University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2013 *Paper* 99

Surrogacy Assessment Using Principal Stratification When Surrogate and Outcome Measures are Multivariate Normal

Anna Conlon* Jeremy M.G. Taylor†
Michael R. Elliott‡

*University of Michigan - Ann Arbor, achern@umich.edu
†University of Michigan - Ann Arbor, jmgt@umich.edu
‡University of Michigan - Ann Arbor, mrelliot@umich.edu

Surrogacy Assessment Using Principal Stratification When Surrogate and Outcome Measures are Multivariate Normal

A.S.C. Conlon[1], J.M.G. Taylor[1] and M.R. Elliott[1,2]

August 21 2012

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109, U.S.A., [2]Survey Methodology Program, Institute for Social Research, Ann Arbor, Michigan, 48109, U.S.A. *email:* achern@umich.edu

Abstract

In clinical trials, a surrogate outcome variable (S) can be measured before the outcome of interest (T) and may provide early information regarding the treatment (Z) effect on T. Most previous methods for surrogate validation rely on models for the conditional distribution of T given Z and S. However, S is a post-randomization variable, and unobserved, simultaneous predictors of S and T may exist. When such confounders exist, these methods do not have a causal interpretation. Using the principal surrogacy framework introduced by Frangakis and Rubin (2002), we propose a Bayesian estimation strategy for surrogate validation when the joint distribution of potential surrogate and outcome measures is multivariate normal. We model the joint conditional distribution of the potential outcomes of T, given the potential outcomes of S and propose surrogacy validation measures from this model. By conditioning on principal strata of S, the resulting estimates are causal. As the model is not fully identifiable from the data, we propose some reasonable prior distributions and assumptions that can be placed on weakly identified parameters to aid in estimation. We explore the relationship between our surrogacy measures and the traditional surrogacy measures proposed by Prentice (1989). The method is applied to data from a macular degeneration study and data from an ovarian cancer study, both previously analyzed by Buyse, *et al.* (2000).

Keywords: Bayesian estimation; Principal stratification; Surrogate endpoints.

1 Introduction

A surrogate endpoint (S) is an intermediate outcome variable occurring in between the treatment (Z) and the outcome of interest (T). The surrogate is usually known to be involved in the mechanism of the disease process and can be measured at an earlier time than the desired outcome. Therefore, there is considerable interest in the use of surrogate markers in clinical trials, as they offer the potential to run trials more cheaply and quickly by extracting information regarding the treatment effect on T through the earlier measured S. Examples of established surrogate markers include blood pressure under anti-hypertensive drug treatment as a surrogate for cardiovascular disease (Weir and Walley, 2006), and three year disease free survival as a surrogate for five year overall survival in colorectal cancer (Sargent et al., 2007). We examine two data examples in the application of our method. The first concerns patients with age-related macular degeneration and considers the use of change in visual acuity at 6 months after starting treatment as a surrogate marker for change in visual acuity at 1 year. The second concerns ovarian cancer and assesses progression free survival as a surrogate for overall survival.

Before a surrogate can be used in practice, it must be shown to be a valid surrogate for the outcome of interest. In a landmark paper, Prentice (1989) proposed a formal definition of surrogacy along with a validation strategy. Prentice's criteria require that S and T be correlated and the treatment effect on T be fully captured by S. Other methods for surrogacy evaluation have since been proposed, including the proportion of treatment effect explained by S (Freedman, Graubard, and Schatzkin, 1992), and individual-level and trial-level surrogacy association measures in meta-analyses (Buyse, et al., 2000).

Some assessments of surrogacy rely on estimating treatment effects by adjusting for a variable that is measured after randomization. However, there may be unmea-

sured confounders in the pathway between the surrogate and final outcome resulting in estimates that will not have a causal interpretation (Rosenbaum, 1984). Therefore, Frangakis and Rubin (2002) (henceforth FR) introduced a definition of a surrogate endpoint, called a "principal surrogate", based on a principal stratification approach. In this framework, each subject has two potential outcomes corresponding to each treatment, denoted $S(Z)$ and $T(Z)$, for $Z = 0, 1$. The principal surrogacy approach looks at the distribution of the potential outcomes of T conditional on principal strata based on the joint distribution of $S(0)$ and $S(1)$. The principal strata are unaffected by treatment, and are thus pre-randomization variables. Treatment effect estimates that condition on these principal strata are therefore causal estimates when treatments are randomly assigned.

The rationale for considering whether the principal stratification approach is appropriate for assessing surrogacy has been discussed in the literature, with some support provided in the discussion by VanderWeele (2011) and by Zigler and Belin (2011). In this approach, the value of S as a surrogate for T is determined by the extent to which the causal effect of treatment on S can reliably predict the causal effect of treatment on T. The rationale for considering principal surrogacy or more generally considering the joint distribution of $S(0), S(1), T(0), T(1)$ is most easily explained in the case where S and T are binary. In this case, the joint distribution of $S(0), S(1), T(0), T(1)$ amounts to a partition of the population into cells with a probability attached to each cell. These probabilities completely characterize the population and from them an assessment of surrogacy can be made. For example, one can consider the fraction of the population for which $T(0)$ is not equal to $T(1)$ amongst those who have $S(0)$ not equal to $S(1)$. Then additionally, this fraction might be contrasted with the fraction of the population for which $T(0)$ is not equal to $T(1)$ amongst those who have $S(0)$ equal to $S(1)$. As we will describe below, other

summary measures that can be obtained from the joint distribution might also be considered. When S and T are continuous, the joint distribution of $S(0), S(1), T(0), T(1)$ again completely characterizes the population, from which summary measures for assessing surrogacy, such as the distribution of $T(1) - T(0)$ given $S(1) - S(0)$, can be obtained. If one accepts that the joint distribution completely characterizes the population, then the challenges are determining what useful summary measures to extract from this distribution, and the estimation of this distribution.

We note that the principal stratification approach to assessing surrogacy uses a causal framework, but the causal framework it uses differs from the framework presented by Pearl (1995) and discussed in Joffe and Greene (2009). In the principal stratification framework, there are only two causal effects, one on S and one on T and we are interested in the association between these two. The other causal framework, while it may also be interesting to consider, does require additional consideration of the effect of S on T, requiring hypothetical manipulations of S. This alternative causal framework is more mechanistic and allows notions of direct and indirect effects of Z on T. We will not pursue it in this paper.

Existing literature on methods for surrogacy assessment using the principal stratification approach has examined settings in which both S and T are binary (Li *et al.* 2010), or in which S is continuous with binary T (Gilbert and Hudgens, 2008; Zigler and Belin, 2011). For a binary S and T, Li, *et al.* (2010) developed an estimation method for the causal quantities associated with the cross classification of the potential outcomes using a log-linear model and Bayesian estimation procedure. Gilbert and Hudgens (2008) (henceforth GH) used the framework of FR to develop an estimand, termed the causal effect predictiveness (CEP) surface for evaluating surrogacy when S is continuous or categorical and T is binary. Work in the PS framework when both S and T are continuous has primarily been discussed in the application

to partial compliance (Bartolucci and Grilli, 2011; Schwartz, *et al.*, 2011). In this context, the joint distribution of the potential outcomes of the intermediate variable, in this case degree of compliance, is modeled either parametrically or semiparametrically with principal causal effects (PCEs) measured by comparisons of the potential outcomes of T conditional on S, where the conditional distributions for $T(0)$ and $T(1)$ are modeled separately.

Here, we consider the entire joint distribution of $(S_i(0), S_i(1), T_i(0), T_i(1))$ and propose estimands to evaluate principal surrogacy when both S and T are continuous and the joint distribution of the potential outcomes is multivariate normal. Once parameter estimates for this distribution are obtained, various causal quantities that may aid in the assessment of S as a surrogate marker for T may be examined. Specific quantities of interest include $E[T(1)-T(0)|S(1), S(0)]$, $P(T(1)-T(0) > 0|S(1), S(0))$, and the correlation between $T(1) - T(0)$ and $S(1) - S(0)$. The use of $cor(T(1) - T(0), S(1)-S(0))$ has been discussed by Wang, *et al.* (2012), who specifically contrast it with the observable correlation between S and T, given the treatment group.

Because some parameters of the joint distribution are not fully identifiable from the data, we use a Bayesian estimation procedure with plausible prior distributions and some reasonable constraints on model parameters to reduce the non-identifiability problem of modeling counterfactual observations and to aid in estimation of the quantities of interest. In order to facilitate the consideration of reasonable constraints we found it convenient to decompose the covariance matrix, Σ of $(S_i(0), S_i(1), T_i(0), T_i(1))$ as $\Sigma = QRQ$ (Barnard et al., 2000), and place constraints on the correlations R, rather than on the covariance terms in Σ. We also explore the relationship between some of the proposed surrogacy assessment quantities and those based on the well known Prentice criteria. In Section 2, we describe the model and possible constraints that could be made to facilitate estimation. In Section 3, we introduce surrogacy mea-

sures based on the potential outcomes framework. Section 4 describes the Bayesian estimation procedure that we use and Section 5 provides simulation results from this procedure. In Section 6 we apply these methods to the macular degeneration data and ovarian cancer data. Section 7 concludes with a discussion.

