Biometrics 69, 561–581
September 2013
DOI: 10.1111/biom.12071

Surrogate Measures and Consistent Surrogates

Tyler J. VanderWeele*

Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
**email:* tvanderw@hsph.harvard.edu

Summary. Surrogates which allow one to predict the effect of the treatment on the outcome of interest from the effect of the treatment on the surrogate are of importance when it is difficult or expensive to measure the primary outcome. Unfortunately, the use of such surrogates can give rise to paradoxical situations in which the effect of the treatment on the surrogate is positive, the surrogate and outcome are strongly positively correlated, but the effect of the treatment on the outcome is negative, a phenomenon sometimes referred to as the "surrogate paradox." New results are given for consistent surrogates that extend the existing literature on sufficient conditions that ensure the surrogate paradox is not manifest. Specifically, it is shown that for the surrogate paradox to be manifest it must be the case that either there is (i) a direct effect of treatment on the outcome not through the surrogate and in the opposite direction as that through the surrogate or (ii) confounding for the effect of the surrogate on the outcome, or (iii) a lack of transitivity so that treatment does not positively affect the surrogate for all the same individuals for whom the surrogate positively affects the outcome. The conditions for consistent surrogates and the results of the article are important because they allow investigators to predict the direction of the effect of the treatment on the outcome simply from the direction of the effect of the treatment on the surrogate. These results on consistent surrogates are then related to the four approaches to surrogate outcomes described by Joffe and Greene (2009, *Biometrics* **65**, 530–538) to assess whether the standard criteria used by these approaches to assess whether a surrogate is "good" suffice to avoid the surrogate paradox.

Key words: Causal inference; Counterfactuals; Principal stratification; Randomized trials; Surrogate outcomes.

1. Introduction

There has been considerable interest in the statistics literature on measures and statistical methods for assessing the adequacy of a surrogate outcome (Prentice, 1989; Freedman, Graubard, and Schatzkin, 1992; Lin, Fleming, and DeGruttola, 1997; Gail et al., 2000; Burzykowski, Molenberghs, and Buyse, 2005; Taylor, Wang, and Thiebaut, 2005; Follmann, 2006; Chen, Geng, and Jia, 2007; Gilbert and Hudgens, 2008; Joffe and Greene, 2009; Wolfson and Gilbert, 2010; Huang and Gilbert, 2011). The use of a surrogate outcome may be desirable in randomized trials if the cost or length of follow-up required to obtain data on the outcome of interest is thought prohibitive. A variety of statistical approaches and measures have been proposed. In a recent article, Joffe and Greene (2009) summarize a number of these statistical approaches from the perspective of causal inference and discuss relations between these approaches.

A smaller literature on surrogate outcomes has considered what is sometimes referred to as the "surrogate paradox." It may be the case that the treatment has a positive effect on the surrogate, that the surrogate and outcome are strongly positively associated and yet that the treatment itself has a negative effect on the outcome! We might refer to such cases as instances of the "surrogate paradox." This was illustrated dramatically in the case of trials evaluating the effect of drug treatment on ventricular arrhythmia, taken as a surrogate for mortality. Ventricular arrhythmia is strongly associated with mortality; several drugs were tested in randomized trials, were found to lower ventricular arrhythmia, and were approved by the Food and Drug Administration. However, in follow-up it became clear that the drugs increased rather than decreased mortality (Moore, 1995; Fleming and DeMets, 1996). One important task then with regard to surrogate outcomes—and the one which will be the focus of this article—is determining when data concerning the effect of treatment on the surrogate can be used to make decisions about the direction of the effect of the treatment on an outcome. In two articles Chen et al. (2007) and Ju et al. (2010) discuss sufficient conditions which, if satisfied by a surrogate, will avoid the surrogate paradox. They refer to surrogates that avoid the surrogate paradox as "consistent surrogates."

There has been little effort to relate these sufficient conditions to the statistical measures and approaches that have been used to assess and measure surrogacy. This article introduces new criteria for consistent surrogates and then revisits the survey of approaches described by Joffe and Greene (2009), evaluating each in light of the surrogate paradox. Sections 2 and 3 summarize the results of Chen et al. (2007) and Ju et al. (2010) on consistent surrogates and then extend their results further to allow for more general settings and to provide a characterization of conditions which are necessary for the surrogate paradox to occur (analogously, are sufficient to avoid it). The conditions and the results of the article are important because they allow investigators to predict the direction of the effect of the treatment on the outcome simply from the direction of the effect of the

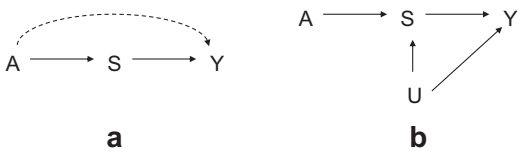

Figure 1. Examples illustrating surrogate outcomes.

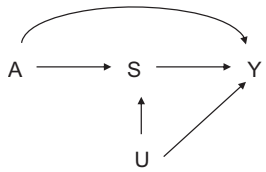

Figure 2. Causal diagram allowing for an effect of A on Y not through the putative surrogate S.

treatment on the surrogate. Section 4 then considers the role and significance of the surrogate paradox for each of the approaches described by Joffe and Greene (2009). Section 5 illustrates the surrogate paradox in the various approaches and Section 6 offers some concluding remarks.

2. Definitions for Surrogates and the Surrogate Paradox

Let A be a treatment of interest that we will assume randomized; let Y be the outcome of interest and let S be a proposed surrogate. Let Y_a and S_a be counterfactual outcomes (or potential outcomes) for Y and S for each individual that would have been obtained if treatment A had, possibly contrary to fact been set to a. Finally let Y_{as} be the counterfactual outcome for each individual that would have been obtained if A had been set to a and if S had been set to s. Contrasts of the form $Y_{as} - Y_{a's}$ are referred to as controlled direct effects (Pearl, 2001). Below we will also describe so-called "natural direct effects" (Robins and Greenland, 1992; Pearl, 2001) but unless otherwise indicated "direct effects" will refer to "controlled direct effects." We restrict our attention to settings in which A, S, Y are measurable for all individuals. We thus do not consider cases in which for some individuals an event Y can occur before S is measured; see Gilbert and Hudgens (2008) and Wolfson and Gilbert (2010) for discussion of these settings.

In what follows we will consider several definitions in the literature concerning surrogate outcomes and discuss how these various definitions are related to the surrogate paradox. In what is now considered a classic article, Prentice (1989) suggested that a surrogate should be such that a test of the null of no effect of the treatment A on surrogate S should serve as a valid test of the null of no effect of the treatment A on outcome Y. Prentice proposed the following two main criteria for assessing this and a variable satisfying such criteria has subsequently been referred to as a "statistical surrogate" (Frangakis and Rubin, 2002).

Statistical Surrogate (Prentice Criteria): S is said to be a surrogate for the effect of A on Y if (i) Y is independent of A conditional on S; (ii) S and Y are correlated.

The criteria are suggested by the diagram in Figure 1a. Suppose there is no controlled direct effect of A on Y, then if there is no effect of A on S it then follows that there will be no effect of A on Y. Moreover, in this diagram if there is no direct effect of A on Y then A will be independent of Y conditional on S. But the relevance of the criteria is less clear if there are unmeasured confounders of S and Y as in Figure 1b. There could be correlation between A and Y conditional on S due to U even if A has no direct effect on Y. The Prentice criterion might only be a reasonable requirement if we could control for the common causes of S and Y, but we will return to these considerations later.

Prompted perhaps in part by these concerns, Frangakis and Rubin (2002) used the potential outcomes framework to propose an alternative criterion to evaluate surrogates and referred to a surrogate that satisfied this criterion as a "principal surrogate."

Principal Surrogate (Frangakis and Rubin, 2002): S is said to be a principal surrogate for the effect of A on Y if for all s, $pr(Y_1|S_1 = S_0 = s) = pr(Y_0|S_1 = S_0 = s)$.

Essentially a principal surrogate requires that whenever the treatment does not change the surrogate ($S_1 = S_0 = s$) there is no difference in the distribution of potential outcomes with versus without treatment. If a surrogate satisfied this property then an effect of A on Y will be present only if an effect of A on S is present. If Y is binary the definition of a principal surrogate is equivalent to $E(Y_1 - Y_0|S_1 = S_0 = s) = 0$, a condition that may be referred to as no principal strata direct effects VanderWeele (2008). This has likewise been referred to as the property of "average causal necessity" (Gilbert and Hudgens, 2008). If Y is not binary, then principal surrogacy as defined above requires the stronger condition $pr(Y_1|S_1 = S_0 = s) = pr(Y_0|S_1 = S_0 = s)$. Lauritzen (2004) proposed a slightly stronger definition related to surrogacy that he referred to as a "strong surrogate":

Strong Surrogate (Lauritzen, 2004): S is a strong surrogate for the effect of A on Y if the causal diagram in Figure 1b is valid.

