Biometrics 68, 226–247
March 2012
DOI: 10.1111/j.1541-0420.2011.01633.x

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

Debashis Ghosh,[1,*] Jeremy M. G. Taylor,[2] and Daniel J. Sargent[3]

[1]Departments of Statistics and Public Health Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48103, U.S.A.
[3]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.
[*]*email:* ghoshd@psu.edu

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.

Key words: 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.

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 endpoints 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 endpoint 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 endpoint, 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 = control; 1 = treatment). We make the assumption that (S, T) is independent of C given Z. The data consist of $(X_{ij}, \delta_{ij}^X, Y_{ij}, \delta_{ij}^Y, Z_{ij})$ $(i = 1, \ldots, I;\ j = 1, \ldots, n_i)$. 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 endpoints, 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 semicompeting 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)

Table 1
Colorectal cancer trial descriptions

	Number of patients			Median followup		
Study	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.

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:

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

and

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

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 (ϵ_1, ϵ_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 $\widehat{\beta}_i$ and $\widehat{\alpha}_i$.
(b) Based on a joint modeling approach, described in Section 3.4, regress $\widehat{\beta}_i$ on $\widehat{\alpha}_i$.

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

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

where $\widetilde{Y}_i(\beta) = Y_i \exp(-\beta' Z_i), i = 1, \ldots, n$. The consistency and asymptotic normality for $\widehat{\beta}$, 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:

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

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 $\alpha \leq \beta$ and $\beta - \alpha$ otherwise. Let $\widehat{\alpha}$ be a zero-crossing of α from setting $U_2(\tilde{\eta}) = 0$, where $\tilde{\eta} = (\alpha, \widehat{\beta})$. 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 $\widehat{\eta}$. 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 $\widehat{\mathbf{V}}$, the estimated variance–covariance matrix of $n^{-1/2}\{U_1(\beta), U_2(\eta)\}$:

(R2) Calculate $n^{-1/2}[U_1(\widehat{\beta} + n^{-1/2}G_1), U_2(\widehat{\eta} + n^{-1/2}\mathbf{G})]$. 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 $n^{-1/2}U_1(\widehat{\beta} + n^{-1/2}G_1)$ on G_1 across the B datasets. Similarly, regress $n^{-1/2}U_2(\widehat{\eta} + n^{-1/2}\mathbf{G})$ on G_1 and G_2 across the B datasets.

(R5) Estimate the variance–covariance matrix of $n^{1/2}(\widehat{\eta} - \eta_0)$ as $\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{V}}\widehat{\mathbf{A}}^{-1}$, where the first row of $\widehat{\mathbf{A}}$ is the slope estimate from the first regression in the previous step, whereas the second row of $\widehat{\mathbf{A}}$ 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 $n^{1/2}(\widehat{\eta} - \eta_0)$. 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 $(\widehat{\alpha}_j, \widehat{\beta}_j), j = 1, \ldots, J$, 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

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

for $j = 1, \ldots, J$. 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 $(\widehat{\alpha}_j, \widehat{\beta}_j), j = 1, \ldots, J$. 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 $(\widehat{\alpha}, \widehat{\beta})$ 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 $(\widehat{\alpha}_j, \widehat{\beta}_j), j = 1, \ldots, J$ and to use the proportion of variance explained by the first eigenvector as our estimate of R^2_{trial}. 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 $\overline{H}(s,t) \equiv \Pr(S > s, T > t)$ be the joint survival distribution of (S, T), a copula model (Nelsen, 1999) decomposes the joint distribution into the marginal components:

$$\overline{H}(s,t) = C_\theta\{\overline{F}_S(s), \overline{F}_T(t)\}, \tag{6}$$

where C is the copula function, θ is a dependence parameter, and $\overline{F}_S$ and $\overline{F}_T$ 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

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

with $\theta \geq 1$. In our setting, where we have to deal with semicompeting 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 θ 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:

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

where $\psi_{ij} = I\{(e_i^T - e_j^T)(e_i^S - e_j^S) > 0\}$ and $\tilde{D}_{ij} = I(S_i \wedge S_j \leq T_i \wedge T_j \leq C_i \wedge C_j), i, j = 1, \ldots, n$. Note that the estimator for the dependence parameter does not require any estimation of the marginal distribution functions. A variance estimate for $\tilde{\theta}$ 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 $\widehat{\theta}$ or $\tilde{\theta}$ assuming a common θ 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 $\hat{\tau} \equiv (\hat{\theta} - 1)/(\hat{\theta} + 1)$. 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 θ across all studies is given by $\tilde{\theta} = 13.14$, 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 $S \leq T$. 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 $\widehat{\theta}$ 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.

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

	Surrogate			True			
Study	Est	SE	Wald	Est	SE	Wald	$\tilde{\theta}^*$
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]$\tilde{\theta}$ estimated using Clayton–Oakes copula model on the wedge; the resampling approach of Ghosh (2009) based on 1000 perturbations revealed the estimates $\tilde{\theta}$ 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

	$Z = 0$		$Z = 1$	
Study	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 θ from Clayton–Oakes model for semicompeting risks data, and SE denotes standard error estimate of estimated value for θ. 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.

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.

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`.

Acknowledgements

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.

References

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.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics* **1,** 49–67.

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.

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16,** 1515–1527.

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.

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1,** 231–246.

Ghosh, D. (2006). Semiparametric inferences for the association parameter with semi-competing risks data. *Statistics in Medicine* **25,** 2059–2070.

Ghosh, D. (2008). Semiparametric inference for surrogate endpoints with bivariate censored data. *Biometrics* **64,** 149–156.

Ghosh, D. (2009). On assessing surrogacy in a single trial setting using a semi-competing risks paradigm. *Biometrics* **65,** 521–529.

Li, Y. and Taylor, J. M. (2010). Predicting treatment effects using biomarker data in a meta-analysis of clinical trials. *Statistics in Medicine* **29,** 1875–1889.PubMed PMID: 20680981.

Lin, D. Y., Robins, J. M., and Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83,** 381–393.

Louis, T. A. (1981). Nonparametric analysis of an accelerated failure time model. *Biometrika* **68,** 381–390.

Nelsen, R. (1999). *An Introduction to Copulas*. New York: Springer.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B* **44,** 414–422.

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

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84,** 487–493.

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

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34,** 541–554.

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J. K., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, J., Francini, G., Grothey, A., O'Connell, M., et al. (2005). 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* **23,** 8664–8670.

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

Tipping, M. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* **61,** 611–622.

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.

Wang, W. and Wells, M. T. (2000). Model selection and semiparametric inference for bivariate censored data (with discussion). *Journal of American Statistical Association* **95,** 62–76.

Ying, Z. (1993). A large-sample study of rank estimation for censored regression data. *The Annals of Statistics* **21,** 76–99.

Zeng, D. and Lin, D. Y. (2008). Efficient resampling methods for nonsmooth estimating functions. *Biostatistics* **9,** 355–363.

Received October 2009. *Revised November* 2010. *Accepted November* 2010.

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

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

where $\eta \in N_r(\eta_0) = \|\mathbf{w} : |\mathbf{w} - \eta_0| < r \parallel$ for sufficiently small $r > 0$, and $\mathbf{A}$ is the asymptotic slope matrix of n^{-1} $\mathbf{U}$ (η_0). This asymptotic linear expansion of $\mathbf{U}$ 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 $(\widehat{\alpha}_j, \widehat{\beta}_j), j = 1, \ldots, J, \mathbf{\Gamma}$ 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 $\sigma^2 \to 0$ 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:

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

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

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

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

for $j = 1, \ldots, J$. In (A.1), $\widehat{\mu}, \widehat{\mathbf{M}}$, and $\widehat{\sigma}^2$ 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.

Discussions

Geert Molenberghs[1,2,*]

[1]I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium
[2]I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium
[*]*email:* geert.molenberghs@uhasselt.be

1. Surrogate Marker Evaluation

Over the most recent couple of decades, the evaluation of surrogate markers and endpoints has received considerable attention. An account is given by Burzykowski, Molenberghs, and Buyse (2005). The original impetus came from Prentice (1989) and Freedman, Graubard, and Schatzkin (1992). Prentice (1989) formally defined surrogacy and offered validation criteria, which we now rather term *evaluation criteria*. His main criterion can informally be described as the requirement that all effect from the treatment, Z say, on the true endpoint, T say, is mediated through a surrogate, S say. The criterion is hard to verify through mere hypothesis testing, which is why Freedman et al. (1992) proposed an estimation-based approach instead, by way of the so-called *proportion of treatment effect explained*, often referred to as PTE or PE. Molenberghs et al. (2002), among others, pointed out fundamental problems with the definition of PE, even though the concept was deemed attractive. A fundamental issue with the original proposals was that surrogate endpoint evaluation is conducted within a single trial. This is why various authors (Daniels and Hughes, 1997; Buyse et al., 2000; Gail et al., 2000) have suggested switching to the so-called metaanalytic framework. The contributions reviewed in Burzykowski et al. (2005) fall predominantly within this framework; the same is true of current work by Ghosh, Taylor, and Sargent. Two types of surrogacy can be considered within a single evaluation effort. First, trial-level surrogacy gauges how well the treatment effect on T is predictable from the treatment effect on S. Because prediction cannot genuinely be assessed without replication in a *learning set* of trials, proceeding in the absence of replication is mission impossible. The second type of surrogacy pertains to the patient level, considering how well a patient's true outcome can be predicted from the surrogate outcome.