2 Potential Outcomes Model

For a randomized trial with treatment assignment Z ($Z = 1$ or 0), continuous surrogate marker S and continuous true endpoint T, each subject i, $i = 1, ..., n$, has two potential outcomes for each of S_i and T_i, denoted by $S_i(Z_i)$ and $T_i(Z_i)$. Only one outcome, corresponding to the received treatment for subject i in each of the pairs $(S_i(0), S_i(1))$ and $(T_i(0), T_i(1))$ can be observed. The joint distribution of $(S_i(0), S_i(1), T_i(0), T_i(1))$ describes the causal associations between Z, S and T. In the continuous setting where $(S_i(0), S_i(1), T_i(0), T_i(1))$ is multivariate normal with mean μ and covariance matrix Σ, we have the following joint distribution:

$$\begin{pmatrix} S_i(0) \\ S_i(1) \\ T_i(0) \\ T_i(1) \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{S_0} \\ \mu_{S_1} \\ \mu_{T_0} \\ \mu_{T_1} \end{pmatrix}, \begin{pmatrix} \sigma^2_{S_0} & \rho_s \sigma_{S_0} \sigma_{S_1} & \rho_{00} \sigma_{S_0} \sigma_{T_0} & \rho_{01} \sigma_{S_0} \sigma_{T_1} \\ & \sigma^2_{S_1} & \rho_{10} \sigma_{S_1} \sigma_{T_0} & \rho_{11} \sigma_{S_1} \sigma_{T_1} \\ & & \sigma^2_{T_0} & \rho_t \sigma_{T_1} \sigma_{T_0} \\ & & & \sigma^2_{T_1} \end{pmatrix} \right)$$

The mean μ and the variances corresponding to the diagonal elements of Σ, along with the correlations between $(S_i(0), T_i(0))$ and $(S_i(1), T_i(1))$ corresponding to ρ_{00} and ρ_{11}, are fully identifiable from the data. Because only one of the counterfactual pairs of outcomes is observed for each subject, ρ_s, ρ_t, ρ_{01}, and ρ_{10} are not identifiable. However, the identifiable correlation parameters together with the requirement that Σ be positive definite places boundary constraints on these non-identified parameters, which, along with other plausible assumptions that we can make, aids in their identifiability. These parameters are therefore considered to be partially identified as opposed to completely unidentified.

We make the standard assumptions of ignorable treatment assignments (Rubin, 1978) and the stable unit treatment value assumption (SUTVA). Ignorable treatment assignment implies that Z is independent of $(S(0), S(1), T(0), T(1))$ and holds for blinded, randomized trials. SUTVA implies that the potential outcomes $(S_i(0), S_i(1), T_i(0), T_i(1))$ are independent of the treatment assignments of other subjects. This allows us to write the potential outcomes for subject i as a function of Z_i rather than of the entire vector of subject treatment assignments.

Other context specific constraints can be added, such as all ρ's ≥ 0, a plausible assumption for most variables S that would be under consideration as a potential surrogate for T, and especially when the identifiable Pearson correlation coefficients, $\hat{\rho}_{00}$ and $\hat{\rho}_{11}$, are positive. Other plausible assumptions are $\rho_{01} < min(\rho_{00}, \rho_{11}, \rho_s, \rho_t)$, and $\rho_{10} < min(\rho_{00}, \rho_{11}, \rho_s, \rho_t)$, indicating a belief that the correlation between the surrogate response and final outcome response in opposite treatment arms is less than the correlation between the surrogate response and final outcome response within the same treatment arm, or the correlation between the surrogate responses or final treatment responses across treatment arms.

3 Assessing Surrogacy Using Potential Outcomes Framework

3.1 Definitions of Surrogacy

Because S is a post-randomization variable, unobserved simultaneous predictors of both S and T may exist. In this case, methods of surrogacy assessment that require conditioning on S do not result in causal estimates (Rosenbaum, 1984). When baseline covariates account for all common causes of S and T, surrogacy measures that condition on S will be causal. However, the assumption of no unmeasured confounders of S and T is untestable, potentially leading to noncausal estimates (Gilbert,

et al., 2009). Therefore, FR proposed a definition of principal surrogacy (PS), which uses a principal stratification approach to assess the validity of a surrogate marker. This framework focuses on the distribution of $p(T(0), T(1)|S(0), S(1))$. Since $S(1)$ and $S(0)$ are unaffected by treatment assignment, they can be treated as baseline covariates. Quantities estimated from this distribution will therefore always have a causal interpretation. FR proposed two measures of surrogacy, the "associative effect" and the "dissociative effect". A measure of the associative effect is given by $E(T_i(1) - T_i(0)|S_i(1) = S_i(0))$ and a measure of the dissociative effect is given by $E(T_i(1) - T_i(0)|S_i(1) \neq S_i(0))$.

For the multivariate normal distribution, the distribution of $(T(1) - T(0)|S(1) - S(0) = s)$ is normal with mean

$$(\mu_{T_1} - \mu_{T_0}) + \left(\frac{\rho_{11}\sigma_{S_1}\sigma_{T_1} - \rho_{10}\sigma_{S_1}\sigma_{T_0} - \rho_{01}\sigma_{S_0}\sigma_{T_1} + \rho_{00}\sigma_{S_0}\sigma_{T_0}}{\sigma^2_{S_0} + \sigma^2_{S_1} - 2\rho_s\sigma_{S_0}\sigma_{S_1}}\right)(s - (\mu_{S_1} - \mu_{S_0}))$$

and variance

$$\sigma^2_{T_0} + \sigma^2_{T_1} - 2\rho_t\sigma_{T_0}\sigma_{T_1} - \frac{\left(\rho_{11}\sigma_{S_1}\sigma_{T_1} - \rho_{10}\sigma_{S_1}\sigma_{T_0} - \rho_{01}\sigma_{S_0}\sigma_{T_1} + \rho_{00}\sigma_{S_0}\sigma_{T_0}\right)^2}{\sigma^2_{S_0} + \sigma^2_{S_1} - 2\rho_s\sigma_{S_0}\sigma_{S_1}}.$$

The mean can be written as $E[T_i(1) - T_i(0)|S_i(1) - S_i(0) = s] = \gamma_0 + \gamma_1 s$, where

$$\gamma_0 = (\mu_{T_1} - \mu_{T_0}) - \left(\frac{\rho_{11}\sigma_{S_1}\sigma_{T_1} - \rho_{10}\sigma_{S_1}\sigma_{T_0} - \rho_{01}\sigma_{S_0}\sigma_{T_1} + \rho_{00}\sigma_{S_0}\sigma_{T_0}}{\sigma^2_{S_0} + \sigma^2_{S_1} - 2\rho_s\sigma_{S_0}\sigma_{S_1}}\right)(\mu_{S_1} - \mu_{S_0})$$

$$\gamma_1 = \left(\frac{\rho_{11}\sigma_{S_1}\sigma_{T_1} - \rho_{10}\sigma_{S_1}\sigma_{T_0} - \rho_{01}\sigma_{S_0}\sigma_{T_1} + \rho_{00}\sigma_{S_0}\sigma_{T_0}}{\sigma^2_{S_0} + \sigma^2_{S_1} - 2\rho_s\sigma_{S_0}\sigma_{S_1}}\right)$$

The value of γ_0 is then a measure of the "dissociative effect". Values of γ_0 near zero indicate that the causal effect of treatment on the final outcome is near zero when the causal effect of treatment on the surrogate is near zero, a characteristic that a good principal surrogate should possess. When γ_0 is near $(\mu_{T_1} - \mu_{T_0})$, there can be a causal effect of the treatment on the final outcome even if there is no causal effect of the treatment on the surrogate, implying that the treatment affects the outcome through

pathways that do not involve the surrogate. We note, however, that a dissociative effect of zero does not exclude the possibility of these pathways. The value of $\gamma_0 + \gamma_1 s$ is a measure of the "associative effect", providing information on how the causal treatment effect on the outcome changes as the causal effect of the treatment on the surrogate changes. A good principal surrogate should result in a large associative effect, indicating that as the treatment effect on the surrogate increases, the treatment effect on the final outcome increases as well. This does not imply an indirect effect of treatment on the outcome or an effect of S on T, but rather the extent to which the effect of Z on S is associated with an effect of Z on T (VanderWeele, 2011).

GH suggest a refined definition of a principal surrogate endpoint. In their setting with binary T they define two properties, "average causal necessity" (ACN) and "average causal sufficiency" (ACS). ACN is satisfied if $risk_{(1)}(s_1, s_0) = risk_{(0)}(s_1, s_0)$ for all $s_1 = s_0$, where $risk_{(z)}(s_1, s_0) = p(T(Z) = 1 | S(1) = s_1, S(0) = s_0)$. ACS is satisfied if there exists some constant $C \geq 0$ such that $risk_{(1)}(s_1, s_0) \neq risk_{(0)}(s_1, s_0)$ for all $|s_1 - s_0| > C$. GH suggest that a valid surrogate marker should satisfy both ACS and ACN. In our setting of continuous T, we can consider the joint conditional distribution of $(T(0), T(1))$. Specific summaries of this joint distribution which are of major interest include $E[T(1) - T(0) | S(1) - S(0) = s]$ for $s = 0$ and $|s| > C$ for some constant $C \geq 0$, $P(T(1) > T(0) | S(1), S(0))$ and the correlation between $T(1) - T(0)$ and $S(1) - S(0)$. Also of interest is the "causal effect predictiveness (CEP) surface" proposed by GH which considers the entire curve of $E[T(1) - T(0) | S(1), S(0)]$ and provides a measure of the treatment effect on T within subgroups defined by the treatment effect on the surrogate. In terms of expectations, ACN is satisfied if $E[T(1) - T(0) | S(1) - S(0) = 0] = 0$ and ACS is satisfied if $E[T(1) - T(0) | S(1) - S(0) = s] \neq 0$ for all $|s| > C$. In the above setting, this corresponds to $\gamma_0 = 0$ and $\gamma_1 \neq 0$. In terms of the entire conditional distribution of $T(1) - T(0)$ given $(S(1) - S(0))$,

ACN is satisfied if $P(T(1)-T(0)>0|S(1)-S(0)=0)=0.5$ and ACS is satisfied if $P(T(1)-T(0)>0|S(1)-S(0)>0)$ increases as $S(1)-S(0)$ increases. For multivariate normal data this conditional probability is:

$$\Phi_{10}(s)=P(T(1)-T(0)>0|S(1)-S(0)=s)=\Phi\left(\frac{\gamma_0+\gamma_1 s}{\sqrt{\sigma^2_{T_0}+\sigma^2_{T_1}-2\rho_t\sigma_{T_0}\sigma_{T_1}-\gamma_1^2(\sigma^2_{S_0}+\sigma^2_{S_1}-2\rho_s\sigma_{S_0}\sigma_{S_1})}}\right)$$

In the multivariate normal setting, these two metrics of ACN and ACS are closely related. If $\gamma_0=0$, then $\Phi_{10}=0.5$ when $S(1)-S(0)=0$ and if $\gamma_1>0$, then $\Phi_{10}>0.5$ when $S(1)-S(0)>0$. So the conclusion drawn regarding the validity of S as a surrogate will be the same under these two measures.

