Biometrics 66, 523–531
June 2010

DOI: 10.1111/j.1541-0420.2009.01303.x

A Bayesian Approach to Surrogacy Assessment Using Principal Stratification in Clinical Trials

Yun Li,* Jeremy M.G. Taylor, and Michael R. Elliott

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
**email:* yunlisph@umich.edu

Summary. A surrogate marker (S) is a variable that can be measured earlier and often more easily than the true endpoint (T) in a clinical trial. Most previous research has been devoted to developing surrogacy measures to quantify how well S can replace T or examining the use of S in predicting the effect of a treatment (Z). However, the research often requires one to fit models for the distribution of T given S and Z. It is well known that such models do not have causal interpretations because the models condition on a postrandomization variable S. In this article, we directly model the relationship among T, S, and Z using a potential outcomes framework introduced by Frangakis and Rubin (2002, *Biometrics* **58,** 21–29). We propose a Bayesian estimation method to evaluate the causal probabilities associated with the cross-classification of the potential outcomes of S and T when S and T are both binary. We use a log-linear model to directly model the association between the potential outcomes of S and T through the odds ratios. The quantities derived from this approach always have causal interpretations. However, this causal model is not identifiable from the data without additional assumptions. To reduce the nonidentifiability problem and increase the precision of statistical inferences, we assume monotonicity and incorporate prior belief that is plausible in the surrogate context by using prior distributions. We also explore the relationship among the surrogacy measures based on traditional models and this counterfactual model. The method is applied to the data from a glaucoma treatment study.

Key words: Bayesian estimation; Counterfactual model; Randomized trial; Surrogate marker.

1. Introduction

A good surrogate marker S usually has a strong association with the true endpoint T. When T is rare, late-occurring, or costly to obtain, one could use an effective surrogate marker to reliably extract information on the effect of the treatment (Z) on T before T is completely observed. While a typical surrogate marker can be a laboratory measurement and used as a substitute for a clinical endpoint such as CD4 counts for HIV infection and prostate-specific antigen for prostate cancer, an earlier laboratory measurement has also been considered as a surrogate marker for a later measurement. By facilitating the early treatment prediction on the later measurement, the earlier measurement can have enormous potential benefits in reducing trial duration and size, and lowering the trial expense. Some examples of using an earlier measurement as a surrogate for a later one include the interim height for adult height on girls with Turner Syndrome by Venkatraman and Begg (1999) and the earlier vision test result for the later result in a study on patients with age-related muscular degeneration by Buyse and Molenberghs (1998). In the data example to which we will apply our method, we consider the intraocular pressure (IOP) at the 12th month as S and the IOP at the 96th month as T among the glaucoma patients.

Prentice (1989) proposed a formal definition of perfect surrogacy that requires that S fully captures the effect of the treatment on T. To measure less than perfect surrogacy, the proportion of the treatment effect explained by S was proposed by Freedman, Graubard, and Schatzkin (1992) and further extended by Wang and Taylor (2002). However, these measures often require one to utilize models for the distribution of T given S and Z. They often do not have causal interpretations because the models used condition on the postrandomization variable S (Rosenbaum, 1984). Other surrogacy measures include the trial-level and individual-level correlations between S and T in a multiple-trial setting (Buyse et al., 2000) and those based on entropy (Alonso and Molenberghs, 2003).

To allow for a causal interpretation, we directly measure the associations among S, T, and Z in a causal modeling framework through the principal stratification approach introduced by Frangakis and Rubin (2002) (FR). This framework hypothesizes the setting wherein each individual has two potential outcomes, corresponding to the two possible treatment regimes (e.g., $Z = 1$ for treatment and $Z = 0$ for placebo). Here, we use the terms causal, counterfactual, and potential outcomes models exchangeably. When both S and T are binary, the potential outcomes for S and T are denoted by $(S(Z) = 0, 1)$ and $(T(Z) = 0, 1)$ with respect to Z. The approach by FR is to examine the distribution of the potential outcomes of T with respect to Z within each principal stratum, which is defined by each pair of possible realizations of the potential outcomes of S. Since the principal strata cannot be changed by treatment, they can be adjusted for as a prerandomization variable. As such, the association measures and the quantities derived have causal interpretations.

However, there has been little work on estimation methods using this framework. A review paper by Weir and Walley (2006) advocates the need for further research. An exception to this is the paper of Gilbert and Hudgens (2008), where they proposed the use of causal effect (CE) predictiveness to assess surrogacy and their context of an HIV vaccine trial allowed them to assume that $S(0) = 0$. In this article, we relax this assumption and propose a Bayesian estimation method to evaluate the counterfactual probabilities associated with the combinations of different sequences of potential outcomes of S and T for each individual. We incorporate the prior knowledge by imposing appropriate prior distributions and placing some reasonable constraints on the model parameters that allows one to reduce the nonidentifiability problem and possibly increase the precision of the statistical inference.

In Section 2, we describe the glaucoma data example. In Section 3, we introduce the causal model, assumptions, and surrogacy measures. In Section 4, we propose a Bayesian estimation method. In Section 5, we apply the method to the glaucoma data and examine the sensitivity of the priors. In Section 6, we examine the properties of our estimates through simulations. In Section 7, we explore the connections among the surrogacy measures based on conventional models and the counterfactual model. Finally, we provide a discussion.

2. Glaucoma Treatment Study

We will apply the method to the data from the Collaborative Initial Glaucoma Treatment Study (CIGTS) (Musch et al., 1999). Glaucoma is a group of diseases that cause vision loss and is a leading cause for blindness. Elevated IOP in the eyes is a major risk factor of glaucoma. The Advanced Glaucoma Intervention Study (AGIS) demonstrated that when IOP reduction from baseline is substantial, progression of visual field loss can be prevented (Musch et al., 2009). The CIGTS is a randomized trial designed to compare the effects of surgery ($Z = 1$) and medicine ($Z = 0$) on reducing IOP. Patients were enrolled between 1993 and 1997. The IOP (recorded in mmHg) level has been measured at different time points following randomization. Our purpose is to examine the property of the IOP measurement at the 12th month as S and for the IOP at the 96th month as T. Both S and T are defined as 1 if IOP is less than 18 mmHg and 0 otherwise. It is found that eyes with IOP of less than 18 at every time point during at least 6 years of follow-up essentially had no further visual field loss (AGIS, 2000). There are 607 patients enrolled at the baseline. Due to drop out, 345 are measured only at month 12 and 228 have IOP measured at both months 12 and 96.

3. The Setup

3.1 *Potential Outcomes Model*

For each subject i, we have two potential outcomes for each of S_i and T_i, denoted by $S_i(Z_i)$ and $T_i(Z_i)$ with respect to Z_i. The possible realizations of $(S_i(0), S_i(1))$ are (0, 0), (0, 1), (1, 1), and (1, 0) and similarly for $(T_i(0), T_i(1))$. There are 16 counterfactual probabilities that are associated with the combinations of different sequences of potential outcomes for S_i and T_i. These probabilities sum to 1 as the 16 cells are the partitions of a population. Collectively, they completely describe the causal associations among T_i, S_i, and Z_i.