Conceived of another way, S is a strong surrogate for the effect of A on Y if A is an instrument for the effect of S on Y (Lauritzen, 2004). If S is not a strong surrogate then the causal diagram would be that in Figure 2, where if the treatment is randomized, U can be taken as the principal stratum (S_0, S_1) so that Figure 2 makes no assumption about counterfactual distributions beyond that implied by the randomization of A. The variable S will be a strong surrogate for the effect of A on Y if the "controlled direct effects" (Pearl, 2001) are such that $Y_{1s} - Y_{0s} = 0$ for all s. A strong surrogate is also a principal surrogate (Lauritzen, 2004; VanderWeele et al., 2008) but the reverse implication does not hold because principal surrogacy only requires no direct effects when $S_1 = S_0 = s$ and only requires this in distribution, not for all individuals. Note also that a strong surrogate will be a statistical surrogate if there is no common cause of the surrogate and the outcome as in Figure 1a but a strong surrogate need not be a statistical surrogate if there is such a common cause as in Figure 1b.

Chen et al. (2007) introduced one further notion concerning surrogacy which they referred to as a consistent surrogate. Chen et al. (2007) restricted discussion of consistent surrogates to setting which involved a strong surrogate. Below we will generalize Chen et al.'s definition to one which allows for a direct effect of A on Y. Chen et al. (2007) defined a strong surrogate S to be a consistent surrogate for the effect of A on Y if, (a) for a positive average causal effect of S on Y, a non-positive (non-negative) average causal effect of A on S implies a non-positive (non-negative) average causal effect of A on Y, (b) for a negative average causal effect of S on Y, a non-positive (non-negative) average causal effect of A on S implies a non-negative (non-positive) average causal effect of A on Y and (c) a null average causal effect of A on S implies a null average causal effect of A on Y.

If a surrogate is not consistent in this sense then we may have effect reversal: treatment A may have a positive effect on S and S on Y but the effect of A on Y may be negative! Chen et al. (2007) refer to such effect reversal as instances of the "surrogate paradox." Chen et al. (2007) went on further to give an example showing that neither a principal surrogate nor even a strong surrogate necessarily satisfies the properties of a consistent surrogate. Both principal surrogates and strong surrogates are subject to the surrogate paradox. This is somewhat surprising as the notions of a principal surrogate and a strong surrogate are already quite stringent; it is also rather disturbing in that such effect reversal seems to completely undermine the value of a surrogate marker. In the next section we review and extend results concerning sufficient conditions that ensure the surrogate paradox is avoided. First, however, we generalize the notion of a consistent surrogate described by Chen et al. (2007) so as to allow for settings in which the surrogate is not a strong surrogate (i.e., the treatment may have a direct effect on the outcome not through the surrogate) and for settings in which we may not be willing to talk about the "causal effect" of the surrogate on the outcome and may not be willing to envision interventions on the surrogate S.

Consistent Surrogate: S is said to be a a consistent surrogate for the effect of A on Y if (a) when S and Y are positively associated, a non-positive (non-negative) average causal effect of A on S implies a non-positive (non-negative) average causal effect of A on Y, (b) when S and Y are negatively associated a non-positive (non-negative) average causal effect of A on S implies a non-negative (non-positive) average causal effect of A on Y. A surrogate that is not a consistent surrogate is said to exhibit the surrogate paradox.

The focus of the remainder of this article will be on articulating conditions under which the surrogate paradox as defined above is avoided that is, when data on the effect of A on S in conjunction with knowledge that the surrogate and outcomes are strongly correlated can together be used to draw conclusions about the direction of the effect of the treatment A on the outcome Y.

3. Results on Consistent Surrogates to Avoid the Surrogate Paradox

Chen et al. (2007) gave the following sufficient conditions concerning avoiding the surrogate paradox.

Proposition 1. *(Chen et al., 2007): If S is a strong surrogate for the effect of A on Y (i.e., if Figure 1b is a valid causal diagram) then if (a) $E(Y|s,u)$ is non-decreasing in s for all u and (b) $pr(S > s|a,u)$ is non-decreasing in a for all s,u, then $E(Y_a) = E(Y|a)$ is non-decreasing in a.*

Viewed another way, if S is a strong surrogate (no direct effects of treatment on the outcome not through the surrogate) and if conditions (a) and (b) are satisfied then the effect of A on Y will be in the direction expected and the surrogate paradox avoided: $E(Y_a)$ is non-decreasing in a so $E(Y_1) - E(Y_0) \geq 0$. The result remains true if in both conditions (a) and (b), "non-decreasing" is replaced by "non-increasing;" if only one of conditions (a) or (b), "non-decreasing" is replaced by "non-increasing" then the conclusion of Proposition 1 changes to $E(Y_a) = E(Y|a)$ is non-increasing in a. Similar remarks hold for the other propositions below. Note that to avoid the surrogate paradox (i.e., to ensure a consistent surrogate) a non-negative average causal of A on S is not sufficient; rather one needs the effect to be non-negative in the distributional sense that $pr(S > s|a,u)$ is non-decreasing in a for all s,u; this is sometimes referred to as "distributional monotonicity" (VanderWeele et al., 2008; VanderWeele and Robins, 2009, 2010). Note that the assumption that S is a strong surrogate is not a testable assumption. Note also there may be different variables U for which Figure 1b could be a valid causal diagram. The conclusion of Proposition 1 will hold if there is any U such that Figure 1b is a causal diagram and such that conditions (a) and (b) hold. Similar points pertain also to Propositions 2–4 below.

Ju and Geng (2010) generalized the result of Chen et al. (2007) to give a stronger conclusion if condition (a) is also replaced by one of distributional monotonicity.

Proposition 2. *(Ju and Geng, 2010): If S is a strong surrogate for the effect of A on Y (i.e., if Figure 1b is a valid causal diagram) and if (a) $pr(Y > y|s,u)$ is non-decreasing in s for all y,u and (b) $pr(S > s|a,u)$ is non-decreasing in a for all s,u, then $pr(Y_a > y) = pr(Y > y|a)$ is non-decreasing in a.*

Here we get the slightly stronger conclusion that not simply does A increase Y on average but that the effect of A on Y is also distributionally monotonic. In fact, as discussed in the online supplement, both of these results of Chen et al. (2007) and Ju and Geng (2010) follow almost immediately from the theory of signed causal directed acyclic graphs (VanderWeele and Robins, 2009, 2010). Moreover, more general results are possible. The definitions and results above have essentially been concerned with the case in which the effect of A on Y is entirely through S. In most settings, this will likely be unrealistic. A good surrogate may account for a large portion of the effect of A on Y but it is unlikely that the surrogate accounts for all of this effect. Likely there will be an effect of A on Y not through S as in Figure 2. It is shown in the online supplement that the following two results hold; these generalize Chen et al. (2007) and Ju and Geng (2010), respectively, by allowing for an effect of A on Y not through S.

Proposition 3. *In the causal diagram in Figure 2, if (a) $E(Y|a,s,u)$ is non-decreasing in a and s for all u and (b)*

$pr(S > s|a, u)$ is non-decreasing in a for all s, u then $E(Y_a) = E(Y|a)$ is non-decreasing in a.

Similar results hold under non-increasing rather than non-decreasing functional relationships. Proposition 3 has an important and intuitive interpretation. Suppose that in a randomized trial we find a positive average causal effect of A on S and we know that S and Y are strongly positively correlated. This is often the setting encountered with surrogate outcomes. In this setting, under what circumstances might the surrogate paradox arise? When might the effect of A on Y be negative rather than positive? Proposition 3 states that at least one of three things must occur if we are to get this effect reversal. First, there may be a negative direct effect of A on Y not through S (i.e., the first part of assumption (a) that $E(Y|a, s, u)$ is non-decreasing in a may be violated). Second, it may be the case that although S and Y are positively correlated this may not indicate the actual causal relationship of S on Y; the association may be due to confounding by U (i.e., the second part of assumption (a) that once we condition on U, $E(Y|a, s, u)$ is non-decreasing in s may be violated). Third, even if neither of these first two phenomenon occur, it may be the case that even though A positively affects S on average and S positively affects Y, A may not positively affect S for all individuals; it may decrease S and thus decrease Y for some individuals; we may have a lack of transitivity (i.e., assumption (b), the assumption concerning distributional monotonicity which guarantees that this is avoided, may be violated). In summary, if the surrogate paradox is to occur we either need (i) a direct effect of A on Y not through S in the opposite direction or (ii) confounding for the effect of S on Y, or (iii) a lack of transitivity so that A does not positively affect S for all the same individuals for which S positively affects Y. In thinking about whether the surrogate paradox might occur and whether one ought to draw conclusions concerning an outcome of interest from the analysis of the results concerning a surrogate, an investigator could think through each of these three possibilities. Proposition 3 states that at least one of them must occur if the surrogate paradox is to arise.

Proposition 4 below gives a somewhat stronger conclusion concerning distributional monotonicity of the effect of A on Y under somewhat stronger assumptions. Proposition 4 generalizes the results of Ju and Geng (2010) to allow for a direct effect of A on Y. If the outcome Y is binary Propositions 3 and 4 are equivalent.

Proposition 4. *In the causal diagram in Figure 2, if (a) $pr(Y > y|a, s, u)$ is non-decreasing in a and s for all y, u and (b) $pr(S > s|a, u)$ is non-decreasing in a for all s, u then $pr(Y_a > y) = pr(Y > y|a)$ is non-decreasing in a.*

In the next section we will relate these results on consistent surrogates to various statistical and causal approaches to the analysis of surrogate outcomes.