Another potentially useful measure to assess surrogacy is the correlation between $T(1)-T(0)$ and $S(1)-S(0)$, which we denote by ρ_{ST}. It can be shown that ρ_{ST} is given by

$$\rho_{ST}=\frac{\rho_{11}\sigma_{S_1}\sigma_{T_1}-\rho_{10}\sigma_{S_1}\sigma_{T_0}-\rho_{01}\sigma_{S_0}\sigma_{T_1}+\rho_{00}\sigma_{S_0}\sigma_{T_0}}{\sqrt{\sigma^2_{S_0}+\sigma^2_{S_1}-2\rho_s\sigma_{S_0}\sigma_{S_1}}\sqrt{\sigma^2_{T_0}+\sigma^2_{T_1}-2\rho_t\sigma_{T_0}\sigma_{T_1}}}$$

A final way that we consider summarizing the conditional distribution of $T(1)-T(0)$ given $S(1)-S(0)=s$, and hence assessing surrogacy, is through the CEP graph, which is a plot of $E[T(1)-T(0)|S(1)-S(0)=s]$ versus s, which is simply a plot of $\gamma_0+\gamma_1 s$ versus s.

3.2 Relationship Between Principal Surrogacy Measures and Prentice Surrogacy Criteria

The ACN and ACS measures corresponding to conditional expectation can be linked to the original surrogacy definition proposed by Prentice (1989). Prentice's criteria

for a valid surrogate require that

$$f(T|Z) \neq f(T)$$

$$f(S|Z) \neq f(S)$$

$$f(T|S) \neq f(T)$$

$$f(T|S, Z) = f(T|S)$$

In the multivariate normal setting with

$$E[T_i|Z_i] = \theta_0 + \theta_1 Z_i$$

$$E[S_i|Z_i] = \alpha_0 + \alpha_1 Z_i$$

$$E[T_i|S_i] = \mu_0 + \mu_1 S_i$$

$$E[T|S, Z] = \beta_0 + \beta_1 Z + \beta_2 S + \beta_3 SZ$$

the Prentice criteria are satisfied when $\theta_1 \neq 0$, $\alpha_1 \neq 0$, $\mu_1 \neq 0$, $\beta_1 = 0$, $\beta_2 \neq 0$, and $\beta_3 = 0$. Relating these to the parameters in the potential outcomes model we have

$$\theta_1 = \mu_{T_1} - \mu_{T_0}$$

$$\alpha_1 = \mu_{S_1} - \mu_{S_0}$$

$$\mu_1 = \frac{1}{2}\left(\frac{\rho_{00}\sigma_{T_0}}{\sigma_{S_0}} + \frac{\rho_{11}\sigma_{T_1}}{\sigma_{S_1}}\right)$$

$$\beta_1 = (\mu_{T_1} - \mu_{T_0}) - \left(\frac{\rho_{11}\sigma_{T_1}}{\sigma_{S_1}}\mu_{S_1} - \frac{\rho_{00}\sigma_{T_0}}{\sigma_{S_0}}\mu_{S_0}\right)$$

$$\beta_2 = \frac{\rho_{00}\sigma_{T_0}}{\sigma_{S_0}}$$

$$\beta_3 = \frac{\rho_{11}\sigma_{T_1}}{\sigma_{S_1}} - \frac{\rho_{00}\sigma_{T_0}}{\sigma_{S_0}}$$

It can be shown that when

$$\frac{\rho_{11}\sigma_{T_1}}{\sigma_{S_1}} = \frac{\rho_{00}\sigma_{T_0}}{\sigma_{S_0}} \tag{1}$$

and

$$\rho_{00}\rho_s = \frac{1}{2}\left(\rho_{10} + \rho_{01}\frac{\sigma_{S_0}\sigma_{T_1}}{\sigma_{S_1}\sigma_{T_0}}\right) \tag{2}$$

we have $\gamma_1 = \beta_2 = \mu_1$, $\gamma_0 = \beta_1$ and $\beta_3 = 0$. Therefore, under these conditions, the Prentice criteria and the principal surrogacy criteria requiring that both ACN and ACS be met (or $\gamma_0 = 0$ and $\gamma_1 \neq 0$) will reach the same conclusions regarding the validity of S as a surrogate. When the above conditions are not met, conflicting conclusions may be drawn by the Prentice criteria and principal surrogacy criteria. As we regard principal surrogacy to be the main objective in surrogacy assessment, approaching the question of surrogacy using the Prentice criteria in this case may lead to erroneous conclusions.

In any real setting we would not expect the conditions in equations 1 and 2 to be exactly satisfied. However, in many settings we can see that the Prentice criteria and principal surrogacy criteria will reach similar conclusions. Often $\sigma_{S_0} \approx \sigma_{S_1}$, $\sigma_{T_0} \approx \sigma_{T_1}$ and we might expect ρ_{00} to be similar to ρ_{11}, thus equation 1 is approximately satisfied. Similarly we may expect ρ_{01} and ρ_{10} to be similar and hence their average to be less than both ρ_{00} and ρ_s; thus departures from equality in equation 2 may not be large.

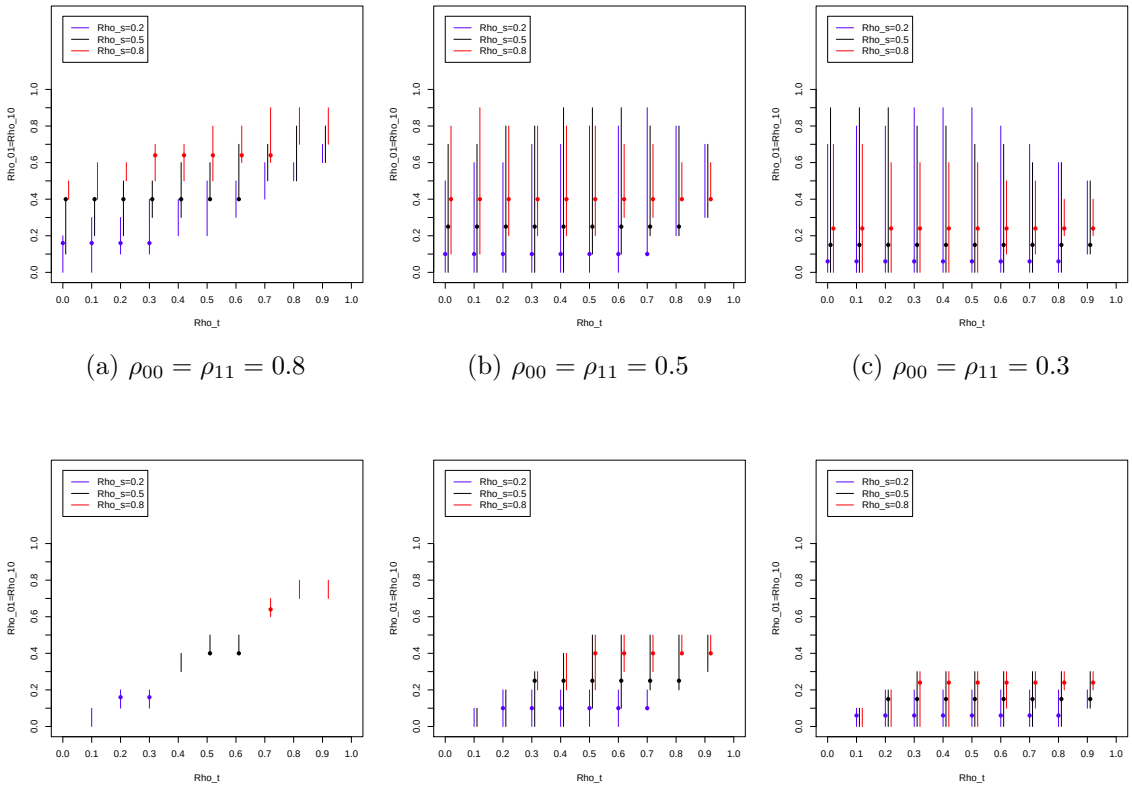
3.3 Parameter Identifiability and Restrictions

