Published in final edited form as:
Stat Med. 2010 December 10; 29(28): 2869–2879. doi:10.1002/sim.4027.

Links between analysis of surrogate endpoints and endogeneity

Debashis Ghosh[1], **Michael R. Elliott**[2], and **Jeremy M. G. Taylor**[2]

[1]Departments of Statistics and Public Health Sciences, Penn State University, University Park, PA

[2]Department of Biostatistics, University of Michigan, Ann Arbor, MI

Summary

There has been substantive interest in the assessment of surrogate endpoints in medical research. These are measures which could potentially replace "true" endpoints in clinical trials and lead to studies that require less follow-up. Recent research in the area has focused on assessments using causal inference frameworks. Beginning with a simple model for associating the surrogate and true endpoints in the population, we approach the problem as one of endogenous covariates. An instrumental variables estimator and general two-stage algorithm is proposed. Existing surrogacy frameworks are then evaluated in the context of the model. In addition, we define an extended relative effect estimator as well as a sensitivity analysis for assessing what we term the treatment instrumentality assumption. A numerical example is used to illustrate the methodology.

Keywords

Clinical Trial; Counterfactual; Nonlinear response; Prentice Criterion; Structural equations model

1 Introduction

There has been great interest in developing and validating surrogate endpoints in clinical research. Surrogate outcomes are measures that have been proposed based on biological considerations within a progression model of disease. One example is CD4 count levels in AIDS; the CD4 count can potentially serve as a surrogate endpoint for death. Another example from cancer studies is using tumor shrinkage as a surrogate endpoint for survival or disease-free survival. Also, high-throughput assays based on genomics and proteomics are generating molecular profiles that investigators may consider as surrogate endpoints in the future. It is thus important to have methods and frameworks for evaluating surrogate endpoints.

Research on surrogate endpoints has been quite intense over the last twenty years. A seminal paper in the evaluation of surrogate endpoints is that of Prentice [1]; he species conditions under which a test of a treatment effect on the surrogate endpoint provides valid inference for testing for the effect of treatment on the true endpoint. Because the so-called Prentice criterion is a difficult one to validate, many authors have focused on alternative measures of assessing surrogacy. A comprehensive discussion of them can be found in the recent monograph by Burzykowski et al. [2].

More recent research on surrogacy has extended in two directions. The first direction is on using information from multiple related trials to assess surrogacy [3,4]. There are two notions of surrogacy here, trial-level and individual-level surrogacy, both of which can potentially be used to evaluate surrogate endpoints.

The second direction of research on surrogacy in the statistical literature has been utilizing ideas of causal inference to the assessment of surrogacy. There are two different approaches to surrogacy that currently appear to dominate the literature. The first approach was suggested in a more general framework by Robins and Greenland [5]. In their work, the surrogate endpoint is an intermediate variable measured after the baseline covariates and before the outcome; this variable is manipulable and can effect the outcome independently of the treatment. Robins and Greenland then describe causal effects of the treatment on the outcome and of the surrogate endpoint on the outcome, termed direct and indirect effects of treatment. This type of model has been advocated more recently by Pearl [6].

Frangakis and Rubin [7] suggested a different counterfactual model and proposed the concept of principal stratification for enabling causal inference. The framework they suggested borrows ideas from the compliance literature. In their framework, potential outcomes are formulated jointly for the surrogate and true endpoints under each of the treatment arms. They define prinicipal strata as contrasts within individuals under different conditions on the counterfactuals for the surrogate endpoint. A major difference between this framework from that of Robins and Greenland [7] is that the surrogate endpoint is only manipulable by manipulating the treatment.

More recently, several authors have extended the two frameworks for assessing surrogacy. Gilbert and Hudgens [8] proposed some estimation and inference procedures for quantifying surrogacy within the principal stratification framework. Li et al. [9] also develop methods for quantifying surrogacy within the frameworks of Robins and Greenland [5] and Frangakis and Rubin [7]. Taylor et al. [10] identified conditions from the two frameworks that yield equivalences with the analysis approach of Freedman et al. [11]. Joffe and Greene [12] give a more general framework that synthesizes and compares various existing surrogacy frameworks.

In this note, we approach the problem of surrogacy from a viewpoint of endogeneity. Endogenous covariates have also been alluded to as internal covariates in survival analysis ([13],§6.3) and have been heavily considered in the econometric literature (e.g., [14], p. 31). By utilizing this framework, we will come to the following conclusions:

1. Thinking of endogeneity of the surrogate is compatible with the framework of Robins and Greenland [5].
2. If treatment assignment can be viewed as an instrumental variable, the relative effect, a measure proposed by Buyse and Molenberghs [15], can be viewed as an instrumental variables estimator for the causal effect of the surrogate endpoint S on the true endpoint T in a linear model setting.
3. If treatment assignment can be viewed as an instrumental variable, under certain conditions, the Prentice criteria yield valid causal inferential results.
4. A general two-stage estimation procedure can be used for estimation of the causal effect of the surrogate on the true endpoint under various distributional assumptions of S and T in the presence of an instrumental variable. The two-stage procedure leads to a different estimator for the relative effect in the nonlinear setting compared to the one proposed in [15].
5. The Prentice criteria are not purely causal in nature.

The organization of the paper is as follows. In Section 2, we describe the data structures and present a structural equation model for describing the relationship between the surrogate and the true endpoint in the population. We then define instrumental variables, which are commonly used in econometrics to handle the endogeneity issue. It turns out that for a

clinical trial, the randomized treatment assignment is a natural candidate for an instrument. This can then be used for estimation via a two-stage procedure, which we outline in Section 2. In Section 3, we relate the proposed structural equation model with several frameworks for surrogacy. We describe extension of the two-stage estimation procedure to noncontinuous outcomes in Section 4. There, a numerical example is used to illustrate the methodology. We conclude with some discussion in Section 5.