4. Consistent Surrogates and Measures of Surrogacy

Joffe and Greene (2009) considered four different approaches that have been proposed to evaluate surrogates or to measure the extent of surrogacy and they derived relations between them under linear model assumptions. Here we will revisit each of these four approaches in light of the results above on consistent surrogates. These four approaches could broadly be described as (i) a "proportion-explained" approach, (ii) an "indirect effects" approach, (iii) a "meta-analytic" approach, and (iv) a "principal stratification" approach. We will consider each in turn. Each of these approaches may tell us something about the role that the surrogate S plays in the relationship between treatment A and outcome Y. Here, however, we will assess whether these approaches help us evaluate whether a surrogate is consistent that is, whether the surrogate paradox is avoided. We will consider the metrics that are used to evaluate surrogacy in each of these four approaches and consider whether these metrics correspond in any way to ensuring that one has a consistent surrogate.

Building on Prentice (1989), Freedman et al. (1992) proposed using a "proportion explained" measure to assess surrogacy. Suppose one were to regress the outcome Y on the exposure A:

$$E(Y|A = a) = \Phi_0 + \Phi_1 a$$

and then regress the outcome Y on the exposure A and the surrogate S:

$$E(Y|A = a, S = s) = \theta_0 + \theta_1 a + \theta_2 s$$

The proportion of the total effect explained by the surrogate is then taken as:

$$(\Phi_1 - \theta_1)/\Phi_1, \tag{1}$$

which is equivalent to $1 - \theta_1/\Phi_1$. Statistical inference for this measure is also described by Lin et al. (1997). The measure does, however, suffer from problems if either Φ_1 is small or if the model for $E(Y|A = a, S = s)$ is not correctly specified (Molenberghs et al., 2002). A similar measure is sometimes used in the setting of "mediation analysis" to assess the proportion of the effect of A on Y mediated by S. In the setting of mediation analysis this measure is problematic because there may be confounding of the effect of S on Y by U; this can occur even if treatment A is randomized since the surrogate S is generally not randomized. Because of this confounding using the proportion in (1) as a measure of mediation can be highly problematic (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2010). However, in the context of surrogacy (rather than mediation) if the goal is simply to assess how much of the effect of A on Y can be predicted by the effect of A on S these concerns about confounding may be less relevant. Even if U is a common cause of S and Y, if because of U, S give important information about Y then S may still be a good surrogate insofar as it may be possible to predict the sign of the effect of A on Y from the sign of the effect of A on S. Although the measure in (1) of the "proportion explained" may thus serve as a useful measure, it is not immune to the surrogate paradox. An example is given below in which the average causal effect of A on S is positive, the average causal effect of S on Y is positive, "proportion explained" is 100%, but the effect of A on Y is negative. This can occur because

it may be the case that A does not positively affect S for the same individuals for which S positively affects Y. Nothing in the "proportion explained" measure guarantees the distributional monotonicity needed to avoid the surrogate paradox. Thus even if a surrogate is judged to be "good" from the standpoint of having a high proportion explained, this does not guarantee that the surrogate is consistent.

The second approach considered by Joffe and Greene (2009) may be referred to as the "indirect effects" approach. This was essentially the approach pursued by Taylor et al. (2005). This approach relies on the counterfactual framework and specifically counterfactual definitions of what are now often called natural indirect effects (Pearl, 2001). The alternative notion of controlled direct effect (Pearl, 2001), although useful for assessing whether there is an effect of the treatment on the outcome not through the surrogate, cannot be employed directly to assess mediation (Robins and Greenland, 1992). The average natural indirect effect is defined as $E(Y_{1S_1} - Y_{1S_0})$ and measures the effect comparing setting the treatment to present with the surrogate set to what it would have been with versus without the treatment (Robins and Greenland, 1992; Pearl, 2001). For it to be non-zero the treatment must have an effect on the surrogate (i.e., S_1 and S_0 must differ) and then the surrogate must have an effect of the outcome (i.e., the change in the surrogate from S_0 to S_1 must have an effect on Y). This is thus sometimes referred to as a "mediated effect." A measure of surrogacy may then be taken as the "proportion mediated" that is, the proportion of the natural indirect effect to the total effect:

$$\frac{E(Y_{1S_1} - Y_{1S_0})}{E(Y_1 - Y_0)}. \tag{2}$$

The conditions for identification and estimation of the natural direct and indirect effect are described elsewhere (Pearl, 2001; Taylor et al., 2005; Joffe and Greene, 2009; VanderWeele and Vansteelandt, 2010; Imai, Keele, and Tingley, 2010) and are beyond the scope of this article. Identification of the natural indirect effect does, however, require control for common causes of the intermediate S and the outcome Y (Pearl, 2001; Joffe and Greene, 2009; VanderWeele and Vansteelandt, 2010). The advantage of this approach to surrogate measures is that, provided the natural indirect effect has been correctly identified and estimated, it gives the actual effect of the treatment on the outcome through the surrogate. Likewise, the natural direct effect, $E(Y_{1S_0} - Y_{0S_0})$, can be used to assess whether there is an effect of the treatment on the outcome not through the surrogate and one could evaluate whether this was in the opposite direction of the direct effect. The natural indirect and direct effects sum to the total effect: $E(Y_{1S_1} - Y_{1S_0}) + E(Y_{1S_0} - Y_{0S_0}) = E(Y_{1S_1}) - E(Y_{0S_0}) = E(Y_1) - E(Y_0)$. Thus, if the natural direct and indirect effects were known this could be useful in diagnosing the surrogate paradox if these two effects were in opposite directions. The difficulties are, however, effectively transferred to the challenge of identifying and consistently estimating the natural indirect effect, $E(Y_{1S_1} - Y_{1S_0})$. The identification conditions needed to identify this natural indirect effect are quite strong (Pearl, 2001; VanderWeele, 2010) which constitutes a disadvantage to this approach. Within the "indirect effects" approach, the criterion generally used to assess whether a surrogate is "good" (whether the proportion mediated is large) unfortunately, however, does not help guarantee that a surrogate is consistent. As will be seen in the illustration below, we can in fact have a high proportion mediated (even 100% mediated) in settings in which S exhibits the surrogate paradox. Although the natural direct and indirect effects themselves (if known) could be useful in diagnosing the surrogate paradox, the proportion mediated criterion itself does not ensure a surrogate is consistent.

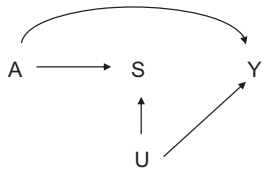

Figure 3. Example of a surrogate S with no effect on the outcome Y.

The "indirect effects" approach, taken as a measure of surrogacy, also suffers from another problem. Consider the causal diagram in Figure 3 in which the surrogate S has no effect on the outcome Y. Now it may be the case that although S has no effect on Y, it may, because of a common cause U, serve as a very good proxy for Y. Knowing about the value of S may be strongly predictive of what will occur with Y potentially for both the treatment and the control arm of a trial. In this case, S could still be a very useful and informative surrogate. However, the natural indirect effect, $E(Y_{1S_1} - Y_{1S_0})$, would be 0 because S has no effect on Y. The measure of surrogacy in (2) would be 0 even though S might be a highly informative surrogate. Whereas the "proportion explained" measure is essentially too liberal for mediation (but may be useful for surrogacy), the "indirect effect" measure is too conservative to assess surrogacy (even though it may be of use in assessing mediation). A good surrogate need not mediate the effect of treatment on the outcome if it is otherwise informative of the effect of treatment on the outcome. Conceived of another way, although confounding is important to consider in evaluating the surrogate paradox, when considering measures of surrogacy it is not always simply a problem to be gotten rid of, but can provide valuable relations between S and Y which may be helpful in predicting the effect of A on Y from the effect of A on S. The "indirect effects" approach by attempting to control for or eliminate confounding essentially misses this potentially important source of information concerning surrogacy. The "indirect effect" measure of surrogacy in (2) may be of use when most of the effect of A on Y is in fact mediated through S and when the confounding between S and Y is weak but in general it eliminates, rather than incorporates, information that may be of importance for assessing the value of a surrogate.

Much of the literature seems to treat the problems of surrogacy and direct/indirect effects as almost interchangeable problems, and certainly the concepts and methods that have been employed have overlapped considerably for surrogacy and mediation. The goals, however, are quite different. In mediation analysis, we are interested specifically in whether there is an effect of treatment on the outcome that operates

through the intermediate. This setting may also be of interest when assessing the properties of a surrogate; but with surrogate outcomes there are settings, as illustrated in Figure 3 above, in which a variable may serve as a very valuable surrogate even if it does not mediate at all the effect of treatment on the outcome. Whereas mediation concerns the pathways by which effects arise, surrogacy concerns principally whether we are able to predict the direction of one effect (of treatment on the outcome) by using another (the treatment on the surrogate). Confounding plays a very different role in questions of mediation versus questions of surrogacy. Whereas it is a problem in assessing mediation, it may be an important source of information in surrogacy. The causal estimands best used to capture mediation and surrogacy also differ. The natural indirect effect (Robins and Greenland, 1992; Pearl, 2001) is arguably the most important counterfactual contrast when assessing mediation. However, as argued above, it may, at least in some settings, be of limited interest in assessing surrogacy. A good surrogate need not mediate the effect. While methods developed for mediation and for surrogacy will undoubtedly inform methodology in the other area, the goals and the questions of each setting should be firmly kept in view in deciding on what concepts, definitions and methods are most relevant.

