence to: Debashis Ghosh.

6. Supplementary Materials: Functions for implementing the proposed methods in R, in conjunction with dynamically compiled code, are available as Supplementary Materials. The code can run on an Apple computer running R64 available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

Many paradigms have been put forward for the assessment of surrogacy; a recent account can be found in the book by Burzykowski, Molenberghs, and Buyse (2005). Historically, the first criterion, or rather set of criteria, for surrogacy was proposed by Prentice (1989). Estimation-based alternatives to the Prentice criterion have been proposed by several authors (Freedman, Graubard, and Schatzkin, 1992; Buyse and Molenberghs, 1998; Wang and Taylor, 2002). More recently, many researchers have begun to look at surrogacy in the multitrial or metaanalytical setting (e.g., Daniels and Hughes, 1997; Gail et al., 2000; Burzykowski et al., 2005, chapters 7, 9, 10).

As in Ghosh (2008, 2009), we focus on the situation where both the surrogate and true endpoints are time to events. Methods for assessing surrogacy based on the proportional hazards (PH) model and the accelerated failure time (AFT) model have been proposed by Burzykowski et al. (2001) and Ghosh (2008). However, if the goal is to have surrogate endpoints that occur sooner than the true endpoint so that the study duration is potentially shorter, then a conceptually appealing framework to consider is that of semicompeting risks data (Fine, Jiang, and Chappell, 2001; Ghosh, 2006, 2009). We define the data structure in Section 2.2; the basic idea is that the region where the time to the surrogate endpoint is less than the time to the true endpoint is the relevant one for making inferences. In the setting of a single study, Ghosh (2009) recently evaluated many existing measures of surrogacy under the semicompeting risks data structure. Given the increasing popularity of the metaanalytic framework described above, it is of interest to extend the semicompeting risk framework to this setting as well. Burzykowski et al. (2001, 2005, chapter 11) describe methods for assessment of surrogate end-points in the bivariate failure time setting, focusing primarily on the PH model. They give a discussion of individual-level surrogacy, which roughly assesses concordance of the outcomes at the individual patient level, versus trial-level surrogacy, which studies the treatment effects on both endpoints across the trials. However, they give less attention to the AFT model and do not account for the semicompeting risks structure.

In this article, we consider statistical models for meta-analysis based on the AFT model for bivariate survival data, incorporating the semicompeting risks data structure. We view the semicompeting risks approach very attractive for surrogacy; this is discussed in Section 2.1. The methodology will be illustrated using data from a colorectal meta-analysis, where the surrogate endpoint is recurrence and the true end-point is death. A differing set of data was previously considered by Sargent et al. (2005). Our novel contributions/insights are the following:

1. Using the semicompeting risks is an appealing paradigm when we wish to consider replacement of the true endpoint by the surrogate endpoint;
2. Development of metaanalytic methods for the AFT model in the presence of semicompeting risks; and
3. Development of a suite of R functions, combined with dynamically loaded C objects, for implementing the proposed methodology.

The structure of the article is as follows. In Section 2, we will describe the motivating study and define the appropriate data structures. We then describe the probability models and

estimation and inference procedures in Section 3. Our proposed methodology is applied to the real dataset in Section 4. Finally, we conclude with some discussion in Section 5.

2. Preliminaries and Background

2.1 Observed Data Structures and AFT Regression Model

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 S denote time to the surrogate endpoint, T the time to the clinical end-point, and C time to independent censoring. Assume that the joint distribution of (S, T, C) is continuous. In this article, we focus on the situation where the true and surrogate endpoints are both times to event. Let Z be the indicator for treatment group ($0 = \text{control}$; $1 = \text{treatment}$). We make the assumption that (S, T) is independent of C given Z . The data consist of $(X_{ij}, \delta_{ij}^X, Y_{ij}, \delta_{ij}^Y, Z_{ij})$ ($i = 1, \dots, I; j = 1, \dots, n_i$). Note that there are two indexes; the first indexes study, whereas the second indexes individuals within a study. We consider the following data structure: for a fixed i , the data are a random sample from $(X, \delta^X, Y, \delta^Y, Z)$, where $X = S \wedge T \wedge C$, $\delta^X = I(S \leq T \wedge C)$, $Y = T \wedge C$ and $\delta^Y = I(T \leq C)$. Note that S is censored by the minimum of T and C and not just by C . By contrast, T is only subject to independent censoring. This type of data structure has been called semicompeting risks in the recent statistical literature (Fine et al., 2001). We refer to this as the “presence of semicompeting risks” paradigm. For the semicompeting risks setting, the following quantities are of potential interest:

- The distribution of the surrogate endpoint in the absence of **both** the true endpoint and independent censoring, potentially adjusting for covariates.
- The distribution of the true endpoint in the absence of independent censoring, potentially adjusting for covariates.
- The correlation between the surrogate and true end-points, adjusting for covariates.

