Published in final edited form as:
Comput Stat Data Anal. 2014 October 1; 78: 1–20. doi:10.1016/j.csda.2014.03.011.

Center-Within-Trial Versus Trial-Level Evaluation of Surrogate Endpoints

Lindsay A. Renfro,
Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Phone: 507-284-3202

Qian Shi,
Division of Biomedical Statistics and Informatics, Mayo Clinic

Yuan Xue,
Department of Statistics, University of Virginia

Junlong Li,
Department of Biostatistics, Harvard School of Public Health

Hongwei Shang, and
HP Labs

Daniel J. Sargent
Division of Biomedical Statistics and Informatics, Mayo Clinic

Lindsay A. Renfro: renfro.lindsay@mayo.edu

Abstract

Evaluation of candidate surrogate endpoints using individual patient data from multiple clinical trials is considered the gold standard approach to validate surrogates at both patient and trial levels. However, this approach assumes the availability of patient-level data from a relatively large collection of similar trials, which may not be possible to achieve for a given disease application. One common solution to the problem of too few similar trials involves performing trial-level surrogacy analyses on trial sub-units (e.g., centers within trials), thereby artificially increasing the trial-level sample size for feasibility of the multi-trial analysis. To date, the practical impact of treating trial sub-units (centers) identically to trials in multi-trial surrogacy analyses remains unexplored, and conditions under which this ad hoc solution may in fact be reasonable have not been identified. We perform a simulation study to identify such conditions, and demonstrate practical implications using a multi-trial dataset of patients with early stage colon cancer.

Keywords

clinical trials; meta-analysis; surrogate endpoints; survival analysis

1 Introduction

1.1 Background and Motivation

Surrogate endpoints are desired in clinical trials where traditional endpoints are too expensive or difficult to obtain, or where substantial follow-up would be required to observe the clinical endpoint (e.g., survival) in a sufficient number of patients to draw meaningful trial conclusions. While numerous methods for evaluating and validating surrogate endpoints have been proposed, recent consensus has supported evaluation of potential surrogates based on patient-level data from multiple similar trials, where surrogate performance is assessed both within trials (i.e., at the patient level) and across trials (trial level). A surrogate endpoint is considered to be *validated* for use in future clinical trials of the same disease setting when both strong patient-level surrogacy and strong trial-level surrogacy are present.

Central to this multi-trial surrogacy evaluation paradigm is the availability of patient-level data from a relatively large number of randomized clinical trials within the disease setting where the surrogate endpoint is proposed for use. Within a comprehensive surrogacy analysis, a strong association between a candidate surrogate endpoint S and a true clinical endpoint T must be present, where this *patient-level surrogacy* is traditionally quantified as a simple correlation where possible, or evaluated through a multi-trial joint model for S and T, such as a copula model (Burzykowski et al., 2001), otherwise. Arguably of equal or greater importance is *trial-level surrogacy*, which may be demonstrated by a strong predictive relationship (e.g., correlation) between treatment effects on S and treatment effects on T. That is, the experimental treatment's observed effects on a valid surrogate endpoint should provide a strong indication of the experimental treatment's (unobserved) effects on the clinical endpoint. In theory, patient-level surrogacy may be established using patient-level data from one or more historically similar clinical trials, through straightforward joint modeling or correlation (where censoring is not an issue) of the two endpoints S and T. On the other hand, a trial-level surrogacy analysis necessitates access to a *collection* of historical clinical trials, the number of which should be sufficient to compute–at a minimum–the correlation of the (estimated) treatment effects on S and T across trials, along with an associated measure of uncertainty (e.g., standard error). For a surrogacy analysis to be truly informative for clinical decision-making, the standard error associated with an estimate of trial-level surrogacy should be sufficiently small to distinguish a strong surrogate from a weak surrogate, which in turn requires a relatively large "trial-level" sample size (see, e.g., Shi et al. (2011)).

In many practical applications, patient-level data from a large number of comparable randomized trials are difficult or impossible to obtain. Challenges may include: reluctance by data owners to share patient-level data with other parties, lack of time, resources, or expertise to successfully define and pool data elements from a large number of disparate trials into a single database, or non-existence of a large number of similar trials within a specified disease setting and class of treatments. Where a surrogacy analysis is desired but one or more of these issues cause only a few (say, one to five) trials to be available for analysis, a common ad-hoc solution is to perform trial-level surrogacy analyses on trial sub-

units, such as centers, investigators, or geographic regions within trials, as if these sub-units were themselves unique trials.

1.2 Published Uses and Explorations of Trial Sub-Units in Surrogacy Evaluation

Published examples estimating trial-level surrogacy using trial sub-units for analysis include: evaluation of time to progression and progression-free survival as surrogates for overall survival in advanced ovarian cancer, where centers within trials are treated as the trial unit (Buyse et al., 2000; Burzykowski et al., 2001; Molenberghs et al., 2002; Tibaldi et al., 2003; Burzykowski and Buyse, 2006); change in visual acuity at 6 months after treatment as a surrogate for change in visual acuity at 12 months in age-related macular degeneration, where centers are treated as trial units (Buyse et al., 2000; Molenberghs, Geys, and Buyse, 2001; Molenberghs et al., 2002; Tibaldi et al., 2003; Alonso et al., 2004, 2006; Pryseley et al., 2007; Abrahantes, Shkedy, and Molenberghs, 2008; Molenberghs et al., 2008); progression-free survival as a surrogate for overall survival in advanced colorectal cancer, with centers as trial units (Burzykowski et al., 2001; Molenberghs et al., 2002; Tibaldi et al., 2003; Burzykowski and Buyse, 2006; Abrahantes, Shkedy, and Molenberghs, 2008); outcomes of the Positive and Negative Syndrome Scale (PANSS) as a surrogates for the Clinician's Global Impression (CGI) scale in schizophrenia, where treating physicians, main investigators, or countries were considered as trial-level replicates (Molenberghs et al., 2002; Renard et al., 2002; Alonso et al., 2002, 2003, 2004b, 2006; Tilahun et al., 2007; Alonso and Molenberghs, 2007; Abrahantes, Shkedy, and Molenberghs, 2008; Molenberghs et al., 2008, 2010); prostate specific antigen (PSA) as a surrogate for overall survival in advanced prostate cancer, where country was used as the trial unit (Renard et al., 2003; Molenberghs et al., 2004); recurrence-free survival as a surrogate for overall survival in colon cancer, with grouped centers treated as the trial unit(Sertdemir and Burgut, 2009); leukemia-free survival as a surrogate for overall survival in maintenance therapy trials for patients with acute myeloid leukemia in complete remission, where countries within a single trial were treated similarly to trials Buyse et al. (2011); pathologic complete response and local control as surrogates for overall survival in advanced rectal cancer, where grouped centers were treated as trial units Bonnetain et al. (2012);and progression-free survival as a surrogate for overall survival in advanced non-small-cell lung cancer, where centers within trials was the unit of assessment of trial-level surrogacy Laporte et al. (2013).

Although use of trial sub-units in place of trials is commonplace among published trial-level surrogacy analyses, the impact of disregarding the subunit-within-trial hierarchy in these convenient substitutions is relatively unexplored. For the case of two normal endpoints S and T, Abrahantes et al. (2004) performed a simulation study to compare trial-level versus center-level surrogacy estimation as a function of other key factors, such as number of trials, equal versus unequal association of treatment effects at the trial and center levels, and relative variability of trial versus center-level effects. They found that when data contains both trial-specific and center-specific treatment effects, and when these treatment effects truly have the same association across trials as within trials, using center as the unit of measurement to assess surrogacy (rather than trial) does not adversely influence results. However, when unequal association of treatment effects across trials versus within trials is assumed, center-level estimation often over-estimated moderate surrogacy and under-

estimated high surrogacy. This observed weakness of naive center-level surrogacy estimation is alleviated when the variability of treatment effects among centers within trials is constrained to a small fraction (1/100) of the variability of treatment effects across trials–a scenario the authors argue is desirable, but seems unlikely to be observed in practice. The practical effects of naive center-level versus trial-level surrogacy evaluation have not been explored to date with non-normal endpoints, such as time-to-event endpoints, which are of particular relevance and importance in settings where surrogates are desired specifically because they occur earlier or more often in the population of interest. In addition, the effects of unit choice on patient-level surrogacy estimation is previously unexplored.

In this paper, we compare the performance of common surrogacy estimation methods when applied to trials versus application to sub-units within trials, and focus on the case of two time-to-event endpoints *S* and *T*. In Section 2, we present an overview of existing joint (patient-level and trial-level) evaluation methods, namely the multi-trial copula modeling approach and weighted least squares of treatment effects from marginal Cox proportional hazards models, and describe the implications of trial-level versus center-level evaluation in analytical terms. In Section 3, we perform a simulation study to quantify the influence of unit choice on trial-level and patient-level surrogacy estimation, considering factors such as sample size (number of trials, number of centers within trials, and number of patients within centers), equal or unequal trial-level versus center-level surrogacy, relative variability of treatment effects among trials versus among centers within trials, patient-level surrogacy, and amount of censoring. We apply the surrogacy evaluation methods to a collection of five randomized clinical trials in colon cancer in Section 4, where both trial-level and center-level evaluations are performed and results compared. We conclude with a discussion in Section 5.

2 Surrogacy Evaluation Methodology and Choice of Units

Here we provide an overview of existing multi-trial surrogacy evaluation methodology for two time-to-event endpoints *S* and *T*, and describe analytical implications when these methods are applied at the trial sub-unit level versus the intended trial level. A similar presentation for two Normally distributed endpoints without censoring was given in Abrahantes et al. (2004), in which both a “full” random effects model and a computationally convenient “reduced” model (originally proposed in Burzykowski et al. (2001)) were described. For brevity, we restrict our discussion to the reduced model as it applies to time-to-event endpoints, the version applied most often in practice. Without loss of generality, we henceforth refer to trial sub-units as “centers,” recognizing that other sub-units such as geographic regions or treating physicians may be considered.

2.1 Models and Notation

Consider a collection of similar historical trials indexed by $i \in \{1, \ldots, N\}$, centers within trials indexed by $j \in \{1, \ldots, N_i\}$ where N_i is the number of centers within trial i, and patients within centers indexed by $k \in \{1, \ldots, n_{ij}\}$ where n_{ij} is the number of patients within center ij. Let z_{ijk} denote the treatment assigned to patient k within center j from trial i, where $Z = 1$ for experimental treatment and $Z = 0$ for control. Denote by s_{ijk} and t_{ijk} the observed surrogate endpoint S and true endpoint T, respectively, of the same patient. We assume that marginal

models (e.g., parametric or Cox propotional hazards models) for S and T depend on trial-specific intercepts μ_{S_i} and μ_{T_i} and trial-specific slopes (treatment effects) a_i and β_i, respectively, given by the trial-level model:

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix}, \quad (1)$$

where $(\mu_S, \mu_T, \alpha, \beta)'$ is a vector of fixed effects and $(m_{S_i}, m_{T_i}, a_i, b_i)'$ is a vector of trial-specific random effects distributed as:

$$\begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad D_r = \sigma_T^2 \begin{pmatrix} d_{SS}/\sigma_T^2 & d_{ST}/\sigma_T^2 & 0 & 0 \\ d_{ST}/\sigma_T^2 & d_{TT}/\sigma_T^2 & 0 & 0 \\ 0 & 0 & 1 & \rho_T \\ 0 & 0 & \rho_T & c \end{pmatrix} \right). \quad (2)$$

The subscript r denotes that covariance D_r is a *reduced* covariance matrix assuming no correlation between trial-specific intercepts and treatment effects. Under the model above, trial-level surrogacy is captured by the quantity $R^2_{trial} = \rho_T^2/c$, where an estimate $\hat{R}^2_{trial}$ close to 1 indicates strong trial-level surrogacy and a value near 0 indicates weak trial-level surrogacy.