Table 1
Probabilities from the counterfactual model with monotonicity assumption

	$(T(0), T(1))$		
$(s(0), s(1))$	(0, 0)	(0, 1)	(1, 1)
(0, 0)	p_{11}	p_{12}	p_{13}
(0, 1)	p_{21}	p_{22}	p_{23}
(1, 1)	p_{31}	p_{32}	p_{33}

3.2 *Assumptions and Identifiability*

Since only one of the potential outcomes is unobserved, the counterfactual model is overparameterized. We make assumptions to assist in the identifiability. In addition to the two standard assumptions, ignorability of treatment assignment (Rubin, 1978) and stable unit treatment value assumption (Rubin, 1980), we also assume monotonicity. Under this assumption, a patient who received $Z = 1$ does not become worse off than that patient if he or she received $Z = 0$. Assume $S = 1$ and $T = 1$ represent better outcomes than $S = 0$ and $T = 0$, respectively. The monotonicity assumption requires that $S_i(1) \geqslant S_i(0)$ and $T_i(1) \geqslant T_i(0)$ for all i; hence, we cannot observe either $(S_i(0) = 1, S_i(1) = 0)$ or $(T_i(0) = 1, T_i(1) = 0)$. The number of free parameters is reduced from 15 to 8 (Table 1). Our data can support only six parameters, as the probabilities $(P(T = t, S = s \mid Z))$ within each treatment group add up to 1; hence, only some of the probabilities or certain combinations are estimable.

3.3 *Surrogacy Measures*

In a traditional model framework, Freedman et al. (1992) proposed the proportion of the treatment effect explained to measure surrogacy in a model that assumes no interaction between S and T. A measure free of this assumption was proposed by Wang and Taylor (2002) as $F_{WT} = \delta\gamma_a/\tau$ where $\delta = P(S = 1 \mid Z = 1) - P(S = 1 \mid Z = 0)$, $\tau = P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0)$ and $\gamma_a = P(T = 1 \mid Z = 0, S = 1) - P(T = 1 \mid Z = 0, S = 0)$. The quantities δ and τ denote the treatment effects on S and T, respectively; and γ_a measures the strength of the association between S and T. Given the effect on T, the larger the effect on S or the stronger the association between S and T, the higher F_{WT} will be. Odds ratios (OR), OR_{g0} and OR_{g1}, measure the associations between S and T in the $Z = 0, 1$ groups, respectively.

In a counterfactual framework, the expressions of CEs are based on the comparisons between two potential outcomes. In a randomized trial, both CEs on S and T are directly estimable. FR proposed the concepts of *associative* and *dissociative* effects. If the CE on T_i is reflected on the change in S_i, the effect is *associative*. Conversely, the effect is *dissociative*. We describe the sequence of the values of the potential outcomes (0, 0), (0, 1), and (1, 1) as "non-responsive," "responsive," and "always responsive." Under the monotonicity assumption, the overall CE on T (CET) is $p_{+2} = p_{12} + p_{22} + p_{32}$,, which measures the fraction of the patients whose T_i's are responsive to the treatment. The associative effect is p_{22} and the dissociative effect is $p_{12} + p_{32}$, where p_{22} refers to the fraction of the patients whose S_i's and T_i's are both responsive to the treatment and the dissociative effect is the fraction of the patients whose T_i's are responsive to the treatment

but whose S_i's are not. To evaluate the degree of surrogacy, Taylor, Wang, and Thiébaut (2005) defined the associative proportion (AP) as $\frac{p_{22}}{p_{12}+p_{22}+p_{32}}$. The dissociative proportion (DP) is $\frac{p_{12}+p_{32}}{p_{12}+p_{22}+p_{32}}$. Two other measures are also quantities of interest: surrogate associative proportion (SAP) $= \frac{p_{22}}{p_{21}+p_{22}+p_{23}}$ and surrogate dissociative proportion (SDP) $= \frac{p_{12}+p_{32}}{p_{1+}+p_{3+}}$, where $p_{j+} = \sum_{k=1}^{3} p_{jk}, j = 1, 3$.

Prentice (1989) defined perfect surrogacy in a traditional model setup that we refer to as perfect statistical surrogacy. FR suggested a definition for perfect principal surrogacy that requires that the CE on T may only exist when that on S exists; i.e., $p_{12} = p_{32} = 0$. When S and T are binary, we argue that with more restrictions, $p_{21} = p_{23} = 0$, it ensures that for every patient, if S_i is responsive, T_i is also responsive; and vice versa. Based on this context, we suggest a new measure, common associative proportion (CAP) $= \frac{p_{22}}{p_{12}+p_{21}+p_{22}+p_{23}+p_{32}}$, to assess the degree of principal surrogacy. When $p_{12} = p_{21} = p_{23} = p_{32} = 0$, S satisfies perfect principal surrogacy and we have CAP = 1; when $p_{22} = 0$, for any individual, no CE on T_i is captured by that on S_i and we have CAP = 0. When CAP = 1, we have SAP = 1, AP = 1, SDP = 0, and DP = 0. The measure CAP is usually smaller than AP and SAP. Unlike F_{WT}, CAP, AP, SAP, SDP, and DP always fall in the range [0, 1].

4. The Methods

4.1 *Observed Data, Complete Data, and Likelihood*

Let r_z denote the number of patients in the $Z = z$ group ($z = 0, 1$) and $r = r_0 + r_1$. Let r_{zst} denote the number of patients for each combination of Z, S, and T. The observed-data likelihood function can be expressed in terms of the counterfactual probabilities as follows:

$$L_{obs} = (p_{11} + p_{12} + p_{21} + p_{22})^{r_{000}} (p_{13} + p_{23})^{r_{001}} (p_{31} + p_{32})^{r_{010}} p_{33}^{r_{011}} \\ p_{11}^{r_{100}} (p_{12} + p_{13})^{r_{101}} (p_{21} + p_{31})^{r_{110}} (p_{22} + p_{23} + p_{32} + p_{33})^{r_{111}}.$$

The complete data consists of all potential outcomes. Let n_{jk} denote the cell count corresponding to the counterfactual probability in the cell (j, k) for the jth row and the kth column of Table 1 for all patients and n_{jk}^z for the treatment group z where j, $k = 1, 2, 3$. The complete data likelihood is

$$\begin{aligned} L_{com} &= p_{11}^{n_{11}^0+n_{11}^1} p_{12}^{n_{12}^0+n_{12}^1} p_{13}^{n_{13}^0+n_{13}^1} p_{21}^{n_{21}^0+n_{21}^1} \\ &\quad \times p_{22}^{n_{22}^0+n_{22}^1} p_{23}^{n_{23}^0+n_{23}^1} p_{31}^{n_{31}^0+n_{31}^1} p_{32}^{n_{32}^0+n_{32}^1} p_{33}^{n_{33}^0+n_{33}^1} \\ &= p_{11}^{n_{11}} p_{12}^{n_{12}} p_{13}^{n_{13}} p_{21}^{n_{21}} p_{22}^{n_{22}} p_{23}^{n_{23}} p_{31}^{n_{31}} p_{32}^{n_{32}} p_{33}^{n_{33}}. \end{aligned}$$

There is a one-to-one or many-to-one correspondence between n_{jk}'s and r_{zst}'s.