2 Structural Model, Background and Instrumental Variables Estimation

Define S to be the surrogate endpoint, T to be the true endpoint, and Z to be a treatment indicator. For the sake of exposition, we will initally assume that S and T are continuous. We propose a structural equations model for associating S and T:

$$T = \alpha_0 = \alpha S + \epsilon_s, \tag{1}$$

where α_0 is an unknown intercept term, α is an unknown regression coefficient, and ε_S is an error term with an unknown distribution. While such a model has a form similar to that of simple linear regression, we take (1) to be a structural model. As described in Pearl [6], setting S and ε_S yields a counterfactual value for T in (1); we can then view α as the causal effect of S on T . Because S take values in $(-\infty, \infty)$, we actually have uncountably many potential outcomes. A structural equations model such as (1) provides a different approach to formulating potential outcomes relative to the framework outlined in Frangakis and Rubin [7]. Note that in model (1), Z appears nowhere in the equation. Thus, treatment group does not figure into interpretation of the structural equation model. The goal of the model is to understand how manipulating the surrogate endpoint impacts the true endpoint in the general population. As we shall see later, Z only figures in the estimation of the parameter α.

We also mention that the target estimand here is the causal effect of the surrogate endpoint on the true endpoint, α. This is quite different from much of the statistical literature on the topic, which attempts to assess the causal effect of treatment in the study population or in a subpopulation, such as those who comply with the treatment. Note that use of model (1) requires the concept of S being manipulable. In the language of Pearl [6], model (1) represents a model in which a manipulation operator is applied to S. In a clinical trial setting, S is observed post-treatment. Thus, model (1) is compatible with the Robins/Greenland framework; it is only compatible with the principal stratification framework of Frangakis and Rubin [7] in the case that $\alpha_1 = \alpha = 0$, i.e. there is no causal effect of S on T so that the issue of the manipulability of the surrogate endpoint is moot.

Endogeneity deals with the fact that in (1), S and ε_S will not be independent. What this means is that the ordinary least squares estimator of T on S will yield a biased estimator of α. To understand what is going on, it is convenient to consider directed acyclic graph models, as presented in Pearl [16]. Briefly speaking, a graphical model consists of vertices representing the variables of interest, and edges that describe associations between them. Statements about causality between variables correspond to arrows in the edges in the direction of the cause to the effect. For our purposes, a very simple graphical model suffices to describe the relationship in (1). In this structural model, S leads to T , but there is confounding of this relationship by U, which represent unobserved confounders. This is called “selection bias” in economics in that there are unmeasured factors that predict S that are independently predictive of T.

What is needed to estimate α in (1) is an instrumental variable or an instrument. An instrumental variable (or instrument) is a random variable W for which the correlation of W and ε_S is zero but for which the correlation of W and T is nonzero. Such a variable can then

be used for estimating α in (1). There are two equivalent characterizations for doing this. The first is given by the following algorithm:

1. Regress S on W.
2. Regress T on $\widehat{S}$, the fitted values from the previous step.

The estimated slope coefficient from the second step is the instrumental variables estimator for α. Note that the estimators for the two steps can be obtained using least squares. We can equivalently replace step 2 with including the residuals from step 1 in addition to S. The algorithm we have described here is known as two-stage least squares [17]. Also note that what is being regressed on in the second step is the estimate of $E(S|W)$. What the instrumental variable does is to serve as a proxy for S while at the same time being independent of ε_S from (1).

It turns out that an alternate characterization of the the instrumental variables estimator of α is given by

$$\widehat{\alpha}_T = \frac{\widehat{cov}(T, W)}{\widehat{cov}(S, W)}, \tag{2}$$

where $\widehat{cov}$ denotes the empirical covariance operator. Formula (2) is a well-known one and dates back to the work of Wright [18]. Because the formula (2) represents a ratio estimator, it will also be the case that the confidence intervals associated with the estimator will tend to be fairly variable.

Let us again use graphical models to illustrate what is going on. Relative to the graph in Figure 1, we have added one more node for W, and by assumptions, directed edges from W to S but neither to U nor T. For W to be a valid instrument, it must be independent of U; in addition, it can only effect T through S. Intuitively, the effect α is inferred indirectly through estimation of the effects of W on S and T. Note that the scientific goal is to assess the causal effect of S on T through model (1). The role of the instrumental variable is to estimate the causal effect in a manner that is consistent for α.

In most clinical trials settings, the treatment assignment, Z, is randomized. This leads to the question of whether or not Z can be used as an instrument for assessing the causal effect in (1). We expect that randomization should balance the distribution of both observed and unobserved covariates so that Z would be uncorrelated with U. However, this assumption is not empirically testable. We also need that Z and S be correlated, which will be empirically verifiable. Finally, we need that the effect of Z on T to be fully mediated through S. If these three assumptions hold, then we term this *treatment instrumenality*. A simple situation where the assumption is violated is if Z were to have effects on T that were not fully mediated through S. If we take W to be Z in the graphical model corresponding to Figure 2, then the situation just described would require another edge from Z to T. Thus, Z has both a direct effect on T as well as an indirect effect through S. For this situation, Z could no longer be an instrumental variable for estimating α in (1). In fact, the structural model (1) would need to be expanded to $T = \beta^* + \alpha^* S + \gamma^* Z + \epsilon_S^*$, assuming no interactions between Z and S. In Section 4.2., we describe an approach to performing sensitivity analysis in order to assess the treatment instrumentality assumption.