In the present setting, we additionally assume existence and knowledge of trial subunits or centers, where trial i is associated with a vector of center effects $(\mu_{S_{ij}}, \mu_{T_{ij}}, a_{ij}, \beta_{ij})'$ given by the center-level model:

$$\begin{pmatrix} \mu_{S_{ij}} \\ \mu_{T_{ij}} \\ \alpha_{ij} \\ \beta_{ij} \end{pmatrix} = \begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} m_{S_{ij}} \\ m_{T_{ij}} \\ a_{ij} \\ b_{ij} \end{pmatrix}, \quad (3)$$

where $(\mu_{S_i}, \mu_{T_i}, a_i, \beta_i)$ was defined in (1) above and $(m_{S_{ij}}, m_{T_{ij}}, a_{ij}, b_{ij})$ is a vector of center-specific random effects distributed as:

$$\begin{pmatrix} m_{S_{ij}} \\ m_{T_{ij}} \\ a_{ij} \\ b_{ij} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad D'_r = \sigma_C^2 \begin{pmatrix} d'_{SS}/\sigma_C^2 & d'_{ST}/\sigma_C^2 & 0 & 0 \\ d'_{ST}/\sigma_C^2 & d'_{TT}/\sigma_C^2 & 0 & 0 \\ 0 & 0 & 1 & \rho_C \\ 0 & 0 & \rho_C & c' \end{pmatrix} \right). \quad (4)$$

From (3)–(4), surrogacy at the center level is given by $R^2_{center} = \rho_T^2/c'$. If we may safely assume independence of the trial-level and center-level random effects $(m_{S_i}, m_{T_i}, a_i, b_i)^T$ and $(m_{S_{ij}}, m_{T_{ij}}, a_{ij}, b_{ij})^T$, it follows that

$$\begin{pmatrix} \mu_{S_{ij}} \\ \mu_{T_{ij}} \\ \alpha_{ij} \\ \beta_{ij} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix}, \quad D_r^* = D_r + D_r' \right). \quad (5)$$

2.2 Surrogacy Estimation

2.2.1 Trial-level surrogacy—When a sufficiently large collection of similar historical trials are available, an estimate of trial-level surrogacy may be obtained from a two-stage procedure introduced by Burzykowski et al. (2001), where estimated trial-specific marginal effects obtained from first-stage modeling may then be used to estimate R^2_{trial} as computed from (1)–(2) in the second stage. Under this reduced model, R^2_{trial} is equivalent to the squared correlation of the trial treatment effects (α_i, β_i), and it is usually estimated from the (log) hazard ratios obtained from fitting trial- and endpoint-specific Cox models (where patient-level association of *S* and *T* is ignored) or from maximum likelihood estimation of a multi-trial copula model for *S* and *T* (which incorporates patient-level association, as described in Section 2.2.2).

Optionally for $\hat{R}^2_{trial}$ computation, trial-specific treatment effects estimates from either the marginal or joint surrogacy modeling approach may be weighted by some function of the contributing trial sizes; weighted least squares regression of the estimated trial treatment effects on the true endpoint onto the corresponding estimated effects on the surrogate is commonly employed, where the squared coefficient of determination from this model serves as an estimate of trial-level surrogacy. Standard errors for $\hat{R}^2_{trial}$ and associated confidence intervals may be obtained via analytical or computational methods.

Although a collection of historical trials data with available center information may be assumed to follow model (5) based on the composition of models (1)–(2) and (3)–(4), in practice, when a sufficient number of trials are available and trial-level surrogacy is the quantity of interest, only the trial-specific treatment effects (α_i, β_i) are traditionally involved in computing trial-level surrogacy, while any center-level effects are often ignored. Abrahantes et al. (2004) considered three-level (trial, center, patient) models for estimation of trial-level surrogacy for the case of two normal endpoints, but they report numerical challenges and describe substantial bias for these methods, specifically in those scenarios likely to be observed in practice. As a result, the more straightforward two-level (trial, patient) modeling procedure described above is widely applied, even in cases where an insufficient number of trials are available.

2.2.2 Patient-level surrogacy—In most comprehensive surrogacy evaluations involving trial-level surrogacy estimation as described in Section 2.2.1, patient-level surrogacy (the association of *S* and *T* at the patient level) is also of interest. In practice, the original endpoints *S* and *T* might be reasonably assumed to arise from a joint model, as non-zero correlation of *S* and *T* is generally expected in the case of a strong surrogate. As previously

mentioned, availability of a large number of trials is favorable but not necessary to evaluate patient-level surrogacy.

While methodology for patient-level surrogacy evaluation is not the focus of this paper, the choice of units in a trial-level surrogacy evaluation may also impact patient-level surrogacy evaluation when joint (e.g., copula) models are utilized. Traditionally for time-to-event endpoints *S* and *T*, a multi-trial copula model is used to simultaneously estimate trial-specific marginal effects (such as the treatment effects subsequently used in a two-stage, trial-level surrogacy evaluation) as well as a measure of patient-level correlation or association of *S* and *T*, as captured by one or more copula association parameters. After estimation of the multi-trial copula model, e.g. via maximum likelihood estimation, the association parameter(s) may be subsequently transformed to yield a measure of patient-level surrogacy, $\hat{R}^2_{indv}$, that is bounded on [0, 1] for ease of interpretation. Similar to the interpretation of $\hat{R}^2_{trial}$, a value of $\hat{R}^2_{indv}$ near 1 indicates strong patient-level surrogacy, while a value near 0 indicates weak patient-level surrogacy. We refer the reader to Burzykowski et al. (2001) for an introduction to patient-level surrogacy modeling using multi-trial copulas.

2.3 Choice of Units in Trial-Level Surrogacy Analyses: Implications

When the number of available trials in a given surrogacy analysis is insufficient for estimation of R^2_{trial} from model (1)–(2), the same model and methodology is commonly applied to trial sub-units. In this case, the quantity actually estimated involves only center-specific intercepts ($\mu_{S_{ij}}$ and $\mu_{T_{ij}}$) and treatment effects (a_{ij} and β_{ij}) without recognition of trial-level effects, and is based on the covariance matrix D^*_r from model (5) instead of D'_r, where

$$D^*_r = \begin{pmatrix} d_{SS}+d'_{SS} & d_{ST}+d'_{ST} & 0 & 0 \\ d_{ST}+d'_{ST} & d_{TT}+d'_{TT} & 0 & 0 \\ 0 & 0 & \sigma^2_T+\sigma^2_C & \rho_T\sigma^2_T+\rho_C\sigma^2_C \\ 0 & 0 & \rho_T\sigma^2_T+\rho_C\sigma^2_C & c\sigma^2_T+c'\sigma^2_C \end{pmatrix}.$$

We refer to trial-level surrogacy estimated in this way as R^2_{naive}, which can be written as a function of R^2_{trial}:

$$\begin{aligned} R^2_{naive} &= \frac{(\rho_T\sigma^2_T+\rho_C\sigma^2_C)^2}{(\sigma^2_T+\sigma^2_C)(c\sigma^2_T+c'\sigma^2_C)} \\ &= \frac{\sigma^2_T(\sigma^2_T+\frac{\rho_C}{\rho_T}\sigma^2_C)^2}{c(\sigma^2_T+\sigma^2_C)(\sigma^2_T+\frac{c'}{c}\sigma^2_C)} \\ &= R^2_{trial} * \frac{(\sigma^2_T+\frac{\rho_C}{\rho_T}\sigma^2_C)^2}{(\sigma^2_T+\sigma^2_C)(\sigma^2_T+\frac{c'}{c}\sigma^2_C)} \\ &= R^2_{trial} * \frac{(1+\frac{\rho_C}{\rho_T}\frac{\sigma^2_C}{\sigma^2_T})^2}{(1+\frac{\sigma^2_C}{\sigma^2_T})(1+\frac{c'}{c}\frac{\sigma^2_C}{\sigma^2_T})}. \end{aligned} \quad (6)$$

Therefore, one possible scenario for $R^2_{center-naive}$ to reasonably approximate R^2_{trial} is given by the joint conditions $\rho_C/\rho_T \to 1$ and $c'/c \to 1$. Note from the trial-level model (1)–(2) and patient-level model (3)–(4) that $corr(\alpha_i, \beta_i) = \rho_T / \sqrt{c}$ and $corr(\alpha_{ij}, \beta_{ij}) = \rho_C / \sqrt{c'}$. Then

$$\frac{\rho_C}{\rho_T} \to 1 \text{ and } \frac{c'}{c} \to 1 \Rightarrow \begin{pmatrix} 1 & \rho_C \\ \rho_C & c' \end{pmatrix} \to \begin{pmatrix} 1 & \rho_T \\ \rho_T & c \end{pmatrix},$$

suggesting that R^2_{naive} is a reasonable approximation for R^2_{trial} when the center-level treatment effects correlation matrix approaches the trial-level treatment effects correlation matrix. Another scenario where estimation of R^2_{naive} may be a reasonable solution to assess trial-level surrogacy is when $\sigma^2_C/\sigma^2_T \to 0$. To this end, it should be noted that $var(\alpha_{ij})/var(\alpha_i) = \sigma^2_C/\sigma^2_T$ and $var(\beta_{ij})/var(\beta_i) = c\sigma^2_C/c'\sigma^2_T$, which is true for all $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, N_i\}$. Then $\sigma^2_C/\sigma^2_T \to 0$ occurs when $var(\alpha_{ij})/var(\alpha_i) \to 0$ and $var(\beta_{ij})/var(\beta_i) \to 0$, or when the variability of center-specific treatment effects within trials is much smaller than the variability of treatment effects across trials, and when this is true for both the surrogate and true endpoints. When it can reasonably be assumed that the ratio of the variances of the surrogate effects to the true effects is equal at the trial and center levels (i.e., that $c = c'$), then it follows from (6) that lower center-level surrogacy than trial-level surrogacy ($\rho_C < \rho_T$) directly results in lower naive trial-level surrogacy performed on centers than trial-level surrogacy performed on trials, i.e. $R^2_{naive} < R^2_{trial}$. Conversely, when surrogacy at the center level within trials is higher than surrogacy across trials ($\rho_C > \rho_T$), then $R^2_{naive} > R^2_{trial}$. This result will not only be confirmed in the simulations of Section 3, but will be evident in the data analysis of Section 4. As we will describe in Section 5, this result may also be exploited to develop practical guidelines for conservative use of centers as units of analysis when use of trials may be untenable.

The theoretical developments of this section suggest two possible scenarios where R^2_{naive} may be reasonably close to R^2_{trial}, thus warranting estimation of R^2_{naive} for convenience in some cases. However, it could be argued that these conditions, namely (1) equality of trial-level and patient-level correlation matrices or (2) much greater variability of treatment effects across trials than within trials, are unlikely to be observed in practice. Therefore, we performed a simulation study to assess the impact of naive trial-level surrogacy estimation performed on centers within trials, focusing on time-to-event endpoints and those combinations of meta-analytic features (e.g., number of trials, number of centers within trials, true underlying trial-level or center-level surrogacy, or relative variability of center versus trial effects) which we deemed likely to be of greatest interest to practitioners.

3 Simulation Study

Because time-to-event endpoints are of primary interest in our own and many other applications, we performed a simulation study to determine the extent to which particular meta-analytic features (e.g., number of trials or centers, underlying trial-level or center-level

surrogacy, or relative variability of treatment effects at each level) influence differences between naive center-level and trial-level surrogacy evaluation.