The third approach considered by Joffe and Greene (2009) may be referred to as the "meta-analytic" approach. It may be applied to subgroups defined across studies (as in traditional meta-analysis) or by creating subgroups based on covariates. Burzykowski et al. (2005), for example, propose using either multiple studies or multiple groups defined by covariates within a study to assess surrogacy. Let Φ_j denote the effect of treatment A on the outcome Y in the jth study/group. Let ϕ_j denote the effect of treatment on the surrogate in the jth study/group. Note that estimation of Φ_j and ϕ_j relies only on the assumption of randomization. To assess surrogacy visually, we could plot estimates of Φ_j against estimates of ϕ_j. For a good surrogate, we would hope to find (i) a monotonic relationship between ϕ_j and Φ_j, (ii) when $\phi_j = 0$ then $\Phi_j = 0$ and (iii) in a (possibly non-parametric) regression of estimates of Φ_j on estimates of ϕ_j we should not find much variability around the regression line. If the relationship between Φ_j and ϕ_j is approximately linear we could run a linear regression of estimates of Φ_j on estimates of ϕ_j and use the R^2 in this regression

$$R^2 = Corr(\Phi_j, \phi_j), \tag{3}$$

as a measure of surrogacy. For this approach to work, however, there must of course be variation in Φ_j and ϕ_j and there must be multiple studies or subgroups in which to estimate effects. Let us now turn to the question of the relation of the meta-analytic approach to the surrogate paradox and the notion of a consistent surrogate. The meta-analytic approach does not give a criterion that ensures the absence of the surrogate paradox, but it can help diagnose and circumvent it. With the meta-analytic approach, if sample sizes are sufficiently large and estimates and modeling assumptions sufficiently precise, an investigator will be able to identify which studies or subgroups are subject to effect reversal (the surrogate paradox) and, for such subgroups, avoid the use of the surrogate. The meta-analytic approach does not give a criterion for avoiding the surrogate paradox but may be of use in detecting groups for which the surrogate is not consistent.

The fourth approach to surrogacy considered by Joffe and Greene (2009) is that of "principal stratification." This approach builds on the initial insights of Frangakis and Rubin (2002) and was developed more fully by Follmann (2006), Gilbert and Hudgens (2008), Wolfson and Gilbert (2010) and Huang and Gilbert (2011). Using notions of principal stratification (i.e., conditioning on the joint counterfactual (S_0, S_1)), Gilbert and Hudgens (2008) define as a measure of surrogacy what they call the "causal effect predictiveness surface" given by:

$$CEP(s_1, s_0) = E(Y_1 - Y_0 | S_1 = s_1, S_0 = s_0). \tag{4}$$

If we knew $CEP(s_1, s_0)$ then we would know for each principal stratum $(S_1 = s_1, S_0 = s_0)$ what the effect of treatment would be. For a binary outcome, the notion of principal surrogacy of Frangakis and Rubin (2002) is simply that $CEP(s_1, s_0) = 0$ for $s_1 = s_0$. For example, suppose the surrogate is binary. The effects $CEP(0, 0)$ and $CEP(1, 1)$ are sometimes referred to as "dissociative effects" and $CEP(1, 0)$ (or $CEP(0, 1)$) as an "associative effect." Principal surrogacy requires that the dissociative effects are zero: $CEP(0, 0) = CEP(1, 1) = 0$ that is, that when the treatment does not change the surrogate, the treatment will not change the outcome. Principal surrogacy is often taken as a criterion for a "good surrogate." The notion is theoretically appealing. Unfortunately, as already indicated above, a principal surrogate does not prevent the surrogate paradox (Chen et al., 2007). A principal surrogate need not be a consistent surrogate. This is also illustrated in the example below. If we knew the causal predictive surface $CEP(s_1, s_0)$ for each principal stratum $(S_1 = s_1, S_0 = s_0)$ then this could potentially be useful in diagnosing the surrogate paradox. For example, if we knew we had a principal surrogate (i.e., $CEP(0, 0) = CEP(1, 1) = 0$) and if we also had monotonicity of the effect of A on S so that the principal stratum $(S_1 = 0, S_0 = 1)$ was empty, then the direction of the average treatment effect of A on Y would be of the same sign as $CEP(1, 0)$. However, the criterion of "principal surrogacy" alone (which itself may be difficult to assess) does not ensure a consistent surrogate. Accordingly, Gilbert and Hudgens (2008) modify the definition of a principal surrogate from that of Frangakis and Rubin (2002) to also require what they call 1-sided average causal sufficiency that, for a binary outcome, $S_1 > S_0$ implies $P(Y_1 = 1 | S_1 = s_1, S_0 = s_0) > P(Y_0 = 1 | S_1 = s_1, S_0 = s_0)$. If a surrogate S has the properties of causal necessity and 1-sided average causal sufficiency, it is straightforward to verify that S cannot exhibit the surrogate paradox. This modified criteria could then be used for diagnosing the surrogate paradox.

Unfortunately, like the "indirect effects" approach, the "principal stratification" approach also requires strong assumptions for identification of the causal predictiveness surface. Moreover, even when assumptions have been made to identify effect measures, one still does not know which individuals fall into which strata and thus the measures are difficult to use in making decisions prospectively about which individuals should or should not be treated. Notions of surrogacy based on principal stratification are theoretically

appealing but difficult to identify in practice. Alternative designs and additional assumptions (Follmann, 2006; Huang and Gilbert, 2011) can help with identification of these effects; alternatively, Follmann (2006) and Huang and Gilbert (2011), have argued that an alternative estimand that conditions only on S_1 and ignores S_0 may be easier to identify from data and still of interest, though, as with others, the value of such alternative estimands in ensuring a consistent surrogate is unclear.

In summary, none of the approaches to surrogate outcomes is entirely immune to the surrogate paradox. For the "proportion explained," "indirect effects" and "principal stratification" approaches, none of the standard criterion guarantee a consistent surrogate. The "proportion explained" may be 100% and yet the surrogate paradox may still arise. Likewise the "proportion mediated" using the ratio of the natural indirect effect to the total effect may be 100% and again the surrogate paradox may arise. Finally, a surrogate may be a "principal surrogate" but not a consistent surrogate—the surrogate paradox may still be present. The "meta-analytic" approach does not provide a criterion to avoid the surrogate paradox but it can be useful in diagnosing it. Likewise in the "indirect effects" approach if the natural direct and indirect effects were known, these could be useful in diagnosing the surrogate paradox if it were due to the direct and indirect effects being in opposite directions; and in the principal stratification approach, if the causal predictiveness surface were known this could likewise be useful in diagnosing the surrogate paradox. Unfortunately, however, both the "indirect effects" approach and the "principal stratification" approach suffer from issues of lack of identification; strong assumptions are in general needed to identify these effects, though alternative study designs or sensitivity analysis techniques can sometimes be useful. In light of the aforementioned issues concerning the problems with the surrogate paradox and difficulties in identification, the "meta-analytic" approach may offer the most promise for assessing surrogate outcomes and for making policy and treatment decisions. The approach in principle relies only on randomization assumptions and does not consider effects that require stronger assumptions to identify; moreover, it allows for easier diagnosis of effect reversal manifested in the surrogate paradox. Nonetheless, it is not without its disadvantages as the sample size requirements for effective implementation may be prohibitively large (Gail et al., 2000). Wu et al. (2011) have also recently proposed some empirical criterion to assess consistent surrogate but sample size requirements may likewise make practical implementation difficult.