4.2 *The Model*

Let $S^* = 1, 2, 3$ denote the ordered categories of ($S(0)$, $S(1)$): (0, 0), (0, 1), and (1, 1) and $T^* = 1, 2, 3$ denote the ordered categories of ($T(0)$, $T(1)$). For convenience, we reparametrize the p_{jk}'s and use a log-linear model for n_{jk}^z. For simplicity, we assume equal allocation, i.e., $E(n_{jk}^z) = \mu_{jk}$. The model is specified as

$$\log \mu_{jk} = \lambda + \lambda_{jS} + \lambda_{kT} + \lambda_{jk}, \tag{1}$$

where λ_{jS} and λ_{kT} denote the row and column effects, respectively and λ_{jk} denote their interaction. For identifiability of the log-linear model, we use the constraints ($\lambda_{2S} = \lambda_{2T} = \lambda_{j2} = \lambda_{2k} = 0$) which lead to nice and simple expressions for the following log odds ratios in the four 2 × 2 subtables in the four corners of Table 1:

$$\begin{aligned} \log(OR_1) &= \log\left((\mu_{11}\mu_{22})/(\mu_{12}\mu_{21})\right) = \lambda_{11}, \\ \log(OR_2) &= \log\left((\mu_{12}\mu_{23})/(\mu_{13}\mu_{22})\right) = -\lambda_{13}, \\ \log(OR_3) &= \log\left((\mu_{21}\mu_{32})/(\mu_{22}\mu_{31})\right) = -\lambda_{31}, \\ \log(OR_4) &= \log\left((\mu_{22}\mu_{33})/(\mu_{23}\mu_{32})\right) = \lambda_{33}. \end{aligned}$$

The parametrization allows us to exploit the associations between the ordered variables S^* and T^*. A positive association between them implies that λ_{11} and λ_{33} are positive and λ_{13} and λ_{31} are negative. Conditional on the total counts, we can express the counterfactual probabilities using the parameters in (1) as:

$$p_{jk} = \frac{\exp(\lambda_{jS} + \lambda_{kT} + \lambda_{jk})}{\sum_j \sum_k \exp(\lambda_{jS} + \lambda_{kT} + \lambda_{jk})}. \tag{2}$$

To estimate the parameters, we adopt a Bayesian approach. We treat the unobserved potential outcomes as missing data and apply imputation techniques.

4.3 *Prior Specifications*

In clinical trials, the selection of the variable to use as S will be based on prior scientific knowledge. The surrogate marker S is often closely related to T, possibly because the marker is in the causal pathway leading to T. Hence, we assume ($S_i(0)$, $S_i(1)$) is more likely to agree with ($T_i(0)$, $T_i(1)$) than not; and S^* and T^* are ordered. That is, when S is nonresponsive (responsive), T is also more likely to be nonresponsive (responsive). Similarly it is unlikely that a person will be nonresponsive in S_i and always responsive in T_i.

The parameters, λ_{1S}, λ_{3S}, λ_{1T}, and λ_{3T}, are identifiable but the others are less so. We have chosen $N(u, v^2)$ as the prior distributions for λ_{1S}, λ_{3S}, λ_{1T}, λ_{3T}, λ_{11}, λ_{13}, λ_{31}, and λ_{33}. The prior for $\exp(\lambda)$ is G (a, b) where G denotes the gamma distribution and $G(a, b)$ is parameterized such that the expected value is ab and the variance is ab^2. We choose noninformative values of $a = 0.001$ and $b = 1000$ and let $v^2 = 9/4$. To incorporate our prior belief and encourage but not force the ordering restriction, we choose $u = 0.7$ for λ_{11} and λ_{33}, and $u = -0.7$ for λ_{13} and λ_{31}, which would suggest moderate positive associations between the potential outcomes of S and T. When the ordering restriction and positive association are not considered, we let $u = 0$. The characteristics of our prior choices on the log-linear model parameters are similar to what Garrett and Zeger (2000) discovered in their work for the logistic regression. The distributions of the probabilities induced by these priors are relatively flat, not overly skewed, and appropriate for our study setting with wide 95% percentile ranges. On the other hand, if vague priors such as normal priors with zero means and very large variances are placed on the log-linear model parameters, they would induce priors on the probabilities whose distributions have point masses concentrated at either zeros or ones (King and Brooks, 2001). When there are nonidentifiable quantities, vague priors can give the posterior distributions undesirable properties and push them toward being overly skewed and nonnormal as observed by Green and Park (2003).

4.4 *Estimation Procedure*

We use data augmentation (Little and Rubin, 2002) to estimate the parameters. Let $r_{obs} = \{r_{000}, r_{001}, r_{010}, r_{011}, r_{100}, r_{101}, r_{110}, r_{111}\}$ and $\theta = (\lambda, \lambda_{jS}, \lambda_{kT}, \lambda_{jk})$. The complete data cell counts are denoted by $n_{com} = \{n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, n_{33}\}$. To implement this procedure, we iterate the following I-step and P-step.

I-step: This step consists of distributing the observed counts into the cells in Table 1. Given θ^{l-1} and r_{obs}, we impute $n_{11}^{0l}, n_{12}^{0l}, n_{21}^{0l}, n_{22}^{0l}, n_{12}^{1l}, n_{13}^{0l}, n_{21}^{1l}, n_{22}^{1l}, n_{23}^{1l}, n_{32}^{1l}, n_{33}^{1l}$, and n_{31}^{0l} where θ^{l-1} denotes all the parameter estimates from the $(l-1)$th iteration, n_{11}^{0l} is the draw of the count that contributes to n_{11}^{1} from r_{000} from the lth iteration, n_{12}^{1l} is the draw of the count that contributes to n_{12}^{1} from r_{101} from the lth iteration, and so on. Let $\omega_1^{l-1} = p_{11}^{l-1} + p_{12}^{l-1} + p_{21}^{l-1} + p_{22}^{l-1}$ and $\omega_2^{l-1} = p_{22}^{l-1} + p_{23}^{l-1} + p_{32}^{l-1} + p_{33}^{l-1}$.