Given the identified parameters, the positive definite restriction on R, and plausible assumptions about correlation values, we can gain some insight into the possible ranges, or "identification regions" (Gustafson, 2010) for the partially identified parameters and examine scenarios within this space which lead to different surrogacy conclusions. Under the restriction that all ρ's are non-negative, and the simplifying assumptions that $\rho_{01} = \rho_{10}$, $\rho_{11} = \rho_{00}$, and $\sigma_{S_0} = \sigma_{S_1} = \sigma_{T_0} = \sigma_{T_1}$, the top half of Figure 1 displays the possible ranges for $\rho_{01} = \rho_{10}$ across different values of ρ_s and ρ_t for a given $\rho_{11} = \rho_{00}$, where ρ_{11} and ρ_{00} are the identifiable Pearson correlation

coefficients between $S_i(1)$ and $T_i(1)$, and $S_i(0)$ and $T_i(0)$, respectively. The length of the identification region for ρ_{01} and ρ_{10} is smallest when ρ_{11} and ρ_{00} are large. For all values of ρ_{11} and ρ_{00}, the length of the identification region for ρ_{01} and ρ_{10} decreases as ρ_s and ρ_t increase. The bottom half of Figure 1 provides ranges for these parameters under the additional restriction that $\rho_{01} < min(\rho_{00}, \rho_{11}, \rho_s, \rho_t)$. This restriction greatly reduces the range of possible values for the partially identified parameters, and has implicit effects on the possible ranges for γ_0 and γ_1. Under these restrictions, γ_1 must be greater than 0, implying that ACS always holds. In this scenario where ACS always holds, poor principal surrogates can be characterized by large values of γ_0, implying that the treatment can effect the outcome without effecting the surrogate. Alternatively, a poor surrogate would have a small value of γ_1, implying that there is still a positive, but weak association between causal effects on the surrogate and causal effects on the outcome. These restrictions seems reasonable, as S is typically known to somehow be associated with or a relevant aspect of the disease process, so even if it is not a valid principal surrogate from an ACN and ACS perspective, we expect there to be at least a small association of treatment effects on S with treatment effects on T. The solid points in each figure are parameter values under which the Prentice criteria and PS criteria are in agreement. In this restricted space the deviation between the Prentice criteria and the PS criteria are less than in the unrestricted space, however we see that scenarios can arise in which the Prentice criteria lead to incorrect conclusions regarding the validity of a principal surrogate.

4 Estimation Procedure

A Bayesian approach is used to estimate parameters. Unobserved potential outcomes are treated as missing data and imputed from the appropriate posterior distribution at each iteration of the Markov chain. The covariance matrix Σ is decomposed as QRQ,



(a) $\rho_{00} = \rho_{11} = 0.8$ (b) $\rho_{00} = \rho_{11} = 0.5$ (c) $\rho_{00} = \rho_{11} = 0.3$

(d) $\rho_{00} = \rho_{11} = 0.8$ (e) $\rho_{00} = \rho_{11} = 0.5$ (f) $\rho_{00} = \rho_{11} = 0.3$

Figure 1: Identification Regions of Unidentified Parameters
Plots (a), (b), and (c): under restriction ρ's ≥ 0
Plots (d), (e), and (f): under restriction ρ's ≥ 0, $\rho_{01} < min(\rho_{00}, \rho_{11}, \rho_s, \rho_t)$
Solid points: PS criteria and Prentice criteria in agreement

where Q is the diagonal matrix of standard deviations and R is the correlation matrix. Assuming a priori independence, this allows us to factor the prior distribution $p(\mu, \Sigma)$ as $p(\mu)p(R)p(Q)$ and to place non-informative priors on the fully identified parameters μ, Q, ρ_{00}, and ρ_{11}. Specifically, the prior for μ is $N_4(0, \Sigma_0)$, where $\Sigma_0 = diag(10^6)$, and the prior for each diagonal element of Q is $p(\sigma_j) \propto 1$, for $j = (S(0), S(1), T(0), T(1))$. We place marginal priors on each of the correlation parameters in R and explore the use of four different prior assumptions. For each of these there is the additional

assumption that R must be positive definite. The four priors are

(a) Jointly uniform prior such that for each of the six correlations $p(\rho) \sim Unif(-1, 1)$

(b) Jointly uniform prior such that for each of the six correlations $p(\rho) \sim Unif(0, 1)$

(c) All $\rho's \geq 0, \rho_{01} < min(\rho_{00}, \rho_{11}, \rho_s, \rho_t)$, and $\rho_{10} < min(\rho_{00}, \rho_{11}, \rho_s, \rho_t)$

(d) Beta priors such that:

- $p(\rho_{11}) \sim Unif(0, 1)$
- $p(\rho_{00}) \sim Unif(0, 1)$
- $p(\rho_{10})$ and $p(\rho_{01}) \sim Beta(3\alpha_0, 3 - 3\alpha_0)$ such that $P(\rho_{01}, \rho_{10} \leq min(\hat{\rho_{00}}, \hat{\rho_{11}})) = 0.80$
- $p(\rho_s)$ and $p(\rho_t) \sim Beta(3\alpha_1, 3 - 3\alpha_1)$ such that $P(\rho_s, \rho_t \geq max(\hat{\rho_{10}}, \hat{\rho_{01}})) = 0.80$

where $\hat{\rho}_{00}$ and $\hat{\rho}_{11}$ are the Pearson correlation coefficients estimated from the observed data. Prior assumption (a) is a non-informative prior on all of the correlations. Under scenario (b), all correlations are constrained to be positive, a plausible assumption especially when $\hat{\rho}_{00}$ and $\hat{\rho}_{11}$ are positive. In scenario (c), in addition to the positivity assumption, we restrict ρ_{01} and ρ_{10} to be smaller than the other four correlation parameters. This seems reasonable as ρ_{01} and ρ_{10} are measures of the correlation between the surrogate response and final outcome response in opposite treatment arms, which is unlikely to be larger than the correlation between the surrogate response and final outcome response within the same treatment arm, or the correlation between the surrogate responses or final treatment responses across treatment arms. Finally, prior assumption (d) places similar restrictions on the correlations as assumption (c), but is a little bit more flexible as ρ_{01} and ρ_{10} are only assumed to be smaller than the other correlations with a probability of 0.8. Appendix A provides density plots of the Beta priors when $\hat{\rho}_{00}$ and $\hat{\rho}_{11}$ are equal to 0.8, 0.5, and 0.3.

Posterior estimates of the unobserved potential outcomes, parameter values, and

the causal quantities of interest, γ_0, γ_1, $\Phi_{10}(0)$, ρ_{ST}, and the CEP curve at the points $(\mu_{S_1} - \mu_{S_0}) \pm 2SD(S(1) - S(0))$, where SD$(S(1) - S(0))$ is the standard deviation of $(S(1) - S(0))$, are obtained using the Gibbs sampler. Each component of Q and R are drawn one at a time. When drawing each element of R, the range of possible values must first be determined in order to satisfy the positive definite requirement, given that the other correlations are held fixed. The range of values corresponding to a positive definite matrix are those in the interval determined by the roots of the quadratic equation that result from solving $|R| = 0$. The specific equations solved to obtain parameter ranges are provided in Appendix B.

As the posterior distributions for the components of Q and R can not be easily sampled from, draws are made using the griddy Gibbs sampler (Ritter and Tanner, 1992). Details of the Gibbs sampler are provided in Appendix C.

5 Simulations

We conduct simulations to evaluate the performance of the above methods of surrogacy assessment. We consider the scenarios where under the true parameter values of the simulated data, surrogate validity is the same (S is valid, or S is invalid) under both the Prentice criteria and PS criteria. We also consider the two cases where, under the true parameter values of the simulated data, the surrogacy conclusions drawn using the Prentice criteria would reach a different conclusion from that drawn using the PS criteria (S valid under Prentice but not under PS, and S valid under PS but not under Prentice). In this paper we interpret the results from the perspective that principal surrogacy is the correct approach. We investigate whether the wrong conclusions would be reached if the Prentice criteria were used instead, and whether it is easier to validate a principal surrogate depending on whether or not the Prentice criteria are also satisfied. Table 1 provides details of the four simulations considered.

Table 1: Simulation Models

	(1) S is a valid principal surrogate; does not satisfy Prentice criteria	(2) S is not a valid principal surrogate; satisfies Prentice criteria	(3) S is not a valid principal surrogate; does not satisfy Prentice criteria	(4) S is a valid principal surrogate; satisfies Prentice criteria
ρ_s	0.5	0.5	0.2	0.4
ρ_{00}	0.7	0.5	0.2	0.8
ρ_{01}	0.15	0.45	0.04	0.32
ρ_{10}	0.15	0.45	0.04	0.32
ρ_{11}	0.7	0.5	0.2	0.8
ρ_t	0.18	0.5	0.3	0.4
σ^*	1	1	1	1
γ_0	0	0.8	1.1	0
γ_1	1.1	0.1	0.2	0.8
ρ_{ST}	0.86	0.1	0.21	0.8
β_1	0.8	0	1.1	0
β_2	0.7	0.5	0.2	0.8
β_3	0	0	0	0
$\rho_{00}\rho_s$	0.35	0.25	0.04	0.32
$\frac{1}{2}\left(\rho_{10}+\rho_{01}\frac{\sigma_{S_0}\sigma_{T_1}}{\sigma_{S_1}\sigma_{T_0}}\right)$	0.15	0.45	0.04	0.32

*$\sigma = \sigma_{S_0} = \sigma_{S_1} = \sigma_{T_0} = \sigma_{T_1}$

We first use these four models to explore the sensitivity of the estimation to the plausible prior restrictions on R that we might make. We simulate 200 data sets under each of the above mentioned priors for the four different surrogacy scenarios. Table 2 provides the posterior means and standard deviations of the Bayesian estimates and means of the posterior standard deviations ($P\bar{S}D$). The identified parameters are not sensitive to changes in the prior specifications while, as expected, the unidentified parameters are quite sensitive to prior assumptions. In all four scenarios, the standard deviation of the Bayesian estimates is smaller than $P\bar{S}D$ for the unidentified parameters. Table 3 provides the means and standard deviations of the Bayesian estimates and $P\bar{S}D$ for the quantities of interest from the Prentice model, and the causal quantities of interest, γ_0, γ_1, ρ_{ST}, $\Phi_{10}(0)$ and the CEP curve at $(\mu_{S_1} - \mu_{S_0}) \pm 2SD(S(1) - S(0))$. There is very little bias in estimating β_1, β_2, and β_3, while there is some bias in estimating γ_0, γ_1, ρ_{ST}, $\Phi_{10}(0)$ and the CEP points, as these are functions of unidentified parameters. The estimation performed using Beta priors appears to provide the best estimation for the unidentified parameters across

these four models. While this prior does not always perform best in terms of bias, it has on average better coverage of the parameters across the different scenarios than the other models, and therefore better power to determine the validity of S.