3.1 Data Generation

Simulated data were generated according to a three step procedure. First, trial effects (α_i, β_i) were generated from (1)–(2) assuming (α, β) = (0, 0), c = 1, and with ρ_T derived from R^2_{trial} and σ^2_T according to the scenarios in Table 1. For the purposes of data generation, trial-specific intercepts μ_{S_i} and μ_{T_i} were fixed at 0, but the estimation process allowed for non-zero intercepts. Next, given the generated trial effects, center-level effects were generated from (3)–(4) assuming c' = 1 and ρ_C derived from R^2_{center} and σ^2_C according to Table 1, assuming center-specific intercepts ($\mu_{S_{ij}}$, $\mu_{T_{ij}}$) = (0, 0). Finally, using the generated center-level effects (α_{ij}, β_{ij}) and random variates (u_{ijk}, v_{ijk}) drawn from a Clayton copula model $C(u, v) = \{u^{1-\gamma} + v^{1-\gamma} - 1\}^{1/(1-\gamma)}$ with patient-level surrogacy defined by Kendall's $\tau = (\gamma - 1)/(\gamma + 1)$, correlated event times (s_{ijk}, t_{ijk}) were produced via application of the inverse-CDF method. Specifically, S and T were assumed to follow marginal Weibull distributions parameterized according to

$$f(x|r,\lambda)=(r/\lambda)(x/\lambda)^{r-1}\exp(-(x/\lambda)^r), \quad r,\lambda,x \geq 0.$$

Allowing unique baseline hazard functions for each endpoint and center within trial, patient-level observations of the (correlated) surrogate and true clinical endpoints were generated from:

$$S_{ijk} \sim \text{Weibull}(r^S_{ij}, \lambda^S_{ijk}) \quad (7)$$

$$\text{and} \quad T_{ijk} \sim \text{Weibull}(r^T_{ij}, \lambda^T_{ijk}), \quad (8)$$

where r^S_{ij} and r^T_{ij} are center-within-trial-specific shape parameters (set equal to 2 for data generation), and regressors z_{ijk} and corresponding generated treatment coefficients α_{ij} and β_{ij} were introduced through center-within-trial and patient-specific scale parameters $\lambda^S_{ijk}=\exp(-\mu^S_{ij}-\alpha_{ij}z_{ijk}/r^S_{ij})$ and $\lambda^T_{ijk}=\exp(-\mu^T_{ij}-\beta_{ij}z_{ijk}/r^T_{ij})$. This model parameterization was chosen for comparability of its treatment coefficients with those resulting from Cox proportional hazards models.

3.2 Scenarios

For each scenario given in Table 1, 10,000 iterations were performed. To investigate the "marginal" effects of particular trial or data settings (such as number of trials, underlying surrogacy at each level, and censoring rate) on trial versus center-level surrogacy evaluation, we performed simulations for a set of "high-level" scenarios presented in the upper part of Table 1. For these scenarios, simulation parameters of interest were varied one at a time, while holding all other parameters fixed at ideal settings. Here, we define "ideal" as the best-case settings theoretically understood to produce the least bias and smallest variability (e.g.,

from prior simulation studies such as Shi et al. (2011) or Renfro et al. (2012)). Specifically, the "ideal" scenario (scenario 1) assumes high surrogacy at the patient, center, and trial levels, with $R^2_{trial}=R^2_{center}=\tau=0.90$. Further assumed are availability of 15 trials with 20 centers per trial and 100 patient per center, larger variability of the trial-specific treatment effects around their mean ($\sigma^2_T=0.50$) than center-level effects around their trial-specific means ($\sigma^2_C=0.05$), and no censoring (e.g., each patient contributes observed events *S* and *T* to model estimation procedures).

For comparison against the ideal scenario, high-level scenarios (1 through 20) were created to study the marginal effects of the following: an increase (scenario 2) or decrease (scenario 3) in the number of available trials, reduced trial-level surrogacy (scenarios 4–5), reduced center-level surrogacy (scenarios 6–7), equally small (scenario 8) or large (scenario 9) variability of trial and center level effects, greater variability among center level effects than trial effects (scenario 10), reduced patient-level surrogacy (scenarios 11–12), increased rate of censoring (scenarios 13–14), reduced number of centers within trials (scenarios 15–16), and reduced (scenario 17), mixed (scenario 18), or increased (scenarios 19–20) number of patients within centers.

Upon completion of the high-level simulations, we performed additional simulations for a set of "focused" scenarios, where settings were specifically chosen to represent those practical situations where trial-level surrogacy analyses might be performed instead on centers: namely, a small number of available trials (1, 3 or 5), a variable number of centers within trials, and a mix of total accrual by center. For simplicity and based on the results of the high-level scenarios, assumptions consistent across the focused scenarios included: strong patient-level and trial-level surrogacy, a mix of sample sizes across centers, no censoring of *S* or *T*, and larger variability of treatment effects at the center level versus trial level.

In Figure 1, we display a random set of generated trial-specific treatment effects (α_i, β_i) along with corresponding generated center-specific effects (α_{ij}, β_{ij}) under the four scenarios comprised of $\sigma^2_T \in \{0.05, 0.50\}$ and $\sigma^2_C \in \{0.05, 0.50\}$ in the simple case where $R^2_{trial}=R^2_{center}=0.90$. If one now assumes that these (generated) treatment effects are trial-specific and center-specific *estimates* to be used in the second stage of a trial-level (and naive center-level) surrogacy estimation, the possible implications of the relative sizes of σ^2_T and σ^2_C become readily apparent. In part because centers and other trial sub-units by definition have (much) smaller sample sizes than their original trials, it is reasonable to expect the observed variability of estimated center-level treatment effects to be larger in practice than the variability observed across trials for trial-level effects. Thus, for the focused scenarios intended to reflect situations where naive center-level estimation might be performed, we consider only the case where $\sigma^2_T=0.05$ and $\sigma^2_C=0.50$.

3.3 Estimation Methods

For all scenarios in Table 1, two different modeling approaches using patient-level data were performed: trial-specific marginal Cox proportional hazards models ignoring patient-level association between *S* and *T*, and a multi-trial Clayton copula model capturing (*S*, *T*) association through a single copula association parameter. Specifically for the later approach, *S* and *T* were assumed to follow correlated Weibull distributions with trial-specific hazard functions according to the copula model and (7)–(8) described in Section 3.1. For both approaches, treatment effects estimates obtained at the first stage were used to estimate trial-level surrogacy at the second stage, according to (1)–(2). Patient-level surrogacy was also investigated, both assuming trial-specific surrogacy τ_i and equal-trial surrogacy τ by transformation of the estimated Clayton copula association parameter to the scale of Kendall's τ. In all cases, estimation of trial-level surrogacy R^2_{trial} was weighted by trial size.

Because use of trials versus centers as the units of interest in surrogacy analyses is the central exploration of this paper, all trial-level surrogacy estimation was separately performed using trials (when $N > 1$) and centers as the unit of evaluation. Although each dataset was generated from assumed "true" levels of surrogacy at the trial and center levels (usually $R^2_{trial}=R^2_{center}=0.90$), in reality, a given set of correlated treatment effects generated for a single iteration rarely adhered to these values. For this reason, throughout this paper, we compare trial-level estimates $\hat{R}^2_{trial}$ and the corresponding naive center-level estimates $\hat{R}^2_{naive}$ to both the underlying "true" trial-level surrogacy R^2_{trial} of interest and the generated R^2_{trial} obtained by squaring the correlation of the generated trial treatment effects.

3.4 Results

3.4.1 Trial-level surrogacy—Table 2 and Figures 2–4 present the results of trial-level surrogacy analyses performed using trials versus centers as the unit of analysis. We note that for these simulations, comparing a surrogacy estimate $\hat{R}^2_{trial}$ against its corresponding generated value of R^2_{trial} is more informative than comparisons against "true" R^2_{trial}, as the correlation of generated treatment effects (α_i, β_i) can vary widely from iteration to iteration. Thus, we emphasize bias and MSE results for comparisons against generated surrogacy whenever possible (i.e., when $N > 1$), but emphasize comparisons against true R^2_{trial} otherwise (when $N = 1$). For completeness, we report both comparisons for all scenarios, and note that the conclusions generally agree.

Among the high-level scenarios, we find varying degrees of negative bias under naive center-level estimation compared to trial-level estimation when true trial-level surrogacy is high (i.e., all scenarios except 4 and 5), and some degree of positive bias when center-level surrogacy is high but true trial-level surrogacy is moderate (scenario 4) or low (scenario 5). We note these results are consistent with the findings of Abrahantes et al. (2004) for Normally distributed endpoints. When many trials are available ($N = 30$ under scenario 3), and even when only a few trials are available, such as $N = 5$ under scenario 2, trial-level surrogacy analyses performed on trials show better estimation performance than those

performed on centers. When trial-level surrogacy is truly high but center-level surrogacy is moderate (scenario 6) or low (scenario 7), performing trial-level surrogacy estimation on centers results in substantial negative bias. Equal variation of treatment effects at the center and trial levels (versus larger variation at the trial level than the center level, as in the ideal scenario) results in negative bias for naive center-level estimation when variability at both levels is low ($\sigma_T^2=\sigma_C^2=0.05$ under scenario 8), but both estimation approaches (trials versus centers as units) perform well when variability at both trial and center levels is high ($\sigma_T^2=\sigma_C^2=0.50$ under scenario 9). When larger variability exists among treatment effects at the center level than the trial level, treating centers as the units of trial-level surrogacy estimation produces similar results to trials (scenario 10). Decreases in patient-level surrogacy cause negative bias for trial-level surrogacy estimation when trial-level surrogacy is truly high and analyses are performed on centers as units, as evidenced for scenarios 11 and 12, while presence of censoring has minimal effects under either approach (scenarios 13 and 14). When a relatively large (N = 15) number of trials are available, decreasing the number of centers within trials has minimal effects on estimation (scenarios 15 and 16). However, even when a large number of trials and centers within trials are available, a reduced number of patients within centers results in substantial negative bias for trial-level surrogacy estimation when treating centers as the unit of analysis (scenarios 17 and 18). For confirmation, scenarios 19 (with n_{ij} = 500) and 20 (with n_{ij} = 1000) were performed, showing that the small amount of negative bias present even for the ideal scenario 1 vanishes when the center-level sample size is extremely large (n_{ij} > 100).

Among the focused scenarios, where a small number of available trials ($N \in \{1, 3, 5\}$) is assumed, we find additional differences in estimation performance for trials versus centers as the units of analysis in trial-level surrogacy estimation. When only 1 trial with 5, 10, or 20 centers is available (scenarios 21, 22, and 23, respectively), performing a trial-level surrogacy analysis on trials is impossible. In this case, performing an analysis on trial sub-units is perhaps the best that can be done, and in our simulations, where a mixture of center-level sample sizes ranging from 10 to 50 is assumed, such an analysis seems reasonable. Specifically, for an increased number of centers within a single trial, naive-center level surrogacy estimation performs increasingly well at estimating the true underlying value of trial-level surrogacy, R^2_{trial}. When 3 trials are available with 5, 20, or a mixture of 5, 10, or 20 centers per trial (scenarios 24, 25, and 26, respectively), performing trial-level surrogacy analyses on trials (versus centers) remains the least biased approach, though with the disadvantage of increased MSE. When 5 trials with 5, 20, or a mixture of 5, 10, or 20 centers is available (scenarios 27–29), the results are similar.

In general, trial-level surrogacy estimates returned by the Cox versus copula modeling approaches are similar, as they should be. Where substantial differences occur, the copula approach offers some advantages. Specifically, the Cox approach results in a small to moderate number of outlying trial-level surrogacy estimates when centers are the unit of analysis, as can be seen when comparing estimates $\hat{R}^2_{naive}$ to generated values of R^2_{trial} for scenarios 1, 4, 6, 7, 9, and 11–14 (comparisons to true underlying R^2_{trial} obscure these outliers). In these cases, it is possible that some instances of numerical instability of the

marginal Cox models (e.g., as a result of the smaller sample sizes associated with centers as units) are mitigated by estimation using the parametric copula model in its place. These differences between methods indeed become more exaggerated when the sample size per center is reduced or varied, as is the case in Scenarios 17, 18, and 23–29. When the sample size per center is artificially increased to n_{ij} = 500 or 1000 (scenarios 19–20), the differences between modeling approaches cease to exist. These results motivate consideration of a copula estimation method for multi-trial surrogacy evaluation, perhaps as confirmation or results produced by marginal Cox models, when small unit-specific sample sizes (e.g., < 100) are present.