The single-trial and metaanalytic frameworks just described are not the only ones. Joffe and Green (2009) identify four frameworks. The first one is based on conditional independence of observable variables. The second one is rooted in so-called direct and indirect effects. Their third one is our metaanalytic framework. The fourth and final one rests upon principal stratification. They classify the first and second as belonging to the causal-effects paradigm. This means that, for a surrogate to be good, the effect of Z on S, combined with the effect of S on T, allows for prediction of the effect of Z on T. In contrast, Joffe and Green (2009) term the latter two approaches as belonging to a so-called causal-association paradigm, where the effect of Z on S is associated with the effect of Z on T. That said, the metaanalytic framework actually combines both; it provides machinery both for quantifying associations and for making predictions.

The latter property is one of the attractive features of the metaanalytic framework. The price to pay is that, in its basic form, a joint hierarchical model is necessary. Indeed, the

statistical model is for a bivariate outcome, given by S and T, nested within a hierarchy generated by the trial. This setting becomes more elaborate in a number of situations: (1) when at least one of the outcomes is longitudinal, and/or (2) when more than one surrogate endpoint is recorded, and/or (3) when the hierarchy consists of more levels. This, of course, implies complexities at the modeling and computational levels.

In their original papers on the metaanalytic framework, Buyse et al. (2000) and Gail et al. (2000) considered normally distributed outcomes. For this, a linear model can be considered. In subsequent work, various homogeneous and heterogeneous cases have been studied. Homogeneity here refers to settings where the surrogate and true endpoints are of the same type. Attention was given to binary and time-to-event outcomes, in particular. In heterogeneous settings, the surrogate and true endpoint are of a different nature, such as binary/continuous, ordinal/time to event, etc. For example, when tumor response is considered as a potential surrogate for survival in cancer, the latter heterogeneous case is on the table.

A drawback arising from the metaanalytic framework when the outcomes are non-Gaussian, is that surrogacy at the individual level is not necessarily captured by a correlation coefficient.

Further challenges, computational and in terms of the validation measures, arise when the outcomes are longitudinal. Not only do the hierarchies involved become deeper, but also the definition of evaluation measures is not without ambiguity. By construction, a longitudinal surrogate provides a vector of surrogates. If also the true endpoint is longitudinal, then the surrogacy problem becomes congruent to the canonical correlation problem (Alonso et al., 2004; Burzykowski et al., 2005). Various measures have been proposed, several bearing resemblance with the root statistics used in multivariate analysis. Incidentally, this machinery is also present in the authors' approach, explaining the presence of principal components and, more broadly, factor analysis ideas.

To address at the same time the increasing complexity of the modeling involved and the disparity of the evaluation measures, Alonso and Molenberghs (2007) proposed an information-theoretic approach, applicable to a wide variety of settings (normal, binary, categorical, and longitudinal outcomes) and reduces, in the various particular settings, to the quantities previously introduced in the literature. In this way, the set of scattered proposals made earlier are placed within a uniform framework.

2. Proposal of Ghosh, Taylor, and Sargent

The method proposed in this article by a top team of fine researchers is appealing for a variety of reasons, in particular because it brings out a number of novel and important ideas.

First, it is placed within the tradition of metaanalytic evaluation. This is important, because the data hierarchy offers replication both at the patient as well as at the trial level. Arguably, approaches based on a single trial can assess surrogacy only through making strong but untestable assumptions, even though they may be sensible from a substantive point of view or may be driven by the design. The method by Ghosh et al. does not suffer from such drawbacks.

Second, the use of the accelerated failure time model allows for a fully parametric approach to surrogate endpoint evaluation, leading to feasible and straightforward algebraic manipulation and interpretable and intuitive validation measures.

The next three features that will be discussed are related to each other.

Third, the use of semicompeting risks data makes is very natural for the type of data considered and is more in line with the genesis of the data than is otherwise possible in a purely descriptive, pragmatic approach based on bivariate, or multivariate, survival data, for example. Indeed, the fact that S is censored by the minimum of T and C is otherwise not or, at best, implicitly taken into account.

Fourth, the composite nature of the endpoint is taken into account by decoupling actual surrogate from its possibly censored observation. Precisely, S is taken to be time to progression, possibly censored by T. From this, a natural composite is derived: $S \wedge T$. Starting from the couple (S, T) defined in this way ensures that $S \equiv T$ only with probability zero, unlike $S \wedge T \equiv T$, which is true whenever death occurs prior to progression.

Fifth, as a consequence of the above, the wedge region $S \leqslant T$ can be considered without any problem, which is very elegant. The methods for time-to-event data proposed in Burzykowski et al. (2005) neglect this aspect and therefore attribute nonzero probability mass to the region $S > T$. For most realistic settings (e.g., with S progression-free survival and T overall survival), this is not in agreement with what is substantively possible. Although the classical methods could be forced to conform to the wedge, for example, by modeling the pair $(S, T - S)$ instead, such action would render difficult the derivation of surrogacy measures. In contrast, the current approach reconciles natural modeling with elegant derivation of evaluation measures.

It would be of interest to explore whether these modeling ideas can be phrased within the other frameworks, such as the direct/indirect effects and principal stratification paradigms.

Evidently, a number of important issues remain. This is not to be viewed as criticism toward the current approach. Not only are they shared with other proposals, but they also should predominantly be viewed as suggestions for the follow-up research agenda. Models may have issues of identifiability and sensitivity, because of their general complexity and, in particular, the fact that unobservables are included. In the current framework, unobservables take two forms. First, there is the censored nature of the outcomes. Given that S can be censored by both C and T deepens the issues. Second, the concept of latent time to recurrence fits in with this issue. Such unobservables are identifiable only through strong, unverifiable model assumptions and may impact, therefore, on the measures derived and the predictions carried out. This phenomenon has been reported in the context of missing data (Molenberghs et al., 2008) and random-effects models (Verbeke and Molenberghs, 2010).

In this regard, it is entirely reasonable, as the authors point out, to evaluate a potential surrogate with a variety of methods, by way of sensitivity analysis, as was done by Sargent

et al. (2005). In the same vein, the Clayton–Oakes model could be supplemented with alternative copula models, such as Hougaard's copula (Hougaard, 1986).

It remains true that, in spite of the most sophisticated statistical technology available, judgment by a multidisciplinary team cannot be replaced by an automated "decision-theoretic" rule. The quality of a surrogate is a function of statistical evaluation, life years gained, risk analysis, and cost analysis. It is possible, to some extent, to incorporate some of these aspects into the evaluation, formally or informally. The authors discuss the issue of life years gained; Assam et al. (2010) considered a broadly defined cost function.

Sensitivity analysis and exploring the impact of the assumptions and other choices made in the analysis should be a key part of a standard evaluation exercise. For example, the authors' decision to combine various arms into a single one might be subjected to sensitivity assessment. Of course, the decision to combine the treatment arms was taken for illustrative purposes only.

It seems relevant to explore whether the model developed by the authors could be cast into the information theory paradigm, because of the elegance of this paradigm, its broad basis, and the possibility that it might also lead to simplified computations.

It would be of interest to examine further the homogeneity, or lack thereof, of the study-specific measures, as laid out in Tables 2 and 3. Overall though, it emanates from the analysis that the qualitative evidence is strong and that the results can be viewed with a certain amount of comfort.

3. Concluding Reflections

One could organize the surrogate-marker-evaluation endeavor into four major steps: (1) conceptualization; (2) modeling; (3) software; and (4) data. Conceptualization pertains to defining the concepts, selecting one of the frameworks, etc. For specific situations, specific models are chosen, such as the one proposed by the authors—the second step. The third step has been neglected for a long time, at least to some extent. It is therefore great to see that these authors, among a number of others, are making `R` functions available. The availability of broadly useful software is, arguably, the only way to ensure that new methodology is actually used. Finally, data need to be available and this has for a long time been the Achilles heel of surrogate marker evaluation. Surrogate marker evaluation requires large amounts of data, like most other metaanalytic efforts. Oftentimes, data are proprietary to the biopharmaceutical industry and other sponsors. It is in their own interest, though, as well as that of public health, to ensure broad access to data, evidently with due caution regarding privacy.

Acknowledgements

The author gratefully acknowledges financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics* **63,** 180–186.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M. G. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach. *Biometrics* **60,** 845–853.

Assam, P., Tilahun, A., Alonso, A., and Molenberghs, G. (2010). Using earlier measures in a longitudinal sequence as potential surrogate for a later one. *Computational Statistics and Data Analysis* **54,** 1342–1354.

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

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1,** 49–67.

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16,** 1515–1527.

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

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1,** 231–246.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73,** 387–396.

Joffe, M. M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65,** 530–538.

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.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B* **70,** 371–388.

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

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J. K., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., Grothey, A., O'Connell, M., Catalano, P. J., Blanke, C. D., Kerr, D., Green, E., Wolmark, N., Andre, T., Goldberg, R. M., and De Gramont, A. (2005). Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology* **23,** 8664–8670.

Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling* **10,** 391–419.

Edward L. Korn

Biometric Research Branch, National Cancer Institute
Bethesda, Maryland 20892, U.S.A.
email: korne@ctep.nci.nih.gov

1. Introduction

Ghosh, Taylor, and Sargent (2011) (GTS) are to be congratulated for exploring new approaches for assessing trial- and individual-level surrogacy using accelerated failure time (AFT) models and a deconstruction of time-to-event composite endpoints. It has been recognized that an endpoint may be a good individual-level surrogate and a poor trial-level surrogate, and vice versa; for example, see Figure 1 of Korn, Albert, and McShane (2005a). I will focus here on some issues raised by the GTS methodology for trial-level surrogacy. In addition, because the use of composite endpoints is so subject-matter dependent, I will restrict attention to oncology trials as their illustrative application is in oncology.