5. Illustration

To illustrate some of the difficulties with the various approaches considered, especially in the absence of subgroup data required by the meta-analytic approach, consider the following example. Suppose A is randomized, that S has three levels, and that $pr(S_1 = 0, S_0 = 0) = pr(S_1 = 1, S_0 = 1) = pr(S_1 = 2, S_0 = 2) = 0.1$, $pr(S_1 = 1, S_0 = 0) = 0.5$, and $pr(S_1 = 1, S_0 = 2) = 0.2$ and finally suppose $Y = (0.1) * 1(S = 1) + 1(S = 2) + \epsilon_Y$, where ϵ_Y is a standard normal random variable. Here it can be calculated that $E(S_{a=1} - S_{a=0}) = 0.3$, $E(Y_{s=2} - Y_{s=1}) = 0.9$, $E(Y_{s=1} - Y_{s=0}) = 0.1$ but $E(Y_{a=1} - Y_{a=0}) = -0.13$ so that the surrogate paradox is present, with a positive effect of A on S, a positive effect of S on Y, no direct effect of A on Y not through S, but a negative overall effect of A on Y; S is not a good surrogate. If we apply the "proportion explained" approach we get a proportion explained estimate of 100%, suggesting that S is a perfect surrogate. If we apply the "indirect effects" approach, the natural indirect effect and total effect are both -0.13, suggesting 100% mediation and thus that S is a good surrogate, which it is not. The surrogate does, moreover, satisfy Prentice's criteria. Finally, using principal strata, we would have $CEP(0, 0) = CEP(1, 1) = CEP(2, 2) = 0$, implying that S is a "principal surrogate" and, by this criterion, thus a good surrogate. In this example, the associative effect $CEP(S_1 = 1, S_0 = 0) = 0.1$, which is of the opposite sign of the overall effect of treatment on the outcome and of the other associative effect, $CEP(S_1 = 1, S_0 = 2) = -0.9$. If we were to use as a criterion for a "good surrogate" either (i) the proportion explained, or (ii) the ratio of the natural indirect effect to total effect, or (iii) principal surrogacy, then all three of these approaches would suggest that we have a good surrogate, when, in fact, with the surrogate coded as $S \in (0, 1, 2)$, the sign of the effect of the treatment on the surrogate is the opposite of the sign of the effect of the treatment on the outcome, even though the surrogate has a positive effect on the surrogate and even though there is no direct effect of treatment on the outcome not through the surrogate. In this example, failure of transitivity causes the problem. In other examples, unmeasured confounding or the presence of a direct effect may give rise to the surrogate paradox. Note that in this particular example a recoding of S to $(0, 1, 10)$ would resolve the surrogate paradox in that the effect of the treatment on the surrogate would be of the same sign as that of the treatment on the outcome.

6. Concluding Remarks

The surrogate paradox is an important problem. If the effect of the treatment on the surrogate is in the opposite direction of the effect of the treatment on the outcome of interest, policy and treatment decisions may be severely misguided. In the case of ventricular arrhythmia, this very problem resulted in an estimated 50,000 excess deaths (Moore, 1995). In this article, we have reviewed definitions relevant to surrogate outcomes and have specifically considered how these definitions are related to the surrogate paradox, namely that, the effect of the treatment on the surrogate may be positive, the surrogate and outcome strongly positively associated, but the effect of the treatment on the outcome might still be negative. Such effect reversal can arise with what has been defined as "statistical surrogates" (Prentice, 1989), "principal surrogates" (Frangakis and Rubin, 2002) and "strong surrogates" (Lauritzen, 2004). We have reviewed and extended results on sufficient conditions that ensure a surrogate is "consistent" that is, that it avoids the surrogate paradox. These results extend previous literature by showing that there are sufficient conditions that avoid the surrogate paradox even when there is a direct effect of the treatment on the outcome not through the surrogate. The results show that for the surrogate paradox to arise at least one of the following must be present: (i) a direct effect of the treatment on the outcome not through the surrogate, (ii) confounding of the

surrogate–outcome relationship or (iii) a lack of transitivity so that the treatment does not positively change the surrogate for all the same persons for whom the surrogate positively changes the outcome. In the case of the drugs for ventricular arrhythmia (Moore, 1995; Fleming and DeMets, 1996) there was a direct effect of the drugs on mortality not through the surrogate. Other instance of phenomena described in (i), (ii) or (iii) above could likewise give rise to the surrogate paradox in other settings. The conditions and the results of the article are important because they provide simple conditions which allow investigators to predict the direction of the effect of the treatment on the outcome from the direction of the effect of the treatment on the surrogate. We have seen how these notions of consistent surrogates are related to four surrogate assessment approaches described by Joffe and Greene (2009): the "proportion explained" approach (Freedman et al., 1992), the "indirect effects" approach (Taylor et al., 2005), the "meta-analytic" approach (Burzykowski et al., 2005) and the "principal stratification" approach (Frangakis and Rubin, 2002). All potentially suffer from the surrogate paradox. In particular, without imposing further conditions, none of these approaches' criteria to assess whether a surrogate is "good" (e.g., "100% proportion explained," "100% proportion mediated," "principal surrogacy") is sufficient to ensure that the surrogate paradox is avoided. However, a modification of the "principal surrogacy" criterion (Gilbert and Hudgens, 2008) does suffice. The "meta-analytic" approach may also prove useful in in making treatment decisions based on surrogates and circumvents some of the identification issues of other approaches, though sample size requirements (Gail et al., 2000) may make this impractical.

In this article, we have focused on the task of determining when data concerning the effect of treatment on the surrogate can be used to make decisions about the direction of the effect of the treatment on an outcome that is, of assessing whether a surrogate is consistent. We have considered the value of a number of different results and approaches to surrogate outcomes in accomplishing this task. Surrogates may however be useful in other tasks. For example, we might be interested in determining the extent to which we can predict the outcome once we observe the treatment and surrogate; or the extent to which we could use treatment, surrogate and outcome data in one population to predict the effect of treatment on outcomes in another population (or the effect of a different treatment in the same population) for which only data on treatment and surrogate are available. Future research could consider the value of the various approaches considered here (proportion explained, indirect effect, meta-analytic, principal stratification) or other approaches in accomplishing these other tasks and goals for which surrogates may be of use.

7. Supplementary Materials

Web Appendices referenced in Section 3 are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

The author thanks the reviewers, the editor and the associate editor for helpful comments. The research was support by NIH grant ES017876.

References

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

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

Fleming, T. R. and DeMets, D. L. (1996). Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine* **125**, 606–613.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine tirals. *Biometrics* **62**, 1161–1169.

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

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

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R.J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246.

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

Huang, Y., and Gilbert, P. B. (2011). Comparing biomarkers as principal surrogate endpoints. *Biometrics* **67**, 1442–1451.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods* **15**, 309–334.

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

Ju, C. and Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *Journal of the Royal Statistical Society: Series B* **72**, 129–142.

Lauritzen, S. L. (2004). Discussion on causality. *Scandinavian Journal of Statistics* **31**, 189–192.

Lin, D. Y., Fleming, T. R., and DeGruttola, V (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16**, 1515–1527.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* **23**, 607–625.

Moore, T. (1995). *Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster.* New York: Simon and Schuster.

Pearl, J. (2001). Direct and indirect effects. *In Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, 411–420.

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 for direct and indirect effects. *Epidemiology* **3**, 143–155.

Taylor, J. M. G., Wang, Y., and Thiebaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate markers. *Biometrics* **61**, 1101–1111.

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

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21**, 540–551.

VanderWeele, T.J. (2011). Principal stratification: uses and limitations. *International Journal of Biostatistics*, **7**, Article 28: 1–14.

VanderWeele, T. J., Hernán, M. A., and Robins, J. M. (2008). Causal directed acyclic graphs and the direction

of unmeasured confounding bias. *Epidemiology* **19**, 720–728.

VanderWeele, T. J. and Robins, J. M. (2009). The properties of monotonic effects on directed acyclic graphs. *Journal of Machine Learning Research* **10**, 699–718.

VanderWeele, T. J. and Robins, J. M. (2010). Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society, Series B* **72**, 111–127.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis with a dichotomous outcome. *American Journal of Epidemiology* **172**, 1339–1348.

Wolfson, J. and Gilbert, P. (2010). Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics* **66**, 1153–1161.

Wu, A., He, P., and Geng, Z. (2011). Sufficient conditions for concluding surrogacy based on observed data. *Statistics in Medicine* **30**, 2422–2434.

Received July 2011. *Revised December* 2012.
Accepted December 2012.

Discussions

Michael R. Elliott

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
email: mrelliot@umich.edu

Anna Conlon

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

and

Yun Li

Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106, U.S.A.

We thank Vanderweele (2013) for yet another excellent contribution to the causal inference literature, this time in the surrogacy setting. His exploration of the "surrogate paradox" phenomenon is important, especially in light of examples such as the anti-arrhythmic drugs to which he refers, approval of which may have cost thousands of lives (Moore, 1995). Our discussion will explore in more detail the implications of consistent surrogacy for the principal stratification and meta-analytic settings, and consider some extensions suggested by these implications. We retain his notation except where otherwise noted.

1. Principal Stratification Approaches

1.1. *Binary Surrogates and Outcomes*

Table 1 gives the joint distribution of the potential surrogate markers and outcomes in the binary setting. The rows consist of the support for the potential surrogate (S_0, S_1), corresponding to the principal strata (PS). Monotonicity assumes that the treatment is not harmful; assuming that 1 corresponds to a "good" value for the marker/outcome, monotonicity implies $\pi_{41} = \cdots = \pi_{44} = \pi_{4+} = 0$ when assumed for the marker, and $\pi_{14} = \cdots = \pi_{44} = \pi_{+4} = 0$ when assumed the outcome.