3.4.2 Patient-level surrogacy—Table 3 and Figures 5–7 present the results of patient-level surrogacy analyses performed using trials versus centers as the units of trial-level surrogacy analyses. For these scenarios, copula estimation assuming both equal patient-level surrogacy τ across units and unit-specific patient-level surrogacy is reported. For the latter approach, bias and MSE are computed for Table 3 by further averaging unit-specific bias and MSE across units, while the minimum, median, and maximum values of $\hat{\tau}_i$ across units are displayed in box plots in Figures 5–7. While we show all results for completeness, for brevity we will focus on estimation assuming equal τ across units of analysis. As a reminder, equal patient-level surrogacy across trials was assumed throughout data generation and is the method generally used in the literature.

In general, estimation of R^2_{trial} using centers versus trials as the unit of analysis has relatively little impact on patient-level surrogacy, with a few notable exceptions. Under the ideal scenario (scenario 1) assuming high patient-level surrogacy τ = 0.90, τ is estimated with negative bias under both approaches, but with larger bias when centers are used as the unit of analysis. This remains true for any number of trials (scenarios 2 and 3), decreased trial-level surrogacy (scenarios 4 and 5), equally small variability of treatment effects across trials and centers within trials ($\sigma^2_C = \sigma^2_T = 0.05$; scenario 8), decreased patient-level surrogacy (scenarios 11 and 12), presence of censoring (scenarios 13 and 14), a decreased number of centers within trials (scenario 16), and a decreased number of patients within centers (scenarios 17 and 18). However, when underlying center-level surrogacy is reduced (scenarios 6, and 7), or when the variability of center-level treatment effects within trials is large ($\sigma^2_C = 0.50$; scenarios 9 and 10), patient-level surrogacy estimation shows less bias when performed using centers as the unit of analysis versus trials. Similarly, when the number of patients within centers is very large (scenarios 19 and 20), naive center-level estimation can show slight gains in accuracy over trial-level estimation, though these gains are diminished as $n_{ij} \to \infty$. When only one center per trial exists (scenario 15), as expected, the two methods yield identical results. Among the focused scenarios where a small number of trials (N = 1, 3, or 5) are considered, patient-level surrogacy estimation performed on trials versus units as centers generally yield similar results. An exception occurs when unit-specific τ_i are assumed. In this case, substantial negative bias and outlying estimates are frequent among $min(\hat{\tau}_i)$, the minimum values of unit-specific surrogacy $\hat{\tau}_i$ estimated at each iteration, when centers are the unit of analysis. This is in sharp contrast to the good estimation performance observed when unit-specific τ_i are similarly assumed but trial is the

level of analysis (see Figure 7). It could be the case that small center-level sample sizes are insufficient to accurately estimate patient-level surrogacy when this surrogacy is separately estimated for individual centers.

4 ACCENT Data Analysis

Sargent et al. (2005) previously validated 3-year disease free survival (DFS) as a surrogate endpoint for 5-year overall survival (OS) in trials of adjuvant treatment for colon cancer, using a variety of meta-analytic surrogacy measures and patient-level data from more than 20,000 individuals enrolled to 18 randomized clinical trials. In this re-analysis of the ACCENT data, we consider 5 of the original 18 trials where center-level identifiers are available, and separately estimate “trial level” surrogacy using trials versus centers as the unit of analysis. In the present manuscript, for purposes of exposition, both time to recurrence (TTR) and disease-free survival (DFS) are considered as potential surrogates for overall survival, and the primary aim is estimation of trial-level and patient-level surrogacy for each candidate endpoint.

Of the 5 original trials, one contained 3 experimental treatment arms, and thus a total of 7 pairwise comparisons can be made (we will henceforth refer to these comparisons as “trials”). Within each of the 7 two-arm trials, centers were also considered. Due to the high rate of censoring (at least 70%) present for each endpoint, centers containing fewer than 20 patients were randomly combined with larger centers from the same trial containing at least 20 patients. This resulted in a total of 60 centers across the 7 trials, with a range of 1 to 18 centers per trial. Trial-level surrogacy estimation was performed using both marginal Cox proportional hazards models and a multi-trial Clayton copula model assuming equal patient-level association across trials, while patient-level surrogacy estimation was achieved using the copula model. In all cases, estimation of trial-level surrogacy R^2_{trial} weighted by trial (center) size was performed, and standard errors were obtained by applying the delta method to the standard errors associated with the weighted correlations.

4.1 Results: ACCENT Data

Results of the ACCENT analyses are presented in Figure 8 and Table 4. We find that for both PFS and TTR, performing surrogacy analyses on trials versus centers yields substantially higher estimates of trial-level surrogacy, $\hat{R}^2_{trial} > \hat{R}^2_{naive}$. In each of these cases, trial-level surrogacy based on centers may be viewed as a conservative estimate of trial-level surrogacy based on trials, as it takes the large variability among centers (as evidenced in Figure 8) into account that is otherwise ignored at the trial level.

As expected, due to estimation based on an increased sample size (number of units), standard errors for $\hat{R}^2_{trial}$ are substantially smaller under naive center-level estimation than trial estimation, which may be viewed as an advantage if the relative disadvantages of treating centers as trials are taken into account. For both DFS and TTR, patient-level surrogacy estimation did not vary according to units of analysis.

Noting that the full ACCENT analysis presented here is based on a number of trial-level units ($N = 7$) larger than what might be available in practice, we also considered trial-level versus naive center-level surrogacy analyses for DFS and TTR on all possible subsets of the 7 ACCENT trials with size $N = 4$ (35 combinations in total). For each candidate surrogate (DFS and TTR) and trial subset, we plotted the trial-level surrogacy estimated from trials versus the surrogacy estimated from centers, shown in Figure 9.

In all but one case for DFS and in all cases for TTR in Figure 9, naive trial-level surrogacy based on centers as units is lower than trial-level surrogacy based on trials. From (6), assuming $c = c'$, $\hat{R}^2_{naive}$ is guaranteed to underestimate R^2_{trial} when $\rho_C < \rho_T$ (or equivalently, when $R^2_{center} < R^2_{trial}$). This phenomenon was examined in simulation scenarios 6 and 7 and seems to be in effects for DFS and TTR in ACCENT. Among the trials in ACCENT with more than one center, such that trial-specific center-level surrogacy may be computed, we find that center-level surrogacy for DFS ranges from 0.837 to 0.959, while center-level surrogacy for TTR ranges from 0.591 to 0.879. Thus, it is unsurprising that $\hat{R}^2_{naive} < \hat{R}^2_{trial}$ for each endpoint.

5 Discussion

A set of recommendations by an NIH workgroup for the evaluation of surrogate endpoints included establishing databases within disease settings, where potential surrogates can be evaluated most robustly (de Gruttola et al., 2001). The ACCENT database utilized in this paper is one such successful endeavor. However, the large number of trials required to utilize the gold standard surrogacy evaluation methodologies are in practice a substantial challenge to assemble, and as such, a common ad-hoc solution is performance of trial-level surrogacy analyses on trial sub-units. Here, we investigated the effects of unit choice (trial versus centers) on multi-trial surrogacy evaluations where time-to-event endpoints are of interest.

Based on the results of the simulations and data analyses of this paper, we provide the following recommendations. First, when trial-level surrogacy (versus center-level surrogacy) is truly the quantity of interest and a sufficient number of trials (e.g., $N > 10$) are available, surrogacy analyses using the original trials as the units of evaluation should be performed as the primary surrogacy analysis. When only a moderate number (e.g., 5 to 9) trials are available for analysis, the practitioner may wish to perform surrogacy analyses at both levels, with greater emphasis given to the trial-level analysis and supporting evidence provided by the center-level analysis. When estimated surrogacy is sufficiently strong in each case, one might conclude that the the surrogate endpoint is promising; otherwise, additional exploration or justification of the surrogate for future use may be required.

When only 3 or 4 trials are available for analysis, we recommend proceeding as follows. First, it can be shown (e.g., via simulation) that a correlation coefficient based on too few units is likely to be negatively biased, with absolute bias increasing with the strength of correlation. Because of this, and due to the relationship between R^2_{trial} and R^2_{naive} derived in equation (6), we suggest performing surrogacy analyses on three separate levels: (1)

estimation of R^2_{trial} using trials as units, (2) estimation of R^2_{naive} using centers (across trials) as units, and (3) estimation of $R^2_{center,i}$, computed across centers within each trial $i \in \{1, \ldots, N\}$ for trials with more than one center (to gain some idea of center-level surrogacy). Based on the relative surrogacy results at each level (which we have grouped into 3 possible situations below), some general conclusions may be drawn:

1. If $\hat{R}^2_{trial} > max(\hat{R}^2_{center,i})$ for all or most $i \in \{1, \ldots, N\}$, it should also be true from (6) that $\hat{R}^2_{trial} > \hat{R}^2_{naive}$. In this case, $\hat{R}^2_{naive}$ may be viewed as a conservative estimate of $\hat{R}^2_{trial}$, the quantity of interest. If $\hat{R}^2_{naive}$ is sufficiently high, the surrogate *S* may be viewed as promising to replace *T* in a future trial.

2. If $\hat{R}^2_{trial} > \hat{R}^2_{center,i}$ for approximately half of $i \in \{1, \ldots, N\}$ and $\hat{R}^2_{trial} > \hat{R}^2_{center,i}$ for the other trials, then $\hat{R}^2_{trial}$ and $\hat{R}^2_{naive}$ should be similar in magnitude. If so, *S* may be considered a reasonable surrogate for *T* when both $\hat{R}^2_{trial}$ and $\hat{R}^2_{naive}$ are sufficiently high.

3. When $\hat{R}^2_{trial} > \hat{R}^2_{center,i}$ for all or most $i \in \{1, \ldots, N\}$, it is likely, based on equation (6) and simulation scenarios 4 and 5 presented in Section 3, that $\hat{R}^2_{trial}$ will be some degree lower than $\hat{R}^2_{naive}$. In this case, the relative influence of two possible factors, (1) negative bias caused by $\hat{R}^2_{trial}$ computation with too few units and (2) higher center-level correlation ρ_C than trial-level correlation ρ_T, cannot be distinguished. For this reason, we caution against use of *S* in future trials based on even promising values of $\hat{R}^2_{naive}$, as $\hat{R}^2_{naive}$ is possibly inflated relative to the quantity of interest R^2_{trial}, which is immeasurable in practice and cannot be assumed to be bounded above another measurable quantity, as was possible for situation (1).

When fewer than 3 trials are available for analysis, even rough estimation of R^2_{trial} using trials as units (for comparison against R^2_{naive} and $R^2_{center,i}$) is impossible. In this case, we caution against performing center-level surrogacy evaluations alone, for at least two reasons. First, (6) demonstrates that promisingly high estimates $\hat{R}^2_{naive}$ are potentially inflated relative to the quantity of interest R^2_{trial}, and this occurs particularly when center-level surrogacy is high. This could lead not only to overconfidence in the surrogacy analysis, but also–and more critically–to a change of endpoint in future trials of the disease setting that might be unjustified. Second, leave-one-out prediction of the effect of treatment on the true endpoint in a ”future” trial given the observed effects on the surrogate endpoint across historical trials–a central component of most published multi-trial surrogacy evaluations–cannot reliably be performed on centers in place of trials, as the location and variability of center-specific effects generally do not represent those of trial effects. For these reasons, when patient-level data from only 1 or 2 trials are available, we advocate for continued use of the true clinical endpoint *T* in future clinical trials, at least until a multi-trial surrogacy evaluation is truly possible.

When any multi-trial surrogacy analysis is performed, it is helpful to recall that the respective trial sizes were originally based (at least in part) on powering the same treatment comparisons now being analyzed across trials in the surrogacy analysis. To this end, one should bear in mind that *center*-specific treatment effects utilized in naive surrogacy analyses will be associated with increased variability relative to the trial effects that a multi-trial surrogacy analysis is intended to quantify and predict. Indeed, the primary goal of a multi-trial surrogacy analysis is to confirm that the treatment effect on the true clinical endpoint in a new trial, which may be difficult or costly to estimate, can be well-predicted by the more readily observed effect on a surrogate endpoint. When a collection of paired historical center-level treatment effects are not sufficiently representative of their corresponding trial-level effects, the derived prediction model based on *centers* may fail to represent what may reasonably be expected for the *S, T* relationship in a future *trial*.