3 Comparison with other frameworks

3.1 Prentice criteria: an instrumental variables interpretation

We now seek to relate other frameworks of surrogacy with the model proposed in the last section. We begin with the criteria set forth in Prentice (1989). They are given by the following:

- **a.** $f(S|Z) \neq f(S)$;
- **b.** $f(T|Z) \neq f(T)$;
- **c.** $f(S,T) \neq f(S)f(T)$;
- **d.** $f(T|S, Z) = f(T|S)$.

In words, criterion (a) says that the surrogate endpoint is associated with treatment. Criterion (b) states that the true endpoint is associated with treatment. Criterion (c) is that the surrogate and true endpoints are associated, while the last criterion is that given the surrogate endpoint, treatment and the true endpoint are independent. Popularly, the last criterion (d) is referred to as the Prentice criterion although in fact all four criteria were originally proposed by Prentice.

We now wish to interpret the Prentice criteria in the context of the structural model from the previous section. Of (a)-(d), the only conditions that pertains solely to the model is (c). What this says is that S and T are not independent, which is consistent with (1) provided $\alpha \neq 0$. This assumption implies that there is a nonzero causal effect of S on T to be estimated. The other criteria involve the instrument. If criterion (a) of the Prentice framework is violated, then this will manifest itself in (2) being close to zero so that $\widehat{\alpha}$ will be unstable. If criterion (b) is violated, then the numerator in (2) will be close to zero.

The last criterion, (d), is more problematic. The right-hand side of the equality pertains solely to the model (1). However, the left-hand side involves both the model and the instrumental variable. As previously mentioned, the role of instrumental variable is only to aid in estimation and is orthogonal to the structural model. In effect, (d) ends up mixing criteria and conditions about the instrument with those regarding the model. While criticisms of the work of Prentice have been presented before [19,7,20] , the argument presented here based on the structural equations approach appears to be novel.

In a related vein, based on Figure 2, models for T given S and Z pose problems from a causal inference point of view. One example is the measure proposed in [11], the proportion of treatment effect explained (PTE). This is calculated by fitting the following two regression models:

$$T=\eta_0+\eta_T S+\gamma_T Z+\epsilon_1, \tag{3}$$

and

$$T=\beta_0+\beta_T Z+\epsilon_2, \tag{4}$$

where ε_1 and ε_2 are error terms, (η_0, β_0) are unknown intercept terms, and β_T and β_S are unknown regression coefficients to be estimated. The PTE is given by the following formula:

$$PTE=\frac{\beta_T-\gamma_T}{\beta_T}, \tag{5}$$

Conceptual criticisms of the PTE have been given in [21]. It is not well-defined when there are interactions between *S* and *Z*, and the estimator of the PTE and its associated 95%CI can fall outside the range [0, 1]. Our critique here of the PTE refers to its causal validity. Let us consider Figure 2 again. In particular, conditioning on *S* will induce association between *U* and *Z*. This means that even if *S* fully mediates the effect of *Z* on *T* , it will not be the case that *Z* is independent of *T* given *S* because of the induced association between *U* and *Z*. In graphical model terminology, *S* is a node with converging arrows (one from *Z* and one from *U*), and conditioning on *S* induces confounding. Pearl [6] has termed *S* a collider, and from his results, conditioning on a collider induces confounding. This point has been discussed nicely in Section 2 of Joffe and Greene [12]. If there exists a set of covariates *X* such that *S* is independent of *U* given *X*, then a modification of (3) in which one includes *X* as covariates on the right-hand side will yield valid causal inferences about α.

3.2 Buyse and Molenberghs criteria

Another set of surrogacy criteria was proposed by Buyse and Molenberghs [15]. They consider two quantities, termed the relative effect (RE) and the adjusted association (AA). The relative effect is a ratio of the treatment effects on the true and surrogate endpoints on the true outcome. The adjusted association is the correlation between the true and surrogate endpoints after adjusting for the treatment effect; generically, AA = cor($T|Z$, $S|Z$). In the single-trial setting, [15] consider likely surrogates to be variables with relative effects near one as well as strong adjusted associations. Note that for *RE* = 1 to have an interpretation, this requires T and S to be measured on the same scale.

Of these two quantities, the relative effect fits in more cleanly with (1). Suppose we fit the following regression models for the two endpoints as a function of treatment: (4) and

$$S=\beta_{S0}+\beta_S Z+\epsilon_3, \tag{6}$$

where ε_3 is an error term, and (β_{S0}, β_S) are unknown regression coefficients to be estimated. If treatment instrumentality holds, then (2) will equal the estimator of relative effect. The proof of this follows from the fact that the population version of the relative effect, based on (4) and (6), is defined to be

$$\mathrm{RE}=\frac{\beta_T}{\beta_S}. \tag{7}$$

Given the linear regression forms for (4) and (6), the regression coefficients can be viewed as population versions of correlation coefficients between predictor and response. An empirical estimator of the relative effect then follows from replacing the quantities in (7) by their data-based counterparts, which yields (2).

While model (1) appears to be very similar in concept to the adjusted association, there is in fact a big difference between the two. Assume as before that the variances of *S* and *T* are equal and further that they are equal to one. Then the two-stage least squares algorithm shows that the estimator of α in (1) is given by the correlation between *T* and the residual from the regression of *S* on *Z*. By contrast, the adjusted association measure would calculate

the correlation between the residuals of the two regressions (4) and (6). It is the extra adjustment for Z in the first regression model that distinguishes the two-stage least squares from AA. Thus, while the RE uses information from the two regressions (4) and (6) that leads to estimation of a parameter with a causal interpretation, the AA does not.