2. Choice of the Time-to-Event Outcome

GTS recommend not using composite time-to-event endpoints. However, in addition to overall survival (time to death from any cause), the use of composite outcomes is common in oncology trials of experimental drugs and typically preferred over an endpoint that censors a patient's data at death. For trials involving metastatic disease, progression-free survival (time to the tumor progression or death) is preferred to time to progression (time to tumor progression with deaths without progression censored). One of the reasons for this is that tumor progression does not account for all the effects of the experimental treatment on patient survival, including early and late deaths due to toxicity of the treatment (Green, Benedetti, and Crowley, 2003, pp 44–45). In addition, as progression is defined as a certain amount of tumor growth on an imaging scan, it is possible to have a patient with progressive disease die before their progression is documented or to have a patient die from their disease but not have a progression. Finally, in terms of measuring patient benefit, an estimated progression-free survival curve tells one what the probability is that patients will be alive and not progressed at time points after randomization. What does an estimated time-to-progression curve tell one—what is the probability that a patient will not have progressed at time points if they have not died before then? In the adjuvant treatment setting where patients have had their tumors eliminated through surgery or radiation, disease-free survival (time to disease recurrence or death) is generally preferred to time to disease recurrence (deaths censored) for similar reasons.

In some settings where the proportion of deaths unrelated to the cancer or treatment is expected to be large, it may make sense to censor unrelated deaths. However, in these situations, it is important to acquire any additional information to help decide if the death is actually unrelated. For example, in the National Lung Screening Trial, a randomized trial of lung cancer screening modalities for older current and former heavy smokers, a panel of independent experts (blinded to the randomization arm) followed a detailed algorithm involving additional records to ascertain whether a death was due to lung cancer (or indirectly from the screening) (National Lung Screening Trial Research Team, 2011).

All of the above is irrelevant if one is using the surrogate endpoint only as surrogate for the definitive endpoint and not to directly measure the treatment benefit. This presumes that the proposed surrogate endpoint is indeed a better surrogate than the usual composite endpoint. The argument by GTS that disease-free survival is not useful as a surrogate for overall survival because it includes deaths needs further explanation; one would think that having something in common with the definitive endpoint would make a better surrogate than not having something in common. Regardless of this argument, metaanalyses involving real applications should settle the issue. However, when used as a definitive primary endpoint assessing clinical benefit (rather than the surrogate endpoint), the usual composite endpoint is more appropriate than censoring deaths. For example, part of the National Surgical Breast and Bowel Project B-27 Trial evaluated preoperative chemotherapy for operable breast cancer (Bear et al., 2006). In this trial, disease-free survival was one of the primary endpoints (and is thought to represent direct clinical benefit to the patient in this setting) and pathological complete response (at the time of surgery) was considered as a possible surrogate. I assume that GTS would agree with the use of the composite endpoint in this application.

Part of a time-to-event outcome is the censoring indicator. With two endpoints there would typically be two censoring indicators. For example, a patient's time-to-progression value would be censored at their last clinic visit but their survival time may be censored at a later time if they continue to be followed for survival. It is not clear how the GTS methodology, which involves artificial censoring of the observations, accommodates this practical consideration.

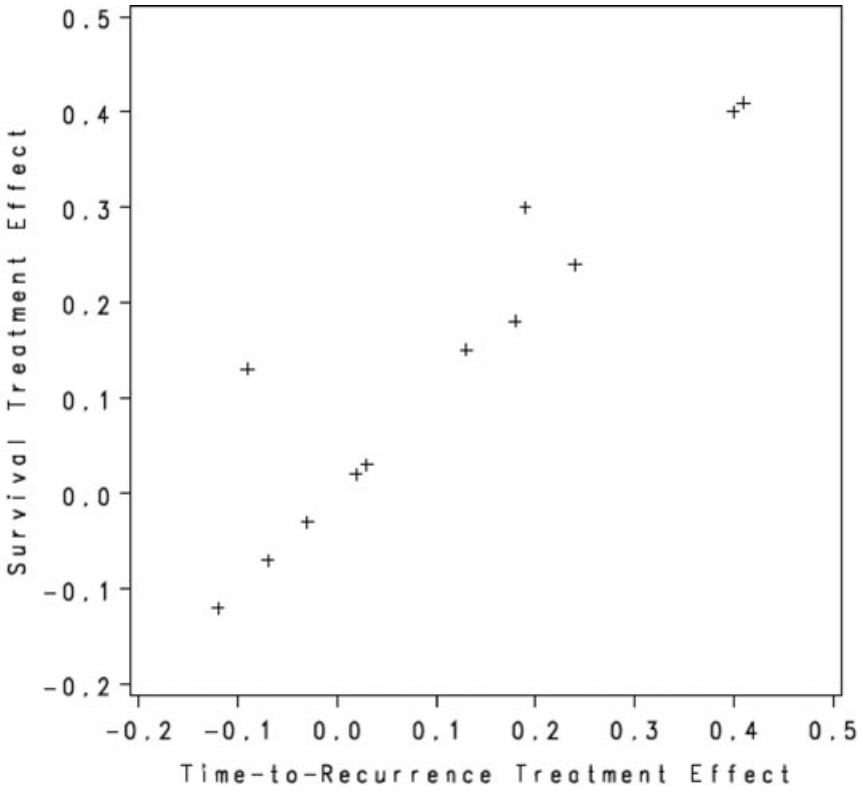

Figure 1. OS trial-level treatment effect versus TTR trial-level treatment effect (coordinates for plotted points are from Table 2 of GTS).

3. Choice of Treatment-Effect Parameter and Across-Trial Modeling of These Parameters

Use of the proportional hazards (PH) model and its associated log-rank statistic is the most popular analysis strategy for analyzing oncology clinical trials. The AFT model has not been used much because historically only a parametric version of the model was readily available. With software available for the semi-parametric version of the model ((1) and (2) of GTS), this constraint is gone. (With intercepts left off the models, the statement that the error terms have mean zero appears incorrect.) One would think that the choice between the models would be based on whichever model fit the data better. However, GTS state their reason for using AFT models rather than PH models is the availability of methods with AFT modeling to account for dependent censoring. This rationale does not apply to the treatment-arm comparison for a definitive endpoint like overall survival. For the choice AFT versus PH models for defining the surrogate treatment-effect parameter, one should choose the model that empirically leads to the best surrogate treatment-effect parameter; there is no a priori reason the same modeling approach need be used for the surrogate and definitive-variable treatment-effect parameters.

GTS's modeling across trials of the treatment-effect parameters could use further explanation. First, randomized trials frequently have more than one experimental arm along with the control arm and pooling all the experimental arms together in a trial as GTS have done can result in a large loss of information. In a standard metaanalysis, the induced correlation between experimental-versus-control treatment comparisons that share a common control arm is easily accommodated. Is there some difficulty with multiarm trials using the proposed methodology? Furthermore, restricting consideration to two-armed trials, I would have expected the covariance matrix of $(\hat{\alpha}_j, \hat{\beta}_j)'$ (equation (5) of GTS) to be $D + \Sigma_j$, where Σ_j is the standard estimated covariance matrix of $(\hat{\alpha}_j, \hat{\beta}_j)'$ conditional on $(\alpha_j, \beta_j)'$ estimated using individual-level data from each trial at a time, and D is the unknown covariance matrix of the $(\alpha_j, \beta_j)'$. Typically, the estimated Σ_j are assumed to be known parameters for the purposes of estimating D; we discuss elsewhere setting the off-diagonal terms of Σ_j to zero to improve the mean square error of the estimated D (Korn et al., 2005a). When some or all of the trials involve two-sided questions (A versus B) rather than one-sided questions (experimental versus standard), then some additional care is required in the modeling (Freedman, 2005; Korn, Albert, and McShane, 2005b).

GTS describe an innovative resampling scheme, but for estimating standard errors of trial-level surrogacy parameters, why not just bootstrap the trials (not the individuals and not the trial arms)? If the concern is that there are too few trials for a bootstrap to be reliable, then any more complex method that apparently gives more reliable estimators must be relying heavily on parametric assumptions for the distribution of trial-level effects.

4. The Purpose of a Trial-Level Surrogacy Metaanalysis

One wants to know whether one can use the surrogate variable in place of the definitive outcome in evaluating the results of a randomized trial. In particular, for a new trial where the surrogate results are available, how well can we estimate the treatment effect for the definitive variable? Note that this is not a forecast of the definitive endpoint results for the trial at hand, but is instead an estimate of the true definitive treatment effect with a confidence interval. (Forecasts of trial results with their prediction intervals can be useful in model checking, e.g., leave-one-out crossvalidation prediction of results for trials included in the metaanalysis; see Sargent et al., 2005, for an example.)

Even in this framework, there are two general issues in performing a metaanalysis with this goal. First, the follow-up in the trials in the metaanalysis will typically be much longer than follow-up in the new trial. In fact, if one had the length of follow-up in the new trial as one had in the trials used in the metaanalysis, one would probably not be interested in the surrogate but would just use the definitive endpoint for the new trial. If you really believe your modeling, this is not a problem. A more cautious approach, and the one used by Sargent et al. (2005), is to additionally censor the surrogate variable data used in the metaanalysis so that they have follow-up similar to what one would see in a new trial. I would not censor the definitive endpoint data, unless it goes beyond where the treatment effect is expected to be seen.