In conclusion, multi-trial meta-analytic evaluations of surrogate endpoints for future use in clinical trials remains the gold standard approach, with inherent practical challenges such as required availability of patient-level data from a large number of similar clinical trials within the same disease setting. With careful understanding of the available data and with consideration given to the recommendations provided here, estimates of trial-level surrogacy may be obtained or reasonably approximated even when the number of trials is limited.

References

Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. Journal of the Royal Statistical Society, Series C (Applied Statistics). 2001; 50:405–422.

Shi Q, Renfro LA, Bot BM, Burzykowski T, Buyse M, Sargent DJ. Comparative assessment of trial-level surrogacy measures for candidate time-to-event endpoints in clinical trials. Computational Statistics and Data Analysis. 2011; 55:2748–2757.

Buyse M, Molenberghs G, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics. 2000; 1:49–67. [PubMed: 12933525]

Molenberghs G, Geys H, Buyse M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. Statistics in Medicine. 2001; 20:3023–3038. [PubMed: 11590630]

Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski B, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Controlled Clinical Trials. 2002; 23:607–625. [PubMed: 12505240]

Renard D, Geys H, Molenberghs G, Burzykowski B, Buyse M. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. Biometrical Journal. 2002; 44:921–935.

Alonso A, Geys H, Molenberghs G, Vangeneugden T. Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. Journal of Biopharmaceutical Statistics. 2002; 12:161–178. [PubMed: 12413238]

Renard D, Geys H, Molenberghs G, Burzykowski B, Buyse M, Vangeneugden T, Bijnens L. Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. Journal of Applied Statistics. 2003; 30:235–247.

Tibaldi FS, Abrahantes JC, Molenberghs G, Renard D, Burzykowski B, Buyse M, Parmar M, Stijnen T, Wolfinger R. Simplified hierarchical linear models for the evaluation of surrogate endpoints. Journal of Statistical Computation and Simulation. 2003; 73:643–658.

Alonso A, Geys H, Molenberghs G, Kenward M, Vangeneugden T. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. Biometrical Journal. 2003; 45:931–945.

Alonso A, Molenberghs G, Burzykowski T, Renard D, Geys H, Shkedy Z, Tibaldi F, Abrahantes JC, Buyse M. Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. Biometrics. 2004; 60:724–728. [PubMed: 15339295]

Molenberghs G, Burzykowski T, Alonso A, Buyse M. A perspective on surrogate endpoints in controlled clinical trials. Statistical Methods in Medical Research. 2004; 13:177–206. [PubMed: 15198486]

Alonso A, Geys H, Molenberghs G, Kenward M, Vangeneugden T. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. Biometrics. 2004; 60:845–853. [PubMed: 15606404]

Alonso A, Molenberghs G, Geys H, Buyse M, Vangeneugden T. A unifying approach for surrogate marker validation based on Prentice's criteria. Statistics in Medicine. 2006; 25:205–221. [PubMed: 16220497]

Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharmaceutical Statistics. 2006; 5:173–186. [PubMed: 17080751]

Tilahun A, Pryseley A, Alonso A, Molenberghs G. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. Computational Statistics and Data Analysis. 2007; 51:4152–4163.

Alonso A, Molenberghs G. Surrogate marker evaluation from an information theory perspective. Biometrics. 2007; 63:180–186. [PubMed: 17447943]

Pryseley A, Tilahun A, Alonso A, Molenberghs G. Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. Clinical Trials. 2007; 4:587–597. [PubMed: 18042568]

Abrahantes JC, Shkedy Z, Molenberghs G. Alternative methods to evaluate trial level surrogacy. Clinical Trials. 2008; 5:194–208. [PubMed: 18559408]

Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. Journal of Statistical Planning and Inference. 2008; 138:432–449.

Sertdemir Y, Burgut R. Does the decision in a validation process of a surrogate endpoint change with level of significance of treatment effect? A proposal on validation of surrogate endpoints. Contemporary Clinical Trials. 2009; 30:8–12. [PubMed: 18809512]

Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Statistical Methods in Medical Research. 2010; 19:205–236. [PubMed: 19608602]

Buyse M, Michiels S, Squifflet P, Lucchesi KJ, Hellstrand K, Brune ML, Castaigne S, Rowe JM. Leukemia-free survival as a surrogate end point for overall survival in the evaluation of maintenance therapy for patients with acute myeloid leukemia in complete remission. Haematologica. 2011; 96:1106–1111. [PubMed: 21546500]

Bonnetain F, Bosset JF, Gerard JP, Calais G, Conroy T, Mineur L, Bouche O, Maingon P, Chapet O, Radosevic-Jelic L, Methy N, Collette L. What is the clinical benefit of preoperative chemoradiotherapy with 5FU/leucovorin for T3–4 rectal cancer in a pooled analysis of EORTC 22921 and FFCD 9203 trials: surrogacy in question? European Journal of Cancer. 2012; 48:1781–1790. [PubMed: 22507892]

Laporte S, Squifflet P, Baroux N, Fossella F, Georgoulias V, Pujol J, Douillard J, Kudoh S, Pignon J, Quinaux E, Buyse M. Prediction of survival ben-efits from progression-free survival benefits in advanced non-small-cell lung cancer: evidence from a meta-analysis of 2334 patients from 5 randomised trials. BMJ Open. 2013; 3:1–6.

Abrahantes JC, Molenberghs G, Burzykowski T, Shkedy Z, Alonso A, Re-nard D. Choice of units of analysis and modeling strategies in multilevel hierarchical models. Computational Statistics and Data Analysis. 2004; 47:537–563.

Renfro LA, Shi Q, Sargent DJ, Carlin BP. Bayesian adjusted R^2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. Statistics in Medicine. 2012; 31:743–761. [PubMed: 22161275]

Sargent DJ, Wieand HS, Haller DG, Gray R, Benedetti JK, Buyse M, Labianca R, Seitz JF, O'Callaghan CJ, Francini G, Grothey A, O'Connell M, Catalano PJ, Blanke CD, Kerr D, Green E,

Wolmark N, Andre T, Goldberg R, de Gramont A. 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. 2005; 34:8664–8670. [PubMed: 16260700]

de Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institutes of Health workshop. Controlled Clinical Trials. 2001; 22:485–502. [PubMed: 11578783]

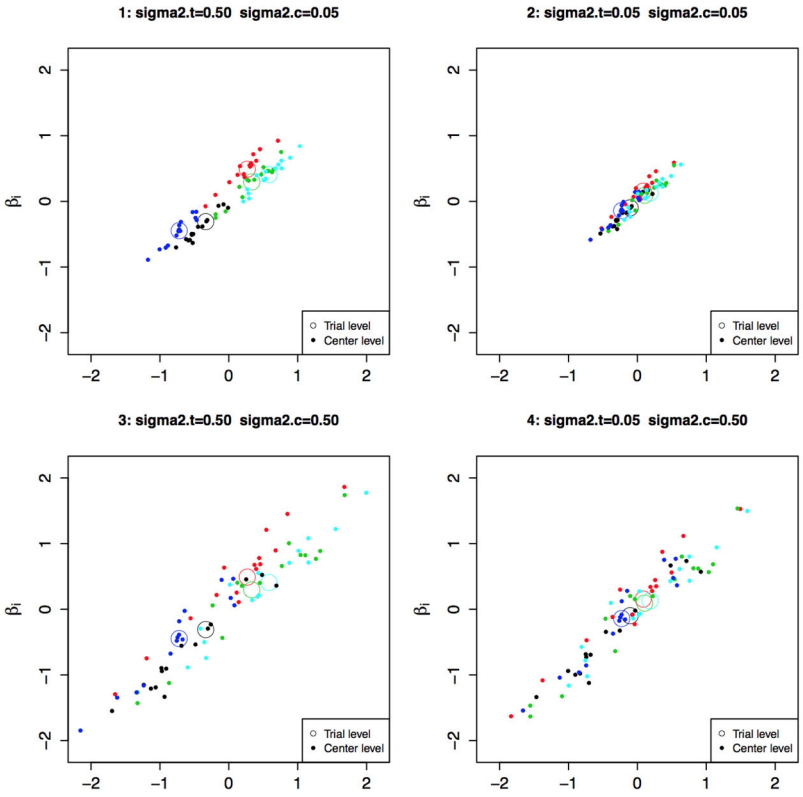

Figure 1.

Generated values of (α_i, β_i) and corresponding (α_{ij}, β_{ij}) where $N = 5$, $N_i = 15$, and $R^2_{trial}=R^2_{center}=0.90$. In each case, generated $R^2_{trial}=0.8994$.

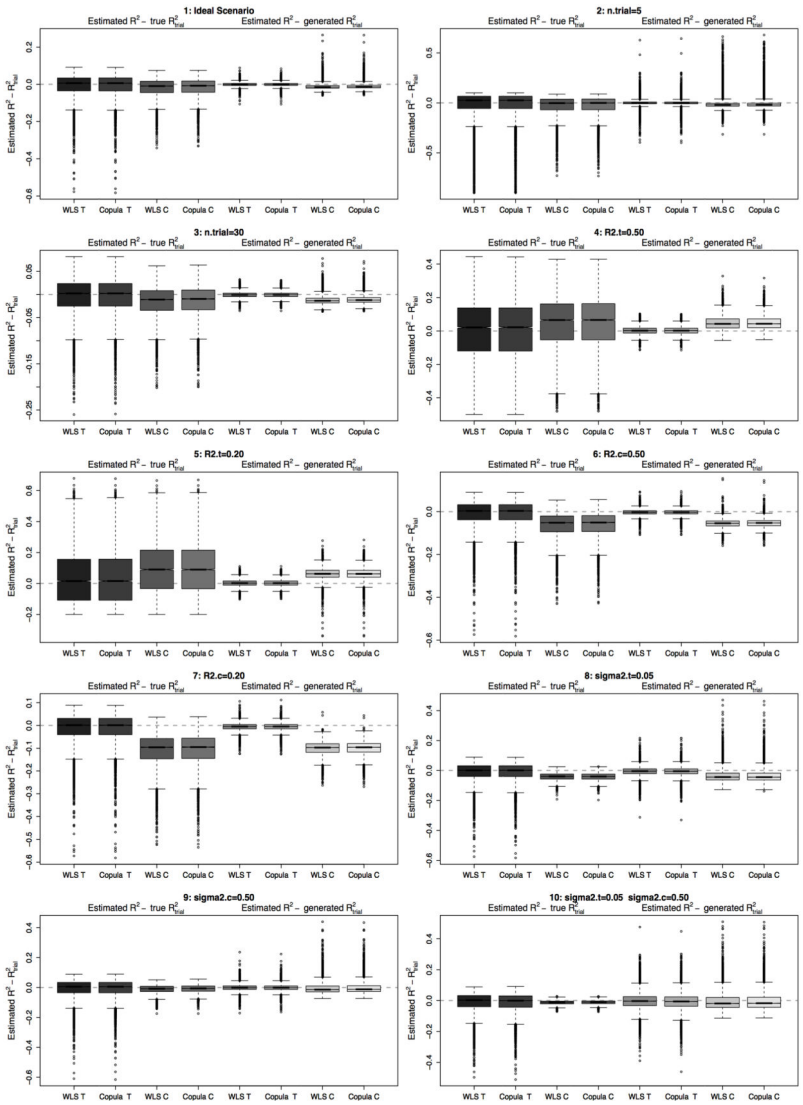

Figure 2.

Boxplots of estimated minus true trial-level surrogacy, by units of analysis (T = trial, C = center), estimation method (Cox, copula), and comparator (true vs. generated R^2_{trial}) for high-level scenarios 1–10.