1. $(n_{11}^{0l}, n_{12}^{0l}, n_{21}^{0l}, n_{22}^{0l}) \sim \text{Multi}(r_{000}, \frac{p_{11}^{l-1}}{\omega_1^{l-1}}, \frac{p_{12}^{l-1}}{\omega_1^{l-1}}, \frac{p_{21}^{l-1}}{\omega_1^{l-1}}, \frac{p_{22}^{l-1}}{\omega_1^{l-1}})$
2. $n_{12}^{1l} \sim \text{Bin}(r_{101}, \frac{p_{12}^{l-1}}{p_{12}^{l-1}+p_{13}^{l-1}})$
3. $n_{13}^{0l} \sim \text{Bin}(r_{001}, \frac{p_{13}^{l-1}}{p_{13}^{l-1}+p_{23}^{l-1}})$
4. $n_{21}^{1l} \sim \text{Bin}(r_{110}, \frac{p_{21}^{l-1}}{p_{21}^{l-1}+p_{31}^{l-1}})$
5. $(n_{22}^{1l}, n_{23}^{1l}, n_{32}^{1l}, n_{33}^{1l}) \sim \text{Multi}(r_{111}, \frac{p_{22}^{l-1}}{\omega_2^{l-1}}, \frac{p_{23}^{l-1}}{\omega_2^{l-1}}, \frac{p_{32}^{l-1}}{\omega_2^{l-1}}, \frac{p_{33}^{l-1}}{\omega_2^{l-1}})$
6. $n_{31}^{0l} \sim \text{Bin}(r_{010}, \frac{p_{31}^{l-1}}{p_{31}^{l-1}+p_{32}^{l-1}})$
7. $n_{11}^{l} = n_{11}^{0l} + r_{100}; \quad n_{12}^{l} = n_{12}^{0l} + n_{12}^{1l};$
   $n_{13}^{l} = n_{13}^{0l} + r_{101} - n_{12}^{1l};$
8. $n_{21}^{l} = n_{21}^{0l} + n_{21}^{1l}; \quad n_{22}^{l} = n_{22}^{0l} + n_{22}^{1l};$
   $n_{23}^{l} = r_{001} - n_{13}^{0l} + n_{23}^{1l};$
9. $n_{31}^{l} = n_{31}^{0l} + r_{110} - n_{21}^{1l}; \quad n_{32}^{l} = r_{010} - n_{31}^{0l} + n_{32}^{1l};$
   $n_{33}^{l} = r_{011} + n_{33}^{1l}$

P-step: Generate θ^l from the posterior distribution, $p(\theta^l \,|\, n_{com}^l)$, where n_{com}^l includes the counts of the complete data obtained in the I-step from the lth iteration.

$$\exp(\lambda^l)\,|\cdot \sim \text{G}\left(n_{1+}^l + n_{2+}^l + n_{3+}^l + a, \frac{1}{\sum_{j=1}^{3}\sum_{k=1}^{3}\left(2\exp\left(\lambda_{jS}^l + \lambda_{kT}^l + \lambda_{jk}^l\right)\right) + \frac{1}{b}}\right),$$

$$\lambda_{1S}^l\,|\cdot \propto V1S \times \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l + \lambda_{3T}^l + \lambda_{13}^l\right)\right) \times \exp\left(-\left(\lambda_{1S}^l - u\right)^2/(2v^2)\right),$$

$$\lambda_{1T}^l\,|\cdot \propto V1T \times \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l + \lambda_{1T}^l + \lambda_{31}^l\right)\right) \times \exp\left(-\left(\lambda_{1T}^l - u\right)^2/(2v^2)\right),$$

$$\lambda_{3S}^l\,|\cdot \propto V3S \times \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l + \lambda_{3T}^l + \lambda_{33}^l\right)\right) \times \exp\left(-\left(\lambda_{3S}^l - u\right)^2/(2v^2)\right)$$

$$\lambda_{3T}^l\,|\cdot \propto V3T \times \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l + \lambda_{3T}^l + \lambda_{33}^l\right)\right) \times \exp\left(-\left(\lambda_{3T}^l - u\right)^2/(2v^2)\right),$$

$$\lambda_{11}^l\,|\cdot \propto \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l + \lambda_{1T}^l + \lambda_{11}^l\right)\right) \times \left(\exp\left(\lambda_{11}^l\right)\right)^{n_{11}^l} \exp\left(-\left(\lambda_{11}^l - u\right)^2/(2v^2)\right),$$

$$\lambda_{13}^l\,|\cdot \propto \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l + \lambda_{3T}^l + \lambda_{13}^l\right)\right) \times \left(\exp\left(\lambda_{13}^l\right)\right)^{n_{13}^l} \exp\left(-\left(\lambda_{13}^l - u\right)^2/(2v^2)\right),$$

$$\lambda_{31}^l\,|\cdot \propto \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l + \lambda_{1T}^l + \lambda_{31}^l\right)\right) \times \left(\exp\left(\lambda_{31}^l\right)\right)^{n_{31}^l} \exp\left(-\left(\lambda_{31}^l - u\right)^2/(2v^2)\right),$$

$$\lambda_{33}^l\,|\cdot \propto \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l + \lambda_{3T}^l + \lambda_{33}^l\right)\right) \times \left(\exp\left(\lambda_{33}^l\right)\right)^{n_{33}^l} \exp\left(-\left(\lambda_{33}^l - u\right)^2/(2v^2)\right),$$

$$p_{jk}^l = \frac{\exp\left(\lambda_{jS}^l + \lambda_{kT}^l + \lambda_{jk}^l\right)}{\sum_j \sum_k \exp\left(\lambda_{jS}^l + \lambda_{kT}^l + \lambda_{jk}^l\right)},$$

where

$$V1S = \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l + \lambda_{1T}^l + \lambda_{11}^l\right)\right) \times \left(\exp\left(\lambda_{1S}^l\right)\right)^{n_{1+}^l} \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l\right)\right),$$

$$V1T = \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l + \lambda_{1T}^l + \lambda_{11}^l\right)\right) \times \left(\exp\left(\lambda_{1T}^l\right)\right)^{n_{+1}^l} \exp\left(-2\exp\left(\lambda^l + \lambda_{1T}^l\right)\right),$$

$$V3S = \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l + \lambda_{1T}^l + \lambda_{31}^l\right)\right) \times \left(\exp\left(\lambda_{3S}^l\right)\right)^{n_{3+}^l} \exp\left(-2\exp\left(\lambda^l + \lambda_{3S}^l\right)\right),$$

$$V3T = \exp\left(-2\exp\left(\lambda^l + \lambda_{1S}^l + \lambda_{3T}^l + \lambda_{13}^l\right)\right) \times \left(\exp\left(\lambda_{3T}^l\right)\right)^{n_{+3}^l} \exp\left(-2\exp\left(\lambda^l + \lambda_{3T}^l\right)\right),$$

