Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials

Peer-reviewed author version

Shi, Qian; Renfro, Lindsay A.; Bot, Brian M.; BURZYKOWSKI, Tomasz; BUYSE, Marc & Sargent, Daniel J. (2011) Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 55 (9). p. 2748-2757.

DOI: 10.1016/j.csda.2011.03.014
Handle: http://hdl.handle.net/1942/12137

Biometrics 000, 000–000 DOI: 000
000 0000

Comparative Assessment of Trial-level Surrogacy Measures for Candidate Time-to-event Surrogate Endpoints in Clinical Trials

Qian Shi[1,*], Lindsay A. Renfro[2], Brian M. Bot[1], Tomasz Burzykowski[3], Marc Buyse[4] and Daniel J. Sargent[1]

[1]Department of Health Science Research , Mayo Clinic, Rochester, Minnesota, U.S.A.

[2]Department of Statistical Science, Baylor University, Waco, Texas, U.S.A

[3]Center for Statistics, Hasselt University, Diepenbeek, Belgium

[4]International Drug Development Institute, Hasselt University, Diepenbeek, Belgium

email: shi.qian2@mayo.edu

Summary: Various meta-analytical approaches have been applied to evaluate putative surrogate endpoints (S) of primary clinical endpoints (T), however a systematic assessment of their performance is lacking. Existing methods in the meta-analytic framework can be grouped into two types – conventional and model-based trial-level surrogacy (TLS) measures. Conventional TLS assess the association between treatment effects on S and T, including correlation coefficients and R-square measures from weighted linear regression. Model-based TLS included Copula R^2 proposed by Burzykowski et al. (2001) which measures the ability to predict treatment effect on T based on observed treatment effect on putative S. We examined and compared the estimation performance of these frequently used surrogacy measures in a large scale simulation study. The impact of several key factors on the estimation performance was assessed, including the strength of the true surrogacy, the amount of effective information provided by available data, and the range of within trial treatment effect on S and T. The TLS can be estimated accurately and precisely by both types of surrogacy measures when the true surrogacy is strong, number of trials is large, and the range of within trial treatment effects is wide. When one or more factors deviate from the "best" scenarios, both types of TLS measures tend to underestimate the true surrogacy with increased variability. The estimation performance of conventional measures is similar to model-based measures, but with higher computational efficiency. The findings are applied to a large individual patient data pooled analysis in colon cancer.

Key words: clinical trials; meta-analysis; survival analysis; trial-level surrogacy.

1. Introduction

The evaluation and validation of putative surrogate endpoints in clinical trials is a highly relevant, and controversial topic in methodological and applied statistics. More than a dozen different statistical surrogacy evaluation methods (Weir and Walley, 2006) and numerous publications (Lassere, 2008) have appeared since Prentice (1989) published his milestone paper which laid out the foundation for the validation of putative surrogate endpoint. However, due to the challenges arising from the methodology and commonly encountered data limitations, no single approach to the assessment of potential surrogate endpoint has been fully accepted across statisticians, clinical trialists, and regulatory authorities.

In recent years, surrogacy evaluation based on multiple trials, as initiated with a Bayesian random effects meta-analysis proposed by Daniels and Hughes (1997), has increasingly become the preferred method. In 1998, Buyse and Molenberghs (1998) were the first to formulate surrogacy estimates at both the patient and the trial level based on single trial data. Two years later, this paradigm evolved into meta-analytic framework (Buyse et al., 2000). The shift from assessing surrogacy only at the patient level (usually within a single-trial setting) to assessment at both patient level, and more-importantly at the trial level (Molenberghs et al., 2002) has resulted in a broadly accepted requirement for a surrogate endpoint: the treatment effect observed on a valid surrogate endpoint (substitute) should reliable and precisely predict the treatment effect on the primary endpoint (Biomarkers Definitions Working Group, 2001). Such two-level surrogacy measures based on meta-analytic procedures have been investigated for various types of endpoints (Burzykowski et al., 2005). The Copula R^2_{trial} proposed by Burzykowski et al. (2001) is perhaps the most popular trial level surrogacy measures in cancer clinical trials when both endpoints are survival outcomes. Shi and Sargent (2009) summarized applications evaluating potential surrogate endpoints in colorectal, prostate, and breast cancer.

Within the meta-analytic framework, there remain several proposed trial-level surrogacy measures which are less mathematically sophisticated but more clinically intuitive. These correlation and coefficient of determination type surrogacy measures have been utilized in the research of Adjuvant Colon Cancer Endpoints (ACCENT) Group and other individual investigators (Shi and Sargent, 2009). The ACCENT group has assembled individual patient data from large randomized phase III clinical trials in adjuvant colon cancer conducted worldwide. Their work led to the acceptance of disease-free survival (DFS) as a surrogate endpoint for overall survival (OS) for fluorouracil-based regimens in adjuvant colon cancer studies by United States Food and Drug Administration (FDA)(Sargent et al., 2005, 2007). In Sargent et al. (2005)'s work, surrogacy of DFS was evaluated not only based on Copular R^2_{trial}, but also by the simple measures based on regression or correlation type of analysis. The broad agreement across these measures strengthened the evidence of DFS as a surrogate for OS in the adjuvant colon cancer settings (Green et al., 2008). We will refer to these simple methods as conventional approaches, and the Copula R^2_{trial}, which is based on joint modeling, as a model-based approach from this point forward.