The second issue concerns therapies that are given to the patient after the surrogate endpoint is observed. If the same effective second-line treatments are used in both treatment arms, then the treatment effect as measured by the definitive outcome will tend to be attenuated as compared to if ineffective or no second-line treatments were used (e.g., the hazard ratio for survival will be closer to one). However, this attenuated effect is the correct one to consider for measuring benefit of the experimental treatment to the patients (Korn and Freidlin, 2010; Korn, Freidlin and Abrams, 2011). On the other hand, if patients crossover from the control treatment arm to the experimental treatment arm at (so that the treatment arms are receiving different second-line treatments), then what the definitive-outcome treatment effect would have been without the crossovers is the one of interest. This suggests that the trials included in the metaanalysis should not have allowed crossovers (although the new trial may have them), and any effective second-line treatments that are being used in the new trial were also used in the trials in the metaanalysis.

5. Application Revisited

Sargent et al. (2005) performed a comprehensive metaanalysis of 18 trials. GTS use a subset of 10 of these trials, added 2 more (including C06 that apparently had a noninferiority design), and combined experimental trial arms for trials that had more than one experimental arm. In addition, GTS excluded patients with zero event times, suggesting a more general problem with AFT modeling—should a prolongation of an event from 1 to 3 days be considered as effective as a prolongation of an event from 1 to 3 years? Because of these considerations, I consider the GTS analyses as a demonstration of statistical methodology and do not make any comments about the clinical relevance of their results to cancer trials.

Table 1
Study-specific log hazard ratios with their standard errors

Study	TTR	OS	Disease-free survival
C01	0.04 ± 0.12	-0.02 ± 0.09	-0.05 ± 0.10
C02	0.07 ± 0.14	0.14 ± 0.10	0.14 ± 0.11
C03	0.33 ± 0.11	0.26 ± 0.09	0.25 ± 0.09
C04	0.15 ± 0.08	0.12 ± 0.07	0.09 ± 0.08
C05	0.10 ± 0.08	0.03 ± 0.07	0.06 ± 0.07
C06	-0.03 ± 0.10	0.02 ± 0.10	0.01 ± 0.09
C07	0.17 ± 0.08	0.15 ± 0.09	0.19 ± 0.08
INT-0035	0.47 ± 0.10	0.28 ± 0.10	0.36 ± 0.09
NCCTG 784852	0.42 ± 0.18	0.17 ± 0.17	0.30 ± 0.17
NCCTG 874651	0.17 ± 0.17	0.24 ± 0.15	0.30 ± 0.15
NCCTG 894651	-0.06 ± 0.12	-0.07 ± 0.11	-0.03 ± 0.11
NCCTG 914653	-0.03 ± 0.12	-0.10 ± 0.11	-0.06 ± 0.10

Note: Hazard ratios are for control over experimental treatment so that log hazard ratios >1 represent the experimental treatment doing better.

The GTS analyses have some strange properties that could use some clarification. First, the standard errors for the overall survival (OS) treatment effects ("True" in Table 2 of GTS) are remarkably small. In particular, they are very much smaller than the standard errors for the time-to-recurrence (TTR) treatment effects ("Surrogate" in Table 2 of GTS). This is surprising given that 80% of the deaths were preceded by recurrence. In addition, what can explain a 10-fold difference in the TTR standard errors for C07 and C01 when there is only a 2-fold difference in the OS standard errors? Secondly, as noted by GTS, many of the OS and TTR estimates are essentially identical. Figure 1 is a plot of the estimates. It is hard to understand how the stated reasons (correlation of event times, constraint on the event times) can explain how this plot can be consistent with the given standard errors.

To better understand these data, we consider standard PH modeling as was used by Sargent et al. (2005) for OS and disease-free survival (DFS) and TTR composite endpoints. GTS have kindly supplied me with the estimated log hazard ratios and their standard errors (Table 1). The estimates and their standard errors appear nonanomalous. The correlation across the 12 trials between the OS log hazard ratio and the OS AFT treatment effect is 0.981, again suggesting a problem with the standard errors in Table 2 of GTS.

Unlike Sargent et al. (2005), GTS provide no indication of how a TTR treatment effect for a new trial would translate into an estimated OS effect.

6. Discussion

The biggest issue concerning trial-level surrogacy has nothing to do with statistical methodology, but is whether it is reasonable to extrapolate from the trials in the metaanalysis to the new trial at hand. If the new trial has treatments with a mechanism of action different than the trials in the metaanalysis, or if new second-line treatments have become available, the results of the metaanalysis may not apply. When possible, it would seem advisable for definitive randomized

clinical trials to follow patients for survival even if a surrogate endpoint is going to be used to report preliminarily the trial results.

Additional References (not in GTS)

Bear, H. D., Anderson, S., Smith, R. E., Geyer Jr., C. E., Mamounas, E. P., Fisher, B., Brown, A. M., Robidoux, A., Margolese, R., Kahlenberg, M. S., Paik, S., Soran, A., Wickerham, D. L., and Wolmark, N. (2006). Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *Journal of Clinical Oncology* **13,** 2019–2027.

Freedman, L. (2005). Commentary on assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine* **24,** 183–185.

Ghosh, D., Taylor, J. M. G., and Sargent, D. J. (2011). Meta-analysis for surrogacy: Accelerated failure time models and semi-competing risks modeling. *Biometrics*, doi: 10.1111/j.1541-0420.2011.01633.x.

Green, S., Benedetti, J., and Crowley, J. (2003). *Clinical Trials in Oncology*, 2nd edition. Boca Raton, Florida: Chapman & Hall.

Korn, E. L. and Freidlin, B. (2010). Causal inference for definitive clinical end points in a randomized clinical trial with intervening nonrandomized treatments. *Journal of Clinical Oncology* **28,** 3800–3802.

Korn, E. L., Albert, P. S., and McShane, L. M. (2005a). Assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine* **24,** 163–182.

Korn, E. L., Albert, P. S., and McShane, L. M. (2005b). Rejoinder to commentary by Dr Freedman of 'Assessing surrogates as trial endpoints using mixed models'. *Statistics in Medicine* **24,** 187–190.

Korn, E. L., Freidlin, B., and Abrams, J. S. (2011). Overall survival as the outcome of randomized clinical trials with effective subsequent therapies. *Journal of Clinical Oncology* (published online before print May 9, 2011). doi:10.1200/JCO.2011.34.6056.

National Lung Screening Trial Research Team (2011). The National Lung Screening Trial: Overview and study design. *Radiology* **258,** 243–253.

Vance W. Berger, Grant Izmirlian,* and Diana Knoll

National Cancer Institute, Biometry Research Group
Executive Plaza North, Suite 3131, 6130 Executive
Boulevard, MSC 7354, Bethesda
Maryland 20892-7354, U.S.A.
**email:* izmirlig@mail.nih.gov

In an interesting analysis of currently applied methods and new developments in the field of surrogacy, Ghosh, Taylor, and Sargent (2011), emphasize the importance of additional work and better observations in the research field of surrogate marker validation. Considering time-to-event endpoints, they look at common validation methods, reveal drawbacks, and try to give possible solutions. Given the popularity of surrogate endpoints for decreasing trial costs and duration (Molenberghs, Geys, and Buyse 2001; Berger, 2004), it is important to analyze all aspects of these potential replacements of the clinical endpoint. We note that we agree with much of what was said, but in our brief communication, we will limit ourselves to our few points of disagreement with the authors.

For full disclosure, we note our general predisposition toward straightforward analyses that allow the data to stand on their own, without the need for unverifiable assumptions, at least if we confine our attention to the evaluation of treatments, as we intend to do. Matters such as the association between the surrogate and the true endpoint are certainly interesting academically speaking, and, moreover, the authors did note that the insights generated are complementary to those found by existing methods. This is fair, but still, we find that even the simpler question of whether a treatment works or not is sufficiently complicated that valid approaches to address it are quite elusive. At least this is the case if by "valid" we mean truly valid, and not merely accepted as such. So we set our sights lower, and confine ourselves to the issue of the role surrogate endpoints can or should play in the (valid) evaluation of treatments.

Our first point of contention concerns the treatment of death as a censoring time for recurrence, as if there is some true, unobserved, latent recurrence time that would have occurred at some point after death, if only death had not made things so inconvenient for the researchers. However, now, with the miracle of modern statistics, the patient need not worry. Though still just as dead, at least now the patient can find some consolation in the fact that we can detect that this hypothetical recurrence time has been pushed back in to the more distant future. Admittedly, our caricature may not be fair, as this may be a reasonable approach for academic interest. However, as noted, we limit ourselves to the perspective of the patient who wants to know only if the new treatment is better than the old one. And how shall we define "better" for this patient? Do we really want to ask the patient to trust us that there is a better afterlife if we can just push back that posthumous recurrence time? An argument was made (Fisher, 1999) in the context of a carvedilol trial that death trumps all other considerations. Though we did not find this argument convincing in the context in which it was initially offered (it does not justify changing the primary endpoint after the data are in), we do find it convincing in the context we now discuss, namely, that recurrence has no meaning once the patient has died. Hence, we find ourselves more in agreement

with Sargent et al. (2005) that the proper formulation of the problem is in terms of disease-free survival.