In thinking about the surrogate endpoint, (a) refers to the pure surrogate endpoint. In the colorectal cancer example to follow in Section 2.2., (a) refers to the distribution of time to recurrence in the absence of death and censoring. This requires formulation of a latent time to recurrence for all individuals, even those who die without recurrence. This issue is one that has been subject to debate in the competing risks literature (Prentice et al., 1978).

Finally, we note that in fact the semicompeting risks analysis uses time to recurrence as the surrogate endpoint, which is different from the disease-free survival (DFS) composite endpoint used in Sargent et al. (2005). In particular, subjects who die without recurrence are treated as (dependently) censored in the semicompeting risks framework but as having an event in the Sargent et al. analyses.

2.2 Motivating Study: Meta-analysis of Colorectal Cancer Clinical Trials

Using several surrogacy methods, Sargent et al. (2005) assessed DFS as a potential surrogate endpoint for overall survival (OS) in colon cancer adjuvant trials. They were interested in assessing the effects in trials that included fluorouracil-based treatment on these outcomes

and performed a pooled analysis of individual-level data on 20,898 patients from 18 phase III trials. They were also interested in determining if DFS at earlier time points would be predictive of OS at a later point in time (e.g., if DFS at 2 years was predictive of OS at 4 years). There was substantial variation in the length of follow-up of patients from the individual trials, so all subjects were censored 8 years from randomization across the individual studies.

Sargent et al. (2005) presented a comparison of a variety of surrogate endpoint methods with goal of having robust findings that were not sensitive to assumptions from any one approach. They found a strong association between disease-free survival assessed after a median follow-up of 3 years with OS assessed after 5 years median followup. The strong correlation between DFS and OS was within individual patients, individual trials and between trials. The measures that they used for this assessment will be described in Section 3. Based on the findings, Sargent et al. (2005) conclude that “DFS can be considered an appropriate primary end point in the setting of clinical trials in adjuvant colon cancer.” Most of their analyses were based on PH modeling following the approach of Burzykowski et al. (2005).

The question that arises here is how to incorporate information on T into S . Sargent et al. (2005) incorporated the information by using the outcome measure DFS, which is in fact a composite endpoint. If S denotes time to colorectal cancer recurrence, and T denotes time to death, the surrogate endpoint proposed by Sargent et al. (2005) is $S \wedge T$. As explained in Sargent et al. (2005), for the study population, 80% of deaths were preceded by recurrence, so the DFS endpoint is mostly dominated by recurrence. However, when death without recurrence is observed in a larger proportion of individuals, DFS becomes more similar to time to death, so that assessing it as a surrogate for time to death would not be useful. In addition, one can also make the criticism from a biological viewpoint that deaths without recurrence might not be cancer-related deaths so that noncancer- and cancer-related events are being mixed in together in the creation of a composite endpoint.

We now explain how the semicompeting risks paradigm would apply to this setting. The main difference is in the treatment of the time to the surrogate endpoint. For semi-competing risks, the time to recurrence is considered to be dependently censored by the time to death. This approach conceptually formulates a latent time to recurrence for all individuals. It is appealing because it avoids the creation of composite endpoints, focuses on modeling S while incorporating the information on time to death (i.e., the true endpoint) as a dependent censoring mechanism. However, semicompeting risks is also a *joint* modeling framework in that we are also modeling the effect of treatment on time to death. In addition, we will also estimate the association between the time to the surrogate endpoint and time to death. Methods for doing this are presented in Section 3. When one uses composite endpoints, the mixing of the true and surrogate endpoints does not allow understanding the direct effect of Z on S or the association between S and T , conditioning on Z . All of these quantities can be estimated using the semicompeting risks framework; thus, we feel that the semicompeting risks framework can generate insights complementary to those found by existing methods.

In this article, we deal with a subset of the studies used in the meta-analysis by Sargent et al. (2005). In particular, we analyze data from 12 studies that are available; details of the patient

population of the studies are given in Table 1. We excluded patients with stage I colorectal cancer as well those with either zero values for time to recurrence or time to death. There are some interesting features to notice about the table. First, although many of the studies have roughly a 50% breakdown in each of the treatment groups, several studies have noticeably different proportions in the two arms. These are studies C04 and all those from the North Central Cancer Trials Group (NCCTG). For these studies, we have taken multiple treatment arms from the initial study and combined them to form one treatment arm. There is also some heterogeneity in the followup times. The shortest and longest median followup times across the studies differ by an approximate factor of two. Table 1 also summarizes the study-specific results of log-rank tests comparing the two groups for three types of endpoints: time to recurrence, time to death, and time to first event, which corresponds to DFS. For multiple reasons (e.g., differing censoring rule, differing availability of followup, etc.), the results in Table 1 may differ from those in the primary clinical manuscript from each trial.