Gilbert and Hudgens (2008) suggest that a good surrogate should possess two properties: "average causal necessity" (ACN) and "average causal sufficiency" (ACS). ACN requires that the causal effect of treatment on the outcome be zero when the causal effect of treatment on the surrogate is zero (i.e., conditioning on the PS for which $S_0 = S_1$) and ACS requires a non-zero treatment effect on the outcome when there is a non-zero treatment effect on the surrogate (i.e., conditioning on the PS for which $S_0 \neq S_1$). As Vanderweele notes, Frangakis and Rubin (2002) proposed the associative effect $\text{AE} = E(Y_1 - Y_0 \mid S_1 \neq S_0) = \pi_{22} + \pi_{42} - (\pi_{24} + \pi_{44})$, corresponding to ACS in the binary setting, and the dissociative effect $\text{DE} = E(Y_1 - Y_0 \mid S_1 = S_0) = \pi_{12} + \pi_{32} - (\pi_{14} + \pi_{34})$, corresponding to ACN in the binary setting. "Perfect" principal surrogacy is defined by the DE being equal to 0. Such a requirement will almost never be met perfectly, suggesting that "large" values of AE and "small" values of DE correspond to good surrogates from a principal stratification perspective. Since AE + DE equals the overall causal effect $\text{CE} = E(Y_1 - Y_0) = \pi_{+2} - \pi_{+4}$ on the outcome, Taylor, Wang, and Thiebaut (2005) suggested using the "associative proportion" and "disassociative proportion" AP = AE/CE and DP = DE/CE to "standardize" the effects relative to the PS. When monotonicity is assumed for both the marker and the outcome, the "common associative proportion" $\text{CAP} = \frac{\pi_{22}}{\pi_{12}+\pi 21+\pi 22+\pi_{23}+\pi_{32}}$ has been proposed as a criterion (Li, Taylor, and Elliott, 2010), since the ideal surrogate will perfectly associate causal effects of the treatment on the marker with causal effects of the treatment on the outcome. In the absence of monotonicity, simple measures such as AP, DP, and CAP are not as easily interpretable; thus the full joint distribution of the potential surrogate marker should be assessed, with a focus on large values of π_{22}/π_{2+} and π_{44}/π_{4+} and small values of π_{12}/π_{1+} and π_{32}/π_{3+} (8). The focus of principal surrogacy measures has been on causal mechanisms rather than surrogate consistency; nonetheless both mechanism and consistency are important.

Table 1
Joint distribution of binary potential surrogate marker and outcome without monotonicity assumptions

	$(Y_0=0,Y_1=0)$	$(Y_0=0,Y_1=1)$	$(Y_0=1,Y_1=1)$	$(Y_0=1,Y_1=0)$	
$(S_0=0,S_1=0)$	π_{11}	π_{12}	π_{13}	π_{14}	π_{1+}
$(S_0=0,S_1=1)$	π_{21}	π_{22}	π_{23}	π_{24}	π_{2+}
$(S_0=1,S_1=1)$	π_{31}	π_{32}	π_{33}	π_{34}	π_{3+}
$(S_0=1,S_1=0)$	π_{41}	π_{42}	π_{43}	π_{44}	π_{4+}
	π_{+1}	π_{+2}	π_{+3}	π_{+4}	1

The parameters in Table 1 are not identifiable from the data, with only 8 sufficient statistics for the 15 free parameters without assuming monotonicity and 6 sufficient statistics for 8 free parameters assuming monotonicity. Boundary conditions for the cell probabilities and associated principal surrogacy measures can be identified (1); alternatively, a fully Bayesian approach can be used to obtain inference in the absence of a convergent likelihood function (5), as in Li, Taylor, and Elliott (2010) and Elliott, Li, and Taylor (2013). A second advantage of the Bayesian approach is the ability to formally incorporate reasonable assumptions via prior distributions on $\boldsymbol{\pi}$, many of which are themselves consistent with surrogate consistency. Indeed, monotonicity can be viewed as a strong prior that assumes $P(\pi_{4j}=0)=P(\pi_{i4}=0)=1$ for $i,j=1,\ldots,4$. As Vanderweele notes, marker monotonicity combined with perfect surrogacy and one-sided ACS implies surrogate consistency, since marker monotonicity and perfect surrogacy together imply $E(S_1-S_0)=\pi_{2+}>0$ and $E(Y_1-Y_0)=\pi_{22}-\pi_{24}$, which is positive by the definition of one-sided ACS in the binary setting. If monotonicity is not assumed for either the marker or outcome, then $E(Y_1-Y_0)=(\pi_{22}+\pi_{42})-(\pi_{24}+\pi_{44})=(\pi_{22}-\pi_{24})+(\pi_{42}-\pi_{44})$ under perfect surrogacy, and preserving surrogate consistency requires $(\pi_{22}-\pi_{24})>(\pi_{44}-\pi_{42})$, or $\frac{\pi_{2+}}{\pi_{4+}}\geq\frac{\pi_{4|4}-\pi_{2|4}}{\pi_{2|2}-\pi_{4|2}}$, where $\pi_{ij}=\pi_{j|i}\pi_{i+}$. This suggests a reasonable result, namely that, to avoid the surrogate paradox, the ratio of those whose marker is helped to those whose marker is harmed must be greater than than ratio of the conditional "harmed" effect on the outcome in the "harmed" marker stratum to "helped" effect on the outcome in the "helped" stratum. More generally, $P(S_1=1)>P(S_0=1)$ and $P(Y_1=1)>P(Y_0=1)$ imply

$$\pi_{2+}>\pi_{4+},\ \pi_{+2}>\pi_{+4}, \tag{1}$$

while $P(Y_1=1\mid S_1=1)>P(Y_1=1\mid S_1=0)$ and and $P(Y_0=1\mid S_0=1)>P(Y_1=1\mid S_1=0)$ imply

$$\begin{aligned}\frac{\pi_{33}+\pi_{34}+\pi_{43}+\pi_{44}}{\pi_{13}+\pi_{14}+\pi_{23}+\pi_{24}}&>\frac{\pi_{3+}+\pi_{4+}}{\pi_{1+}+\pi_{2+}},\\ \frac{\pi_{22}+\pi_{23}+\pi_{32}+\pi_{33}}{\pi_{12}+\pi_{13}+\pi_{42}+\pi_{44}}&>\frac{\pi_{2+}+\pi_{3+}}{\pi_{1+}+\pi_{4+}}.\end{aligned} \tag{2}$$

These conditions are coherent with a strong correlation between the potential surrogate marker values and the potential outcome values, since such a correlation will lead to larger values in the numerator and smaller values in the denominator on the left hand side of (2) because the numerators include terms from the diagonal or near-diagonal elements of Table 1, while the denominator contain terms from the off-diagonal elements of Table 1; while (1) while tend to increase the denominators on the right hand side of (2) relative to the numerators. These relationships are also consistent with less restrictive constraints than monotonicity, such as "stochastic monotonicity" $(\pi_{2j}>\pi_{4j},\ j=1,2,3)$ or positive odds ratios among the 2×2 cells $\frac{\pi_{ij}/\pi_{i,j+1}}{\pi_{i+1,j}/\pi_{i+1,j+1}}$, constraints that have been utilized in (7) and (4). Finally, while these quantities in (1) and (2) are estimable from the observed data, we can use the Bayesian approach to assess the posterior probability that conditions (1) and (2) are met, and determine the degree to which the consistency results of (1) and (2) occur in concert with large values of AP and small values of DP for the joint posterior distributions of these quantities.

1.2. *Gaussian Surrogates and Outcomes*

Assuming multivariate normality for marker and outcome in the continuous setting, we have the following joint distribution (Conlon, Taylor, and Elliott, 2013):

$$\begin{pmatrix}S_0\\S_1\\Y_0\\Y_1\end{pmatrix}\sim N\left(\begin{pmatrix}\mu_{S_0}\\\mu_{Y_1}\\\mu_{Y_0}\\\mu_{Y_1}\end{pmatrix},\begin{pmatrix}\sigma^2_{S_0}&\rho_s\sigma_{S_0}\sigma_{S_1}&\rho_{00}\sigma_{S_0}\sigma_{Y_0}&\rho_{01}\sigma_{S_0}\sigma_{Y_1}\\&\sigma^2_{S_1}&\rho_{10}\sigma_{S_1}\sigma_{Y_0}&\rho_{11}\sigma_{S_1}\sigma_{Y_1}\\&&\sigma^2_{Y_0}&\rho_t\sigma_{Y_1}\sigma_{Y_0}\\&&&\sigma^2_{Y_1}\end{pmatrix}\right). \tag{3}$$