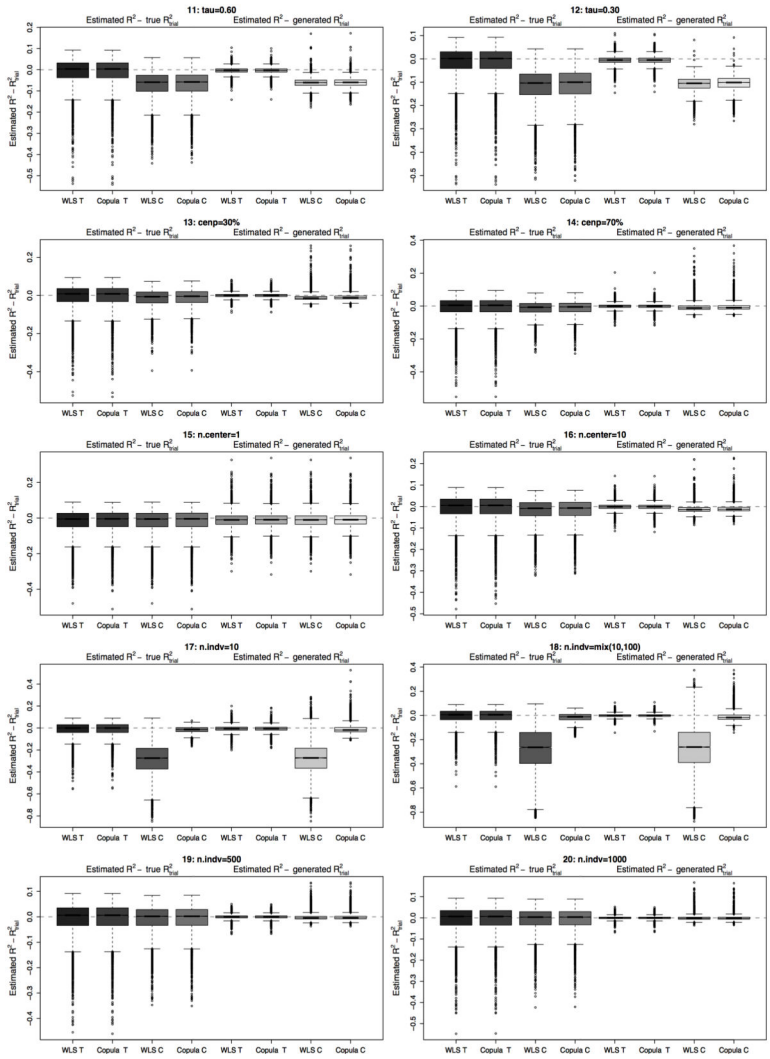

Figure 3.

Boxplots of estimated minus true trial-level surrogacy, by units of analysis (T = trial, C = center), estimation method (Cox, copula), and comparator (true vs. generated R^2_{trial}) for high-level scenarios 11–20.

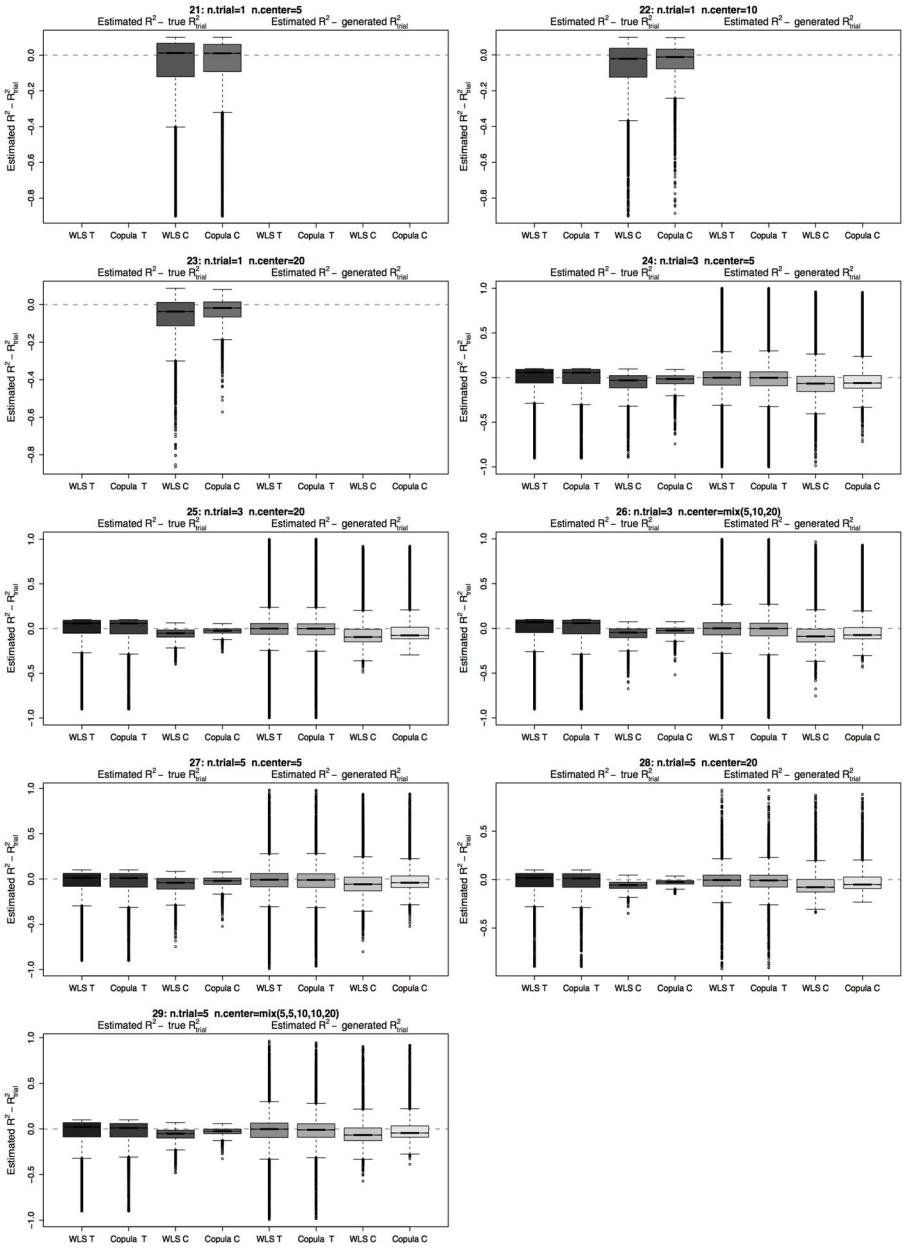

Figure 4.

Boxplots of estimated minus true trial-level surrogacy, by units of analysis (T = trial, C = center), estimation method (Cox, copula), and comparator (true vs. generated R^2_{trial}) for focused scenarios 21–29.

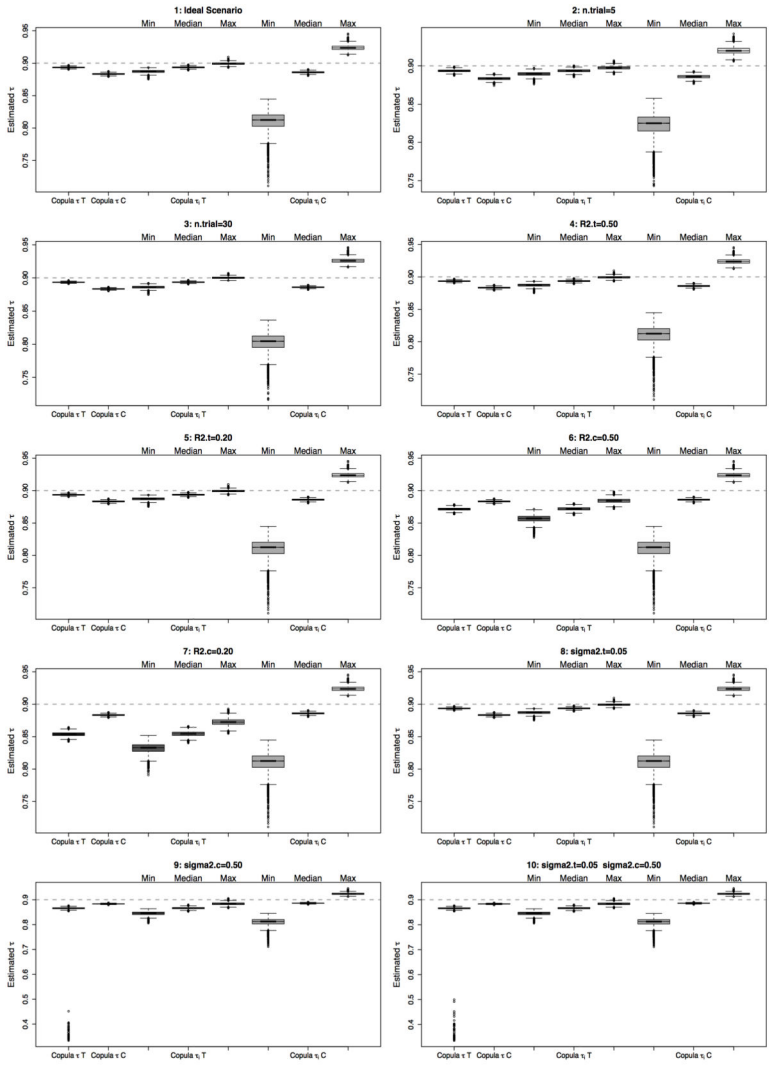

Figure 5.

Boxplots of estimated patient-level surrogacy, by units of analysis (T = trial, C = center), and assuming equal patient-level surrogacy τ or unit-specific patient-level surrogacy τ_i across units, for high-level scenarios 1–10.

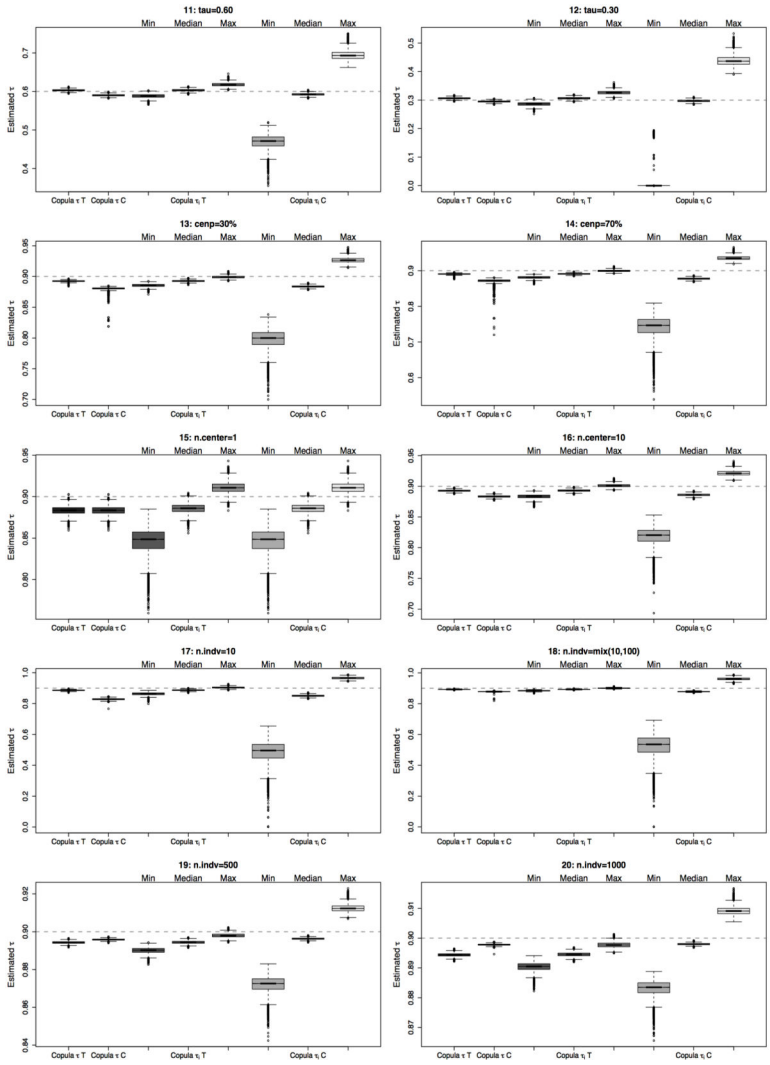

Figure 6.

Boxplots of estimated patient-level surrogacy, by units of analysis (T = trial, C = center), and assuming equal patient-level surrogacy τ or unit-specific patient-level surrogacy τ_i across units, for high-level scenarios 11–20.