Both the conventional and the model-based trial level surrogacy measures quantify the ability to predict the treatment effect on true endpoint based on observed treatment effect on the surrogate endpoint through hierarchical estimating procedures, i.e. by first estimating the treatment effects on both endpoints within each trial, then performing regression analyses or random effect modeling using the estimated treatment effects across trials. However, a key difference between two approaches is that the surrogate and true endpoints are jointly modeled in a structure of bivariate distribution at the individual patient level for model-based measures, whereas the estimation procedures of conventional measures ignores the generic correlations between endpoints on the same patient. Estimating both level surrogacy simultaneously introduces computational challenges. In practice, the computational burden

of the model-based approaches has been a barrier for the investigators to apply the method. If simple regression-based approaches could provide good performance with available standard functions or procedures in the statistical software this could have important practical advantages.

The performance of both existing model-based and conventional surrogacy metrics is likely affected by many trial characteristics. It is presently unknown whether the commonly-used measures of surrogacy can provide reliable and unbiased estimates regardless of the magnitude of the true surrogacy of a candidate endpoint. The effective amount of information provided by the data – as determined by the number of trials, trial sizes, and percentage of complete observations – is also likely to affect the surrogacy estimation, but the relative influence of these factors remains poorly understood. Other factors likely to influence surrogate evaluation include the range and location of within-trial treatment effects and the adequacy of distributional assumptions.

In order to address these various issues, we performed a large-scale simulation study to evaluate the estimation performance of the trial level surrogacy measures under a variety scenarios, and to compare the conventional marginal and model-based bivariate surrogacy measures. This work is directly motivated by an ongoing collaboration in colon cancer, the previously mentioned ACCENT group, using data from 18 randomized trials including 20,898 patients. The data set used includes trials from 1977 – 1999, and included only trials using flourouracil-based regimens (that was the only effective agent at the time when trials were conducted).

From 1997 to 2002, six new large colon cancer trials were conducted, of which four used new agents in addition to flourouracil. This data is now available. The critical issue of whether DFS remains a valid surrogate endpoint for OS in the presence of these new agents and for more contemporary patients is highly relevant for ongoing drug development. As the

ACCENT group endeavored to explore the surrogacy assessment for limited number of new trials with a smaller range of treatment effects, a large-scale simulation-based comparative evaluation of popular surrogacy measure under less-than-ideal scenarios became necessary and important.

The remainder of this article is as follows. We review each of the trial level surrogacy measures for the time-to-event endpoints in Section 2. In Section 3, we introduce a two-stage data generation procedure used to create the multi-trial datasets. Numerical results of the simulation study are reported in Section 4, and head-to-head comparison between an analysis of the new trials in ACCENT database and a simulation with scenario similar are given in Section 5. Discussion and final remarks are given in Section 6, along with further challenges and technical issues of evaluating potential surrogate endpoint and motivation for creation of a simulation engine for testing surrogate endpoint evaluation methods.

2. Measures of Trial-level Surrogacy

2.1 *Surrogate endpoints in Oncology*

In oncology, overall survival has been widely used as the primary outcome to evaluate novel agents, and is often considered as the "true" endpoint, as it is clearly defined, simple to measure, and easy to interpret. However, in early stage disease settings, extended patient follow-up is required to observe a sufficient number of deaths to achieve the desired statistical power to demonstrate the treatment effect. In addition, the treatment effect measured on OS might be contaminated by the impact of active second line therapies. Considerable interest lies in replacing OS with an earlier endpoint, such as time to recurrence (TTR) or disease-free survival (DFS) in early disease settings or progression-free survival (PFS) in advanced disease settings. These outcomes have been investigated as potential surrogate endpoints in many oncology applications (Shi and Sargent, 2009). Here, we focus on the scenario that the

both true and surrogate endpoints are time-to-event endpoints, as is typically the case in cancer clinical trials.

2.2 *Surrogacy measures under investigation*

Before describing each of the surrogacy measures of interest, we first introduce the notation used throughout the article. Suppose there are total of I trials ($i = 1, \ldots, I$) and in the ith trial n_i patients ($j = 1, \ldots, n_i$) are enrolled. Let T_{ij} and S_{ij} be the time-to-event variables that denote the true and surrogate endpoints respectively, and let X_{ij} be an indicator variable for treatment, with value of 1 for experimental arm and 0 for control arm.