3.3 Rubin Causal Model framework

In their foundational paper on principal stratification, Frangakis and Rubin [7] define principal strata as being formed by the joint distribution of a post-randomization adjustment variable under all possible treatment assignments, thus manufacturing a latent pre-randomization variable whose values must be independent of randomized assignment. They propose applying this methodology to the analysis of surrogacy data, where the values of the surrogate under each treatment assignment form the principal strata. Defining $S(Z)$ as the potential surrogate and $T(Z, S(Z))$ as the potential outcome associated with the potential surrogate value under a two-arm treatment $Z = \{0, 1\}$, Frangakis and Rubin define associative effects as comparisons between $T_i(1)$ and $T_i(0)$ among subjects for whom the treatment has a causal effect on the surrogate: $S_i(1) \neq S_i(0)$; and disassociative effects as comparisons between $T_i(1, S_i(1))$ and $T_i(0, S_i(0))$ among subjects for whom the treatment has no causal effect on the surrogate: $S_i(1) = S_i(0)$. Frangakis and Rubin argued that a "principal surrogate" was one in which dissassociative effects were 0, i.e. $S_i(1) - S_i(0) = 0$ implies $T_i(1, S_i(1)) - T_i(0, S_i(0)) = 0$.

It is difficult to relate the instrumental variable approach to the principal stratification approach in a general form, because principal stratification sacrifices the manipulability of S to condition on the pre-randomization values of $(S(0), S(1))$ in order to make meaningful causal interpretation. Nonetheless, IV estimators can be interpretated as principal stratification estimators under certain restrictions, as earlier work by Angrist, Imbens, and Rubin [22] discuss. In particular, if we formulate a linear structural model for S and T (in contrast to log S and log T from (1) and (2)), we can express the instrumental variable estimator for α as

$$\widehat{\alpha}_T = \frac{\Sigma_i T_i Z_i / \Sigma_i Z_i - \Sigma_i T_i (1 - X_i) / \Sigma_i (1 - Z_i)}{\Sigma_i S_i Z_i / \Sigma_i Z_i - \Sigma_i S_i (1 - X_i) / \Sigma_i (1 - Z_i)}.$$

This formula assumes that treatment instrumentality holds. We then have that, under randomization,

$$\widehat{\alpha}_T = \frac{1/n_1 \Sigma_{i:Z_i=1} T_i(1) - 1/n_0 \Sigma_{i:Z_i=1} T_i(0)}{1/n_1 \Sigma_{i:Z_i=1} S_i(1) - 1/n_0 \Sigma_{i:Z_i=1} S_i(0)},$$

or the ratio of the treatment effect on the outcome to the treatment effect on the surrogate. Following Angrist et al. [22], we assume that the surrogate is dichotomous and it does not perform "worse" under the treatment than under the control: $S_i(1) \geq S_i(0)$, the monotonicity assumption. If we further assume the exclusion restriction – that the effect of the treatment on the outcome must be entirely through the surrogate ($T(1, S(1) = s) = T(0, S(0) = s)$, and thus that $T(Z, S(Z)) = T(S(Z))$) – we have

$$\begin{aligned} E(T_i(1) - T_i(0)) &= \sum_{s=0}^{1} \sum_{s'=0}^{1} E\left(T_i(1, S(1) = s) - T_i\left(0, S(0) = s'\right) | S(1) = s, S(0) = s'\right) \times P\left(S(1) = s, S(0) = s'\right) \\ &= E(T_i(1, S(1) = 1) - T_i(0, S(0) = 0) | S(1) = 1, S(0) = s) P(S(1) = 1, S(0) = 0) \end{aligned}$$

where the first equality follows from the law of total probability and the second from the monotonicity and exclusion restriction assumptions. Thus $\widehat{\alpha}_T$ is a consistent estimator of

$$\frac{E(T_i(1)-T_i(0))}{E(S_i(1)-S_i(0))} = \frac{E(T_i(1,S(1)=1)-T_i(0,S(0)=0)|S(1)=1,S(0)=0)P(S(1)=1,S(0))}{P(S(1)=1,S(0))} = E(T_i(1,S(1)) - T_i(0,S_i(0))|S_i(1)=1,S(0)=0).$$

This is what Angrist et al. [22] term the local average treatment effect. In the special case where the surrogate is a (dichotomous) measure of compliance to treatment, this can be interpreted as the causal effect of the treatment among subjects who comply with their treatment under the assumptions of monotonicity (subjects never take the opposite of their treatment assignment) and the exclusion restriction (the effect of treatment assignment is only through actually taking the treatment). In the more general surrogacy setting, it can be interpreted as the causal effect of the treatment among subjects for whom the treatment changes the surrogate measure, subject to the monotonicity and exclusion restriction assumptions relevant to the particular setting.

Li et al. [9] explore principal stratification approach to assessing surrogacy in the special case of dichotomous surrogates and outcomes. Li et al. assume monotonicity for both the surrogate and the final outcome of interest: $S_i(1) \geq S_i(0)$ and $T_i(1, S(1)) \geq T_i(0, S(0)) \forall i$. Under this constraint, the potential surrogate and outcome form a 3 × 3 table, where π_{jk} is the proportion of the population for whom $S(0) = S(1) = 0$ for $j = 1$, $S(0) = 0$, $S(1) = 1$ for $j = 2$, and $S(0) = S(1) = 1$ for $j = 3$; and similarly for k and $T(Z, S(Z))$ (see Table 1). As in Taylor et al. [10], Li et al. define the associative proportion AP≡ $\pi_{22}(\pi_{12} + \pi_{22} + \pi_{32})^{-1}$ as the fraction of the population response to treatment that is also responsive to the surrogate and the disassociative proportion DP≡ $(\pi_{12} + \pi_{32})[\pi_{12} + \pi_{22} + \pi_{32}]^{-1}$ as the fraction of the population response to treatment that is not responsive to the surrogate. Under the proposed definition of principal surrogacy in [7], AP=1 and DP=0. Li et al. [9] argued, however, that a good principal surrogate should also consider the effect of the treatment on the outcome within the principal strata in which the treatment as a causal effect on the surrogate ($S(0) = 0$, $S(1) = 1$). Thus a good surrogate should have relatively few patients where the treatment affects the surrogate but not the outcome of interest, suggesting the common associative proportion CAP= $\pi_{22}(\pi_{12} + \pi_{21} + \pi_{22} + \pi_{23} + \pi_{32})^{-1}$ as a better measure of surrogacy.