Our wish to adjust for the dependent censoring by the true endpoint complicates estimation and inference procedures. Another unique feature of the semicompeting risks problem is that we wish to utilize information on the region $S \leq T$. This is called the “wedge region” and seems pertinent for our context. For surrogacy, if we deal with the goal of finding an endpoint to replace the true endpoint, then we do wish to make the wedge restriction. Practically speaking, if the surrogate endpoint occurs, then this will trigger some type of intervention (here, treatment for recurrence) that may alter the association between S and T . By restricting to $S \leq T$, we do not have to worry about the change in association between S and T due to the intervention.

We will be fitting the following models to each study:

$$\log T = \beta Z + \varepsilon_1, \quad (1)$$

and

$$\log S = \alpha Z + \varepsilon_2, \quad (2)$$

where α and β are scalar regression coefficients and ε_1 and ε_2 are mean-zero error terms. These are known as AFT models, which directly model the failure times and are a useful alternative to PH model. If we make no constraints on the ordering and censoring of S and T , then models (1) and (2) can be treated as marginal models and estimated separately. However, this cannot be done when there are semicompeting risks because of the constraint on the joint distribution of $(\varepsilon_1, \varepsilon_2)$. This reinforces the notion that we are fitting a joint model to the data in the setting of semicompeting risks.

3. Proposed Methodology

3.1 General Algorithm for Trial-Level Surrogacy

We suggest the following two-step strategy for estimation:

- a. Estimate β_i and α_i for each study using methods for the AFT model. Call the resulting estimators $\hat{\beta}_i$ and $\hat{\alpha}_i$.
- b. Based on a joint modeling approach, described in Section 3.4, regress $\hat{\beta}_i$ on $\hat{\alpha}_i$.

We refer to steps (a) and (b) as within-trial estimation and between-trial estimation.

Measures of within-trial surrogacy will be calculated using the output in (a). Based on the output of (b), we will be in a position to calculate between-study measures of surrogacy. We will describe methods of estimation for each of these steps in turn.

3.2 Within-Trial Estimation in the Presence of Semicompeting Risks

We first consider the treatment effect on the true endpoint. Because T is only subject to independent censoring by C , we can still use the ordinary log-rank estimating function (Louis, 1981) for estimation. This is given by

$$U_1(\beta) = \sum_{i=1}^n \delta_i^Y \left[Z_i - \frac{\sum_{j=1}^n I\{\tilde{Y}_j(\beta) \geq \tilde{Y}_i(\beta)\} Z_j}{\sum_{j=1}^n I\{\tilde{Y}_j(\beta) \geq \tilde{Y}_i(\beta)\}} \right], \quad (3)$$

where $\tilde{Y}_i(\beta) = Y_i \exp(-\beta' Z_i)$, $i = 1, \dots, n$. The consistency and asymptotic normality for $\hat{\beta}$, the zero-crossing of $U_1(\beta)$, was given in Ghosh (2008).

However, for the estimation of α , we must take into account the dependent censoring of S by T . The dependent censoring is adjusted for through use of an artificial censoring technique (Lin, Robins, and Wei, 1996). We artificially trim the transformed surrogate endpoint time by a factor that allows for valid comparison between the two treatment groups. Define $\eta = (\alpha, \beta)$. This leads to the following estimating equation for estimation of η :

$$U_2(\eta) = \sum_{i=1}^n \tilde{\delta}_i^X(\eta) \left[Z_i - \frac{\sum_{j=1}^n I\{\tilde{X}_j(\eta) \geq \tilde{X}_i(\eta)\} Z_j}{\sum_{j=1}^n I\{\tilde{X}_j(\eta) \geq \tilde{X}_i(\eta)\}} \right], \quad (4)$$

where $\tilde{X}_i(\eta) = \{S_i \exp(-\alpha Z_i) \wedge T_i \exp(-\beta Z_i - d) \wedge C_i \exp(-\beta Z_i - d)\}$, $\tilde{\delta}_i^X(\eta) = I\{S_i \exp(-\alpha Z_i) \leq T_i \exp(-\beta Z_i - d) \wedge C_i \exp(-\beta Z_i - d)\}$ and $d = 0$ if $\alpha = \beta$ and $\beta - \alpha$ otherwise. Let $\hat{\alpha}$ be a zero-crossing of α from setting $U_2(\tilde{\eta}) = 0$, where $\tilde{\eta} = (\alpha, \hat{\beta})$. The role of d is to artificially censor observations in one of the treatment arms depending on the relative magnitudes of the treatment effects on the true and surrogate endpoints. Note that although U_1 is a function of β only, U_2 depends on both α and β .