We investigate how well the estimation procedure is able to identify the validity of S as a surrogate marker under Beta priors. Table 4 provides an estimate of the proportion of times that S would be considered a good principal surrogate based on the proposed measures. For γ_0 and γ_1 this means that 0 is in the 95% credible interval for γ_0, and outside of the 95% credible interval for γ_1. For $\Phi_{10}(0)$ this means that 0.5 is in the 95% credible interval. For ρ_{ST}, we look at the proportion of times that its credible interval is outside of 0, and for the CEP curve we look at the proportion of times that the 95% credible intervals at the points $(\mu_{S_1}-\mu_{S_0})+2SD(S(1)-S(0))$ and $(\mu_{S_1}-\mu_{S_0})-2SD(S(1)-S(0))$ do not overlap (denoted by $CEP^U_{-2SD} < CEP^L_{+2SD}$). Table 4 also provides an estimate of the proportion of times that S would be a valid surrogate based on the Prentice criteria (0 in the 95% confidence interval for $\hat{\beta}_1$, and $\hat{\beta}_3$ and 0 outside of the 95% confidence interval for $\hat{\beta}_2$ in a regression on the observed data) for the four simulation scenarios considered. The entire CEP curve, shown in Figure 2, is also used to visually assess principal surrogacy and the expected treatment effect on T at relevant values of $S(1)-S(0)$. We explored additional models under each of the four surrogacy scenarios to gain a better understanding of how our estimation procedure performs across the parameter space. The results presented appear to be characteristic of most models that would fit into each of the four scenarios.

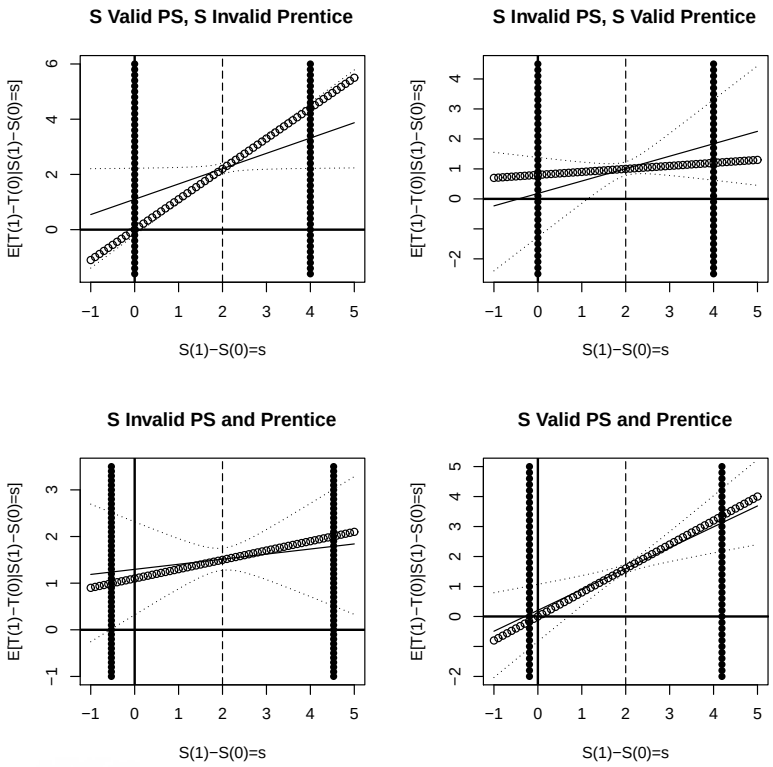
Our estimation procedure for γ_0 and γ_1 reaches the correct conclusion regarding surrogate validity when principal surrogacy is unmet, regardless of whether or not the Prentice criteria are met under the true parameters. We correctly identify S as a poor principal surrogate 99% of the time in the scenario in which S is invalid under the Prentice criteria, and 85% of the time when S is valid under the Prentice criteria.

In comparison, the Prentice criteria incorrectly determine S to be a valid surrogate 26% and 92% of the time, respectively, in these two scenarios. When S is in truth a good principal surrogate, our procedure most reliably determines surrogate validity when the Prentice criteria would also conclude that S is a good surrogate. In this scenario, we correctly identify S as a valid principal surrogate 94% of the time, while the Prentice criteria conclude S to be a good surrogate 95% of the time. When S is a good principal surrogate but the Prentice criteria show S to be invalid, our estimation procedure and the Prentice approach have a similar ability to detect surrogacy, with neither approach providing reliable surrogacy conclusions. In the scenario considered, our estimation procedure correctly identified S as a good surrogate 37% of the time, while the Prentice approach correctly identified S as a good surrogate 52% of the time.

We note that by basing surrogacy assessment on the criteria that $\gamma_0 = 0$ and $\gamma_1 \neq 0$, we do not avoid the problem that is inherent in the Prentice criteria of proving a null hypothesis, namely that certain parameters assume the value of 0. Therefore, in addition to these quantities, we can also examine the other proposed estimands to aid in validating S as a surrogate. The tests of $\rho_{ST} = 0$ and $CEP^U_{-2SD} < CEP^L_{+2SD}$ have similar power to correctly determine surrogacy. When S is a poor principal surrogate, the two quantities reject surrogacy 83% and 85% of the time, respectively, when the Prentice criteria are in truth satisfied, and 99% of the time when they are not. In the two scenarios where S is a good principal surrogate, these two quantities improve upon the γ_0, γ_1 criteria, correctly determining surrogacy a majority of the time, with greater power to detect surrogacy in the scenario where the Prentice criteria are also met. Principal surrogacy is correctly identified 57% and 55% of the time, respectively, when the Prentice criteria are not met and 94% and 93% of the time, respectively, when the Prentice criteria are also met. In contrast, while the criterion

of $\Phi_{10}(0) = 0.5$ being included in the 95% credible interval does reasonably well at determining surrogacy when the Prentice criteria and PS criteria are in agreement, it is unable to reliably distinguish good principal surrogates from poor ones with the two criteria disagree.

Figure 2: Simulation results: CEP curves



KEY: ∘ True line, − Mean CEP, · CEP 95% CI, −− $(\mu_{S_1} - \mu_{S_0})$, • $(\mu_{S_1} - \mu_{S_0}) \pm 2SD(S(1) - S(0))$

6 Application

6.1 Early Change in Visual Acuity as a Surrogate for Later Change in Visual Acuity in a Trial of Age-related Macular Degeneration

We apply our estimation method to a clinical trial for 183 patients with age-related macular degeneration. This data set was considered in a previous paper by Buyse,

et al. (2000) where a meta-analysis surrogate validation strategy was used. These data come from a multicenter trial comprised of 36 different centers. The number of patients per center ranges from 2 to 18. In this example, we have a binary treatment indicator (Z_i) equal to 0 for placebo and 1 for the treatment, interferon-α. The surrogate marker (S_i) is change in visual acuity at 6 months after starting treatment and the final endpoint (T_i) is change in visual acuity at 1 year. We first check the Prentice criteria, subtracting off the Best Linear Unbiased Predictor (BLUP) estimates from S_i and T_i to account for random center effects. We have:

$$\hat{\theta}_1 = -3.34(SE = 2.13, P = 0.12)$$

$$\hat{\alpha}_1 = -2.03(SE = 1.90, P = 0.29)$$

$$\hat{\mu} = 0.65(SE = 0.07, P < 0.0001)$$

$$\hat{\beta}_1 = -2.67(SE = 1.94, P = 0.17), \hat{\beta}_2 = 0.69(SE = 0.09, P < 0.0001),$$

$$\hat{\beta}_3 = -0.11(SE = 0.14, P = 0.44)$$

As θ_1 and α_1 are not statistically significant, the Prentice criteria are not met. Using our Bayesian estimation procedure with Beta priors for the correlation parameters, we get the following posterior estimates for the principal surrogacy parameters of interest:

$$\gamma_0 = -1.62(-5.49, 2.16)$$

$$\gamma_1 = 0.60(-0.24, 1.43)$$

As γ_1 contains 0 within its 95% credible interval, we conclude that change in visual acuity at 6 months is not a valid principal surrogate for change in visual acuity at 12 months. The average Pearson correlation, ρ_{ST}^{-} of $T_i(1) - T_i(0)$ and $S_i(1) - S_i(0)$ was 0.48 (-0.16, 0.92), also indicative of a poor principal surrogate. This is in agreement

with the conclusion reached by Buyse, *et al.* (2000) in their analysis. Figure 3(a) shows a plot of the (CEP) curve, where $CEP = E[T(1) - T(0)|S(1) - S(0) = s]$ with a 95% credible interval for each value of s. The middle dashed line indicates the posterior mean of $\mu_{S_1} - \mu_{S_0}$, and the outer two dashed lines show the posterior means of $\mu_{S_1} - \mu_{S_0} \pm 2SD_{S(1)-S(0)}$, where $SD_{S(1)-S(0)}$ is the standard deviation of $S(1) - S(0)$, given by $\sqrt{\sigma^2_{S_0} + \sigma^2_{S_1} - 2\rho_s\sigma_{S_0}\sigma_{S_1}}$. The plot shows that 0 is contained within the credible interval at almost all values of s, indicating that there could be large effects of treatment on the surrogate with no expected effect of treatment on the outcome. Similarly, when there is no treatment effect on S, there could still be a treatment effect on T.