"$\cdot$" represents all the rest of the parameters, $n_{j+}^l = \sum_{k=1}^3 n_{jk}^l, n_{+k}^l = \sum_{j=1}^3 n_{jk}^l$, and so on. For $\exp(\lambda)$, the conditional draws can be made directly from the gamma distribution using the Gibbs sampler. For λ_{1S}, λ_{3S}, λ_{1T}, λ_{3T}, λ_{11}, λ_{13}, λ_{31}, and λ_{33}, we use the Metropolis–Hastings algorithm and the proposal distribution is normal with mean as the current value and variance adjusted to give an acceptance rate of approximately 40%. The mixing behavior of the MCMC sampler for a nonidentifiable model can be rather slow and poor (Gelfand and Sahu, 1999). In our case the Markov chain does not move quickly and the sample autocorrelation is high. For the CIGTS study, a burn-in period of 200,000 iterations is needed for the MCMC samples to stabilize. After burn-in, we sample every 100th MCMC iteration from the posterior distribution to reduce the autocorrelation and obtain samples with a good mixing property. The sensitivity toward the initial values is evaluated by comparing parameter estimates from five chains, on which we obtained the Gelman–Rubin Statistic ($\hat{R}$) (Gelman et al., 2004). Generally, $\hat{R} < 1.2$ is considered sufficient. For all estimates in the CIGTS data, we have min $\hat{R} \approx 0.9999$ and max $\hat{R} \approx 1.0007$.

Table 2

Bayesian estimates for the counterfactual model for glaucoma data. PSD = posterior standard deviation; AP = associative proportion; DP = dissociative proportion; SAP = surrogate associative proportion; SDP = surrogate dissociative proportion; CAP = common associative proportion; CET = causal effect on T. Prior specifications: $a = 0.001$, $b = 1000$, $u = 0.7$, *and* $v^2 = 9/4$.

Parameter	Mean	Median	PSD	95% CI
p_{11}	0.101	0.099	0.028	(0.054, 0.160)
p_{12}	0.033	0.030	0.019	(0.006, 0.077)
p_{13}	0.051	0.048	0.028	(0.007, 0.112)
p_{21}	0.051	0.048	0.024	(0.012, 0.101)
p_{22}	0.047	0.044	0.024	(0.011, 0.101)
p_{23}	0.170	0.170	0.042	(0.085, 0.252)
p_{31}	0.048	0.045	0.026	(0.006, 0.100)
p_{32}	0.063	0.061	0.030	(0.011, 0.128)
p_{33}	0.437	0.437	0.043	(0.356, 0.524)
AP	0.328	0.321	0.108	(0.133, 0.550)
DP	0.672	0.679	0.108	(0.447, 0.866)
SAP	0.174	0.168	0.073	(0.051, 0.334)
SDP	0.132	0.130	0.051	(0.039, 0.240)
CAP	0.126	0.119	0.057	(0.041, 0.239)
CET	0.143	0.144	0.050	(0.048, 0.238)

5. Application to Glaucoma Data

5.1 *The Results*

We apply the estimation method to the Glaucoma data on 228 patients in the CIGTS for whom S, T, and Z are completely observed. The observed counts are: $r_{000} = 28$, $r_{001} = 29$, $r_{010} = 14$, $r_{011} = 55$, $r_{100} = 11$, $r_{101} = 8$, $r_{110} = 10$, and $r_{111} = 73$. In Table 2, we report the means, medians, and 95% credible intervals (CI) from the posterior distributions of the counterfactual probabilities and surrogacy measures. We choose $a = 0.001$, $b = 1000$, $u = 0.7$, $v^2 = 9/4$. The posterior means and medians are similar. The estimated CET ($\hat{p}_{+2}$) has its mean(95% CI) as 0.14(0.05,0.24). Without the counterfactual model, we estimate CET directly from the observed data as $\hat{P}(T = 1 \mid Z = 1) - \hat{P}(T = 1 \mid Z = 0) = 0.13$ with its 95% confidence interval (0.014,0.25). The similarity between the two CET estimates suggests the goodness of fit of the counterfactual model and the slight difference may result from the prior assumptions. The mean(95% CI) for AP is 0.328(0.133, 0.550). It shows that about one-third of the CE on T is reflected by that on S; however, the wide CI implies that AP is quite variable. SAP is estimated as 0.174(0.051, 0.334) indicating among the patients whose S_i's are responsive to Z_i, only about 17% of their T_i's are also responsive. As expected, CAP is smaller than either AP or SAP and estimated as 0.126(0.041, 0.239) showing that S is far from satisfying the perfect principal surrogacy. In a conventional model setup, the estimated proportion of treatment effect explained, $\hat{F}_{WT}$, has the mean of 0.732 and median of 0.588 with its 95% bootstrap confidence interval of (0.17, 2.51). The correlation coefficients between S and T in the medicine and surgery groups are 0.304 and 0.441, respectively. The estimated OR and its 95% confidence interval between S and T in the medicine group is $\mathrm{OR}_{g0} = 3.79(1.73, 8.30)$ and that in the surgery group is $\mathrm{OR}_{g1} = 10.04(3.26, 30.93)$. It indicates that the IOP at the 12th month is a good surrogate for that at the 96th month in a conventional model setting, although the association between the CE on S and that on T is small.

5.2 *Sensitivity of Priors*

In Figure 1, we evaluate identifiability by plotting the prior and posterior distributions against each other (Garrett and Zeger, 2000), where $u = 0.7$ and $v^2 = 9/4$. Generally, the more substantial the average overlap and the more similarity between the prior and posterior is, the less identifiable the parameter is likely to be. We find that p_{11}, p_{33} and the CEs are more identifiable than p_{13}, p_{21}, p_{31}, and p_{32}. The counterfactual surrogacy measures are moderately identifiable. OR_2 and OR_3 appear to be least identifiable.

To further assess the extent of the impact of the priors on the posterior distributions on the counterfactual probabilities and surrogacy measures, we vary u of the prior $N(u, v^2)$ and fix v^2. Then, we vary v^2 but fix u. The results are listed in the Web Appendix. When we change u, we observe bigger changes in the posterior means than the posterior standard deviations (PSD). Relative to those when $u = 0.7$, with $u = 0$ or $u = 1.4$, the extent of the changes in the posterior means is less than 6% for most of the probabilities and surrogacy measures except for p_{31} and p_{13}. When we change v^2, we observe more changes in PSDs than in the posterior means. Compared with those when $v^2 = 9/4$, with $v^2 = 1$ or $v^2 = 4$, the changes in PSDs are generally less than 15%. Overall, the quantities of interest are not overly sensitive to the prior specifications.