Ghosh (2009) proposed a resampling method for estimating the variance–covariance matrix of $\hat{\eta}_j$. Such a resampling is needed because consistent estimation of the variance of the regression coefficients requires complicated nonparametric estimation of the density of the errors and its derivative. We take an alternative approach based on resampling recently proposed by Zeng and Lin (2008). The algorithm works as follows:

- (R1) Generate observations $\mathbf{G} \equiv (G_1, G_2)$, a bivariate observation with mean zero vector and variance–covariance matrix $\hat{\mathbf{V}}$, the estimated variance–covariance matrix of $n^{-1/2}\{U_1(\beta), U(\eta)\}$:
- (R2) Calculate $n^{-1/2}[U_1(\hat{\beta} + n^{-1/2}G_1), U_2(\hat{\eta} + n^{-1/2}\mathbf{G})]$. Note that U_2 depends on both α and β , which requires the inclusion of both G_1 and G_2 in the resampling algorithm.
- (R3) Repeat steps (R1) and (R2) B times.
- (R4) Regress $n^{-1/2}U_1(\hat{\beta} + n^{-1/2}G_1)$ on G_1 across the B datasets. Similarly, regress $n^{-1/2}U_2(\hat{\eta} + n^{-1/2}\mathbf{G})$ on G_1 and G_2 across the B datasets.
- (R5) Estimate the variance-covariance matrix of $n^{1/2}(\hat{\eta} - \eta_0)$ as $\hat{\mathbf{A}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{A}}^{-1}$, where the first row of $\hat{\mathbf{A}}$ is the slope estimate from the first regression in the previous step, whereas the second row of $\hat{\mathbf{A}}$ are slope estimates from the second regression in the previous step.

Modifying the arguments in Zeng and Lin (2008), we show in Appendix A that this algorithm provides a consistent estimator of the variance–covariance matrix of $n^{1/2}(\hat{\eta} - \eta_0)$. In our experience, we have found this algorithm to be much faster than that of Ghosh (2009).

3.3 Between-Trial Estimation: Joint Modeling Algorithm

Buyse et al. (2000) developed measures of between- and within-trial association using R^2 measures. A surrogate for which the individual study-specific R^2 is one is a perfect surrogate at the individual level. This says that the surrogate and true endpoints are perfectly correlated within an individual subject. A surrogate endpoint for which the R^2 between the study-specific treatment effects on the surrogate and true outcome is one is a perfect surrogate at the trial level. Buyse et al. (2000) and Gail et al. (2000), among others, have argued that the between-trial association is a far more important characteristic of the surrogate than the within-trial measure for predicting the treatment effect in the new trial, although this is not the case when T is partially observed from the new trial (Li and Taylor, 2010). Given estimates $(\hat{\alpha}_j, \hat{\beta}_j)$, $j = 1, \dots, J$, we now seek to estimate the relationship between the two regression coefficients. As discussed in chapter 11 of Burzykowski et al. (2005), one can either assume that there is no error in the estimated regression coefficients or that there is error. The assumption made at this stage dictates the appropriate method of analysis used. We adopt the approach of assuming error in estimation, which is always present in practice.

We note that there is in fact an approximate equivalence between the models being fit in Section 11.3 of Burzykowski et al. (2005) with principal components analysis (PCA) techniques. Formally, the trial-level model of Burzykowski et al. (2005, Section 11.3) can be formally expressed as

$$\begin{pmatrix} \hat{\alpha}_j \\ \hat{\beta}_j \end{pmatrix} \sim N(\mu, \Sigma + \mathbf{M}\mathbf{M}'), \quad (5)$$

for $j = 1, \dots, J$, where Σ is an unknown 2×2 variance matrix, and $\mathbf{M}$ is an unspecified 2×1 vector. This model can be interpreted in terms of latent factor models for the covariance matrix of $(\hat{\alpha}_j, \hat{\beta}_j)$, $j = 1, \dots, J$. We make the simplifying assumption that $\Sigma = \sigma^2 \mathbf{I}$. However, this model is nonidentifiable from the observed data. By arguing as in Tipping and Bishop (1999), the principal components of the covariance matrix of $(\hat{\alpha}, \hat{\beta})$ is an approximate maximum likelihood estimator in this model. It is approximate in the sense that we must let σ^2 approach zero. Based on Tipping and Bishop's result, our proposal is to calculate the principal components of the estimated variance–covariance matrix of $(\hat{\alpha}_j, \hat{\beta}_j)$, $j = 1, \dots, J$ and to use the proportion of variance explained by the first eigenvector as our estimate of R_{trail}^2 .

We then assess variability based on either a nonparametric bootstrap, which resamples subjects conditional on treatment group, or a model-based bootstrap using the estimated coefficients in (5). Further details are given in Appendix B. In the data example presented here, both bootstrap approaches used gave similar values for the 95% CI (data not shown).