Table 1 summarizes the key features of four trial-level surrogacy measures assessed in the current simulation study. We use the notation R^2_{trial} with different prefixs to distinguish four measures. All four measures are estimated based on a general two-stage estimation procedure. At the first stage, the treatment effects on both endpoints, i.e. Hazard Ratios (HRs) or the regression coefficients associated with the treatment indicator ($ln(HR)$s) in current setting, were estimated based on individual patient data within each of the trials. At the second stage, trial-level surrogacy was estimated based on treatment effects obtained for each trial and endpoint at the first stage.

[Table 1 about here.]

The Copula R^2_{trial} is a typical model-based trial-level surrogacy measure, introduced by Burzykowski et al. (2001) for time-to-event endpoints. Briefly, a Clayton's copula bivariate survival model with Weibull baseline hazard is used to estimate the treatment effects within each trials, i.e. the regression coefficients, on both endpoints at the first stage. Since the copula association parameter captures the association between two failure time variables, the individual-level surrogacy can be estimated simultaneously. At the second stage, a estimate of Copula R^2_{trial} is then given by the coefficient of determination from the random effects modeling of the regression coefficients returned from the first-stage. The commonly used

version of Copula R^2_{trial} is a reduced version which ignores the random intercepts and the estimation variability of the within trial treatment effects. This version of Copula R^2_{trial} is indeed the square of the Pearson correlation coefficient between $ln(HR)$s on S and T (Burzykowski et al., 2001).

The other R^2-type of estimators of the trial-level surrogacy presented in Table 1 are based on conventional marginal regression models at the first stage. For each, two independent Cox proportional hazard (PH) models are used to estimate the HRs on both endpoints within each trials, and the resulting treatment effect estimates are then used to estimate the trial-level surrogacies. Specifically, Spearman R^2_{trial} is estimated from the squared Spearman correlation of the hazard ratios associated with S and T across trials, Pearson R^2_{trial} is similarly estimated from the squared Pearson correlation coefficient of the treatment effects across trials, and WLS R^2_{trial} is estimated by the coefficient of determination from the weighted regression of treatment effects for T onto treatment effects for S across trials. Here, regression weights are given by the sample size for each trial.

In order to produce comparable results, the treatment effects were kept in the regression coefficient form (i.e. $ln(HR)$) and correlation coefficients were squared to align with coefficients of determination. Standard errors of the three conventional measures were estimated from 1,000 bootstrapped samples per dataset.The standard error for Copula R^2_{trial} was estimated based on the delta method. Both data generation and surrogacy estimation were performed in R.

3. Simulations

3.1 *Two-stage Data Generation*

A hierarchical two-stage data generation process was developed to simulate the multi-trial clinical trial datasets. The trial-specific intercepts and treatment effects (slopes) for both

S and T were generated from a multivariate normal distribution at the trial level. These simulated coefficients were used to generate the time-to-event S_{ij} and T_{ij} for jth patient in ith trial at the individual patient level, as detailed below.

Let $(\mu_{S_i}, \mu_{T_i}, \alpha_i, \beta_i)$ denote the regression coefficient vector for ith trial. This random vector can be simulated from the following multivariate normal (MVN) distribution which is equivalent to the random-effect model described by Buyse et al. (2000) and Burzykowski et al. (2001),

$$\begin{bmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{bmatrix}, \Sigma_{tiral} = \begin{bmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{bmatrix} \right),$$

where μ_S and α (μ_T and β) denote the means of intercepts and slopes across all the trials for S (T). The reduced Copula R^2_{trial} assumes the prediction of the treatment effect (slope) on T for a new trial based on observed slope on S from the same new trial is independent of random intercepts, and thus can be formulated as $R^2_{trial} = d^2_{ab}/d_{aa}d_{bb}$. Therefore, it is straightforward to control the true value of Copula R^2_{trial} based on the preceding relationship.

A joint Weibull survival model was constructed to generate the correlation between the uncensored time-to-event S and T for each patient at the individual level. Cowles (2004) showed that a Weibull distribution can be expressed as a scaled mixture of a half-normal distribution (i.e. of standard normal distribution truncated to the positive real line) and an exponential random variable with rate parameter equal to one. Two separate sets of independent exponential random variables are simulated for S and T. When only one set of truncated normal variates are generated to form two time-to-event outcomes, there will be strong correlation between S_{ij} and T_{ij}. When two separate sets of truncated normal variates

are used, S_{ij} and T_{ij} will be uncorrelated. This gives two scenarios – strong and weak individual level surrogacy for the generated S. In the scenarios involving censoring, independent and non-informative censorship was assumed. The censoring time was generated by independent uniform distribution and the desired censoring rates controlled by the parameter values of the uniform distribution.

Prior to performing our primary simulations, we conducted pilot simulations to confirm the desired data characteristics were precisely and consistently represented in our generated datasets. The two-stage data generation procedure described above performs well to produce the scenarios considered.

3.2 *Simulation Settings*

Considering TTR as the candidate surrogate endpoint and OS as the true endpoint, levels explored for each factor of interest are listed in table 2, below.

[Table 2 about here.]