6.2 Progression Free Survival Time as a Surrogate for Overall Survival Time in an Ovarian Cancer Trial

Our second data application is to data from a randomized trial in advanced ovarian cancer. This trial along with 3 others were analyzed by Buyse, *et al.* (2000) using a meta-analytic validation method, with the center in which patients were treated in each trial as the unit of analysis. In the trial we examine, a total of 274 women were treated for ovarian cancer in two treatment arms. Of these patients, 201 experienced a cancer progression prior to death, and 43 died without recurrence. The remaining 30 patients were censored for death and were not considered in the analysis. Again, we have a binary treatment with 126 subjects in the control arm and 118 in the treatment arm. The surrogate marker is progression free survival (PFS) time, in months and the final endpoint is overall survival (OS) time, in months. As both of these outcomes were right skewed, a log-transformation was taken to normalize the data. Estimates of parameters used to assess the validity of the Prentice criteria are

as follows:

$$\hat{\theta}_1 = 0.07(SE = 0.13, P = 0.58)$$

$$\hat{\alpha}_1 = 0.17(SE = 0.14, P = 0.23)$$

$$\hat{\mu} = 0.90(SE = 0.02, P < 0.0001)$$

$$\hat{\beta}_1 = -0.26(SE = 0.17, P = 0.13), \hat{\beta}_2 = 0.88(SE = 0.03, P < 0.0001),$$

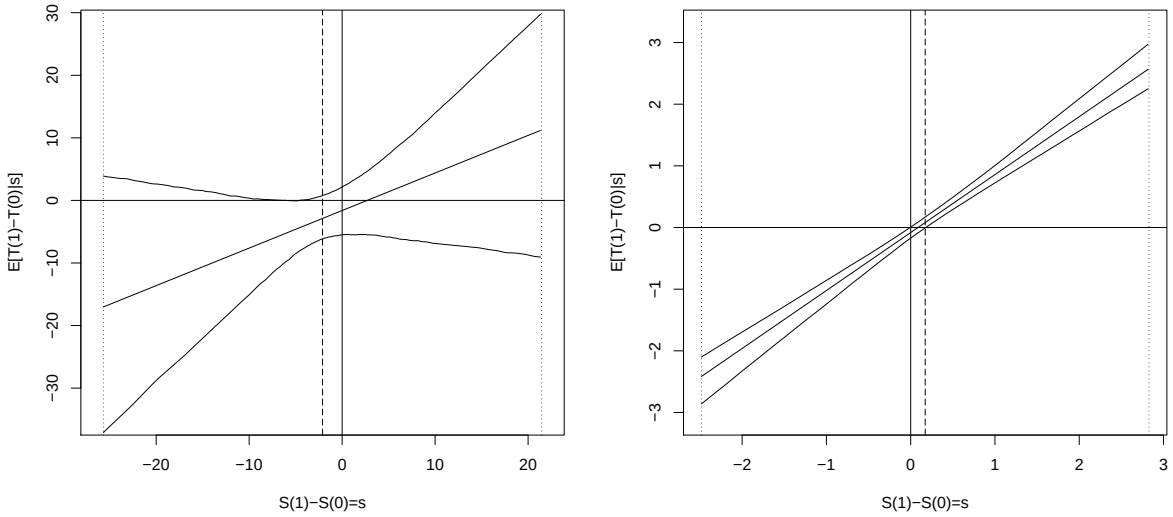
$$\hat{\beta}_3 = 0.05(SE = 0.04, P = 0.27)$$

As in the macular degeneration data, θ_1 and α_1 are not statistically significant, and the Prentice criteria are therefore unmet. We obtain the following posterior estimates for the causal quantities of interest using our method with Beta priors on the unidentified parameters:

$$\gamma_0 = -0.08(-0.17, 0.007)$$

$$\gamma_1 = 0.94(0.83, 1.09)$$

The 95% credible interval for γ_0 (barely) contains 0 while the 95% credible interval for γ_1 does not and ρ_{ST}^{-} was 0.93 (0.87, 0.98). We therefore conclude that progression free survival time is a marginally valid principal surrogate for overall survival. This agrees with the findings of Buyse, *et al.* (2000), who concluded that progression free survival could be used as a surrogate for overall survival in advanced ovarian cancer. Figure 3(b) provides a plot of the CEP curve and 95% credible interval at each $S(1) - S(0) = s$, with both S and T on the log scale. The middle and two outer dashed lines indicate the posterior mean of $\mu_{S_1} - \mu_{S_0}$, and the posterior means of $\mu_{S_1} - \mu_{S_0} \pm 2SD_{S(1)-S(0)}$, respectively. The plot shows that when there is no treatment effect on S, there is little or no expected treatment effect on T, and as the treatment effect on S increases, the treatment effect on T is also expected to increase.



(a) early change in visual acuity as a surrogate for late change in visual acuity

(b) PFS time as a surrogate for OS time

Figure 3: CEP Curves for Data Examples

7 Discussion

In this article, we develop a method for the assessment of surrogate markers within the principal surrogate framework. We assume a multivariate normal distribution for the potential surrogate outcomes and potential final outcomes and derive quantities that may be useful in determining the validity of a surrogate marker. Through our model setup, context specific assumptions can be incorporated into the prior distributions of unidentified parameters to aid in estimation. The estimation procedure can be extended to scenarios where T is partially missing, or to the multiple trial setting.

We compare some of the proposed quantities for surrogate validation to the original validation criteria put forth by Prentice and show that, in many settings, we might expect the Prentice and principal surrogacy criteria to be in agreement. Based on our simulation study, it appears that when principal surrogacy is present, it is most accurately determined in cases where the Prentice criteria would also correctly

identify surrogacy. When principal surrogacy is not present, it can be determined both when the Prentice criteria are able to correctly identify S as invalid and when the Prentice criteria incorrectly deem S to be valid. We note that even with the use of informative priors to aid in the estimation of the partially identified parameters, the coverage rates in many cases are not ideal.

Each of the proposed quantities have merits and drawbacks in terms of their ability to characterize surrogacy. The proposed γ_0 and γ_1 quantities are easily interpretable, but proving that γ_0 is equal to 0, a necessary condition for a valid surrogate, is difficult to do in practice. The correlation measure, ρ_{ST}, captures the causal correlation between the treatment effect on the surrogate and the treatment effect on the outcome, but fails to capture the concept of ACN. The CEP graph provides a way to estimate expected treatment effects on T when treatment effects on S are at relevant clinical values, but does not offer a single summary of the value of S as a surrogate. Finally, the Φ_{10} quantity provides information about the entire conditional distribution, as opposed to just the expectation, but is more difficult to estimate and seems to have poor properties. While no single parameter estimate can completely assess principal surrogacy, a variety of measures that consider the distribution of the causal effect of treatment on the outcome conditional on the causal effect of treatment on the surrogate can be used in combination to provide evidence as to whether or not S is a valid surrogate for T.

Due to the nonidentifiability of some parameters in our model, certain assumptions on the relationships between nonidentifiable associations were made and informative priors were used for unidentified parameters to aid in estimation. The use of other priors or other context specific assumptions about parameters could be made. Zigler and Belin (2011) also explore the effects of various model assumptions in a principal surrogacy estimation procedure. They use a Bayesian estimation approach for the

CEP surface when S is continuous and T is binary. In their procedure, priors are placed on the regression coefficients of the CEP surface, and an independence assumption is made for $T(1)$ and $T(0)$ conditional on the surrogate and other baseline covariates.

Previous work on principal surrogates has focused on binary endpoints (Li *et al.*, 2010) or a categorical or continuous surrogate outcome with a binary or continuous final endpoint with the conditional distributions of pairs of potential outcomes estimated separately (Gilbert and Hudgens, 2008). Qin *et al.* (2008) used a principal stratification approach in the assessment of a continuous surrogate with a time to event outcome. Extensions to other common data types, such as the setting where both the surrogate and final outcome are time to event endpoints, may be possible through the use of multivariate copula models.

Acknowledgements

The authors are grateful to the Ovarian Cancer Meta-analysis Project (Coordinator: Marc Buyse, ScD) and to the Pharmacological Therapy for Macular Degeneration Study Group for permission to use their individual patient data. They would also like to thank Marc Buyse and Tomasz Burzykowski for providing the data. This research was supported by NCI Grants R01 CA12910201, T32 CA-83654 and R01 MH078016.

References

Barnard, J., McCulloch, R. and Meng, X-L (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica* **10,** 1281–1311.

Bartolucci, F., and Grilli, L. (2011). Modeling partial compliance through copu-

las in a principal stratification framework. *Journal of the American Statistical Association* **106,** 469–479.

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

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

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

Gilbert, P.B., and Hudgens, M.G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64,** 1146–1154.

Gilbert, P., Qin, L., and Self, S. (2009). Response to Andrew Dunning's comment on "Evaluating a surrogate endpoint at three levels, with application to vaccine development". *Statistics in Medicine* **28,** 716–719.