What, then, are we to make of the allegation and criticism that the disease-free survival time is actually a composite endpoint? The characterization of the disease-free survival time as a composite endpoint is not dissimilar to the characterization (Berger, 2000, p. 1321) of an approximate p-value, $p(A)$, as the sum of two components, the exact p-value, $p(E)$ and the error (so to speak), $p(A) - p(E)$. One who perversely clings to the indefensible notion that an approximation is to be preferred to the very quantity it is trying to approximate might just as easily call the difference $D = p(E) - p(A)$ the error, and then proceed to express $p(E)$ as the sum $p(A) + D$. Does this characterization of either p-value as a sum of two quantities mean that any p-value we might ever present is a composite endpoint by virtue of being expressible as a sum of two components? This seems to be a highly suspect allegation. And at least here the usual rules of arithmetic apply, as we observe (or can directly compute) both $p(A)$ and $p(E)$.

The same cannot be said for the two component endpoints (survival time and recurrence time). When death comes first, we do not observe the recurrence time, so the minimum of these two certainly is not computed as the minimum of the two; it cannot be. Rather, it is directly observed as the time to the first event, which in some cases (when death comes first) will turn out to be the only event of interest. So it is not clear that the time to first event, $\min(S, T)$, is any more of a composite endpoint than is the survival time (T) itself, which may be expressed as the sum of two components, $T = \min(S, T) + I(S < T)(T - S)$. However, even if we do concede that $\min(S, T)$ is a composite endpoint, how much of an indictment is this of the endpoint? Are composite endpoints really to be avoided? Although it has been shown that relying solely on composite endpoints could lead to flawed conclusions and could tempt researchers to influence their findings in a favorable direction (Montori et al., 2005), it is also clear that composite endpoints are highly desirable in some contexts (Berger, 2002); therefore, if an argument is to be compelling that they are undesirable in this case, then a much stronger argument will be needed.

Beyond our disagreement with (1) the treatment of the recurrence time as censored when death comes first, and (2) criticizing composite endpoints with a broad brush, we pondered over the insinuation that a cross-trial coefficient of correlation in excess of 0.90 should be a high enough aggregate correlation to ensure validity of a surrogate endpoint. We then began to wonder how validity in terms of hypothesis testing holds for a particular coefficient of correlation. Let Δ_S, Δ_T be the normalized test statistics based upon data on the surrogate and upon the clinical endpoints, respectively. Under the global null hypothesis, the vector (Δ_S, Δ_T) is asymptotically distributed as a bivariate normal with components of mean zero, variance one and correlation, ρ_{Δ_S,Δ_C}. Suppose that we are benchmarking a potential surrogate endpoint by conducting a metaanalysis of trials of a given "family" of agents in which both the surrogate and clinical endpoints are measured. Suppose that it is our intention to use the resulting information, namely, our estimate of ρ_{Δ_S,Δ_C}, to ensure that in a future trial of an agent from the same family in which only the surrogate endpoint is measured, that a surrogate endpoint based test of the null hypothesis at level Err_S^I will correspond to a hypothetical clinical endpoint based test of the null hypothesis at level Err_C^I. Certainly if $\rho_{\Delta_S,\Delta_C} = 1$ then the test statistics either reject or accept their corresponding null hypotheses together and the probabilities of type I error, Err_S^I and Err_C^I, are identical. Of course this is never the case. When the correlation coefficient, ρ_{Δ_S,Δ_C}, is less than one, then we have suddenly an additional error probability that must be controlled, but is seldom talked about as far as we are aware. This is the complement of:

$$\psi = \mathbb{P}_{H_0^G}\{\Delta_C > z_C \mid \Delta_S > z_S\}, \tag{1}$$

where H_0^G in the above is the global null hypothesis. This is the conditional probability, under H_0^G, that the clinical endpoint based test of the null hypothesis at level Err_C^I is rejected given that the surrogate endpoint based test of null hypothesis at level Err_S^I is rejected. Having introduced this concept, the first question that arises is: what value of ψ should be considered reasonable? Naturally our first instinct is to consider what is already considered reasonable and attempt to extrapolate upon that. We wish that there were no false positives in all clinical research, but in order not to throw the baby out with the bathwater, we should be content with garbage a proportion Err_C^I of the time. Applying this logic recursively to the setting of tolerable false clinical positives among surrogate positives, the value $\psi = 1 - Err_C^I$ or better should be used. We list in Table 1 below, values of ρ_{Δ_S,Δ_C}, which correspond to stipulated values of ψ (0.95 or 0.99) and Err_S^I (0.0025, 0.001, 0.01, 0.02, 0.03, 0.04, 0.05), when Err_C^I is fixed at 0.05. Note that, for example, if we want a surrogate endpoint based test of the null hypothesis at level $Err_S^I = 0.05$

Table 1

Correlation coefficient between surrogate and clinical endpoint test statistics, ρ_{Δ_S,Δ_C}, that is required given stipulated level of significance in surrogate endpoint test of null hypothesis, Err_S^I, and stipulated conditional probability, under the global null hypothesis, that clinical endpoint test of null hypothesis is rejected given that the surrogate endpoint test of null hypothesis is rejected, ψ, when the level of significance in clinical endpoint test of null hypothesis is fixed at 0.05*. All hypothesis tests are two sided.*

ψ	Err_S^I	$Crit_S$	ρ_{Δ_S,Δ_C}
0.95	0.0025	3.0233	0.8633
0.95	0.0050	2.8070	0.8941
0.95	0.0100	2.5758	0.9268
0.95	0.0200	2.3263	0.9604
0.95	0.0300	2.1701	0.9792
0.95	0.0400	2.0537	0.9912
0.95	0.0500	1.9600	0.9986
0.99	0.0025	3.0233	0.9116
0.99	0.0050	2.8070	0.9350
0.99	0.0100	2.5758	0.9584
0.99	0.0200	2.3263	0.9803
0.99	0.0300	2.1701	0.9912
0.99	0.0400	2.0537	0.9972
0.99	0.0500	1.9600	0.9999

to "guarantee," at least $\psi = 95\%$ of the time, a corresponding level $Err_C^I = 0.05$ clinical endpoint based test of the null hypothesis, then we require the test statistics to have correlation $\rho_{\Delta_S,\Delta_C} = 0.9986$. Of course this is too tall of an order! If, as indicated by the comments on the example in Ghosh et al. (2011), a correlation of $\rho_{\Delta_S,\Delta_C} = 0.90$ is more realistic, then to ensure that the level of the clinical endpoint based test of the null hypothesis is valid at $Err_C^I = 0.05$ then we must set the level of the surrogate endpoint based test of the null hypothesis to $Err_S^I = 0.005$ (that is one tenth of the usual 0.05, not a typo). We can only guess what goes on in practice.

In summary, we have reservations regarding (1) the preference for disease recurrence, censored by death being preferable to disease-free survival and (2) the criticism of composite endpoints in general. These more substantive limitations aside, however, we applaud the author's work introducing a methodology for the survival analytic setting whereby an association parameter between the surrogate and clinical time to event can be tied to the correlation parameter between the associated Wald statistics in a log-rank type test. Beyond that, we are grateful to have this opportunity to comment in a larger context on the validation of surrogate endpoints. Specifically we have highlighted the need to attach the concept of validity to the relationship between surrogate endpoint and clinical endpoint based tests of the null hypothesis. We hope that our discussion of the conditional probability that a clinical endpoint based test rejects the null hypothesis at a determined level given that the surrogate endpoint based test rejects the null hypothesis at a fixed level and its relation to the intertest-statistic correlation coefficient provides some insight. Certainly its implications in terms of meaningful effect sizes for the clinical endpoint and surrogate endpoints are areas that we intend to explore in the future.

Beyond our commentary on what we consider to be quite elegant methodology for validating candidate surrogate endpoints in the setting of a metaanalysis of trials in which both surrogate and clinical endpoints are measured, we take one more opportunity to comment on the use of surrogate endpoints in general. By the principle of minimum energy, which predicts the path connecting mice with their ultimate destination, the "cheese" as it were–and keep in mind we are not insinuating that any entity involved would intentionally "cheat," the existence of such a path of least resistance precludes its omission from a discussion such as this one. Consider the possibility that an agent, ξ, failed in a test of efficacy on clinical endpoint, C, but is known to have strong efficacy for surrogate endpoint, S. Is it not possible to find agents η_1, $\eta_2, \ldots, \eta_k$ in the same chemical family (creative synthetic organic chemistry should allow one to stick a familial ligand on somewhere), which all show strong efficacy for both the clinical and for the surrogate endpoint. For this reason, should not there always be a calibration sample on which both surrogate and clinical endpoints are measured?

Surrogate endpoints serve a useful purpose, in that they can greatly reduce the cost and duration of a trial, thereby allowing these resources to be diverted to other useful activities. However, there is a fundamental problem, or impossibility theorem, that governs their use. If a treatment has already been studied with trials using the clinical endpoint, then it is unclear what added benefit accrues from studying it again with a surrogate endpoint. So let us confine our attention to the framework of validating surrogate endpoints so that they can be used in future studies of future treatments that have not, and will not, be studied with the clinical endpoints. Moreover, let us consider also the usual paradigm of validating (or attempting to validate) surrogate endpoints so that they can be used in future studies, in lieu of clinical endpoints. In this context, we recognize the need to test the new treatment; that is, we do not rely on prior information regarding other treatments, even if in the same class, to state that we do not need to study this treatment at all. And yet we seem to lose sight of this enlightened and cautious realization that each treatment is unique when we assume that structural relations among variables that have been observed in the studies of some treatments will continue to hold true when this new treatment is studied.