With several trial characteristics above to consider, we assess their relative impact on the estimation performance of the surrogacy measures of interest by holding all other factors fixed at the "best-case scenario" as described by the following: true trial-level surrogacy of $R^2 = 0.90$, no censoring, 50 trials, 2000 patients per trial, strong patient-level correlation between S and T, and a wide range of hazard ratios covering one. For each scenario, 500 multi-trial datasets were generated, and all four surrogacy measures and their standard errors were computed. Other parameters involved in the data generation were fixed at the values derived from ACCENT database.

4. Simulation Results

4.1 *Simulations based on Weibull Data Generation*

The results of the simulation assessment according to each of the main factors were presented in Table 3. Bias and mean squared error (MSE) were calculated to evaluate the accuracy and precision of the surrogacy measures of interest, respectively. Additionally, the coverage of the 95% confidence interval estimates was reported.

Under the best-scenario both model-based and conventional surrogacy measures show sufficient precisions and satisfactory coverage of the 95% confidence intervals. Although a systematic underestimation of the true trial-level surrogacy was observed even in this case, the magnitude of the bias is minor. As each factor was set farther from the level given by the optimal scenario, the estimation performances of all four surrogacy measures decreased in a monotonic fashion. Comparing to other measures, trial level surrogacy estimates based on Spearman's correlation coefficient presents the worst overall estimation performance, at times demonstrating nearly twice the bias and MSE of the other estimators. The other two conventional surrogacy measures, Pearson and WLS R^2_{trial} achieve performance similar to the model-based Copula R^2_{trial}. It is interesting to note that they perform slightly better than Copula R^2_{trial} in some settings, such as a high rate of censoring (70%) or a small range of hazard ratios across trials. Although deviating from the best-scenario affects estimation performance to a different degree across factors, we observed unacceptably low coverage of 95% confidence interval for all surrogacy estimators when the worst scenario presents for each factor.

[Table 3 about here.]

Considering each factor individually, fixing all other factors at their ideal levels, we notice the direction of bias shifts from underestimation to overestimation as the true trial-level surrogacy decreases. The magnitude of the bias remains relatively small when the other

factors remain at their best-case levels. Although the coverage of 95% confidence intervals decreases with true trial-level surrogacy, it remains satisfactory ($> 92.5\%$) in the moderate case with $R^2_{trial} = 0.60$. Furthermore, we find that MSE increases as true trial-level surrogacy decreases. For 18 or more trials, all of the surrogacy measures perform well in terms of bias, MSE and coverage. Performance is substantially decreased, however, when only 6 trials are available. Specifically, bias is nearly doubled, MSE increases nearly ten-fold, and coverage drops to around 85%. Both censoring rate and trial size have an impact on estimation performance similar to that observed for number of trials. As the effective sample size decreases (higher censoring rate or smaller trials), bias and MSE increase while coverage decreases. In particular, when censoring is approximately 70% in trials with 2000 patients, or when trials contain only 500 subjects without censoring, performance of each surrogacy measure in terms of bias and coverage is particularly severe.

When 70% of OS times are censored in trials with 2000 patients, Copula R^2_{trial} actually demonstrates the worst performance of all the surrogacy measures. In this scenario, mean bias is -0.093 and coverage is only 60% for the Copula R^2_{trial}, compared to estimated mean bias of -0.057 and coverage of 86% for both Pearson and WLS R^2_{trial}. For most of the settings we consider, the number of patients remains the same across trials, and therefore quantities associated with Pearson and WLS R^2_{trial} are equal. When trials vary in size, as is the usually case in practice, WLS R^2_{trial} performs better than both Pearson and Copula R^2_{trial} in terms of bias, MSE, and coverage. In Table 3, we see that the range of treatment effects across trials has a severe negative impact on trial-level surrogacy estimation; specifically, the coverage of 95% confidence intervals becomes extremely low when the range of the treatment effects is small, while bias and MSE increase dramatically. Furthermore, this negative impact is worst for Copula R^2_{trial}. Even with a moderate range of treatment effects across trials, Copula R^2_{trial} underestimates the truth by 0.112 on average with a corresponding coverage of 46%. Finally,

the patient-level correlation between S and T has only a slight effect on the estimation performance for all four surrogacy measures.

4.2 *Simulations Based on a Log-normal Data Generation*

The estimation of Copula R^2_{trial} assumes trial and endpoint-specific Weibull baseline hazard functions. The simulation results presented in Section 4.1 were based on data generated from Weibull distributions at the patient level, so it is reasonable to expect that the copula and proportional hazards models used in the first stage of surrogacy estimation provided a good fit to the data. In practice, however, clinically observed time-to-event data may be poorly represented by monotone or proportional hazard functions. Thus, we extend our simulation study to investigate the impact of model misspecification, by generating new trials of data from correlated log-normal distributions. We provide detailed results in Table 4. We repeat the strategy of holding all other factors fixed at the best case scenario, and focus on estimation performance under variations of true trial-level surrogacy and rate of censoring.