6. Simulation Study

We conduct a simulation study to examine the frequentist properties of the estimates. We simulate 100 data sets under the parameter specification: $\lambda_{1S} = 0.15$, $\lambda_{1T} = -0.3$, $\lambda_{3S} = 0.3$, $\lambda_{3T} = -0.7$, $\lambda_{11} = 0.5$, $\lambda_{13} = -0.8$, $\lambda_{31} = -0.5$, and $\lambda_{33} = 0.8$. We vary u, v^2, and λ: (u = true, $v^2 = 9/4$), ($u = 0$, $v^2 = 1/64$), and $\lambda = 2$, 3.5, 7 where "true" refers to the true parameter value and λ controls the sample size. The simulation results from (u = true, $v^2 = 9/4$, $\lambda = 3.5$) are listed in Table 3 and the others are in the Web Appendix. The quantity $\mathrm{SD}(\overline{Est})$ refers to the standard deviation of Bayesian estimates and $\overline{PSD}$ is the mean of PSDs. Both posterior means and medians have very little bias. For the less identifiable parameters, $\mathrm{SD}(\overline{Est})$ is usually smaller than $\overline{PSD}$. When the 5–95 percentile ranges of the priors include all the true values, we consistently observe over-coverage regardless of the sample size. However, when the 5–95 percentile ranges of the priors do not include the true values, we may observe extreme under-coverage or over-coverage. As the sample size increases, the performance typically becomes better as the influence of the priors becomes smaller, but we do not usually have nominal large-sample coverage rates. On the other hand, regardless of priors, for the identifiable parameters, the coverage rates usually approach the nominal levels as the sample size increases. These findings are different from the situations for the identifiable models where Bayesian CIs can usually asymptotically match frequentist coverage; however, they are consistent with the literature for nonidentifiable models (Gustafson, 2005; McCandless, Gustafson and Levy, 2007).

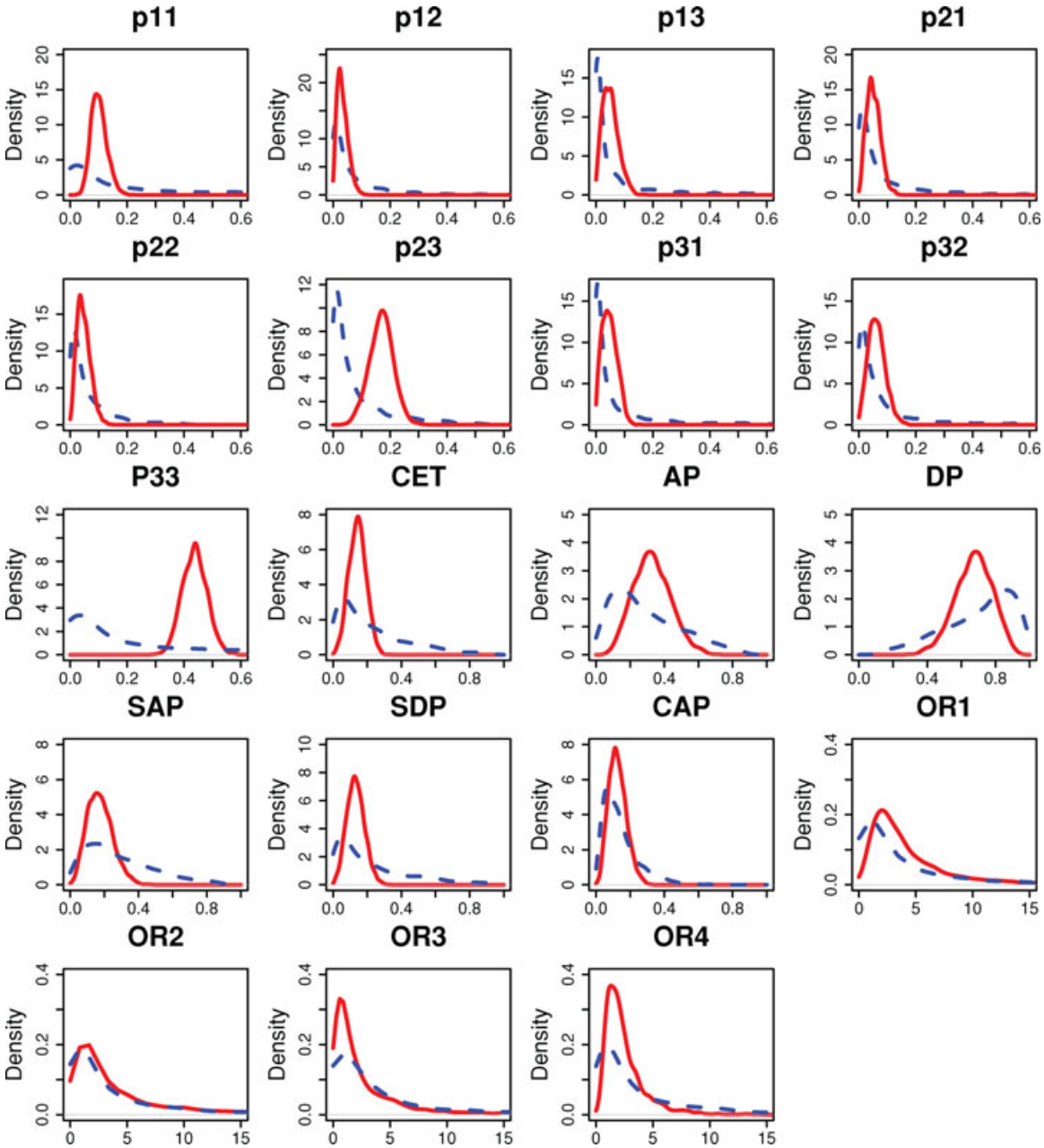

Figure 1. Prior and posterior distributions on selected quantities of interest. AP = associative proportion; DP = dissociative proportion; SAP = surrogate associative proportion; SDP = surrogate dissociative proportion; CAP = common associative proportion; CET = causal effect on T. Dashed lines for the prior distributions and solid lines for the posterior distributions. This figure appears in color in the electronic version of this article.