3.4 Assessing Individual Level Surrogacy: General Algorithm

Let us first consider the simpler case of two survival endpoints S and T , where T does not censor S . Letting $\bar{H}(s, t) \equiv \Pr(S > s, T > t)$ be the joint survival distribution of (S, T) , a copula model (Nelsen, 1999) decomposes the joint distribution into the marginal components:

$$\bar{H}(s, t) = C_{\varepsilon}\{\bar{F}_S(s), \bar{F}_T(t)\}, \quad (6)$$

where C is the copula function, ε is a dependence parameter, and $\bar{F}_S$ and $\bar{F}_T$ are the marginal survivor functions of S and T . The most commonly chosen copula model for multivariate survival data is called the Clayton–Oakes model (Clayton, 1978; Oakes, 1986) and is given by

$$C_{\varepsilon}(u, v) \equiv \{u^{1-\varepsilon} + v^{1-\varepsilon} - 1\}^{1/(1-\varepsilon)}$$

with $\varepsilon \geq 1$. In our setting, where we have to deal with semi-competing risks (T censors S), we still work with Clayton–Oakes type of model. Fine et al. (2001) formulated the Clayton–Oakes model only for the semicompeting risks setting where the joint distribution of (S, T) is restricted to the upper wedge. They also presented a closed-form estimator of ε using modified weighted version of concordance estimating function method of Oakes (1982, 1986) along with an asymptotic variance estimator.

We calculate dependence parameters on a study-specific basis, so for this purpose we consider data from one study at a time. Define $e^T \equiv e^T(\beta) = \log T - \beta Z$ and $e^S \equiv e^S(\alpha) = \log S - \alpha Z$ to be the population residuals corresponding to (1) and (2). The following estimator is used:

$$\tilde{\epsilon} = \frac{\sum_{i < j} \tilde{D}_{ij} \psi_{ij}}{\sum_{i < j} \tilde{D}_{ij} (1 - \psi_{ij})},$$

where $\psi_{ij} = I\{(e_i^T - e_j^T)(e_i^S - e_j^S) > 0\}$ and $\tilde{D}_{ij} = I(S_i \wedge S_j - T_i \wedge T_j - C_i \wedge C_j)$, $i, j = 1, \dots, n$.

Note that the estimator for the dependence parameter does not require any estimation of the marginal distribution functions. A variance estimate for $\tilde{\epsilon}$ can be obtained using U-statistic theory (van der Vaart, 2000) and is described in Ghosh (2009). For multiple studies, we construct a metaestimator by estimating $\hat{\epsilon}$ or $\tilde{\epsilon}$ assuming a common ϵ across studies using a Mantel–Haenszel type estimator proposed in Ghosh (2008). Note that we can convert results to Kendall's tau using the formula $\hat{\tau} \equiv (\hat{\epsilon} - 1)/(\hat{\epsilon} + 1)$. This conversion presumes a Clayton–Oakes model for the entire region for (S, T) over the entire positive quadrant, with S regarded as being defined in theory for every patient, but subject to dependent censoring by T .

4. Application to Colorectal Cancer Data

We now apply the proposed methodologies to the colorectal cancer data described in Section 2. The data are analyzed using a semicompeting risks framework based on the time to recurrence and death. The results are given in Table 2.

The estimator of ϵ across all studies is given by $\epsilon = 13.14$, corresponding to a Kendall's tau of 0.86. The results for the treatment effects on death are, in general, qualitatively similar to those presented in Table 1. In addition, for many of the studies, the estimates of the two regression coefficients are effectively identical. This is because of the extremely strong correlation between the two event times and because of the fact that we are imposing the constraint $S \geq T$. Table 3 summarizes the within-treatment arm estimates of the dependence between the two endpoints using the Fine et al. (2001) estimation procedure with the semicompeting risks copula dependence parameter.

Converting the results in Table 3 for $\hat{\epsilon}$ to Kendall's tau shows them to all be at least 0.8. If one were to calculate Wald statistics based on the estimates in Table 3 divided by their standard errors, their corresponding p-values (assuming a standard normal distribution under the null hypothesis) would be miniscule. However, we also find that imposing the constraint leads to greater evidence of trial-level surrogacy. Now we have a trial-level R^2 of 0.96, with an associated 95% confidence interval of (0.93, 0.99). Figure 2 of Sargent et al. (2005), based on a composite endpoint analysis with a somewhat different set of data, reports a trial-level R^2 of 0.90. When we apply Sargent et al.'s (2005) composite endpoint approach to the dataset we are using here, we obtain a trial-level R^2 of 0.89, with an associated 95% CI of (0.86, 0.92). We note two issues in comparing the two datasets. First, there is the nonoverlap

in the datasets used. Second, they use a different surrogate endpoint, namely the composite endpoint time to first event.