As Hume (1896) noted, "probability is founded on the presumption of a resemblance betwixt those objects, of which we have had experience, and those, of which he have had none; and therefore it is impossible that this presumption can arise from probability. The same principle cannot be both the cause and effect of another." What, then, can justify the assumption that structural relations among variables will remain intact? It seems difficult to find a convincing answer to this vexing problem, and burying it under the carpet represents nothing more than the shell game of hiding the uncertainty, or parlaying pseudocertainty in an arena less likely to be scrutinized (the structural relation of the variables) into pseudocertainty (with the appearance of true certainty) in an area that is scrutinized routinely (the effectiveness of treatments). However, the foundation of the argument is no Archimedes fixed point, and conclusions that follow from premises can be no more certain than the premises on which they are based. So if the assumption (of stable structural relations among variables) is needed to establish that a treatment is effective, then one has to wonder just how effective the treatment really is. Hence, our primary contention is that surrogate endpoints should be used in conjunction with clinical endpoints, and not as replacements for them. That said, the models proposed by the authors are a useful tool that may well have a far reaching impact in the future of clinical research.

References

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine* **19,** 1319–1328.

Berger, V. W. (2002). Improving the information content of categorical clinical trial endpoints. *Controlled Clinical Trials* **23,** 502–514.

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

Fisher, L. D. (1999). Carvedilol and the Food and Drug Administration (FDA) approval process: The FDA paradigm and reflections on hypothesis testing. *Controlled Clinical Trials* **20,** 16–39.

Ghosh, D., Taylor, J. M. G., and Sargent, D. J. (2011). Meta-analysis for surrogacy: Accelerated failure time models and semi-competing risks modeling. *Biometrics*, in press.

Hume, D. (1896). *A Treatise of Human Nature*, Reprint. Oxford: Clarendon Press.

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed

discrete and continuous outcomes. *Statistics in Medicine* **20,** 3023–3038.

Montori, V. M., Permanyer-Miralda, G., Ferreira-González, I., et al. (2005). Validity of composite end points in clinical trials. *British Medical Journal* **330,** 594–596.

Sargent, D. J., Wieand, H. S., Haller, D. G., et al. (2005). Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology* **23,** 8664–8670.

John O'Quigley

Laboratoire de Statistique Théorique et Appliquée
Université Paris VI, France
email: john.oquigley@upmc.fr

and

Philippe Flandre

UMR-S943, Institut National de la Santé et de la Recherche Médicale, 75013 Paris, France

1. Introduction

The contribution of Ghosh, Taylor, and Sargent (GTS) adds a new angle to the surrogate endpoint literature by moving slightly away from the established paradigm put in place by Buyse and Molenberghs (1998) and Buyse et al. (2000), and built on substantially since by themselves and coworkers. This established paradigm is based essentially on the use of multivariate linear models and two endpoints, one of which is the main endpoint of interest, usually survival time, and the other an alternative endpoint, which may potentially serve as a surrogate endpoint in place of survival time. GTS appeal to a semicompeting risk framework that allows them to model the region where the (potential) surrogate endpoint occurs before survival time itself. Some of the tools developed within the joint endpoint literature, for example, methods for evaluating trial level surrogacy, both between and within, are given analogues in the semicompeting risks setting.

GTS focus interest on the "wedge region" where the surrogate endpoint is less than or equal to the survival endpoint, something which is not always made explicit in other approaches, but which is natural in that, to play a useful role as a surrogate, we need the event to occur earlier. The classical illness–death model where illness does not censor for death but death censors for illness (Fix and Neyman, 1951) has the same structure as the semicompeting risk model of Fine, Jiang, and Chappell (2001) and it would be helpful to underline any perceived essential differences between these models. Xu, Kalbfleisch, and Tai (2010) note that, apart from references to latent times (an abstract and not an operational consideration), the illness–death model and the semicompeting risk model describe the same probabilistic structure. Just as it seems preferable not to pay attention to the idea of becoming ill once you have died, by focusing on the wedge region, GTS make the same argument, in a less direct way, that, once the subject has died, it may make sense to give no consideration to the latent occurrence of surrogate events.

In the following section, we contrast some recent techniques for assessing surrogate endpoints in the survival setting, as they relate to the composite endpoint approach and the semicompeting risks approach. In Section 3, we make the case that an analysis of surrogacy can be carried out using a proportional hazards model with time-dependent effects. This is not the current view and, in their response, the authors may take the opportunity to spell out more clearly why they do not favor such an approach.

2. Composite Endpoints versus Semicompeting Risks

The joint endpoint approach will typically take the surrogate outcome, S, and the survival outcome, T, and consider some joint model for these. Taking the minimum of S and T has been suggested as a composite endpoint although this reduces to a single endpoint under a different definition, the most obvious example being disease-free survival. The models that have seen most success are the multivariate normal model for which many useful features become available to us, including partial and multiple correlation coefficients. Having fit such a model, then certain aspects of the parameterization are of immediate interest, notably the degree of association between T and S and, in particular, the way in which this association manifests itself in terms of the conditional distribution of T given some realized value of S. The accelerated failure time model proposed by GTS cannot, of itself, be considered essentially a different model from the one of Buyse et al. (2000). We know that S and T are necessarily positive variables, the multinormal model then necessarily an approximation assigning probabilities close to zero for negative values of S and T. By taking the logarithm of T, which is linearly regressed on the treatment indicator as well as other potential covariates, GTS are not doing anything fundamentally

different from Buyse et al. (2000) apart from opening up the possibility of working with different error distributions.

However, the question of different error distributions is not studied by GTS and their equation (3) is a familiar score type estimating equation that is valid when the censoring variable is independent of survival. The investigation of the surrogate endpoint itself differs from that of Buyse and colleagues in that they use a technique developed by Lin, Robins, and Wei (1996). This technique allows for the comparison of two survival curves in the presence of dependent censoring. Obviously, restrictions are needed and, as pointed out by GTS, they will often be satisfied in studies on surrogate endpoints. Nonetheless, the procedure of Lin et al. (1996) is itself onerous and inference requires much more work. In particular, we need some consistent estimate of the variance of the regression coefficients, which requires estimation of the density of the errors as well as their derivatives. This is a notoriously difficult problem. In any statistical context, such estimation is a very real challenge, and GTS choose as an operational solution to the problem a resampling technique recently developed by Zeng and Lin (2008). The work of Zeng and Lin (2008) is promising, and their simulations encouraging, but certain of their conjectures seem quite optimistic, especially that the general approach would work in cases where the convergence rate is slower than $n^{\frac{1}{2}}$ or where some nuisance parameters can be infinite dimensional. It would be reassuring to have some further theoretical backing.

Apart from the very much more involved estimating approach of GTS, the differences are not great and we would anticipate that the inferences obtained in one approach would be confirmed by an analysis based on the other. However, although broad inferences would be the same, the quantification of the impact of the surrogate, as measured by some coefficient of association, may differ. The set-up of Buyse et al. (2000) is a more comfortable one in that all the required summary statistics flow from a single unifying model. GTS, while using the same basic model (aside from taking logarithms) and appealing to different estimating equations (which, if consistency can be assumed) means we end up ultimately converging to the same quantities, prefer to lean on other measures for trial and individual level surrogacy. Burzykowski, Molenberghs, and Buyse (2005), also considered these alternative measures, based on a suitable model for the marginal distributions of S and T, structured together via a copula function with dependence parameter θ. These have been studied in the semicompeting risks framework (Fine et al., 2001) and an estimator of θ is available. A variance estimator for this has been described in Ghosh (2009).

The nature of the dependence between T or $\log\ T$ and S or $\log\ S$ for the two approaches is not essentially different and, again, the novelty presented by GTS is an inferential one, in particular making use of the work of Lin et al. (1996) and Zeng and Lin (2008), to obtain the correct estimates in the presence of a particular kind of dependency. Malani (1995) proposed a method for dealing with dependent censoring and, at least in spirit, it appears to be very similar to that of Lin et al., although with a potential advantage in being very much simpler to implement. Her ideas were developed further in the context on improving efficiency based on auxiliary survival endpoints (Flandre and O'Quigley, 1995; Cook and Lawless, 2001). In the semicompeting risk setting we can frame the surrogate endpoint problem in terms of making use of an auxiliary survival endpoint so that the Malani algorithm may be a simpler computational alternative to Lin et al. and may be worthy of further investigation. The use made of copula models by GTS is not entirely satisfactory in that it does not fit in, in any natural way, with the chosen approach to basic modeling. It would seem preferable to obtain association measures of trial level surrogacy or individual level surrogacy within the framework of the same model, as was the case with the approach of Buyse and Molenberghs (1998) and Buyse et al. (2000). An alternative way to dealing with these questions is through the illness–death model of Fix and Neyman (1951) and, leaning on a proportional hazards formulation, this is fairly straightforward.

3. Proportional Hazards Analysis

The illness–death model of Fix and Neyman (1951) can be precisely formulated within the context of a proportional hazards model that includes a time-dependent indicator covariate taking the value zero until illness (occurrence of the surrogate endpoint) and, thereafter, the value one. Treatment is then simply another binary indicator variable in the proportional hazards model. Note that proportional hazards models are readily broadened to deal with nonproportional hazards as well as allowing for the estimate of average effects that arise under nonproportionality (O'Quigley, 2008, chapters 6, 7). These models represent an alternative approach to linear and log-linear models. In Prentice (1989), the idea of a pathway is fundamental and this is mirrored precisely by the illness–death model.