Relating the work of Angrist et al. [22] to Li et al. [9], we see that the IV estimator $\widehat{\alpha}_T = (\pi_{12}+\pi_{22}+\pi_{32})(\pi_{21}+\pi_{22}+\pi_{23})^{-1}$ in this setting of dichotomous outcomes and surrogates. Here we ignore the fact that T is dichotomous, but the IV development assumes T to be continuous. Angrist [23] argues this has little impact on linear estimators of treatment effects such as those considered here, although Bhattacharya et al. [24] argues that this conclusion is highly sensitive to correct model specification. However, making the exclusion restriction assumption conditions the causal effect of the treatment on the outcome to the stratum in which the treatment has a causal effect on the surrogate, in which case $\widehat{\alpha}_T = \pi_{22}(\pi_{21}+\pi_{22}+\pi_{23})^{-1}$. This corresponds to what Li et al. termed the surrogate associative proportion (SAP). Note further that the exclusion restriction implies $\pi_{12} = \pi_{32} = 0$, and thus the SAP corresponds to the CAP under the monotonicity and exclusion restriction assumptions.

3.4 Meta-analysis for surrogacy

Another line of research has been in the setting of multiple trials (e.g., [3]). The authors typically consider the situation where S and T are measured in n trials and only S is measured in the $(n + 1)^{th}$ trial, with the goal being to assess the treatment effect on the true endpoint in the new trial based on all the available data. Hierarchical models are natural to

use in this setting and various measures of surrogacy have been suggested. In the setting of multiple trials suppose for each trial (S_{ij}, T_{ij}, Z_{ij}) represent the surrogate and true endpoints and treatment group, respectively, for individual j in trial i. We assume as before S_{ij} and T_{ij} are continuous. A natural formulation for joint modelling of the endpoints is through a bivariate mixed model [25]:

$$S_{ij}=\alpha_0+\alpha_1 Z_{ij}+\alpha_{0i}+\alpha_{1i}Z_{ij}+\epsilon_{Sij}$$
$$T_{ij}=\gamma_0+\gamma_1 Z_{ij}+r_{0i}+r_{1i}Z_{ij}+\epsilon_{Tij}$$

where

$$\begin{pmatrix}\epsilon_{sij}\\ \epsilon_{tij}\end{pmatrix}\sim MVN\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_{ss} & \sigma_{st}\\ & \sigma_{tt}\end{pmatrix}\right)$$

$$\begin{pmatrix}a_{0i}\\ r_{0i}\\ a_{1i}\\ r_{1i}\end{pmatrix}\sim\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix},\begin{pmatrix}d_{ss} & d_{st} & d_{sa} & d_{sr}\\ & d_{tt} & d_{ta} & d_{tr}\\ & & d_{aa} & d_{ar}\\ & & & d_{rr}\end{pmatrix}\right).$$

and *MVN* denotes the multivariate normal distribution. Using (8), Buyse et al. [25] developed the following measures of between-trial and within-trial association:

$$R^2_{trail}=\frac{(d_{sr}\quad d_{ar})\begin{pmatrix}d_{ss} & d_{sa}\\ d_{sa} & d_{aa}\end{pmatrix}^{-1}\begin{pmatrix}d_{sr}\\ d_{ar}\end{pmatrix}}{d_{rr}},\ R^2_{indiv}=\frac{\sigma^2_{st}}{\sigma_{ss}\sigma_{tt}}$$

In terms of model (1), the meta-analysis setting can be viewed as one in which individual study-specific variables constitute part of U in Figure 1. Of the two measures, the individual-level measure R^2_{indiv} is closer in spirit to α in (1) than R^2_{trial}. By definition, R^2_{indiv} captures the correlation between the surrogate and true endpoints after adjusting for both treatment and study effects. As the name implies, it is supposed to capture the correlation between the surrogate and true endpoints within an individual. Similarly, the causal effect α in (1) represents the contrast in counterfactual values of T associated with a one-unit increase in S within an individual. By contrast, the R^2_{trial} measure assesses correlation between study-specific treatment effects. The relevant population for R^2_{trial} is those of the studies. The spirit of this measure is generally incompatible with causal inference measures.

4 Two-Stage Procedure: Algorithm and Application

4.1 Estimation procedures for causal effects with noncontinuous outcomes

In this section, we describe estimation procedures for the causal effect in (1). As mentioned, the instrumental variables estimator will yield a consistent estimator of the causal effect. Two-stage least squares is a simple algorithm that can be used when T is continuous and uncensored. In practice, however, T is typically not of this type, so other methods are needed.

One situation is *T* being binary. Then a structural linear regression model for *T* given *S* seems undesirable the fitted values from such a model are not necessarily constrained to lie between zero and one. A more desirable class of structural models would be the following:

$$E[T|S] = g\left(\tilde{\alpha}_0 + \tilde{\alpha}\, S\right) \tag{8}$$

where *g* is a monotone and continuous function mapping from the real line to [0, 1], $\tilde{\alpha}_0$ an intercept term, and $\tilde{\alpha}$ is a regression coefficient. For example, if *g* is the inverse of the logistic function, then $\tilde{\alpha}$ represents a causal odds ratio of *T* associated with a one-unit change of *S*. To estimate $\tilde{\alpha}$, we can use the following modification of the two-stage least squares algorithm:

1. Fit the model

$$E[S|Z] = g_S(\gamma_0 + \gamma_S Z) \tag{9}$$

where g_S is an appropriately defined link function by some algorithm, e.g. nonlinear least squares.