[Table 4 about here.]

In general, we find that both the conventional and the model-based surrogacy measures are robust to (mildly incorrect) simplifying assumptions regarding the endpoint and trial-specific hazard functions of the patient-level outcomes. It should be noted that the log-normal data in these simulations, similar to the Weibull data in our previous simulations, were generated using parameter estimates from ACCENT. When censoring is introduced, however, we observe that Copula R^2_{trial} – which continues to assume the patient-level data are marginally distributed as Weibull – becomes sensitive to parametric misspecification. Specifically, the degree of underestimation becomes severe, while decreased true trial-level R^2 corresponds to significantly decreased coverage.

5. Findings from ACCENT Data

ACCENT identified and obtained individual patient data from 6 new phase III adjuvant colon clinical trials testing new biologic agents oxaliplatin and irinotecan combined with 5-FU/LV, and oral fluoropyrimidine regiments. These trials accrued patients between 1997 - 2002, involved 12,676 patients. Since adjuvant therapy for stage II patients remains controversial, we only use data on stage III patients in our study. Sample size within trial varied from 828 to 2264 patients. The censoring rate was generally around 70% for OS in these new ACCENT trials. Hazard ratios of six trials were between 0.84 and 1.07 for OS, and between 0.74 and 1.14 for TTR. These six new ACCENT trials represent a real example where more than one factor with worst scenario, i.e. limited number of trials, high censoring rate, and small range of treatment effects. The estimated HRs based on Copula and Cox PH models for TTR and OS are shown in Figure 1

[Figure 1 about here.]

We performed another set of simulations to assess trial-level surrogacy while mimicking characteristics of stage III patients from the 6 new ACCENT trials. In addition to match with trial sizes, censoring rate, and range of the treatment effects, Weibull baseline hazards were assumed for the correlated event times TTR and OS. All other data generation parameters were fixed at estimates from the new ACCENT trials, considering stage III patients only. In this set of simulations, we consider true trial-level surrogacy equal to 0.90. The results of these simulations are given in Table 5, as well as the estimates of surrogacy based on the true data from 6 new ACCENT studies.

[Table 5 about here.]

As expected, all four trial-level surrogacy measures perform poorly with severe under-estimation and low coverage of 95% confidence intervals. Consistently with the simulation

results, Spearman R^2_{trial} gives the lowest estimate and largest variability among four measures in the real data estimation.

The point estimates of four measures based on ACCENT data range from 0.78 to 0.90. If these methods do tend to underestimate the true surrogacy as suggested by the simulation results, TTR may represent an excellent surrogate for OS in stage III colon cancer patients. However, the large uncertainty illustrated by the extreme wide confidence intervals makes this conclusion less compelling.

6. Discussion

To the best of our knowledge, this is the first large-scale simulation study to assess and compare trial-level surrogacy measures frequently used in practice. Characteristics of the generated data were chosen to capture key aspects of multi-trial analyses where both candidate surrogate and true outcomes are time-to-event in nature. In addition to true trial-level surrogacy, the number of trials, sample size within trials, censoring rate, range of treatment effects across trials, and patient-level correlation between endpoints were selected as important factors for practical consideration. The data generation process developed here facilitated a systematic approach to our study, in which one or more key factors may be varied in a controlled manner while holding all other factors fixed at ideal settings. Four trial-level surrogacy measures were evaluated in terms of their relative estimation performance and robustness to less-than-ideal trial characteristics one might encounter in practice.

Our simulation study demonstrated that true surrogacy can be estimated accurately and precisely when a large number of large trials are available, rate of censoring is low, and a large range of treatment effects exists. Unfortunately in practice this best-case scenario is unlikely to be encountered. When the degree of departure is moderate when each factor considered separately, the loss of precision and underestimation of surrogacy remain somewhat tolerable. It should be note that, unfortunately, frequently several of the worst-case settings considered

separately here are commonly seen in combination in practice. For example as illustrated in Section 5, the data in hand had a limited number of trials with high censoring rates and a small range of the treatment effects across trials.

Overall, we can draw certain useful general conclusions from this work. First, we conclude that the surrogacy measure based on Spearman's correlation coefficient consistently exhibits the worst performance across scenarios of all the measures considered here. The range of treatment effects across trials affects the estimation performance of all four trial-level surrogacy measures substantially. Based on the results presented here, evaluating trial-level surrogacy in meta-analyses with a small range of treatment effects across trials (e.g. excluding trials with negative effects) cannot be practically recommended. In reality, it is nearly impossible to gather multiple trials with exactly the same sample size per trial. When unequal trial sizes are present, it appears that WLS R^2_{trial} performs better than the other measures considered here.