5. Discussion

In this work, we have extended the work of Ghosh (2008, 2009) to the multiple-study framework for assessing surrogacy. Because the semicompeting risks paradigm explicitly builds in the constraint that $S \leq T$, the analyses from the previous section suggest that colorectal cancer recurrence will be a very strong surrogate endpoint for death. A major advantage of the semicompeting risks paradigm is that it allows one to study the time to recurrence separately from time to death. There is no creation of composite endpoints, as in Sargent et al. (2005). Of course, the semicompeting risks approach also requires assuming a latent time to recurrence for all individuals, including those who die without recurrence.

One of the practical issues alluded to in chapter 11 of Burzykowski et al. (2005) regarding the assessment of trial-level surrogacy is whether to use linear regression models versus latent variable models. Given that there is error in the estimates of the treatment effects on the surrogate and true endpoints, we recommend the use of latent variable models in practice.

An important issue brought up by a referee was the sensitivity of the estimation procedure to model misspecification. We note that we have been dealing with a randomized trial setting. Although the estimated treatment effects in the AFT model are unbiased even when important variables are omitted, this is not the case for the PH model. Understanding the behavior of the regression estimators under misspecification is needed. However, another major issue is specification of the dependence models. We have been using Clayton–Oakes models in both the absence and presence of semicompeting risks due to their popularity, but model assessment for copulas is an area that remains in its infancy (Wang and Wells, 2000).

Although the PH model is the most commonly used regression method in survival analysis, we have adopted the AFT model. This is primarily for two reasons: (a) the availability of methods to account for dependent censoring and (b) the availability of computationally simple procedures for dependence estimation. Although methods exist for (b) with the PH model (e.g., Shih and Louis, 1995), there are few methods for accounting for dependent censoring with the PH model. Further work is needed in this area.

Acknowledgments

The first author thanks Dr Morton Brown for useful discussions, and the National Surgical Adjuvant Bowel Project for allowing the use of the data from trials C01–C07. The authors also like to acknowledge the editor, associate editor, and referees whose comments have substantially improved the article. This research is supported by National Institutes of Health grants CA129102, CA25224, and CA12027.

Appendix A

Validity of Proposed Resampling Scheme

Assuming that the true (α, β) line in an interior point of a compact subspace of R^2 , and that $E\{U_2(\alpha, \beta_0)\} = 0$ has a unique solution at zero, we can extend the approach of Ying (1993) to show that

$$n^{-1/2}U(\eta_0) = n^{-1/2}U(\eta_0) + \mathbf{A}n^{1/2}(\eta - \eta_0) + o_P(1),$$

where $\eta \in N_r(\eta_0) = \{\mathbf{w} : \|\mathbf{w} - \eta_0\| < r\}$ for sufficiently small $r > 0$, and $\mathbf{A}$ is the asymptotic slope matrix of $n^{-1/2}U(\eta_0)$. This asymptotic linear expansion of U is what is needed to apply the method of Zeng and Lin (2008).

Appendix B

Application of Tipping and Bishop Results and Bootstrap Resampling Schemes

Based on the model (5), it can be shown by applying the arguments from Tipping and Bishop (1999) that the maximum likelihood estimator of $\mathbf{M}$ for a fixed value of σ^2 is given by

$$\widehat{\mathbf{M}} = \mathbf{U}(\mathbf{\Gamma} - \sigma^2\mathbf{I})^{1/2}\mathbf{R},$$

where $\mathbf{U}$ is a $d \times 2$ matrix whose columns are the eigenvectors of the empirical variance-covariance matrix of $(\hat{\alpha}_j, \hat{\beta}_j)$, $j = 1, \dots, J$, $\mathbf{\Gamma}$ is a 2×2 diagonal matrix of the corresponding eigenvalues, $\mathbf{I}$ is a 2×2 identity matrix and $\mathbf{R}$ is an arbitrary orthogonal matrix. Letting $\sigma^2 \rightarrow 0$ gives the equivalence with PCA.

In the same vein, Tipping and Bishop (1999) show that the maximum likelihood estimator for σ^2 for a fixed dimension d is given by the following:

$$\hat{\sigma}^2 = \begin{cases} (\lambda_1 + \lambda_2)/2, & d = 0 \\ \lambda_2, & d = 1 \\ 0, & d = 2, \end{cases}$$

where $\lambda_1 \geq \lambda_2 \geq 0$ are the eigenvalues corresponding to the empirical variance-covariance matrix of $(\hat{\alpha}_j, \hat{\beta}_j)$, $j = 1, \dots, J$. Note that the cases $d = 0$ and $d = 2$ correspond to degenerate estimators for $\mathbf{M}$ and $\hat{\sigma}^2$. We thus deal with the case of $d = 1$. Recognizing $\hat{\sigma}^2 = \lambda_2$ as the proportion of unexplained variation, or equivalently, $1 - R_{\text{trail}}^2$, leads to the use of the first eigenvector for estimation of R_{trail}^2 .