3.1 *Nesting of Models*

Many authors, including GTS, underline the fact that, outside of the multinormal model, it is generally not possible to nest restricted models within broader models from the same family. For example, if we can assume that a proportional hazards model holds for the combined effects of treatment and a time-dependent covariate representing the surrogate endpoint, then the restricted model, in which only treatment is considered, no longer belongs to the proportional hazards class. More importantly, if the treatment effect is orthogonal to the effect of the surrogate endpoint, then the coefficient for treatment will be the same, whether or not the surrogate variable is included, only in the case of multinormal models. As a result of this many authors talk of bias caused by modeling for all models other than the multinormal one. This advantage, however, is a very relative one and is lost entirely once we view our models as working approximations to some more complex reality rather than being exact. We can use proportional hazards models to estimate average effect when the true hazards are nonproportional (Xu and O'Quigley, 2000), and this seems to be more useful. Briefly, when comparing models, the issue of model bias has almost certainly been overstated.

3.2 *Prentice Criteria*

For the proportional hazards regression model (Cox, 1972) we specify the intensity function as;

$$\lambda\{t \mid Z(t)\} = Y(t)\lambda_0(t)\exp\{\beta Z(t)\}, \tag{1}$$

where $\lambda_0(t)$ is a fixed but unknown "baseline" hazard function, and β is a $p \times 1$ regression parameter to be estimated. The first surrogacy criterion proposed by Prentice (1989) requires the failure rate for T be independent of treatment, conditional on the surrogate variable. This notion is defined by the relation (Prentice, 1989, p. 433)

$$\lambda\{t \mid Z_1(t), Z_2(t)\} = \lambda\{t \mid Z_1(t)\}, \tag{2}$$

where $Z_1(t)$ is a binary surrogate variable and $Z_2(t)$ is the treatment indicator, in general, not depending upon t, although it can be allowed to and would enable us to analyze crossover designs, for example. This criterion ensures that a surrogate for T should be able to capture the dependence of T on treatment. The second criterion considers a model with only the surrogate variable and requires that the surrogate response have some prognostic implication for the true endpoint (Prentice, 1989, p. 434); that is,

$$\lambda\{t \mid Z_1(t)\} \neq \lambda(t) \tag{3}$$

for all t. Conditions (2) and (3) were proposed as operational criteria for surrogate endpoints in clinical trials (Prentice, 1989).

3.3 *Proportion of Treatment Effect (PTE) Explained*

Freedman, Graubard, and Schatzkin (1992) indicated that it would be quite possible to fail to reject the hypothesis, expressed by equation (2), for instance due to a lack of power, and—as in any testing situation—such a failure to reject cannot be taken as confirmation of the hypothesis. In other words, if we carry out the test, a nonsignificant finding does not really allow us to make any useful statement about the value of the variable $Z_1(t)$ in terms of its potential as a surrogate. Again, as in the usual testing situation, a significant result means something more concrete, in this particular context that there is information left in the treatment variable once the surrogate has been accounted for as far as survival differences are concerned. All of the operational difficulty arises because a null hypothesis cannot be "proven," it can only be rejected.

Freedman et al. (1992) saw that testing whether or not any information remains concerning treatment effect, once the surrogate has been accounted for, was not enough. Some quantification of the strength of effects was needed and, as a suggestion, they proposed a measure PTE. The purpose of PTE was to gauge the amount of treatment effect on survival captured by the surrogate endpoint alone. Although making intuitive sense, the coefficient PTE is difficult to interpret precisely and although work has been done on the statistical properties of the measure, and large sample theory (Lin, Fleming, and DeGruttola, 1997), the nature of the population equivalent of PTE is still not clear. It is not always interpretable as a proportion. Flandre and Saidi (1999) give some examples of use of PTE in practice and these suggest that further theoretical investigation is required. Specifically, they found that for the Delta trial which compares AZT alone to AZT plus ddI or AZT plus ddC, the index proportion of treatment explained by RNA levels to week 16 on time to AIDS/death was estimated as 183%. In the case of AZT+ddI and AZT+ddC, the index was higher at 249%.

Perceived difficulties with PTE has prompted other suggestions for addressing the questions raised by Freedman et al. One of the authors of the current article proposed an index called the F-measure (Wang and Taylor, 2002). This measure is not restricted to the interval (0,1) (or 0% to 100% when dealing in percentages) and, in fact reduces to the PTE (see Ghosh, 2008) in particular cases, thereby inheriting the difficulties associated with PTE. Buyse et al. (2000) argued that two measures—rather than the single PTE measure—were needed, the first called the relative effect (the ratio of the overall treatment effect on survival over that on the surrogate endpoint) and a measure of association between survival and the surrogate having accounted for the effect of treatment.

Within the framework of proportional hazards models, rather than using the estimated regression coefficients to obtain a value of PTE, it is possible to work with R^2 measures of explained variation. These allow us to quantify the strength of the effect of the surrogate endpoint alone, as it affects survival, as well as allowing us to quantify the strength of the effect of treatment on survival after having taken into account the impact of the surrogate endpoint. Indeed the Prentice criteria can be reformulated in an equivalent way in terms of R^2. Specifically, we can reexpress equation (2) as; $R^2\ (Z_2(t)|Z_1(t)) = 0$, where we use the notation $R^2(A|B)$ to indicate the partial explained variation of survival with A after having accounted for the effect of B. The second Prentice criterion (equation (3)) can be reexpressed as $R^2(Z_1(t)) > 0$, i.e., the surrogate endpoint has some impact on survival. The theory of explained variation for survival models, in particular the proportional hazards model, has only recently been fully worked out (O'Quigley, 2008). The theory allows us to give a very concrete interpretation to the values of R^2 as estimates of equivalent population quantities which, in the case of the Prentice criteria, directly translate two things: first, the percentage of variation in survival which can be attributed to the surrogate endpoint and, second, the percentage of variation in survival that can be attributed to the treatment after having accounted for any effect of the surrogate endpoint. These are the very quantities we need in an analysis of surrogacy. A parallel theory in terms of explained randomness is also available (O'Quigley, 2008). In practice, numerical results will be close and, indeed, when dealing with normal variates then explained variation and explained randomness coincide.

As an illustration we looked at a study of 219 patients with resected lung carcinoma randomized 3 weeks after surgery (Decroix, Chastang, and Fichet, 1984). The likelihood ratio test for the first Prentice's criterion leads to a p-value of 0.68, which indicates that the effect of stage (playing the role of treatment here) on survival appears to be largely captured by the surrogate endpoint relapse. The evaluation of the second Prentice's criterion leads to a p-value <0.001, which confirms the strong prognostic effect of relapse on the risk of death. In the light of the Prentice criteria, it seems plausible that relapse be a valid surrogate endpoint for the effect of the grouping variable stage on survival. The R^2 measures allow us to say more. Indeed, we have $R^2(Z_2 \mid Z_1(t)) = 0.00024$, where $Z_1(t)$ is the time-dependent indicator for relapse and Z_2 plays the role of the treatment variable, in this case stage. The very small value, 0.00024, in good agreement with the

likelihood ratio test, implies that when the surrogate variable is already in the model, inclusion of the staging variable does not increase, significantly, the predictive ability of the model. The predictive capability of the surrogate endpoint itself is given by $R^2(Z_1(t)) = 0.696$, a quantity which, while differing significantly from zero and thereby addressing Prentice's second criterion, is sufficiently large to suggest itself as a potentially powerful surrogate. Almost 70% of the variability in survival can be explained by the factor relapse.

3.4 *Metaanalysis*

GTS pay particular attention to metaanalysis as a tool to evaluate surrogacy, especially by appealing to the accelerated failure time model and semicompeting risks. In the context of a proportional hazards model, and an illness–death structure, all of these questions can be dealt with in a relatively straightforward way. In the model, alongside the covariate treatment, the time-dependent covariate surrogate endpoint, we would include a group of $J - 1$ indicator variables to designate the J different trials. This set up easily fits in with the ideas of Buyse et al. (2000) on between- and within-trial surrogacy. Also, we can readily test, and quantify via R^2 measures, center effects alongside treatment and surrogacy effects. Furthermore, questions such as homogeneity of treatment effect across centers, homogeneity of surrogacy across centers and homogeneity of treatment effect given the surrogate endpoint across centers are all readily tested using widely available techniques for nested and stratified proportional hazards models, as well as models with time-dependent covariates.

References

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

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

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, J. (2000). The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics* **1,** 49–67.

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

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

Decroix, G., Chastang, C., and Fichet, D. (1984). Adjuvant immunotherapy with nonviable mycobacterium segmatis in resected primary lung carcinoma. *Cancer* **53,** 906–912.

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

Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse and loss of patients. *Human Biology* **23,** 205–241.

Flandre, P. and O'Quigley, J. (1995). A two-stage procedure for survival studies with surrogate endpoints. *Biometrics* **51,** 969–976.

Flandre, P. and Saidi, Y. (1999). Comment on estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **18,** 107–109.

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

Ghosh, D. (2008). Semiparametric inference for surrogate endpoints with bivariate censored data. *Biometrics* **64,** 149–156.

Ghosh, D. (2009). On assessing surrogacy in a single trial setting using a semi-competing risks paradigm. *Biometrics* **65,** 521–529.

Lin, D. Y., Robins, J. M., and Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83,** 381–393.

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.

Malani, H. M. (1995). A modification of the redistribution to the right algorithm using disease markers. *Biometrika* **82,** 515–526.

O'Quigley, J. (2008). *Proportional Hazards Regression*. New York: Springer.

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

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

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics* **1,** 423–439.

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness-death processes and semi-competing risks data. *Biometrics* **66,** 716–725.