It is interesting that performance patterns are very similar between conventional and model-based surrogacy measures. In a few cases, the conventional measures show better estimation properties than the model-based measure, Copula R^2_{trial}, especially when the parametric distributional assumption is violated. Since estimation of Copula R^2_{trial} requires more elaborate and computationally expensive programs, it may be practical to use the conventional surrogacy measures in some real applications. However, the Copula R^2_{trial} measure assessed in this paper is the reduced version proposed by Burzykowski et al. (2001). More importantly, this measure assumes that the treatment effects are estimated without error at the patient level. Burzykowski et al. (2001) also developed the adjusted R^2_{trial} which accounts for error in estimation of the treatment effects. The program created by these authors for estimating adjusted R^2_{trial}, however, suffers convergence issues in many realistic settings.

Development of a more efficient program for estimating the adjusted R^2_{trial} is ongoing and in the future should allow us to extend our simulation study to include the adjusted measure.

Multiple additional challenges and technical issues exist for evaluating potential surrogate endpoints in real applications. In particular, when only a limited number of trials are available, the estimation variability and low coverage of 95% confidence intervals become a major concern. A common quick fix to this problem involves breaking a trial into sub-units such as centers and estimating the surrogacy on the level of the sub-units rather than the true trials, thereby effectively increasing the meta-analytic sample size. However, the tradeoff between gain of precision and loss of accuracy by splitting trials is not transparent. Further study regarding this situation is ongoing.

Within our current study, the two correlated time-to-event endpoints are assumed to be exchangeable. In many real applications, this many not be the case. For example, DFS is defined as time from randomization to the earlier of disease recurrence or death. When DFS is considered as candidate surrogate endpoints for OS, the surrogate endpoint of interest contains the true clinical endpoint in its composition. The surrogate endpoint is constrained to occur earlier than or be censored by the event of the true clinical endpoint. This clearly creates a more complex mathematical relationship between the endpoints. However, the estimation procedures of frequently-used surrogacy measures (e.g. Copular R^2_{trial}), treat the two endpoints as symmetrical. Further work of comparative assessment of existing surrogacy evaluation methods when constraints between endpoints are present is ongoing.

One of the barriers of applying the two-level surrogacy estimation methods is the computation challenge. This encouraged us to develop user friendly software for surrogate endpoint evaluations (available from the first author by request). This is our first step to creating a integrated software suite which can include multiple methods for estimating or evaluating surrogate endpoints. Combined with a sophisticated data generating process, our goal is

to create a simulation engine to test and compare existing and newly developed surrogate evaluation methods.

References

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

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

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

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

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

Cowles, M. K. (2004). Evaluating surrogate endpoints for clinical trials: A Bayesian approach. Technical report, Department of Statistics and Actuarial Science, University of Iowa.

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16,** 1965–1982.

Green, E., Yothers, G., and Sargent, D. J. (2008). Surrogate endpoint validation: statistical elegance versus clinical relevance. *Stat Methods Med Res* **17,** 477–486.

Lassere, M. N. (2008). The biomarker-surrogacy evaluation schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multi-

dimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Stat Methods Med Res* **17,** 303–340.

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. *Control Clin Trials* **23,** 607–625.

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

Sargent, D. J., Patiyil, S., Yothers, G., Haller, D. G., Gray, R., Benedetti, J., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., Grothey, A., O'Connell, M., Catalano, P. J., Kerr, D., Green, E., Wieand, H. S., Goldberg, R. M., de Gramont, A., and Group, A. C. C. E. N. T. (2007). End points for colon cancer adjuvant trials: observations and recommendations based on individual patient data from 20,898 patients enrolled onto 18 randomized trials from the accent group. *J Clin Oncol* **25,** 4569–4574.

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 Gramont, A. D. (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. *J Clin Oncol* **23,** 8664–8670.

Shi, Q. and Sargent, D. J. (2009). Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int J Clin Oncol* **14,** 102–111.

Weir, C. and Walley, R. J. (2006). Statistical evaluation of biomarkers as surrogate endpoints: A literature review. *Statistics in Medicine* **25,** 183–203.