2. Obtain the residuals from step 1, say $\widehat{e_S}$, and fit the model

$$E[T|S, \widehat{e_S}] = g\left(\alpha_0^* + \alpha^* S + \gamma T^{\widehat{e}} Z\right). \tag{10}$$

One can then use $\widehat{\alpha^*}$ as the estimator of $\tilde{\alpha}$ from (8). As demonstrated in [26], this yields a consistent estimator for $\tilde{\alpha}$. In fact, model (8) and the attendant two-stage estimation procedure can be extended to *T* being distributed within the exponential family of distributions for which there exists a generalized linear model relating *S* to a functional of the mean of *T*.

The equivalence between two-stage least squares and the relative effect quantity proposed by [15] breaks down in the nonlinear situation. The ratio of the regression coefficients for treatment effects from nolinear models for *S* given *Z* and *T* given *Z* will in general not coincide with the estimand being estimated the two-stage residual inclusion algorithm given in the previous paragraph. We now describe an extended relative effect measure that is more congruent with the two-stage least squares algorithm. Note that in the linear model case, we can express the estimator of 7 and the instrumental variables estimator as solutions to the following equations:

$$\begin{aligned} \sum_{i=1}^{n} (T_i - \alpha_0 - S_i\alpha) &= 0. \\ \sum_{i=1}^{n} Z_i (T_i - \alpha_0 - S_i\alpha) &= 0. \end{aligned} \tag{11}$$

Notice that the term inside the parentheses inside the summation on the left-hand side of (11) is the population structural model (1) for the *i*th individual. Thus, we can interpret 11 as the inner product between the instrument with the error term for (1). By assumption of the instrumental variable, $E[Z\varepsilon] = 0$. Thus, when the treatment instrumentality assumption holds, (11) defines a valid estimating equation for estimating α. For situations where we wish to fit (8), we employ the following two-step algorithm to estimate the extended relative effect:

1. Calculate the adjusted dependent variable corresponding to a linearization of 8 from a regression of T on S. Call this variable $\tilde{Y}_i, i=1,\ldots,n$.
2. Estimate the extended relative effect as

$$\widetilde{RE}=\frac{\sum_{i=1}^{n} Z_i \tilde{Y}_i}{\sum_{i=1}^{n} Z_i S_i}.$$

Note that the formula for $\widetilde{RE}$ bears more of a resemblance to (2) than to the relative effect formula given in [15].

4.2 Sensitivity Analysis for the Treatment Instrumentality Assumption

In this section, we describe an approach to performing sensitivity analysis for the treatment instrumentality assumption. Our approach is again motivated by consideration of the linear model case and follows the development in [27]. If the correlation between Z and ε_S is not equal to zero, then the IV estimator defined will be biased for estimation of α. We make the assumption that (Z, S, ε_S) are jointly normally distributed with mean zero vector and variance-covariance matrix

$$\begin{bmatrix} \sigma_{ZZ}^2 & \sigma_{ZS} & \sigma_{Ze} \\ \sigma_{ZS} & \sigma_{SS}^2 & \sigma_{Se} \\ \sigma_{Ze} & \sigma_{Se} & \sigma_{ee}^2 \end{bmatrix}.$$

From Lemma 1 of [27], we have that in general, the instrumental variable estimator $\tilde{\alpha}$ is consistent for $\alpha + \delta$, where $\delta = \sigma_{Ze}/\sigma_{ZS}$. When the treatment instrumentality assumption holds, $\sigma_{Ze} = 0$. Our sensitivity analysis approach thus requires σ_{Ze} as a user input. This suggests the following algorithm. First, we calculate the instrumental variable estimator $\tilde{\alpha}$ that makes the treatment instrumentality assumption along with a 95% CI ($(\widehat{\alpha}_L, \widehat{\alpha}_U)$). In the example in the next section, we use the nonparametric bootstrap to construct the 95% CI. Next, we plot the line $\alpha=\widehat{\alpha} - \alpha_{Ze}/\widehat{\alpha}_{Zs}$, where $\widehat{\alpha}_{Zs}$ is the empirical covariance between Z and S. In addition, we superimpose the lines $\widehat{\alpha}_L - \alpha_{Ze}/\widehat{\alpha}_{Zs}$ and $\widehat{\alpha}_U - \alpha_{Ze}/\widehat{\alpha}_{Zs}$. These lines will give the analyst regions of plausible values for the causal effect of S on T while relaxing the treatment instrumentality assumption.

4.3 Numerical example: Glaucoma Data

In this section, we consider data from the Collaborative Initial Glaucoma Treatment Study (CIGTS) [28]. The CIGTS was a randomized trial that compared surgery (Z = 1) to medicine (Z = 0) on reducing intraocular pressure (IOP). Elevated IOP is a major risk factor of glaucoma. The true endpoint was defined to be IOP being greater than 18mmHg eight years after randomization, while the surrogate endpoint was defined to be IOP being greater than 18mmHg one year after randomization. Note that the true and surrogate endpoints are binary.