As in the binary setting, we focus on the distribution of the treatment effect conditional on S_0,S_1. Letting S_1-S_0 correspond to a continuous extension of the the principal stratum concept, the distribution of $(Y(1)-Y(0)|S(1)-S(0)=s)$ is normal with mean $(\mu_{Y_1}-\mu_{Y_0})+(\frac{\rho_{11}\sigma_{S_1}\sigma_{Y_1}-\rho_{10}\sigma_{S_1}\sigma_{Y_0}-\rho_{01}\sigma_{S_0}\sigma_{Y_1}+\rho_{00}\sigma_{S_0}\sigma_{Y_0}}{\sigma^2_{S_0}+\sigma^2_{S_1}-2\rho_s\sigma_{S_0}\sigma_{S_1}})(s-(\mu_{S_1}-\mu_{S_0}))$. Thus we have $E[Y_i(1)-Y_i(0)|S_i(1)-S_i(0)=s]=\gamma_0+\gamma_1 s$, where

$$\begin{aligned}\gamma_0&=(\mu_{Y_1}-\mu_{Y_0})\\&\quad-\left(\frac{\rho_{11}\sigma_{S_1}\sigma_{Y_1}-\rho_{10}\sigma_{S_1}\sigma_{Y_0}-\rho_{01}\sigma_{S_0}\sigma_{Y_1}+\rho_{00}\sigma_{S_0}\sigma_{Y_0}}{\sigma^2_{S_0}+\sigma^2_{S_1}-2\rho_s\sigma_{S_0}\sigma_{S_1}}\right)\\&\quad\times(\mu_{S_1}-\mu_{S_0})\\\gamma_1&=\left(\frac{\rho_{11}\sigma_{S_1}\sigma_{Y_1}-\rho_{10}\sigma_{S_1}\sigma_{Y_0}-\rho_{01}\sigma_{S_0}\sigma_{Y_1}+\rho_{00}\sigma_{S_0}\sigma_{Y_0}}{\sigma^2_{S_0}+\sigma^2_{S_1}-2\rho_s\sigma_{S_0}\sigma_{S_1}}\right).\end{aligned}$$

ACN is then satisfied if $\gamma_0=0$ and ACS is satisfied if $\gamma_1\neq 0$. It is simple to see that, if ACN is satisfied, $E[Y_i(1)-Y_i(0)|S_i(1)-S_i(0)=s]$ will be in the same direction as

$S_1 - S_0 = s$ as long as $\gamma_1 > 0$, thus avoiding the surrogate paradox. If ACN is not perfectly satisfied, then there is a small range of s for which the surrogate paradox can occur. If $\gamma_0 < 0$, then $E[Y_i(1) - Y_i(0)|S_i(1) - S_i(0) = s] < 0$ for $s \in [0, -\gamma_0/\gamma_1]$; conversely, if $\gamma_0 > 0$, then $E[Y_i(1) - Y_i(0)|S_i(1) - S_i(0) = s] > 0$ for $s \in [-\gamma_0/\gamma_1, 0]$. This result indicates that the more nearly ACN is satisfied relative to ACS, the more constricted the surrogate paradox will be.

Note that $\gamma_1 > 0$ is equivalent to $\frac{\rho_{11}\sigma_{S_1}\sigma_{Y_1} + \rho_{00}\sigma_{S_0}\sigma_{Y_0}}{2} > \frac{\rho_{10}\sigma_{S_1}\sigma_{Y_0} + \rho_{01}\sigma_{S_0}\sigma_{Y_1}}{2}$, requiring the mean of the "within treatment arm" covariances between the marker and the outcome to be greater than the mean of the "across treatment arm" covariances between the marker and the outcome, consistent with intuition about a well-behaved surrogate marker. (3) considered variations on this constraint to improve efficiency and repeated measure properties for principal surrogacy assessment in the multivariate normal setting. As in the categorical setting, this model is not fully identifiable: ρ_s, ρ_t, ρ_{01}, ρ_{10} cannot be directly estimated from the data, although the requirement that the joint variance–covariance matrix in (3) be invertible does impose constraints. Use of priors in a Bayesian setting can help insure these constraints hold: for example, tuned beta priors on ρ_{01} and ρ_{10} can be used to ensure that the prior probability for these "across treatment arm" correlations is zero for negative values but also unlikely to be larger than the minimum of the (estimable) "within treatment arm" correlations.

2. Meta-Analytic Approaches

In contrast to the principal stratification approach, which focuses on causal associations between the treatment, marker, and outcome, the meta-analytic approach focuses on prediction of the association between the effect of A on S (ϕ_i) and the effect of A on Y (Φ_i) across multiple trials. While not requiring untestable assumptions as in the direct/indirect effect or principal stratification approaches, the meta-analytic approach comes with its own set of analytical challenges. Large numbers of identical trials need to be available to achieve sufficient statistical power in evaluating surrogate consistency across trials. However, studies are rarely identically repeated due to cost, ethical considerations, and the complexity of disease progression. If similar studies are conducted, many aspects likely vary, such as treatment compositions, patient disease conditions and eligibility criteria. To deal with this heterogeneity, (2) (BMBRG) proposed a bivariate mixed model used to describe the joint distribution of S_{ij} and Y_{ij}:

$$\begin{aligned} S_{ij} &= \alpha_S + \beta_S A_{ij} + a_{Si} + b_{Si} A_{ij} + \epsilon_{Sij}, \\ Y_{ij} &= \alpha_Y + \beta_Y A_{ij} + a_{Yi} + b_{Yi} A_{ij} + \epsilon_{Yij}, \end{aligned} \tag{4}$$

where

$$\begin{pmatrix} \epsilon_{Sij} \\ \epsilon_{Yij} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma = \begin{pmatrix} \sigma_{ss} & \sigma_{st} \\ & \sigma_{tt} \end{pmatrix} \right) \tag{5}$$

and

$$\begin{pmatrix} a_{Si} \\ a_{Yi} \\ b_{Si} \\ b_{Yi} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, D = \begin{pmatrix} d_{ss} & d_{sy} & d_{sa} & d_{sb} \\ & d_{yy} & d_{ya} & d_{yb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix} \right). \tag{6}$$

The treatment effect in the nth trial is $\Phi_n = \beta_Y + b_{Yn}$. Motivated by the fact that the trial level effects for the marker a_{Sn} and b_{Sn} are typically available in advance for those of the outcome, BMBRG showed that Φ_n given a_{Sn} and b_{Sn} follows a normal distribution with conditional mean

$$E(\Phi_n \mid a_{Sn}, b_{Sn}) = \beta_Y + \begin{pmatrix} d_{sb} & d_{ab} \end{pmatrix} \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} a_{Sn} \\ b_{Sn} \end{pmatrix} \tag{7}$$

and conditional variance

$$\text{var}(\Phi_n \mid a_{Sn}, b_{Sn}) = d_{bb} - \begin{pmatrix} d_{sb} & d_{ab} \end{pmatrix} \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}. \tag{8}$$

The trial-level correlation between S and Y is defined by BMBRG as

$$R^2_{\text{trial}} = \frac{\begin{pmatrix} d_{sb} & d_{ab} \end{pmatrix} \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}}{d_{bb}},$$

so that when $R^2_{\text{trial}} = 1$, $\text{var}(\Phi_n \mid a_{Sn}, b_{Sn}) = 0$.

R^2_{trial} has often been used as measure of the quality of the surrogate marker in the meta-analytic setting (see, e.g., Saad et al. 2010). We note this is closely related to Vandeweele's $\text{Corr}(\Phi_i, \phi_i)$, since $E(\Phi_i \mid a_{Si}, b_{Si})$ is linear in $\phi_i = \beta_S + b_{Si}$:

$$\begin{aligned} E(\Phi_i \mid a_{Si}, b_{Si}) &= \beta_Y + \theta_1 a_{Si} + \theta_2 b_{Si} \\ &= \beta_Y - \theta_2 \beta_S + \theta_1 a_{Si} + \theta_2 \phi_i \end{aligned} \tag{9}$$

for $\theta_1 = \frac{d_{sb}d_{aa} - d_{ab}d_{sa}}{d_{ss}d_{aa} - d_{sa}^2}$ and $\theta_2 = \frac{d_{ab}d_{ss} - d_{sb}d_{sa}}{d_{ss}d_{aa} - d_{sa}^2}$. Thus $\text{var}(\Phi_i \mid a_{Si}, b_{Si}) = 0$ implies $\text{Corr}(\Phi_i, \phi_i) = 1$ as long as $\theta_2 > 0$. However, we note that these conditions do not avoid the possibility of the "surrogate paradox," which in the meta-analytic setting we take to mean, as Vanderweele does, effect reversal (signs of Φ_i and ϕ_i differ); if a_{Si} is large and of the reverse sign of β_Y, β_S, and ϕ_i, the signs of Φ_i and ϕ_i might still differ. Testing the surrogate paradox in the meta-analytic setting is equivalent to testing the null hypothesis H_0: $\phi_i > 0 (\text{or } < 0)$ AND $\Phi_i > 0 (\text{or } < 0)$ for all i, versus H_a: $\phi_i > 0 (\text{or } < 0)$ AND $\Phi_i < 0 (\text{or } < 0)$ for at least one i (Gail and Simon 1985; Zelterman 1990). A Bayesian alternative might be to consider to what extent the joint distribution of (ϕ_i, Φ_i) satisfies the above constraints. Consider a plane with the x-axis defined by ϕ_i and y-axis defined by Φ_i, and note that, from (9) $\beta_y - \theta_2\beta_s + \theta_1 a_{Si} = 0$ corresponds to the "necessity" concept while $\theta_2 \neq 0$ corresponds to "sufficiency." In the event that the directionality

of Y and S are the same, quadrants I and III correspond to surrogate consistency, while quadrants II and IV correspond to the region in which the surrogate paradox occurs. Under (4)–(6), the proportion of the CDF of (ϕ_i, Φ_i) that lies in these two quadrants is given by $F_1(0; \beta_S, d_{aa}) + F_1(0; \beta_Y, d_{bb}) - 2F_2((0,0); \boldsymbol{\beta}, \Sigma)$, where $F_k(x, \Theta, \Psi)$ is the CDF of a k-variate normal distribution with mean Θ and variance Ψ evaluated at x, $\boldsymbol{\beta} = (\beta_S, \beta_Y)^{\mathrm{T}}$, and $\Sigma = \begin{pmatrix} d_{aa} & d_{ab} \\ & d_{bb} \end{pmatrix}$. As long as R^2_{trial} is large and β_S and β_Y are of the same sign, the proportion of joint distribution of (ϕ_i, Φ_i) in the region where surrogacy consistency holds will also be large. Also, this approach will also favor markers for which $E(\Phi_i \mid \phi_i = 0) = 0$, a feature corresponding to the "necessity" concept but which is not typically considered in the meta-analytic approach to our knowledge. This approach could be subsetted to focus on subgroups of concern, as Vanderweele also notes, or extended outside of the assumption of multivariate normality by use of copulas or bivariate kernel density estimators.