Gustafson, P. (2010). Bayesian inference for partially identified models. *The International Journal of Biostatistics* **Vol. 6: Iss. 2,** Article 17.

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

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

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82(4),** 669–710.

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

Qin, L., Gilbert, P.B., Follmann, D., and Li, D. (2008). Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. *The Annals of Applied Statistics* **Vol.2, No. 1,** 386–407.

Ritter, C., and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association* **87,** 861–868.

Rosenbaum, P. (1984). The consequences of adjustment for a concomitant variable that has been affected by treatment. *Journal of the Royal Statistical Society, Ser. B* **147,** 656–666.

Rosenbaum, P. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association* **79,** 41–48.

Rubin, D.B. (1978). Bayesian-inference for causal effects-role of randomization. *The Annals of Statistics* **6,** 34–58.

Sargent, D.J., Patiyil, S., Yothers, G., et al. (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. *Journal of Clinical Oncology* **25,** 4569–4574.

Schwartz, S.L., Li, F., and Mealli, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association* **106,** 1331–1344.

VanderWeele, T.J. (2011). Principal Stratification - Uses and Limitations. *The International Journal of Biostatistics* **Volume 7, Issue 1,** Article 28.

Wang, Y., Mogg, R., and Lunceford, J. (2012). Evaluating correlation-based metric for surrogate marker qualification within a causal correlation framework.

Biometrics **68,** 617–627.

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

Zigler, C.M., and Belin, T.R. (2011). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics* doi:10.1111/j.1541-0420.2011.01736.x

Table 2: Simulation Results Under Different Prior Specifications

Identified Parameters

		S Valid PS, S Invalid Prentice			S Invalid PS, S Valid Prentice			S Invalid PS & Prentice			S Valid PS & Prentice		
Parameter	Prior Scenario	True Value	Mean (SD)	$P\bar{S}D$	True Value	Mean (SD)	$P\bar{S}D$	True Value	Mean (SD)	$P\bar{S}D$	True Value	Mean (SD)	$P\bar{S}D$
μ_{s_0}	1[1]	4	4.00(0.07)	0.08	4	4.00(0.08)	0.08	4	4.00(0.09)	0.08	4	4.00(0.08)	0.08
	2[2]		4.00(0.09)	0.08		4.01(0.08)	0.08		4.00(0.08)	0.08		4.00(0.08)	0.08
	3[3]		4.00(0.08)	0.08		4.00(0.08)	0.08		4.01(0.08)	0.08		4.00(0.09)	0.08
	4[4]		3.99(0.08)	0.08		4.01(0.08)	0.08		3.99(0.08)	0.08		4.00(0.08)	0.08
μ_{s_1}	1	6	5.99(0.07)	0.08	6	5.99(0.09)	0.08	6	6.00(0.08)	0.08	6	6.00(0.08)	0.08
	2		6.01(0.08)	0.08		5.99(0.07)	0.08		5.99(0.08)	0.08		6.00(0.08)	0.08
	3		6.00(0.08)	0.08		6.00(0.08)	0.08		6.00(0.08)	0.08		5.99(0.08)	0.08
	4		6.01(0.08)	0.08		5.99(0.08)	0.08		5.99(0.08)	0.08		5.99(0.09)	0.08
μ_{t_0}	1	7.8	7.80(0.08)	0.08	9	9.00(0.08)	0.08	8.5	8.51(0.08)	0.08	8.4	8.40(0.08)	0.08
	2		7.80(0.09)	0.08		9.00(0.08)	0.08		8.50(0.08)	0.08		8.39(0.08)	0.08
	3		7.80(0.08)	0.08		9.00(0.09)	0.08		8.50(0.08)	0.08		8.40(0.08)	0.08
	4		7.78(0.08)	0.08		9.00(0.08)	0.08		8.49(0.07)	0.08		8.40(0.08)	0.08
μ_{t_1}	1	10	9.99(0.08)	0.08	10	9.99(0.08)	0.08	10	10.00(0.09)	0.08	10	10.00(0.08)	0.08
	2		10.00(0.08)	0.08		10.00(0.08)	0.08		9.99(0.08)	0.08		10.00(0.07)	0.08
	3		10.00(0.07)	0.08		10.00(0.08)	0.08		9.99(0.08)	0.08		9.99(0.08)	0.08
	4		10.01(0.09)	0.08		10.00(0.08)	0.08		10.00(0.08)	0.08		10.00(0.09)	0.08
σ_{s_0}	1	1	1.01(0.06)	0.09	1	1.00(0.05)	0.07	1	1.02(0.06)	0.08	1	1.01(0.06)	0.09
	2		1.00(0.06)	0.06		1.00(0.06)	0.06		1.00(0.06)	0.06		1.00(0.06)	0.06
	3		1.00(0.05)	0.06		1.00(0.06)	0.06		1.01(0.06)	0.06		1.00(0.06)	0.06
	4		0.99(0.06)	0.06		1.01(0.06)	0.06		1.00(0.06)	0.06		0.99(0.06)	0.06
σ_{s_1}	1	1	1.01(0.06)	0.09	1	1.00(0.06)	0.08	1	1.01(0.06)	0.07	1	1.01(0.06)	0.09
	2		1.00(0.06)	0.06		1.01(0.05)	0.06		1.01(0.06)	0.06		1.00(0.06)	0.06
	3		1.00(0.05)	0.06		1.01(0.06)	0.06		1.01(0.06)	0.06		0.99(0.06)	0.06
	4		1.00(0.06)	0.06		1.01(0.05)	0.06		1.01(0.05)	0.06		0.99(0.06)	0.07
σ_{t_0}	1	1	1.00(0.06)	0.09	1	1.01(0.06)	0.08	1	1.01(0.06)	0.07	1	1.01(0.06)	0.10
	2		1.00(0.06)	0.06		1.00(0.05)	0.06		1.01(0.06)	0.06		1.00(0.06)	0.06
	3		1.00(0.05)	0.06		1.01(0.06)	0.06		1.02(0.06)	0.06		0.99(0.05)	0.06
	4		1.00(0.06)	0.06		1.01(0.06)	0.06		1.01(0.06)	0.06		0.99(0.05)	0.06
σ_{t_1}	1	1	1.01(0.06)	0.09	1	1.00(0.06)	0.08	1	1.02(0.06)	0.07	1	1.01(0.06)	0.10
	2		1.00(0.06)	0.06		1.00(0.06)	0.06		1.00(0.06)	0.06		1.00(0.06)	0.06
	3		1.00(0.06)	0.06		1.00(0.06)	0.06		1.01(0.05)	0.06		1.00(0.06)	0.06
	4		1.01(0.06)	0.06		1.00(0.06)	0.06		1.01(0.06)	0.06		0.99(0.06)	0.06
ρ_{00}	1	0.7	0.68(0.04)	0.05	0.5	0.48(0.07)	0.06	0.2	0.19(0.07)	0.08	0.8	0.78(0.03)	0.03
	2		0.68(0.04)	0.04		0.48(0.06)	0.06		0.20(0.07)	0.07		0.79(0.03)	0.03
	3		0.69(0.04)	0.04		0.48(0.06)	0.06		0.20(0.07)	0.07		0.78(0.03)	0.03
	4		0.68(0.04)	0.04		0.48(0.06)	0.06		0.20(0.07)	0.07		0.79(0.03)	0.03
ρ_{11}	1	0.7	0.68(0.04)	0.05	0.5	0.48(0.07)	0.06	0.2	0.18(0.08)	0.08	0.8	0.78(0.03)	0.04
	2		0.68(0.05)	0.04		0.49(0.06)	0.06		0.19(0.07)	0.07		0.79(0.03)	0.03
	3		0.68(0.04)	0.04		0.49(0.07)	0.06		0.20(0.07)	0.07		0.78(0.03)	0.03
	4		0.69(0.04)	0.04		0.49(0.06)	0.06		0.20(0.06)	0.07		0.78(0.03)	0.03

Unidentified Parameters

ρ_s	1	0.5	-0.35(0.23)	0.33	0.5	-0.22(0.22)	0.35	0.2	-0.15(0.23)	0.37	0.4	-0.35(0.24)	0.34
	2		0.32(0.08)	0.19		0.39(0.07)	0.22		0.37(0.08)	0.22		0.24(0.07)	0.15
	3		0.34(0.06)	0.13		0.45(0.05)	0.16		0.46(0.06)	0.21		0.22(0.04)	0.10
	4		0.47(0.07)	0.18		0.43(0.06)	0.20		0.34(0.06)	0.21		0.43(0.08)	0.16
ρ_{01}	1	0.15	-0.45(0.21)	0.29	0.45	-0.28(0.21)	0.33	0.04	-0.18(0.23)	0.35	0.32	-0.48(0.22)	0.29
	2		0.32(0.08)	0.19		0.39(0.07)	0.22		0.37(0.06)	0.22		0.24(0.07)	0.15
	3		0.14(0.04)	0.11		0.16(0.03)	0.10		0.06(0.02)	0.04		0.09(0.03)	0.07
	4		0.40(0.08)	0.20		0.28(0.07)	0.19		0.14(0.04)	0.15		0.40(0.09)	0.18
ρ_{10}	1	0.15	-0.37(0.21)	0.32	0.45	-0.28(0.24)	0.34	0.04	-0.16(0.23)	0.35	0.32	-0.39(0.21)	0.33
	2		0.34(0.08)	0.19		0.39(0.07)	0.22		0.37(0.07)	0.22		0.24(0.07)	0.15
	3		0.15(0.04)	0.11		0.16(0.02)	0.11		0.06(0.02)	0.04		0.10(0.03)	0.08
	4		0.42(0.08)	0.19		0.28(0.07)	0.19		0.14(0.03)	0.14		0.42(0.08)	0.17
ρ_t	1	0.18	-0.47(0.19)	0.27	0.5	-0.32(0.23)	0.32	0.3	-0.18(0.22)	0.36	0.4	-0.53(0.18)	0.28
	2		0.31(0.08)	0.19		0.37(0.07)	0.22		0.37(0.07)	0.22		0.24(0.06)	0.16
	3		0.32(0.05)	0.13		0.44(0.05)	0.16		0.46(0.06)	0.21		0.21(0.04)	0.09
	4		0.45(0.07)	0.19		0.42(0.06)	0.20		0.34(0.05)	0.21		0.42(0.08)	0.17