There are two bootstrap schemes that were considered. The first was a nonparametric bootstrap, stratified by treatment group:

1. Resample from (α_j, β_j) with replacement; this creates values $(\tilde{\alpha}_j, \tilde{\beta}_j), j = 1, \dots, J$.
2. Perform a PCA of the $J \times 2$ matrix with j th row $(\tilde{\alpha}_j, \tilde{\beta}_j)$, where the columns are standardized, and obtain the proportion of variance explained from the first eigenvector as described in the previous paragraph.
3. Repeat steps 1–2 B times.
4. Use the 2.5th and 97.5th empirical distribution of the B bootstrapped values of the proportion of variance explained to obtain a 95% CI for the trial-level R^2 .

The other scheme we tried was a bootstrap, which proceeded as follows:

1. Perform a PCA of the $J \times 2$ matrix with j th row $(\hat{\alpha}_j, \hat{\beta}_j)(j = 1, \dots, J)$, where the columns are standardized. Estimate μ in (5) as $\sum_{j=1}^J (\hat{\alpha}_j, \hat{\beta}_j)'$, $\mathbf{M}$ by the eigenvector corresponding to the first eigenvalue from the PCA, and σ^2 by the second eigenvalue.
2. Generate new observations

$$\begin{pmatrix} \alpha_j^* \\ \beta_j^* \end{pmatrix} \sim N(\hat{\mu}, \hat{\sigma}^2 \mathbf{I} + \widehat{\mathbf{M}\mathbf{M}'}), \quad (\text{B1})$$

for $j = 1, \dots, J$. In (A.1), $\hat{\mu}$, $\hat{\mathbf{M}}$, and $\hat{\sigma}^2$ are the estimates from the previous step.

3. Calculate the PCA based on the matrix generated in the previous step, and obtain the proportion of variance explained from the first eigenvector.
4. Repeat steps 1–2 B times.
5. Use the 2.5th and 97.5th empirical distribution of the B bootstrapped values of the proportion of variance explained to obtain a 95% CI for the trial-level R^2 .

References

- Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics*. 2001; 50:405–422.
- Burzykowski, T., Molenberghs, G., Buyse, M. *The Evaluation of Surrogate Endpoints*. New York: Springer-Verlag; 2005.
- Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998; 54:1014–1029. [PubMed: 9840970]
- Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics*. 2000; 1:49–67. [PubMed: 12933525]
- Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978; 65:141–151.
- Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*. 1997; 16:1515–1527. [PubMed: 9249922]

- Fine JP, Jiang H, Chappell R. On semi-competing risks data. *Biometrika*. 2001; 88:907–919.
- Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine*. 1992; 11:167–178. [PubMed: 1579756]
- Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Bio-statistics*. 2000; 1:231–246.
- Ghosh D. Semiparametric inferences for the association parameter with semi-competing risks data. *Statistics in Medicine*. 2006; 25:2059–2070. [PubMed: 16196081]
- Ghosh D. Semiparametric inference for surrogate endpoints with bivariate censored data. *Biometrics*. 2008; 64:149–156. [PubMed: 17651457]
- Ghosh D. On assessing surrogacy in a single trial setting using a semi-competing risks paradigm. *Biometrics*. 2009; 65:521–529. [PubMed: 18759839]
- Li Y, Taylor JM. Predicting treatment effects using biomarker data in a meta-analysis of clinical trials. *Statistics in Medicine*. 2010; 29:1875–1889. [PubMed: 20680981]
- Lin DY, Robins JM, Wei LJ. Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*. 1996; 83:381–393.
- Louis TA. Nonparametric analysis of an accelerated failure time model. *Biometrika*. 1981; 68:381–390.
- Nelsen, R. *An Introduction to Copulas*. New York: Springer; 1999.
- Oakes D. A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*. 1982; 44:414–422.
- Oakes D. Semiparametric inference in a model for association in bivariate survival data. *Biometrika*. 1986; 73:353–361.
- Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*. 1989; 84:487–493.
- Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*. 1989; 8:431–440. [PubMed: 2727467]
- Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34:541–554. [PubMed: 373811]
- Sargent DJ, Wieand HS, Haller DG, Gray R, Benedetti JK, Buyse M, Labianca R, Seitz JF, O'Callaghan J, Francini G, Grothey A, O'Connell M, et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: Individual patient data from 20,989 patients on 18 randomized clinical trials. *Journal of Clinical Oncology*. 2005; 23:8664–8670. [PubMed: 16260700]
- Shih J, Louis TA. Inference on the association parameter in copula models for bivariate survival data. *Biometrics*. 1995; 51:1384–1399. [PubMed: 8589230]
- Tipping M, Bishop CM. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*. 1999; 61:611–622.
- van der Vaart, A. *Asymptotic Statistics*. Cambridge: Cambridge University Press; 2000.
- Wang Y, Taylor JM. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*. 2002; 58:803–812. [PubMed: 12495134]
- Wang W, Wells MT. Model selection and semi-parametric inference for bivariate censored data (with discussion). *Journal of American Statistical Association*. 2000; 95:62–76.
- Ying Z. A large-sample study of rank estimation for censored regression data. *The Annals of Statistics*. 1993; 21:76–99.
- Zeng D, Lin DY. Efficient resampling methods for non-smooth estimating functions. *Biostatistics*. 2008; 9:355–363. [PubMed: 17925303]