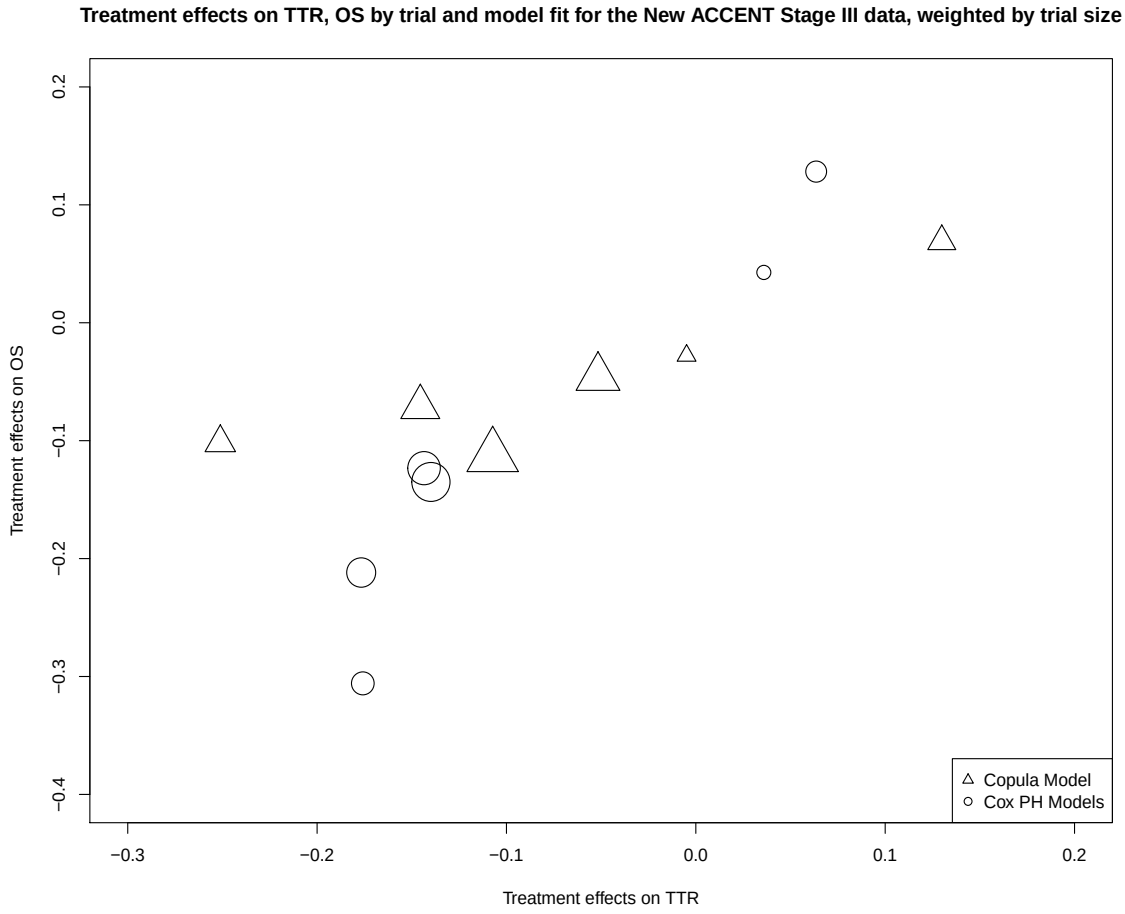
Treatment effects on TTR, OS by trial and model fit for the New ACCENT Stage III data, weighted by trial size
Treatment effects on OS
Treatment effects on TTR
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
Copula Model
Cox PH Models

Figure 1.

Table 1
Trial-level Surrogacy Measures

Surrogacy measure estimator	Spearman R^2_{trial}	Pearson R^2_{trial}	WLS R^2_{trial}	Copula R^2_{trial}
Estimation method at trial level	Spearman Corr	Pearson Corr	Weighted Least Square	Pearson Corr
Treatment effect estimated by	Cox PH	Cox PH	Cox PH	Copula
Joint modeling at indv. level	No	No	No	Yes

Table 2
Factors of interest and corresponding levels considered in the simulation study.

Factors	**Levels Explored in Simulations**
Trial-level R^2 for TTR, OS	0.2, 0.6, 0.9
Censoring rate for OS	0%, 30%, 70%
Number of trials	6, 18, 36 (50 in other simulations)
Sample sizes within trials (equal)	500, 1000, 2000
Unequal trial sizes	50% each of (500, 2000), 33.3% each of (500, 1000, 2000)
Individual-level correlation	Weak (correlation near 0), Strong (correlation near 0.70)
Range of hazard ratios across trials	Approximate ranges: Large (0.5 to 2.0), Moderate(0.7 to 1.6), Small (0.9 to 1.1)