1: No restrictions on ρ
2: $\rho \geq 0$
3: $\rho \geq 0$ and $\rho_{10}, \rho_{01} < \rho_s, \rho_t, \rho_{00}, \rho_{11}$
4: Beta priors

Table 3: Simulation results: Bias, variability and coverage rate of surrogacy parameters

		S Valid PS, S Invalid Prentice				S Invalid PS, S Valid Prentice				S Invalid PS & Prentice				S Valid PS & Prentice			
Parameter	Prior Scenario	True Value	Mean (SD)	$P\bar{S}D$	95% Coverage	True Value	Mean (SD)	$P\bar{S}D$	95% Coverage	True Value	Mean (SD)	$P\bar{S}D$	95% Coverage	True Value	Mean (SD)	$P\bar{S}D$	95% Coverage
β_1	1	0.8	0.81(0.42)	0.51	0.99	0	0.06(0.60)	0.56	0.93	1.1	1.15(0.59)	0.59	0.96	0	0.05(0.36)	0.47	0.99
	2		0.82(0.45)	0.44	0.96		0.02(0.51)	0.52	0.96		1.18(0.54)	0.54	0.96		0.03(0.37)	0.37	0.95
	3		0.84(0.44)	0.43	0.94		0.04(0.56)	0.52	0.93		1.07(0.51)	0.55	0.98		-0.01(0.35)	0.07	0.96
	4		0.78(0.43)	0.44	0.94		0.004(0.51)	0.52	0.95		1.10(0.54)	0.54	0.96		0.04(0.37)	0.38	0.95
β_2	1	0.7	0.68(0.06)	0.07	0.96	0.5	0.49(0.08)	0.08	0.94	0.2	0.19(0.08)	0.08	0.96	0.8	0.78(0.05)	0.07	0.97
	2		0.69(0.06)	0.06	0.97		0.49(0.07)	0.07	0.96		0.21(0.07)	0.08	0.97		0.79(0.05)	0.05	0.96
	3		0.69(0.06)	0.06	0.96		0.49(0.08)	0.07	0.93		0.20(0.07)	0.07	0.94		0.78(0.05)	0.05	0.93
	4		0.68(0.06)	0.06	0.97		0.49(0.07)	0.07	0.94		0.20(0.07)	0.07	0.95		0.79(0.05)	0.05	0.97
β_3	1	0	0.004(0.08)	0.10	0.99	0	-0.008(0.12)	0.11	0.95	0	-0.0007(0.11)	0.11	0.96	0	-0.004(0.07)	0.09	0.98
	2		0.001(0.08)	0.09	0.96		0.002(0.10)	0.10	0.95		-0.01(0.10)	0.10	0.96		0.001(0.07)	0.07	0.94
	3		-0.002(0.08)	0.08	0.95		-0.001(0.11)	0.10	0.92		0.004(0.10)	0.11	0.97		0.007(0.07)	0.07	0.96
	4		0.010(0.08)	0.08	0.96		0.006(0.10)	0.10	0.95		0.003(0.10)	0.10	0.96		-0.002(0.07)	0.07	0.96
γ_0	1	0	0.54(0.23)	0.38	0.72	0.8	-0.33(0.33)	0.54	0.27	1.1	0.81(0.41)	0.66	0.97	0	-0.25(0.21)	0.33	0.96
	2		1.09(0.26)	0.52	0.49		0.64(0.28)	0.70	1		2.07(0.28)	0.65	0.69		0.12(0.17)	0.37	1
	3		0.54(0.15)	0.31	0.70		-0.30(0.23)	0.50	0.01		0.82(0.22)	0.45	1		-0.19(0.11)	0.20	0.92
	4		1.10(0.26)	0.63	0.60		0.18(0.26)	0.67	0.96		1.30(0.20)	0.48	0.99		0.20(0.22)	0.51	1
γ_1	1	1.1	0.83(0.11)	0.19	0.68	0.1	0.66(0.16)	0.26	0.27	0.2	0.34(0.20)	0.32	0.97	0.8	0.92(0.10)	0.16	0.94
	2		0.55(0.12)	0.26	0.48		0.28(0.13)	0.35	1		-0.29(0.12)	0.32	0.64		0.74(0.08)	0.18	1
	3		0.83(0.06)	0.15	0.63		0.65(0.10)	0.24	0		0.34(0.10)	0.22	1		0.90(0.04)	0.09	0.91
	4		0.55(0.12)	0.31	0.60		0.41(0.12)	0.33	0.97		0.11(0.08)	0.23	1		0.70(0.11)	0.25	1
ρ_{ST}	1	0.86	0.77 (0.06)	0.11	0.97	0.1	0.60 (0.11)	0.18	0.26	0.21	0.31 (0.16)	0.26	0.99	0.8	0.85 (0.04)	0.08	0.95
	2		0.53(0.10)	0.20	0.54		0.17 (0.10)	0.29	1		-0.27 (0.10)	0.27	0.63		0.73(0.06)	0.12	1
	3		0.81(0.04)	0.09	1		0.62 (0.06)	0.15	0		0.31 (0.07)	0.14	1		0.89 (0.02)	0.05	0.73
	4		0.52(0.09)	0.23	0.68		0.38(0.09)	0.26	0.97		0.10(0.07)	0.20	1		0.66(0.08)	0.17	0.99
$\gamma_0 + \gamma_1(\mu_{S_1} - \mu_{S_0} + 2SD_{S(1)-S(0)})$	1	4.4	4.86 (0.39)	0.57	0.90	1.2	2.97 (0.49)	0.67	0.27	2.0	2.47 (0.52)	0.80	0.96	3.35	4.59 (0.37)	0.57	0.31
	2		3.45 (0.28)	0.50	0.46		1.37 (0.26)	0.64	1		0.89 (0.25)	0.61	0.56		3.40 (0.24)	0.38	0.99
	3		4.08 (0.21)	0.32	0.93		2.31 (0.20)	0.39	0		2.10 (0.18)	0.28	1		3.81 (0.18)	0.24	0.49
	4		3.31 (0.26)	0.55	0.36		1.82 (0.23)	0.58	0.96		1.74 (0.19)	0.45	1		3.02 (0.27)	0.45	0.96
$\gamma_0 + \gamma_1(\mu_{S_1} - \mu_{S_0} - 2SD_{S(1)-S(0)})$	1	0	-0.47 (0.38)	0.57	0.87	0.8	-1.00 (0.47)	0.67	0.25	0.99	0.51 (0.52)	0.79	0.95	-0.15	-1.39 (0.40)	0.57	0.31
	2		0.96 (0.30)	0.50	0.44		0.63 (0.27)	0.64	1		2.11 (0.26)	0.61	0.58		-0.19 (0.24)	0.38	0.99
	3		0.31 (0.22)	0.32	0.91		-0.30 (0.23)	0.39	0.02		0.88 (0.18)	0.28	1		-0.62 (0.17)	0.24	0.45
	4		1.14 (0.26)	0.54	0.32		0.18 (0.25)	0.58	0.96		1.29 (0.20)	0.45	0.97		0.18 (0.25)	0.45	0.96
$\Phi_{10}(0)$	1	0.5	0.69 (0.08)	0.13	0.72	0.79	0.40 (0.09)	0.16	0.24	0.83	0.70 (0.10)	0.16	0.95	0.5	0.39 (0.08)	0.14	0.96
	2		0.84 (0.06)	0.13	0.49		0.70 (0.08)	0.20	1		0.95 (0.02)	0.07	0.69		0.55 (0.08)	0.18	1
	3		0.77 (0.07)	0.15	0.70		0.38 (0.08)	0.19	0.16		0.78 (0.06)	0.15	1		0.36 (0.07)	0.14	0.92
	4		0.85 (0.07)	0.16	0.60		0.56 (0.09)	0.23	0.96		0.86 (0.04)	0.10	1		0.58 (0.10)	0.23	1

1: No restrictions on ρ
2: $\rho \geq 0$
3: $\rho \geq 0$ and $\rho_{10}, \rho_{01} < \rho_s, \rho_t, \rho_{00}, \rho_{11}$
4: Beta priors

Table 4: Simulation results: Principal Surrogacy Assessment

Model	1	2	3	4
Truth				
PS satisfied	Yes	No	No	Yes
Prentice satisfied	No	Yes	No	Yes
Estimation Results				
$\gamma_0 = 0$ Not Rejected, Reject $\gamma_1 = 0$	0.37	0.15	0.01	0.94
Reject $\rho_{ST} = 0$	0.57	0.17	0.01	0.94
$CEP^U_{-2SD} < CEP^L_{+2SD}$	0.55	0.15	0.01	0.93
$\Phi_{10}(0) = 0.5$ Not Rejected	0.60	1	0.20	1
Prentice Criteria Not Rejected	0.52	0.92	0.26	0.95