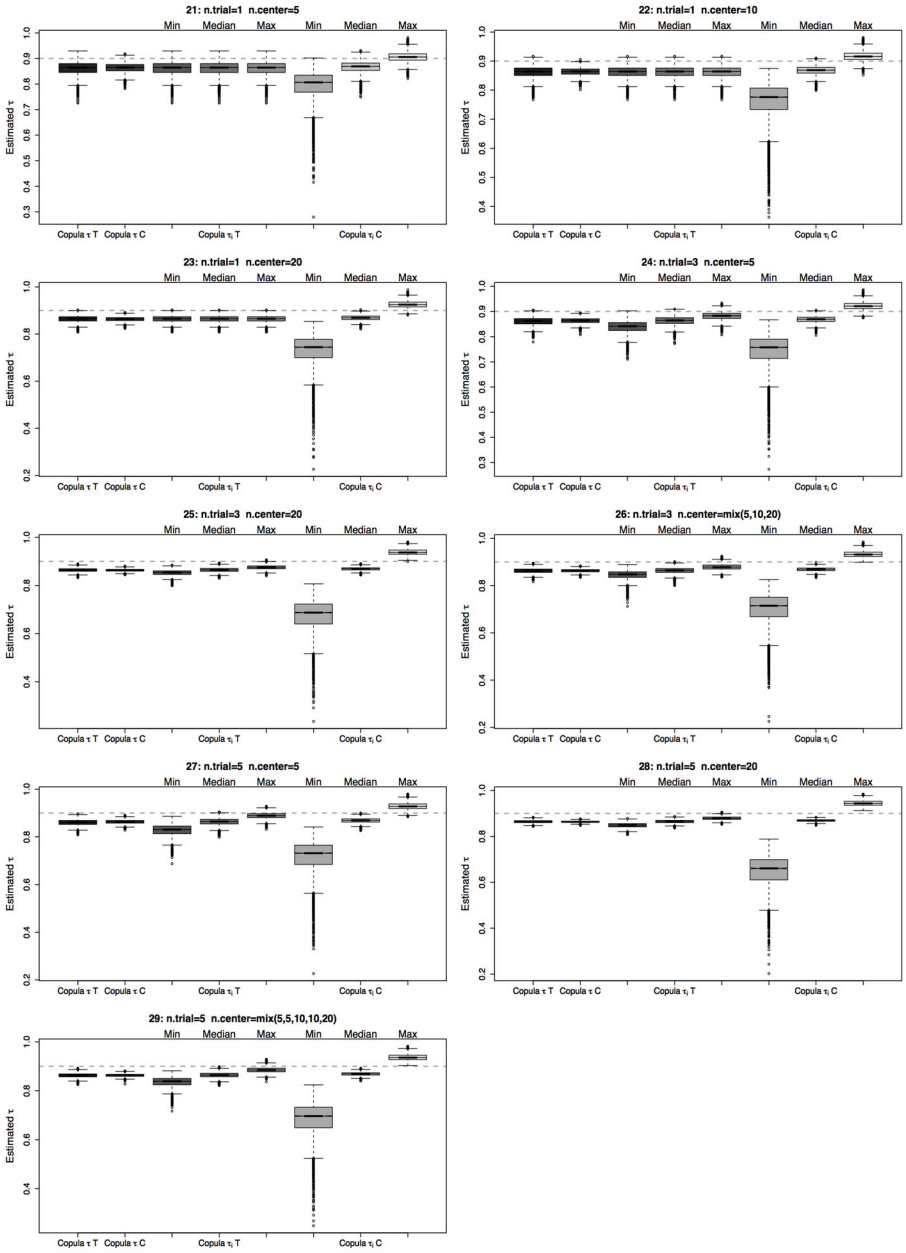

Figure 7.

Boxplots of estimated patient-level surrogacy, by units of analysis (T = trial, C = center), and assuming equal patient-level surrogacy τ or unit-specific patient-level surrogacy τ_i across units, for focused scenarios 21–29.

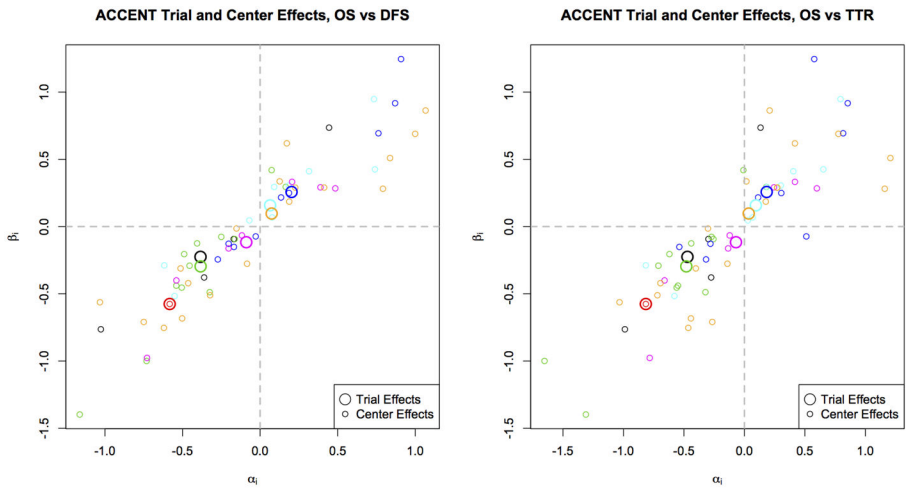

Figure 8.

Copula estimated trial-specific and center-specific treatment effect pairs for OS versus DFS (left) and TTR (right). Common colors relate centers to their parent trials. Cox estimates (not shown) are similar.

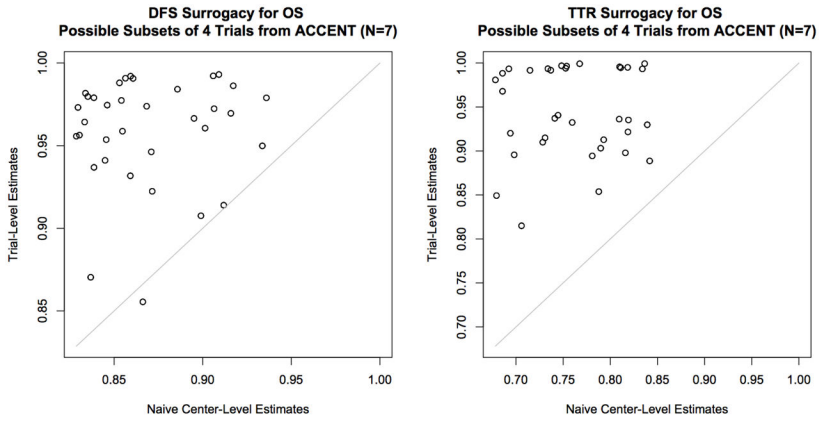

Figure 9.

Cox estimated trial-level surrogacy $\hat{R}^2_{trial}$ versus naive center-estimated surrogacy $\hat{R}^2_{naive}$ for all possible subsets of 4 ACCENT trial units among 7 total ACCENT trial units, presented for DFS surrogacy (left) and TTR surrogacy (right).

Table 1

Simulation scenarios

Scenario	N	R^2_{trial}	R^2_{center}	σ^2_T	σ^2_C	τ	% Censoring	# Centers	# Patients
High-level scenarios									
1	**15**	**0.90**	**0.90**	**0.50**	**0.05**	**0.90**	**0**	**20**	**100**
2	**5**	0.90	0.90	0.50	0.05	0.90	0	20	100
3	**30**	0.90	0.90	0.50	0.05	0.90	0	20	100
4	15	**0.50**	0.90	0.50	0.05	0.90	0	20	100
5	15	**0.20**	0.90	0.50	0.05	0.90	0	20	100
6	15	0.90	**0.50**	0.50	0.05	0.90	0	20	100
7	15	0.90	**0.20**	0.50	0.05	0.90	0	20	100
8	15	0.90	0.90	**0.05**	0.05	0.90	0	20	100
9	15	0.90	0.90	0.50	**0.50**	0.90	0	20	100
10	15	0.90	0.90	**0.05**	**0.50**	0.90	0	20	100
11	15	0.90	0.90	0.50	0.05	**0.60**	0	20	100
12	15	0.90	0.90	0.50	0.05	**0.30**	0	20	100
13	15	0.90	0.90	0.50	0.05	0.90	**30%**	20	100
14	15	0.90	0.90	0.50	0.05	0.90	**70%**	20	100
15	15	0.90	0.90	0.50	0.05	0.90	0	**1**	100
16	15	0.90	0.90	0.50	0.05	0.90	0	**10**	100
17	15	0.90	0.90	0.50	0.05	0.90	0	20	**10**
18	15	0.90	0.90	0.50	0.05	0.90	0	20	**mix(10,100)**
19	15	0.90	0.90	0.50	0.05	0.90	0	20	**500**
20	15	0.90	0.90	0.50	0.05	0.90	0	20	**1000**
Focused scenarios									
21	**1**	0.90	0.90	**0.05**	**0.50**	0.90	0	**5**	**mix(10,20,30,40,50)**
22	**1**	0.90	0.90	**0.05**	**0.50**	0.90	0	**10**	**mix(10,20,30,40,50)**
23	**1**	0.90	0.90	**0.05**	**0.50**	0.90	0	**20**	**mix(10,20,30,40,50)**
24	**3**	0.90	0.90	**0.05**	**0.50**	0.90	0	**5**	**mix(10,20,30,40,50)**
25	**3**	0.90	0.90	**0.05**	**0.50**	0.90	0	**20**	**mix(10,20,30,40,50)**

Scenario	*N*	R^2_{trial}	R^2_{center}	σ^2_T	σ^2_C	τ	% Censoring	# Centers	# Patients
26	**3**	0.90	0.90	**0.05**	**0.50**	0.90	0	**mix(5,10,20)**	**mix(10,20,30,40,50)**
27	**5**	0.90	0.90	**0.05**	**0.50**	0.90	0	**5**	**mix(10,20,30,40,50)**
28	**5**	0.90	0.90	**0.05**	**0.50**	0.90	0	**20**	**mix(10,20,30,40,50)**
29	**5**	0.90	0.90	**0.05**	**0.50**	0.90	0	**mix(5,10,20)**	**mix(10,20,30,40,50)**

Table 2

Bias and MSE of trial-level surrogacy estimates using trials ($\hat{R}^2_{trial}$) or centers ($\hat{R}^2_{naive}$) as the unit of analysis, presented according to estimation approach (Cox versus copula) and trial-level surrogacy comparator ("true" versus generated R^2_{trial}).