Table 1
Colorectal cancer trial descriptions

Study	Number of patients			Median followup		
	Tx	Control	Time (years)	Survival	Recurrence	Disease free survival
C01	375	349	8	0.07	0.13	0.23
C02	344	342	9.41	1.68	0.25	1.72
C03	522	518	11.87	8.29	9.86	7.12
C04	691	1386	11.68	2.94	2.99	2.81
C05	1069	1059	9.07	0.17	1.47	0.31
C06	770	779	8	0.04	0.07	0.01
C07	1209	1200	5.39	2.88	4.57	6.00
INT-0035	469	457	8.08	8.54	20.22	14.64
NCCTG 784852	126	121	6.08	1.03	5.24	3.29
NCCTG 874651	153	255	7.97	2.55	0.96	4.18
NCCTG 894651	225	685	7.58	0.38	0.24	0.07
NCCTG 914653	439	434	7.87	0.81	0.08	0.31

Note: The table shows two-sample unweighted log-rank statistics for comparing the two treatment groups for each study. For the sake of reference, under the null hypothesis, the statistic should have an approximate chi-squared distribution with one degree of freedom. For all three endpoints (survival, recurrence, disease-free survival), censoring is defined to be the difference between the date of the end of the study and date of patient randomization. In the recurrence endpoint analyses, it is possible for subjects to die but to be censored at a later time.

Table 2
Semicompeting risks study-specific regression results for colorectal cancer meta-analysis

Study	Surrogate				True			
	Est	SE	Wald		Est	SE	Wald	$\tilde{\epsilon}^*$
C01	-0.03	0.30	-0.10		-0.03	0.04	-0.75	12.04
C02	0.13	0.15	0.83		0.15	0.04	4.07	10.63
C03	0.41	0.16	2.50		0.41	0.03	11.67	7.87
C04	0.18	0.11	1.58		0.18	0.02	7.27	10.97
C05	0.03	0.11	0.30		0.03	0.02	1.34	13.90
C06	0.02	0.12	0.12		0.02	0.03	0.52	13.93
C07	-0.09	0.03	-3.63		0.13	0.02	5.77	12.01
INT-0035	0.40	0.17	2.33		0.40	0.04	10.94	9.26
NCCTG 784852	0.24	0.33	0.73		0.24	0.07	3.41	5.94
NCCTG 874651	0.19	0.11	1.73		0.30	0.06	5.48	9.38
NCCTG 894651	-0.07	0.21	-0.32		-0.07	0.04	-1.84	11.91
NCCTG 914653	-0.12	0.22	-0.52		-0.12	0.04	-3.04	12.14

Note: These are the results from fitting AFT models to the true and surrogate endpoints. ¹ $\tilde{\epsilon}$ estimated using Clayton–Oakes copula model on the wedge; the resampling approach of Ghosh (2009) based on 1000 perturbations revealed the estimates $\tilde{\epsilon}$ to be highly significant at significance level 0.05 (data not shown). Surrogate section denotes estimates for α and attendant standard error estimators, and True section denotes estimates for β and attendant standard error estimators. Although the estimates and standard errors have been rounded to two significant figures, the Wald statistics were computed based on original values before rounding. The endpoints used here are time to recurrence and time to death. The former endpoint is subject to censoring both by time to death and independent censoring, whereas the latter endpoint is subject to dependent censoring.

Table 3
Study-specific association between S and T in the colorectal cancer meta-analysis

Study	Z = 0		Z = 1	
	Est	SE	Est	SE
C01	15.05	0.11	20.94	0.18
C02	13.81	0.10	19.12	0.16
C03	13.64	0.07	19.25	0.12
C04	16.06	0.07	19.43	0.04
C05	19.92	0.05	19.52	0.04
C06	22.97	0.09	17.47	0.06
C07	15.16	0.02	20.45	0.04
INT-0035	10.68	0.05	21.18	0.14
NCCTG 784852	10.16	0.17	17.13	0.44
NCCTG 874651	12.19	0.22	19.07	0.22
NCCTG 894651	15.16	0.16	16.63	0.05
NCCTG 914653	15.46	0.10	18.73	0.13

Note: Est refers to treatment group-specific estimates of dependence parameter ρ from Clayton–Oakes model for semicompeting risks data, and SE denotes standard error estimate of estimated value for ρ . The endpoints used here are time to recurrence and time to death. The former endpoint is subject to censoring both by time to death and independent censoring, whereas the latter endpoint is subject to dependent censoring.