7. Surrogacy Measures in the Counterfactual and Conventional Models

7.1 *Perfect Statistical Surrogacy and Perfect Principal Surrogacy*

The perfect statistical surrogacy requires that T and Z are conditionally independent given S. In the causal framework, when CAP = 1, S satisfies perfect principal surrogacy. For S to be meaningful, we require that $p_{22} > 0$. When CAP = 1, we have $p_{12} = p_{21} = p_{23} = p_{32} = 0$, and thus $P(T = 1 \mid S = 0, Z = 0) = \frac{p_{13}}{p_{11}+p_{22}+p_{13}}$, $P(T = 1 \mid S = 0, Z = 1) = \frac{p_{13}}{p_{11}+p_{13}}$, $P(T = 1 \mid S = 1, Z = 0) = \frac{p_{33}}{p_{31}+p_{33}}$, and $P(T = 1 \mid S = 1, Z = 1) = \frac{p_{33}+p_{22}}{p_{31}+p_{33}+p_{22}}$. We consider two scenarios when CAP = 1. Scenario (1): when $p_{13} = p_{31} = 0$, $(S(0), S(1)) = (T(0), T(1))$, $S(Z) = T(Z)$; that is, S and T are identical. In this trivial scenario, S satisfies both perfect principal surrogacy and perfect statistical surrogacy. Scenario (2): when p_{13} and p_{31} are nonzero, T and Z are not conditionally independent given S; as such, S does not satisfy perfect statistical surrogacy. However, this situation seems less plausible. A marker tends to be chosen as S because

Table 3

Bias, standard deviation (SD) of posterior estimates, mean of posterior standard deviations ($\overline{PSD}$), and coverage rates from 100 *simulations. Posterior estimates (Est) are either posterior medians or posterior means. PSD = posterior standard deviation; AP = associative proportion; DP = dissociative proportion; SAP = surrogate associative proportion; SDP = surrogate dissociative proportion; CAP = common associative proportion; CET = causal effect on T. Prior specifications: $G(0.001, 1000)$ for $\exp(\lambda)$ and $N(true, 9/4)$ for all other parameters where "true" refers to the true parameter values. The parameter specification: $\lambda_{1S} = 0.15$, $\lambda_{1T} = -0.3$, $\lambda_{3S} = 0.3$, $\lambda_{3T} = -0.7$, $\lambda_{11} = 0.5$, $\lambda_{13} = -0.8$, $\lambda_{31} = -0.5$, $\lambda_{33} = 0.8$ and $\lambda = 3.5$. $E(r) = 565$.*

		Prior Distributions		Est: Median		Est: Mean			
	TRUE	2.5%	97.5%	Bias	SD	Bias	SD	$\overline{\text{PSD}}$	95% Coverage
p_{11}	0.166	0.001	0.815	0.003	0.022	0.003	0.022	0.022	95
p_{12}	0.136	0.001	0.476	−0.001	0.021	−0.001	0.022	0.027	98
p_{13}	0.030	0.000	0.520	−0.001	0.009	0.002	0.008	0.020	100
p_{21}	0.087	0.002	0.235	−0.001	0.014	0.000	0.013	0.032	100
p_{22}	0.117	0.002	0.309	−0.007	0.020	−0.005	0.020	0.033	100
p_{23}	0.058	0.001	0.301	−0.003	0.012	−0.002	0.011	0.021	99
p_{31}	0.071	0.000	0.618	−0.001	0.019	0.000	0.017	0.032	100
p_{32}	0.158	0.002	0.495	−0.002	0.022	−0.002	0.021	0.035	100
p_{33}	0.175	0.001	0.892	0.004	0.022	0.005	0.022	0.023	96
AP	0.285	0.024	0.711	−0.011	0.037	−0.008	0.036	0.074	100
DP	0.715	0.288	0.975	0.011	0.037	0.008	0.036	0.074	100
SAP	0.447	0.090	0.611	−0.011	0.041	−0.009	0.040	0.104	100
SDP	0.399	0.020	0.777	−0.006	0.042	−0.006	0.041	0.051	100
CAP	0.211	0.022	0.401	−0.010	0.024	−0.005	0.024	0.060	100
CET	0.412	0.025	0.711	−0.007	0.041	−0.007	0.041	0.038	95

there is a strong biological mechanistic evidence that it is linked to T. A likely positive association between $(S(0), S(1))$ and $(T(0), T(1))$ implies that p_{31} and p_{13} are likely smaller than other probabilities. Hence, p_{13} and p_{31} would be zeros or close to zeros when $p_{12} = p_{21} = p_{23} = p_{32} = 0$. The fact that perfect principal surrogacy precludes perfect statistical surrogacy holds only in this implausible scenario.

7.2 *Surrogacy Measures Under Two Hypothetical Examples*

In a conventional model setup, under the monotonicity assumption, we can express the elements of F_{WT} using the counterfactual probabilities as follows: $\delta = p_{21} + p_{22} + p_{23}$, $\tau = p_{12} + p_{22} + p_{32}$ and $\gamma_a = \frac{p_{33}}{p_{31}+p_{32}+p_{33}} - \frac{p_{13}+p_{23}}{p_{11}+p_{12}+p_{21}+p_{22}+p_{13}+p_{23}}$. Similarly, for the odds ratios, we have $\text{OR}_{g0} = \frac{(p_{11}+p_{12}+p_{21}+p_{22})p_{33}}{(p_{13}+p_{23})(p_{31}+p_{32})}$ and $\text{OR}_{g1} = \frac{p_{11}(p_{22}+p_{23}+p_{32}+p_{33})}{(p_{12}+p_{13})(p_{21}+p_{31})}$.

To better understand the surrogacy measures in both traditional and counterfactual model settings with respect to the underlying causal associations, we calculate the surrogacy measures in two hypothetical examples (Table 4). In Example 1, when the CE on T is the same across three principal strata, CAP, SAP, and AP are relatively small indicating a small causal association between S and T; however, the large values in F_{WT}, OR_{g0}, and OR_{g1} show that S is closely related to T in a conventional model setup.

In Example 2, all surrogacy measures indicate a close relationship between S and T in both traditional and counterfactual model framework. In general, when p_{11} and p_{33} are relatively large compared with the off-diagonal probabilities in the same rows and columns, S is highly associated with T in a traditional model setup. When p_{22} is relative large compared with the off-diagonal probabilities in the same row and column, S is closely associated with T in a counterfactual framework. Although a thorough investigation of the critical values and the variability of the counterfactual surrogacy measures and their connections with F_{WT} and ORs is beyond the scope of this article, it would be very useful as future research.

Table 4

Two hypothetical numerical examples. AP = associative proportion; DP = dissociative proportion; SAP = surrogate associative proportion; SDP = surrogate dissociative proportion; CAP = common associative proportion; CET = causal effect on T. Example 1: $AP = 1/3$, $SAP = 0.20$, $DP = 2/3$, $SDP = 0.20$, $CAP = 0.14$, $F_{WT} = 1.00$, $OR_{g0} = 16$, $OR_{g1} = 16$*; Example* 2: $AP = 0.77$, $SAP = 0.77$, $DP = 0.23$, $SDP = 0.10$, $CAP = 0.63$, $F_{WT} = 0.80$, $OR_{g0} = 90$, $OR_{g1} = 157$.