		$\hat{R}^2_{trial}$ **vs True** R^2_{trial}		$\hat{R}^2_{naive}$ **vs True** R^2_{trial}		$\hat{R}^2_{trial}$ **vs Gen** R^2_{trial}		$\hat{R}^2_{naive}$ **vs Gen** R^2_{trial}	
Scenario		**Cox**	**Copula**	**Cox**	**Copula**	**Cox**	**Copula**	**Cox**	**Copula**
					High-Level Scenarios				
1	Bias	−0.0077	−0.0077	−0.0185	−0.0171	−0.0008	−0.0007	−0.0116	−0.0102
	MSE	0.0037	0.0037	0.0027	0.0027	0.0001	0.0001	0.0004	0.0004
2	Bias	−0.0261	−0.0262	−0.0321	−0.0308	−0.0007	−0.0007	−0.0066	−0.0053
	MSE	0.0237	0.0237	0.0109	0.0110	0.0013	0.0013	0.0047	0.0046
3	Bias	−0.0039	−0.0038	−0.0153	−0.0139	−0.0007	−0.0007	−0.0121	−0.0107
	MSE	0.0015	0.0015	0.0013	0.0013	0.0000	0.0000	0.0002	0.0002
4	Bias	0.0035	0.0035	0.0498	0.0503	0.0023	0.0023	0.0486	0.0491
	MSE	0.0333	0.0334	0.0262	0.0264	0.0006	0.0005	0.0040	0.0039
5	Bias	0.0393	0.0393	0.0994	0.0991	0.0031	0.0031	0.0631	0.0628
	MSE	0.0320	0.0320	0.0382	0.0382	0.0005	0.0005	0.0054	0.0053
6	Bias	−0.0101	−0.0101	−0.0622	−0.0610	−0.0032	−0.0031	−0.0553	−0.0540
	MSE	0.0039	0.0039	0.0073	0.0072	0.0002	0.0002	0.0034	0.0033
7	Bias	−0.0127	−0.0127	−0.1078	−0.1067	−0.0058	−0.0057	−0.1008	−0.0997
	MSE	0.0042	0.0042	0.0165	0.0164	0.0003	0.0003	0.0110	0.0108
8	Bias	−0.0123	−0.0128	−0.0414	−0.0418	−0.0053	−0.0058	−0.0345	−0.0349
	MSE	0.0041	0.0041	0.0023	0.0024	0.0010	0.0010	0.0032	0.0032
9	Bias	−0.0081	−0.0083	−0.0110	−0.0091	−0.0010	−0.0012	−0.0039	−0.0020
	MSE	0.0038	0.0038	0.0008	0.0008	0.0006	0.0006	0.0016	0.0016
10	Bias	−0.0111	−0.0141	−0.0127	−0.0114	−0.0038	−0.0068	−0.0054	−0.0041
	MSE	0.0039	0.0041	0.0003	0.0003	0.0032	0.0035	0.0034	0.0034
11	Bias	−0.0104	−0.0105	−0.0689	−0.0682	−0.0035	−0.0035	−0.0620	−0.0612
	MSE	0.0040	0.0040	0.0084	0.0083	0.0002	0.0002	0.0043	0.0042
12	Bias	−0.0128	−0.0127	−0.1149	−0.1113	−0.0060	−0.0059	−0.1081	−0.1045
	MSE	0.0041	0.0041	0.0181	0.0173	0.0003	0.0003	0.0125	0.0118

Scenario		$\hat{R}^2_{trial}$ **vs True** R^2_{trial}		$\hat{R}^2_{naive}$ **vs True** R^2_{trial}		$\hat{R}^2_{trial}$ **vs Gen** R^2_{trial}		$\hat{R}^2_{naive}$ **vs Gen** R^2_{trial}	
		Cox	**Copula**	**Cox**	**Copula**	**Cox**	**Copula**	**Cox**	**Copula**
13	Bias	−0.0053	−0.0052	−0.0151	−0.0133	−0.0005	−0.0004	−0.0103	−0.0084
	MSE	0.0035	0.0035	0.0023	0.0023	0.0001	0.0001	0.0005	0.0004
14	Bias	−0.0093	−0.0093	−0.0143	−0.0125	−0.0015	−0.0015	−0.0065	−0.0047
	MSE	0.0038	0.0038	0.0020	0.0019	0.0002	0.0002	0.0007	0.0006
15	Bias	−0.0188	−0.0180	−0.0188	−0.0180	−0.0118	−0.0110	−0.0118	−0.0110
	MSE	0.0045	0.0045	0.0045	0.0045	0.0021	0.0020	0.0021	0.0020
16	Bias	−0.0074	−0.0074	−0.0177	−0.0164	−0.0015	−0.0015	−0.0118	−0.0104
	MSE	0.0037	0.0037	0.0028	0.0028	0.0002	0.0002	0.0005	0.0004
17	Bias	−0.0138	−0.0138	−0.2828	−0.0164	−0.0067	−0.0067	−0.2757	−0.0093
	MSE	0.0042	0.0042	0.1001	0.0011	0.0007	0.0007	0.0955	0.0017
18	Bias	−0.0073	−0.0072	−0.2732	−0.0160	−0.0014	−0.0013	−0.2673	−0.0101
	MSE	0.0037	0.0037	0.1111	0.0014	0.0002	0.0002	0.1075	0.0013
19	Bias	−0.0072	−0.0072	−0.0086	−0.0084	−0.0003	−0.0003	−0.0017	−0.0014
	MSE	0.0037	0.0037	0.0027	0.0027	0.0001	0.0001	0.0001	0.0001
20	Bias	−0.0078	−0.0078	−0.0077	−0.0076	−0.0000	−0.0000	0.0001	0.0002
	MSE	0.0038	0.0038	0.0028	0.0028	0.0001	0.0001	0.0001	0.0001
				Focused Scenarios					
21	Bias	–	–	−0.0728	−0.0530	–	–	–	–
	MSE	–	–	0.0511	0.0349	–	–	–	–
22	Bias	–	–	−0.0710	−0.0396	–	–	–	–
	MSE	–	–	0.0295	0.0129	–	–	–	–
23	Bias	–	–	−0.0662	−0.0330	–	–	–	–
	MSE	–	–	0.0169	0.0057	–	–	–	–
24	Bias	−0.0405	−0.0429	−0.0666	−0.0346	−0.0105	−0.0129	−0.0367	−0.0047
	MSE	0.0509	0.0519	0.0208	0.0077	0.0886	0.0893	0.0618	0.0508
25	Bias	−0.0372	−0.0419	−0.0628	−0.0300	−0.0072	−0.0119	−0.0329	−0.0000
	MSE	0.0498	0.0524	0.0080	0.0023	0.0760	0.0776	0.0491	0.0453
26	Bias	−0.0393	−0.0442	−0.0634	−0.0304	−0.0112	−0.0162	−0.0353	−0.0024
	MSE	0.0557	0.0550	0.0107	0.0032	0.0902	0.0869	0.0524	0.0468

		$\hat{R}^2_{trial}$ **vs True** R^2_{trial}		$\hat{R}^2_{naive}$ **vs True** R^2_{trial}		$\hat{R}^2_{trial}$ **vs Gen** R^2_{trial}		$\hat{R}^2_{naive}$ **vs Gen** R^2_{trial}	
Scenario		**Cox**	**Copula**	**Cox**	**Copula**	**Cox**	**Copula**	**Cox**	**Copula**
27	Bias	−0.0443	−0.0488	−0.0649	−0.0324	−0.0186	−0.0232	−0.0393	−0.0068
	MSE	0.0295	0.0311	0.0138	0.0047	0.0447	0.0459	0.0323	0.0251
28	Bias	−0.0376	−0.0430	−0.0628	−0.0291	−0.0138	−0.0192	−0.0391	−0.0054
	MSE	0.0273	0.0285	0.0063	0.0016	0.0316	0.0324	0.0245	0.0211
29	Bias	−0.0500	−0.0497	−0.0638	−0.0304	−0.0213	−0.0211	−0.0352	−0.0017
	MSE	0.0380	0.0322	0.0088	0.0026	0.0505	0.0442	0.0279	0.0239

Table 3

Bias and MSE of patient-level surrogacy estimates assuming equal τ across trials or trial-specific τ_i, $i \in 1, \ldots, N$, and using trials or centers as the unit of analysis. Where trial-specific estimates are assumed, bias and MSE are computed by further averaging across the units of analysis (trials or centers).

		Trial Level		Center Level	
Scenario		**Copula τ**	**Copula τ_i**	**Copula τ**	**Copula τ_i**
		High-level scenarios			
1	Bias	−0.0066	−0.0065	−0.0167	−0.0163
	MSE	0.0000	0.0001	0.0003	0.0006
2	Bias	−0.0065	−0.0065	−0.0167	−0.0163
	MSE	0.0000	0.0001	0.0003	0.0006
3	Bias	−0.0066	−0.0065	−0.0167	−0.0163
	MSE	0.0000	0.0001	0.0003	0.0006
4	Bias	−0.0066	−0.0065	−0.0167	−0.0163
	MSE	0.0000	0.0001	0.0003	0.0006
5	Bias	−0.0066	−0.0065	−0.0167	−0.0163
	MSE	0.0000	0.0001	0.0003	0.0006
6	Bias	−0.0286	−0.0285	−0.0167	−0.0163
	MSE	0.0008	0.0009	0.0003	0.0006
7	Bias	−0.0463	−0.0462	−0.0167	−0.0163
	MSE	0.0022	0.0023	0.0003	0.0006
8	Bias	−0.0066	−0.0065	−0.0167	−0.0163
	MSE	0.0000	0.0001	0.0003	0.0006
9	Bias	−0.0381	−0.0343	−0.0167	−0.0163
	MSE	0.0033	0.0013	0.0003	0.0006
10	Bias	−0.0378	−0.0343	−0.0167	−0.0163
	MSE	0.0031	0.0013	0.0003	0.0006
11	Bias	0.0030	0.0031	−0.0100	−0.0088
	MSE	0.0000	0.0001	0.0001	0.0016
12	Bias	0.0065	0.0066	−0.0047	−0.0062
	MSE	0.0001	0.0002	0.0000	0.0038
13	Bias	−0.0075	−0.0075	−0.0197	−0.0189

Scenario		Trial Level		Center Level	
		Copula τ	Copula τ_i	Copula τ	Copula τ_i
	MSE	0.0001	0.0001	0.0004	0.0008
14	Bias	−0.0095	−0.0093	−0.0317	−0.0269
	MSE	0.0001	0.0001	0.0012	0.0017
15	Bias	−0.0166	−0.0163	−0.0166	−0.0163
	MSE	0.0003	0.0006	0.0003	0.0006
16	Bias	−0.0073	−0.0073	−0.0167	−0.0163
	MSE	0.0001	0.0001	0.0003	0.0006
17	Bias	−0.0143	−0.0142	−0.0719	−0.0652
	MSE	0.0002	0.0003	0.0052	0.0106
18	Bias	−0.0072	−0.0072	−0.0217	−0.0407
	MSE	0.0001	0.0001	0.0005	0.0056
19	Bias	−0.0057	−0.0057	−0.0042	−0.0041
	MSE	0.0000	0.0000	0.0000	0.0001
20	Bias	−0.0056	−0.0056	−0.0022	−0.0022
	MSE	0.0000	0.0000	0.0000	0.0000
		Focused scenarios			
21	Bias	−0.0390	−0.0390	−0.0364	−0.0408
	MSE	0.0022	0.0022	0.0017	0.0043
22	Bias	−0.0375	−0.0375	−0.0366	−0.0408
	MSE	0.0018	0.0018	0.0015	0.0043
23	Bias	−0.0360	−0.0360	−0.0367	−0.0408
	MSE	0.0015	0.0015	0.0014	0.0042
24	Bias	−0.0394	−0.0390	−0.0367	−0.0408
	MSE	0.0018	0.0022	0.0015	0.0043
25	Bias	−0.0360	−0.0359	−0.0369	−0.0409
	MSE	0.0014	0.0015	0.0014	0.0043
26	Bias	−0.0373	−0.0377	−0.0369	−0.0410
	MSE	0.0015	0.0018	0.0014	0.0043
27	Bias	−0.0397	−0.0392	−0.0369	−0.0410
	MSE	0.0017	0.0023	0.0014	0.0043

Scenario		Trial Level		Center Level	
		Copula τ	Copula τ_t	Copula τ	Copula τ_t
28	Bias	−0.0362	−0.0361	−0.0369	−0.0409
	MSE	0.0013	0.0015	0.0014	0.0043
29	Bias	−0.0373	−0.0377	−0.0369	−0.0410
	MSE	0.0015	0.0019	0.0014	0.0043

Table 4

Estimates of trial-level surrogacy using trials (R^2_{trial}) versus centers (R^2_{naive}) as units of analysis, under marginal Cox and joint Clayton copula models weighted by trial size, with associated standard errors (SE). Patient-level surrogacy τ estimated from Clayton copula models.

	R^2_{trial}		R^2_{naive}	
Estimate	**Marginal Cox**	**Joint Copula**	**Marginal Cox**	**Joint Copula**
DFS Surrogacy				
Estimate	0.965	0.959	0.866	0.862
SE	0.164	0.178	0.089	0.090
$\hat{\tau}$	–	0.669	–	0.670
TTR Surrogacy				
Estimate	0.951	0.943	0.766	0.768
SE	0.194	0.207	0.111	0.111
$\hat{\tau}$	–	0.638	–	0.630