For this analysis, we first begin by using analyses based on the identity link function for (1). In this situation, the relative effect estimator of [15] and the two-stage least squares algorithms give exactly the same numerical answer. The estimate of the relative effect is 0.048 with an associated 95% CI of (0.05, 0.93). The confidence interval was constructed using the nonparametric bootstrap, stratified by treatment group. Finally, we show a plot of

the sensitivity analysis using the approach described in Section 4.2. Given that the covariance between S and Z is only 0.06, the range of values for Cov(Z, ε_S) shown on the x-axis for Figure 1 is approximately 20 times larger. The key thing to note is that the range of the y-axis is almost the same as the ratio of the range of Cov(Z, ε_S) to Cov(Z, S), which is 16.67. This implies that the magnitude of the causal effect is on the order of the Cov(Z, ε_S)/Cov(Z, S) so that the results are extremely sensitive to validity of the treatment instrumentality assumption for this example.

Our second set of analyses involve using the logistic link. In situation, the estimator of [15] will in general be quite different from the two-stage estimation procedure described in Section 4.1. We assume a logistic regression model for analysis of Z on S as well as for the effect of Z on T . This yields an estimated log odds ratio of 1.26 for treatment on the surrogate endpoint and 0.65 for treatment on the true endpoint. Combining, this gives a relative effect estimate of 0.51 for the estimated relative effect of [15]. A 95% confidence interval for the relative effect is given by (0.05, 1.03).

Next, we fit the instrumental variable estimator using the two-step estimation procedure. Assuming a logit link in step one of the algorithm, the estimate of the relative effect using the two-stage estimation procedure is 2.4 with a 95% confidence interval of (−0.82, 5.88). Notice that in this example, the variability is much wider using the two-stage estimation procedure relative to that from [15]. If we calculate the extended relative effect estimator described at the end of Section 4.1., we obtain an estimator of 1.88 with an associated 95% CI of (1.29, 2.58). The wide discrepancy of these analyses, in conjunction with the sensitivity analysis described for the linear model setup, suggest the sensitivity of results to the treatment instrumentality assumption.

5 Discussion

In this article, we have proposed a simple structural equations model for modelling the causal effect of the surrogate endpoint on the true endpoint. Based on this model, one can view the randomization of treatment as an instrumental variable by which the causal effect in (1) can be made. This approach leads to clarification of previously studied approaches to surrogacy. While the relative effect measure is a ratio of causal effects of treatment under certain assumptions, in this work we show that it in fact estimates a causal parameter within the framework of Pearl [16,6] under the assumption of treatment instrumentality. There is a very general algorithm (two-stage least squares) available for instrumental variables estimation, the example shows that it can be more or less variable on a real dataset than the relative effect estimator of Buyse and Molenberghs [15].

In addition, we have developed an extended estimator of relative effect that coincides more with the two-stage least squares algorithm relative to the measure proposed by Buyse and Molenberghs [15]. While the two approaches are equivalent in the situation when the linear model assumption holds for (1), they diverge when the structural model is nonlinear. Another new tool is a sensitivity analysis approach to assess the treatment instrumentality assumption. Based on our arguments and the real data example, we find that the causal effect estimate will be quite sensitive to the validity of the treatment instrumentality assumption.

An underlying theme of this manuscript is the importance of being able to formulate appropriate causal pathways in considering the effects of a treatment on a surrogate and true endpoint. Provided one can do this, then this will dictate what variables should be measured and what analyses should be done. This underscores the necessity in developing strong

mechanistic understanding of the biological role of surrogate endpoints in the disease process, which was mentioned in the discussion of Prentice [1].

The endogeneity viewpoint taken in this paper also allows for a new causal viewpoint towards surrogacy. Provided that one is able to believe in manipulability of the surrogate endpoint, the approach leads to relative effect-type estimators as appropriate instrumental variables-based estimator of causal effects in the model (1). A more general two-stage algorithm for estimation was also presented that allows for estimation of causal effects under a variety of conditions on the distribution of S and T , as well as their relationship.

While the problem of endogeneity has been heavily considered in the econometrics literature, its application of instrumental variables to biostatistical problems have not been fully exploited. We hope that the ideas presented here lead to some cross-fertilization in research between the two disciplines. Already, this is starting to be seen in articles such as Stukel et al. [29], Sobel [30] and Joffe et al. [31]

Acknowledgments

This research is supported by NIH R01-CA129402.

References

1. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in Medicine. 1989; 8:431–440. [PubMed: 2727467]
2. Burzykowski, T.; Molenberghs, G.; Buyse, M. The Evaluation of Surrogate Endpoints. Springer-Verlag; New York: 2005.
3. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. Statistics In Medicine. 1997; 16:1965–1982. [PubMed: 9304767]
4. Gail M, Pfeiffer R, Houwelingen HCV, Carroll RJ. On meta-analytic assessment of surrogate outcomes. Biostatistics. 2000; 1:231–246. [PubMed: 12933506]
5. Robins JM, Greenland S. Identifiability and exchangeability of direct and indirect effects. International Journal of Epidemiology. 1992; 3:143–155.
6. Pearl, J. Causality: Models, Reasoning and Inference. Cambridge University Press; 2001.
7. Frangakis CE, Rubin DB. Principal stratification in casual inference. Biometrics. 2002; 58:21–29. [PubMed: 11890317]
8. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. Biometrics. 2008; 64:1146–1154. [PubMed: 18363776]
9. Li Y, Taylor JMG, Elliott MR. A Bayesian Approach to Surrogacy Assessment Using Principal Stratification in Clinical Trials. Biometrics. 2009 published online August 10, 2009, doi: 10.1111/j.1541-0420.2009.01303.x.
10. Taylor JMG, Wang Y, Thiebaut R. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. Biometrics. 2005; 61:1102–1111. [PubMed: 16401284]
11. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic disease. Statistics in Medicine. 1992; 11:167–178. [PubMed: 1579756]
12. Joffe MM, Greene T. Related causal frameworks for surrogate outcomes. Biometrics. 2008; 65:530–538. [PubMed: 18759836]
13. Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. Wiley; New York: 2002.
14. Lancaster, T. The Econometric Analysis of Transition Data. Cambridge University Press; 1990. 1990
15. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics. 1998; 54:1014–1029. [PubMed: 9840970]
16. Pearl J. Causal diagrams for empirical research (with discussion). Biometrika. 1995; 82:669–710.