Potential Outcomes	$(T(0), T(1))$			
$(S(0), S(1))$	(0, 0)	(0, 1)	(1, 1)	Marginal
Example 1				
(0, 0)	0.267	0.066	0.001	0.334
(0, 1)	0.133	0.066	0.133	0.332
(1, 1)	0.001	0.066	0.267	0.334
Example 2				
(0, 0)	0.310	0.030	0.005	0.345
(0, 1)	0.030	0.240	0.040	0.310
(1, 1)	0.005	0.040	0.300	0.345

8. Discussion

This article examines the association between the effect of Z on S and that on T, as if we had observed both outcomes of S and T corresponding to two treatment options for every patient. Different from those based on the traditional models, the associations between $(S(0), S(1))$ and $(T(0), T(1))$ can not be changed by the treatment assignment and always have causal interpretations. The traditional models also ignore the fact that the effect of Z on T may occur to the patients who are inherently never-responsive or always-responsive in S regardless of the treatment received, however, the counterfactual model teases out the effect of Z on T in each subgroup of subjects defined by their responsiveness in S to the treatment received. The causal framework used here is similar in spirit to that used in the compliance literature (Balke and Pearl, 1997; Imbens and Rubin, 1997) where the main interest is to estimate the CE of a treatment for the compliers.

We use a log-linear model to directly model the association between the potential outcomes of S and T through the odds ratios of $(S(0), S(1))$ and $(T(0), T(1))$. We believe that there is an ordering in the sequence of the potential outcomes of (0, 0), (0, 1), and (1, 1). With our model setup, the scientific assumptions can be conveniently incorporated through the prior distributions for the odds ratios, for which there is little information from the observed data. The proposed estimation method can be readily extended to the settings when T is partially missing or when there are multiple trials. Besides the log-linear model, we also fit a multinomial model with Dirichlet priors. Although it is easier computationally, the model is less flexible and the impact of the priors on the estimable quantities such as the treatment effect on T is much larger than the log-linear model. Like the multinomial models or logistic regressions for contingency tables, the probabilities based on the log-linear model are required to be positive and as such we cannot test whether S is a perfect principal surrogate. Nonetheless, in practice, it is almost certain that no surrogate exists that either satisfies perfect principal surrogacy or perfect statistical surrogacy.

We adopt the framework proposed by FR. Robins and Greenland (1992) (RG) proposed another counterfactual framework that allows one to manipulate S. It requires additional probabilities to describe the likelihood of how T changes by changing S. This framework has been used by Chen, Geng, and Jia (2007) and Taylor et al. (2005) to study the surrogacy consistency. RG defined direct and indirect effects where the indirect effect is the part of the effect that Z affects T by affecting S and direct effect is the part not through this pathway. The relationships between the direct/indirect effect proposed by RG and the associative/dissociative effect by FR are explored in depth by VanderWeele (2008) and Joffe and Greene (2009). While the elaboration of the relationships is beyond the scope of this article, we know that if S is in the causal pathway between Z and T, p_{22} is large. On the other hand, a very high p_{22} only shows that the CE of Z on S is highly associated with that on T but it does not necessarily imply that Z affects T by affecting S.

One of the key assumptions is monotonicity that is useful and necessary to reduce the number of parameters to have a more identifiable counterfactual model. If this assumption is correctly specified, we expect our estimates to be more efficient and less biased than those based on the conventional model. However, this assumption requires that every single patient would have done at least as well as that when she or he receives $Z = 1$ relative to that when she or he receives $Z = 0$. It is perhaps true for most of the patients but not usually satisfied for all patients. For example, in the CIGTS study, it is conceivable that some patients may be better if they received medicine instead of surgery, even though the average effect of surgery is consistently better. Assessing the impact of the violations of the monotonicity assumption would be an important extension.

We assumed that missingness is ignorable, and it will be useful to conduct sensitivity analysis to investigate this assumption. It will also be useful to calculate the nonparametric bounds free of the prior assumptions and quantify the ranges of the counterfactual probabilities in our context (Balke and Pearl, 1997). Extensions to other data types are possible. Some work has been done by Gilbert and Hudgens (2008) whose proposal of CE predictiveness surface as a surrogacy measure can be applied to different types of outcomes and by Gallop et al. (2009) who considered a normally distributed outcome with a binary mediator that can be easily adapted to the surrogacy setting. Nonetheless, the majority of the literature has focused on binary endpoints. We advocate the need for research on more complex data structure, for example, two failure-time endpoints in oncology, where the settings are more common and can be more important.

9. Supplementary Materials

The Web Appendix referenced in Sections 5.2 and 6 is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors are grateful for the helpful comments made by the reviewers. They would also like to thank Dr Brenda Gillespie for providing the CIGTS data. This research was supported by National Institutes of Health Grants MH078016 and CA129102.

References

AGIS Investigators. (2000). The Advanced Glaucoma Intervention Study (AGIS): 7: The relationship between control of intraocular pressure and visual field deterioration. *American Journal of Ophthalmology* **130,** 429–440.

Alonso, A. and Molenberghs, G. (2003). Surrogate marker evaluation from an information theory perspective. *Biometrics* **63,** 180–186.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92,** 1171–1176.

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

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1,** 49–68.

Chen, H., Geng, Z., and Jia, J. (2007). Criteria for surrogate endpoints. *Journal of the Royal Statistical Society, Series B* **69,** 919–932.

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

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

Gallop, R., Small, D., Lin, J., Elliott, E. R., Joffe, M. M., and Ten Have, T. (2009). Mediation analysis with principal stratification. *Statistics in Medicine* **28,** 1108–1130.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics* **56,** 1055–1067.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* **94,** 247–253.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. New York: Chapman and Hall.

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

Green, P. E. and Park, T. (2003). A Bayesian hierarchical model for categorical data with nonignorable nonresponse. *Biometrics* **59,** 886–896.

Gustafson P. (2005). On model expansion, model contraction, identifiability, and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science* **20,** 111–140.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* **25,** 305–327.

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

King, R. and Brooks, S. P. (2001). Prior induction in log-linear models for general contingency table analysis. *The Annals of Statistics* **29,** 715–747.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd edition. New York: Wiley.

McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine* **26,** 2331–2347.

Musch, D. C., Gillespie, B. W., Lichter, P. R., Niziol, L. M., and Janz, N. K., and CIGTS Study Investigators. (2009). Visual field progression in the Collaborative Initial Glaucoma Treatment Study: The impact of treatment and other baseline factors. *Ophthalmology* **116,** 200–207.

Musch, D. C., Lichter, P. R., Guire, K. E., Standardi, C. L., and CIGTS Investigators. (1999). The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology* **106,** 653–662.

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

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology* **3,** 143–155.

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

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

Rubin, D. B. (1980). Randomization analysis of experimental-data—the Fisher randomziation test—comment. *Journal of the American Statistical Association* **75,** 591–593.

Taylor, J. M. G., Wang, Y., and Thiébaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61,** 1102–1111.

VanderWeele, T. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters* **78,** 2957–2962.

Venkatraman, E. S. and Begg, C. B. (1999). Properties of A nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics* **55,** 1171–1176.

Wang, Y. and Taylor, J. M. G. (2003). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **58,** 803–812.

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

Received August 2008. *Revised April* 2009.
Accepted May 2009.