Zeng, D. and Lin, D. Y. (2008). Efficient resampling methods for nonsmooth estimating functions. *Biostatistics* **9,** 355–363.

Rejoinder

Debashis Ghosh,
Jeremy M. G. Taylor,
and Daniel J. Sargent

1. Introduction

We would first like to express our appreciation to coeditor David Zucker and the Associate Editor for organizing this discussion. We also thank the discussants for their comments on our article. They have raised many excellent points, and in our response, we only deal with a subset of them.

Geert Molenberghs (M) and John O'Quigley and Philippe Flandre (OF) accurately describe the methodology in our article as joint regression and association modeling of the surrogate and true endpoints in which a constraint is placed on the type of data that are used (the "wedge" region). As OF noted, this constraint leads to the multistate model of Fix and

Neyman (1951). This data structure complicates the standard estimation procedures that were developed by Burzykowski, Molenberghs, and Buyse (2005, Ch. 11). However, much of the model formulation is very similar to what was described there. The constraints in our approach can be viewed as a different model for the error distribution. Our focus is not on predictions, as advocated by Edward Korn (K), partly because it is very hard with censored data to estimate the intercept parameter in a linear model well without making strong assumptions (Ying, Jung, and Wei, 1995). K is suspicious of the standard errors in our semicompeting risks analysis, but our application of the methodology to data from Ghosh (2009) yielded essentially identical answers to those reported there (data not shown). An implication of the artificial censoring strategy we propose here is that we are throwing away information on recurrences. Consequently, the standard errors for the treatment effects on the surrogate endpoint will increase in our approach relative to approaches that do not throw away that information (e.g., the analyses in Table 1 of K's discussion). An implication of the semicompeting risks approach will be that the magnitude of the treatment effect on the surrogate endpoint will be less than or equal to that on the true endpoint because of the wedge contstraint.

Vance Berger, Grant Izmirlian, and Diana Knoll (BIK) and K criticize us with respect to composite endpoints. There are two issues here. The first is whether or not composite endpoints should be used for assessing treatment effects in clinical trials. BIK and K strongly advocate for composite endpoints such as disease-free survival in oncology trials. Since disease-free survival is arguably a meaningful clinical endpoint, we agree with BIK and K's point if the goal is simply to understand the treatment effect. However, a second goal is attempting to understand the association between the surrogate endpoint with the true endpoint. As we discussed in the article, this is problematic if the surrogate endpoint is a composite endpoint that uses information on the true endpoint. In the context of the motivating colorectal cancer example, we are arguing that recurrence and death are separate processes. One can interpret our modeling strategy as a model for the process that gave rise to the data, rather than a model for the observed data. In modeling the biology, in this context, it is useful to recognize that recurrence is not a spontaneous event. It occurs because the cancer is regrowing and reaches a size where it is detected. From this perspective, there is some rationale for considering when the cancer would have grown to such a size to be detected had not the patient died from something else. K says patients might die from their disease without having progression or having it observed. That is context dependent and pretty rare in the cancer clinical trials we analyze.

OF and K advocate the use of proportional hazards (PH) models in their discussions. Since we were focusing on estimation using the wedge constraint, PH models were not available to us. The recent work of Xu, Kalbfleisch, and Tai (2010), discussed by OF, allows for proportional hazards models for S and T in the semicompeting risks setting. The type of R^2 that OF describe comes from a comparison of models for $T \mid Z$ and $T \mid S$, Z. It is not at all straightforward to calculate this quantity here because of two reasons. The first is that including S as a covariate, in conjuction with the constraint that $S < T$, will complicate estimation. Second, provided one could develop a valid method for estimation in the model for $T \mid S$, Z with $S < T$, calculating an R^2-type measure poses its own issues. Guidance for constructing such measures would come from previous proposals to create likelihood ratio-type statistics from estimating equations (e.g., Li, 1993).

Table 1
R^2 values for recurrence in the colorectal cancer data

Study	R^2
C01	0.51
C02	0.38
C03	0.55
C04	0.52
C05	0.48
C06	0.42
C07	0.38
INT-0035	0.52
NCCTG 784852	0.57
NCCTG 874651	0.54
NCCTG 894651	0.56
NCCTG 914653	0.53

Note: The method of Nagelkerke (1991) was used to calculate R^2. In particular, two PH models were compared. Both had age (log transformed), stage, and treatment as covariates; one included recurrence as a time-dependent covariate, the other did not. The baseline hazard function was modeled using a Weibull distribution.

OF were interested in the R^2 values for our example. We show them in Table 1 for the colorectal cancer data in which parametric Weibull PH models are fit, along with adjustment for stage, age (log transformed), and treatment. The method of Nagelkerke (1991) for calculating R^2 was used. The values range between 0.38 and 0.56, compared with the R^2 value of 0.69 that OF obtained in their example. The question remains of how to set guidelines for using the R^2 value in deciding whether to use the surrogate marker.

M makes a push for performing sensitivity analyses in our modeling procedures. We agree this is an important task and area for future research. He also asks about the potential for causal interpretations of the parameters that we have estimated. Using the structural modeling framework of Pearl (2001), we (Ghosh, Elliott, and Taylor, 2010) have recently shown that the relative effect (i.e., ratio of the two regression coefficients) can be interpreted as a causal parameter in the linear case. There has been recent work on framing the surrogacy problem in the potential outcomes framework (Gilbert and Hudgens, 2008; Li, Taylor, and Elliott, 2010). Attempting to incorporate the semicompeting risks data structure into the potential outcomes framework is more challenging. Suppose we define the potential outcomes $\{S_i^*(1), S_i^*(0), T_i^*(1), T_i^*(0)\}$, $i = 1, \ldots, n$, where $\{S_i^*(Z), T_i^*(Z)\}$ denotes the joint potential outcome for time to the surrogate and true endpoints, respectively, for the ith individual if assigned treatment Z, $Z = 0/1$. Then causal estimands are defined to be within-individual contrasts in T^* and S^*. Frangakis and Rubin (2002) defined the concept of principal stratification, in which within-individual contrasts for T^* are considered conditional on S^*. The problem with the semicompeting risks approach is that S^* might not be well defined if the person experiences

the true endpoint but not the surrogate endpoint. This has been referred to by Zhang and Rubin (2003) as "truncation by death." While the potential outcomes framework might not allow for well-conceptualized causal estimands with semicompeting risks data, this is not the only model for causality that exists in the literature. In particular, econometricians work with so-called structural selection models (Abbing and van den Berg, 2003), and such a modeling framework might allow for better incorporation of semicompeting risks data. Of course, "causal estimand" has a different meaning using these models relative to the potential outcomes framework. This research is currently under investigation.

We broadly agree with much of what the discussants proposed regarding trial-level meta-analysis. K advocates prediction, but as noted before, that is not straightforward in our modeling scheme with censoring present. In our example, we combined treatment arms despite well-documented evidence of heterogeneity in the different groups. We note that of 14,246 initial subjects, there are 56 and 48 subjects with time to recurrence and time to death equaling zero, all of which are censored, so this represents a very small percentage of observations.

The lively discussion of our article has led us to consider a compromise between the association framework proposed here with another view of surrogates, termed auxiliary variables, that might lead to greater consensus. If T is missing, auxiliary variable methods would impute the value of T based on the value of S. In this way, the composite endpoint of disease-free survival (DFS) can be viewed as an imputation strategy by replacing missing values of T with S. From the perspective of auxiliary variables this is clearly biased, but in this setting this might be reasonable for two reasons. First, S and T are highly correlated so we might expect S to be a good prediction of T. Second, DFS as an endpoint has a clinically meaningful interpretation. Thinking of surrogate markers as auxiliary information would seem to be a strategy that could keep the discussants such as K and BIK happy because we still use the real endpoint if it is available but would allow for information in S to be utilized. If S was only weakly related to T then there would be little gain in efficiency. By contrast, if S was strongly related to T then there are potential gains in efficiency.

In closing, we would like to stress that if the goal is to identify surrogate endpoints that occur before the true endpoint so that trials can be done more quickly, then this will necessitate accepting a greater level of uncertainty. There are aspects based on the semicompeting risks framework that allow for this, but by no means is this the only type of methodology available. The question then becomes how much are you willing to lean on the knowledge from biology and data from other trials to help control this uncertainty. How to do this is where the role of statistics is crucial.

Acknowledgements

The authors would like to thank Anna Conlon for the R^2 calculations in Table 1. This research is supported by National Institutes of Health grants CA129102, CA25224, and CA12027.

References

Abbing, J. H. and van den Berg, G. J. (2003). The nonparametric identification of the treatment effects in duration models. *Econometrica* **71,** 1491–1517.

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

Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology* **23,** 205–241.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Ghosh, D. (2009). On assessing surrogacy in a single-trial setting using a semi-competing risks paradigm. *Biometrics* **65,** 521–529.

Ghosh, D., Elliott, M. R., and Taylor, J. M. (2010). Links between surrogate endpoints and endogeneity. *Statistics in Medicine* **29,** 2869–2879.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating causal effect predictiveness of candidate surrogate end-points. *Biometrics* **65,** 1223–1232.

Li, B. (1993). A deviance function for the quasi likelihood method. *Biometrika* **80,** 741–753.

Li, Y., Taylor, J. M. G., and Elliott, M. R. (2010). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66,** 523–531.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78,** 691–692.

Pearl, J. (2001). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Xu, J., Kalbfleisch, J. D., and Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* **66,** 716–725.

Ying, Z., Jung, S. H., and Wei, L. J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association* **90,** 178–184.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics* **28,** 353–368.