17. Hausman JA. Specification tests in econometrics. Econometrica. 1978; 46:1251–1271.
18. Wright S. Correlation and causation. Journal of Agricultural Research. 1921; 20:557–585.
19. Begg CB, Leung DHY. On the use of surrogate endpoints in randomized trials (with discussion). Journal of the Royal Statistical Society, Series A. 2000; 163:15–28.
20. Berger VW. Does the Prentice criterion validate surrogate endpoints? Statistics in Medicine. 2004; 23:1571–1578. [PubMed: 15122737]
21. Wang Y, Taylor JM. A measure of the proportion of treatment effect explained by a surrogate marker. Biometrics. 2002; 58:803–812. [PubMed: 12495134]
22. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association. 1996; 91:444–455.
23. Angrist, JD. Instrumental variables estimation of average treatment effects in econometrics and epidemiology. National Bureau of Economics Research; 1991. Working Paper 115
24. Bhattacharya J, Goldman D, McCaffrey D. Estimating probit models with self-selected treatments. Statistics in Medicine. 2006; 25:389–413. [PubMed: 16382420]
25. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics. 2000; 1:49–67. [PubMed: 12933525]
26. Terza JV, Basu A, Rathouz P. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of Health Economics. 2008; 27:531–543. [PubMed: 18192044]
27. Ashley R. Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. Journal of Applied Econometrics. 2009; 24:325–337.
28. Musch DC, Lichter PR, Guire KE, Standardi CL, CIGTS Investigators. The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. Ophthalmology. 1999; 106:653662.
29. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. Journal of the American Medical Association. 2007; 297:278–85. [PubMed: 17227979]
30. Sobel ME. Identification of causal parameters in randomized studies with mediating variables. Journal of Educational and Behavioral Statistics. 2008; 33:230–251.
31. Joffe M, Small D, Brunelli S, Ten Have T, Feldman H. Extended instrumental variables estimation for overall effects. International Journal of Biostatistics. 2008; 4(1) Article 4.

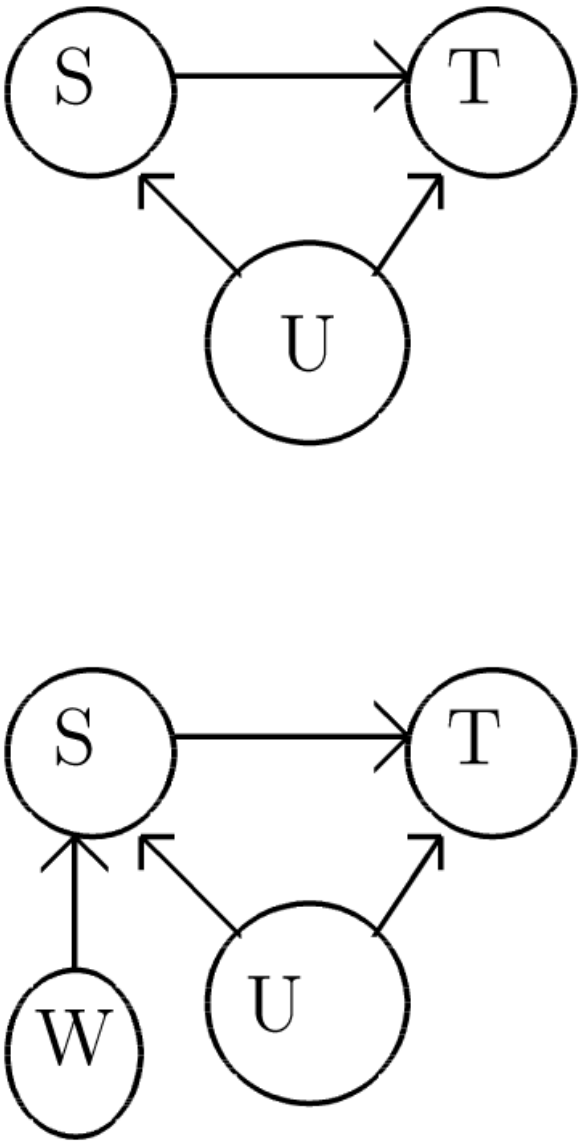

Figure 1.
Graphical model describing structural equations model (1). S denotes the surrogate endpoint, T the true endpoint, and U denotes unmeasured variables.

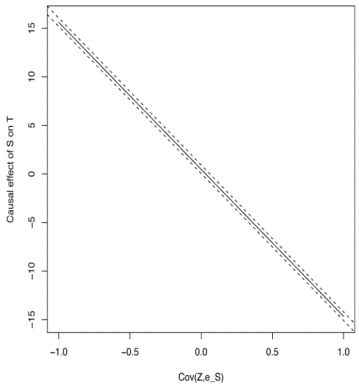

Figure 2.
A modified version of Figure 1, with variable W , representing the instrumental variable, included.

Table 1

Joint distribution of counterfactual surrogate S and outcome Y under monotonicity assumption

		$T(0, S(0)), T(1, S(1))$			
		(0,0)	**(0,1)**	**(1,1)**	
$S(0), S(1)$	(0,0)	π_{11}	π_{12}	π_{13}	π_{1+}
	(0,1)	π_{21}	π_{22}	π_{23}	π_{2+}
	(1,1)	π_{31}	π_{32}	π_{33}	π_{3+}
		π_{+1}	π_{+2}	π_{+3}	1