Table 3
Impact of Main Factors on the Estimation Performances

Surrogacy Measures	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
True Trial-level Surrogacy									
	R^2 = 0.90			*R^2 = 0.60*			*R^2 = 0.20*		
Spearman R^2_{trial}	-0.045	0.976	0.004	-0.045	0.952	0.012	-0.004	0.896	0.011
Pearson R^2_{trial}	-0.021	0.940	0.002	-0.014	0.926	0.009	0.013	0.882	0.011
WLS R^2_{trial}	-0.021	0.940	0.002	-0.014	0.926	0.009	0.013	0.882	0.011
Copula R^2_{trial}	-0.023	0.944	0.002	-0.014	0.938	0.009	0.014	0.886	0.011
Number of Trials									
	I = 36			*I = 18*			*I = 6*		
Spearman R^2_{trial}	-0.052	0.986	0.006	-0.072	0.994	0.013	-0.126	0.796	0.059
Pearson R^2_{trial}	-0.023	0.972	0.002	-0.023	0.946	0.004	-0.044	0.862	0.026
WLS R^2_{trial}	-0.023	0.972	0.002	-0.023	0.946	0.004	-0.044	0.862	0.026
Copula R^2_{trial}	-0.025	0.976	0.002	-0.025	0.946	0.004	-0.045	0.824	0.026
Censoring Rate									
	0%			*30%*			*70%*		
Spearman R^2_{trial}	-0.045	0.976	0.004	-0.055	0.984	0.005	-0.087	0.884	0.011
Pearson R^2_{trial}	-0.021	0.940	0.002	-0.027	0.952	0.002	-0.057	0.864	0.005
WLS R^2_{trial}	-0.021	0.940	0.002	-0.027	0.952	0.002	-0.057	0.864	0.005
Copula R^2_{trial}	-0.023	0.944	0.002	-0.051	0.902	0.005	-0.093	0.600	0.011
Trial Size (Number of subjects per trial)									
	2000			*1000*			*500*		
Spearman R^2_{trial}	-0.045	0.976	0.004	-0.064	0.952	0.007	-0.098	0.824	0.013
Pearson R^2_{trial}	-0.021	0.940	0.002	-0.040	0.914	0.003	-0.071	0.746	0.007
WLS R^2_{trial}	-0.021	0.940	0.002	-0.040	0.914	0.003	-0.071	0.746	0.007
Copula R^2_{trial}	-0.023	0.944	0.002	-0.042	0.918	0.003	-0.076	0.726	0.008
Trial Size Mixing									
	2000			*(500, 1000, 2000)*			*(500, 2000)*		
Spearman R^2_{trial}	-0.045	0.976	0.004	-0.069	0.930	0.007	-0.075	0.932	0.008
Pearson R^2_{trial}	-0.021	0.940	0.002	-0.044	0.896	0.004	-0.049	0.870	0.004
WLS R^2_{trial}	-0.021	0.940	0.002	-0.032	0.954	0.003	-0.036	0.956	0.003
Copula R^2_{trial}	-0.023	0.944	0.002	-0.048	0.866	0.004	-0.052	0.864	0.004
Range of Treatment Effect									
	Large			*Moderate*			*Small*		
Spearman R^2_{trial}	-0.045	0.976	0.004	-0.078	0.900	0.009	-0.240	0.154	0.066
Pearson R^2_{trial}	-0.021	0.940	0.002	-0.048	0.884	0.004	-0.207	0.072	0.049
WLS R^2_{trial}	-0.021	0.940	0.002	-0.048	0.884	0.004	-0.207	0.072	0.049
Copula R^2_{trial}	-0.023	0.944	0.002	-0.112	0.460	0.015	-0.214	0.052	0.052
Individual Level Correlation									
	Strong			*Weak*					
Spearman R^2_{trial}	-0.045	0.976	0.004	-0.070	0.926	0.008			
Pearson R^2_{trial}	-0.021	0.940	0.002	-0.046	0.886	0.004			
WLS R^2_{trial}	-0.021	0.940	0.002	-0.046	0.886	0.004			
Copula R^2_{trial}	-0.023	0.944	0.002	-0.045	0.902	0.004			

Table 4
Impact of Distributional Assumption on the Estimation Performance

	$R^2 = 0.90$			$R^2 = 0.60$			$R^2 = 0.20$		
Surrogacy Measures	Bias	Cov.	MSE	Bias	Cov.	MSE	Bias	Cov.	MSE
Log-normal distribution, Censoring rate = 0%									
Spearman R^2_{trial}	-0.026	0.992	0.002	-0.037	0.930	0.012	-0.005	0.908	0.009
Pearson R^2_{trial}	-0.003	0.952	0.001	0.001	0.918	0.008	0.013	0.910	0.009
WLS R^2_{trial}	-0.003	0.952	0.001	0.001	0.918	0.008	0.013	0.910	0.009
Copula R^2_{trial}	-0.007	0.954	0.002	0.002	0.916	0.008	0.018	0.924	0.011
Log-normal distribution, Censoring rate = 30%									
Spearman R^2_{trial}	-0.032	0.990	0.002	-0.033	0.942	0.012	-0.003	0.914	0.010
Pearson R^2_{trial}	-0.009	0.946	0.001	-0.000	0.902	0.009	0.013	0.920	0.010
WLS R^2_{trial}	-0.009	0.946	0.001	-0.000	0.902	0.009	0.013	0.920	0.010
Copula R^2_{trial}	-0.058	0.844	0.002	-0.135	0.750	0.031	-0.109	0.554	0.019

Table 5
Head-to-head comparison between real data estimation and simulations for 6 new ACCENT trials in stage III colon cancer patients

	Real Data Estimation			Simulation Results ($R^2 = 0.90$)		
Surrogacy Measures	Estimate	SE	95% CI	Average	Coverage	MSE
Spearman R^2_{trial}	0.784	0.266	(0.070, 0.975)	0.442	NA	0.298
Pearson R^2_{trial}	0.903	0.080	(0.365, 0.989)	0.494	0.648	0.248
WLS R^2_{trial}	0.884	0.092	(0.294, 0.987)	0.499	0.670	0.244
Copula R^2_{trial}	0.811	0.152	(0.110, 0.978)	0.510	0.696	0.231