A final practical matter is that when treatment effects are small or uncertain—often part of motivations to conduct new trials—it becomes even more challenging for surrogacy consistency to hold. Furthermore, we must be reminded that even if the surrogacy consistency is not rejected through statistical tests, we will need to make extrapolations that the results based on current trials (subgroups) are applicable to future studies. Taking account of surrogate consistency when designing trials involving surrogate markers requires the new trial to have sufficient power to confirm prior observations on surrogacy inconsistencies, whether overall or for certain subgroups. Other strategies include collecting some information on Y in the new trial in addition to the information on S. It has been shown that when a new study only has A and S available but not Y, the statistical power can be extremely poor. On the other hand, a small fraction of information on Y can improve the statistically power tremendously (6), particularly when $R^2_{\text{indiv}} = \sigma^2_{st}/\sigma_{ss}\sigma_{tt}$ is high. For example, when 30 per cent of Y are observed, the lost information due to missingness is almost completely recovered from S when $R^2_{\text{indiv}} = 0.9$. Since randomized trials often recruit subjects sequentially, it is quite possible to consider collecting some information on Y to improve statistical power.

References

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

Conlon, A. S. C., Taylor, J. M. G., and Elliott, M. R. (2013). Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. University of Michigan Department of Biostatistics Working Paper #99. Available at: http://www.bepress.com/umichbiostat/

Elliott, M. R., Li, Y., and Taylor, J. M. G. (2013). Accommodating missingness when assessing surrogacy via principal stratification. *Clinical Trials*, in press.

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

Gail, M., and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patent subsets. *Biometrics*, **41**, 361–372.

Gilbert, D. B., Hudgens, M. G. (2008). Evaluating Candidate Principal Surrogate Endpoints. *Biometrics*, **64**, 1146–1154.

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

Li, Y. and Taylor J. M. G. (2010). Predicting treatment effects using surrogate markers in a meta-analysis of clinical trials. *Statistics in Medicine* **29**, 1875–1889.

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

Li Y., Taylor J. M. G., Elliott M. R., and Sargent D. (2011). Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics* **12**, 478–492.

Moore, T. (1995). Deadly Medicine: Why Tens of Thousands for Patients Died in America's Worst Drug Disaster. New York: Simon and Schuster.

Saad, E. D., Katz, A., Hoff, P. M., and Buyse, M. (2010). Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Annals of Oncology* **21**, 7–12.

Taylor, J. M. G., Wang, Y., and Thiebault, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker, *Biometrics*, **61**, 1101–1111.

Zelterman, D. (1990). On tests for qualitative interactions. *Statistics and Probability Letters*, **10**, 59–63.

Marshall M. Joffe

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6021, U.S.A.

email: mjoffe@mail.med.upenn.edu

Tyler VanderWeele (2013) (henceforth VW) has written a useful paper, which advances the ideas of the surrogate paradox to imperfect surrogates. This paper and earlier papers on the paradox (Chen, Geng, and Jia, 2007; Ju and Geng, 2010) correctly require the field to reevaluate the definition of and criteria for a good surrogate outcome. The paper additionally

proposes sufficient conditions for avoiding the paradox which are more relaxed than those previously published. Unfortunately, these conditions are not directly applicable to proxy or noncausal surrogates, the most relevant class of surrogates for applied settings.

This commentary will thus focus on two main issues:

1. The importance of avoiding the so-called surrogate paradox, a point which was not adequately explained in either this paper or the paper that introduced the notion of the paradox; and
2. How the approach presented here applies to noncausal or proxy surrogates, which do not themselves affect the outcome of clinical interest.

In discussing these issues, we will consider the usefulness of causal ideas in the proposing and evaluating surrogate outcomes.

1. The Importance of the Surrogate Paradox

In his seminal paper, Prentice (1989) defined a surrogate endpoint as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." More recent papers have proposed alternative definitions; for example, Joffe and Greene (2009) define a surrogate endpoint as "an outcome for which knowing the effect of treatment on the surrogate allows prediction of the effect of treatment on the more clinically relevant outcome." We consider why these definitions are inadequate, as both VW and earlier papers on the paradox consider its problematic nature almost self-evident; in fact it is not.

In the case of the Prentice definition, the problem with a putative surrogate meeting the definition yet subject to the paradox is nearly self-evident. Under this definition and the associated criteria, one could correctly reject the null hypothesis of no effect using the surrogate outcome, yet conclude that the treatment is beneficial when it is in fact harmful. One might rephrase the definition by including directionality explicitly. Thus, for a superiority trial, one might define a surrogate outcome as "a response variable for which a test of the null hypothesis of no *positive* relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." As shown in VW as well as in earlier papers (Chen, Geng, and Jia, 2007; Ju and Geng, 2010), the associated operational criteria are not sufficient under the new definition.

It is less self-evident why the surrogate paradox makes inadequate the prediction-based definition discussed by Joffe and Greene (2009). There are several possible ways to predict the effect of treatment on the primary outcome based on its effect on the surrogate. Consider, for example, one based on the Prentice criterion:

$$\begin{aligned} E(Y_a) &= \sum_s E(Y_a|S_a = s)Pr(S_a = s) \\ &= \sum_s E(Y|S = s, A = a)Pr(S = s|A = a) \qquad (1) \\ &= \sum_s E(Y|S = s)Pr(S = s|A = a), \end{aligned}$$

where the second equality follows from randomization and the third if the so called Prentice criterion involving conditional independence ($Pr(Y|S = s, A = a) = Pr(Y|S = s)$) holds ((1) may be viewed as a version proposed by Prentice as applied to continuous outcomes instead of hazards). If the criteria hold, we can in fact identify the outcome distribution among the treated $Pr(Y_1)$ and controls $Pr(Y_0)$ and their contrast using (1) even if the surrogate paradox holds. Consider the illustration in Section 5 of VW (2013); we can compute $E(Y_1) = \sum_s E(Y|S = s)Pr(S = s|A = a) = 0 * 0.1 + 0.1 * 0.8 + 1 * 0.1 = 0.18$ and $E(Y_0) = 0.31$ and thus the risk difference is correctly characterized as -0.13. Thus, we can correctly predict the effect of treatment on the Y from its effect on S. Why is this inadequate?

To answer this, consider how judgments based on surrogate outcomes are typically made. They are based on combining information from a trial with a surrogate outcome and knowledge from sources external to that trial about the association of the putative surrogate with the clinical outcome. Typically, one would not expect the numerical relationships between the surrogate and the clinical outcome derived from other sources to apply precisely in the new setting (e.g., had the current study been extended long enough to have sufficient numbers of events to study the clinical outcome directly). Thus, it is usually expected that the directions of the relationships between the variables will continue to hold in the new setting even if more precise quantification is elusive. As such, formulae such as (1) are of limited use, and judgments are typically made on the basis of directions of associations, which may be more transportable across settings (Pearl and Barenboim, 2011). The criteria proposed by VW are based on such judgments, and so are in concert with the style of reasoning used in these settings and are an advance on earlier criteria.

VW's conditions for avoiding the surrogate paradox involve causal judgments rather than the merely associational ones implicit in or the Prentice criteria. Basing the criteria on causal judgments has advantages and disadvantages. Because S is not randomized and U may not be fully observed, the conditions in Propositions 3 and 4 may not be verifiable from knowledge of the joint distribution of the observable variables; thus, these conditions may not be useful for validation or evaluation of a putative surrogate from a single trial in which A, S, and Y are all measured. In contrast, the conditional independence of the Prentice criteria is approximately verifiable, although the sample sizes needed to verify rough conditional independence can be huge, especially as one demands close approximation to that independence.

Nonetheless, making judgments about potential surrogacy based on causal knowledge is an important advance. The causal judgments required to justify the assumptions underlying Propositions 3 and 4 may be based on external knowledge of causal relations among variables, and so will, at least sometimes, be based on fundamental building blocks of our knowledge (Pearl, 2009). In contrast, failure to reject conditional independence in the Prentice criteria may be less stable across settings, as, in small samples, approximate conditional independence in one setting may be the result of chance cancellations due to different paths (e.g., $A \rightarrow S \leftarrow U \rightarrow Y$ and $A \rightarrow Y$ in Figure 2 of VW) and may not hold in another setting.

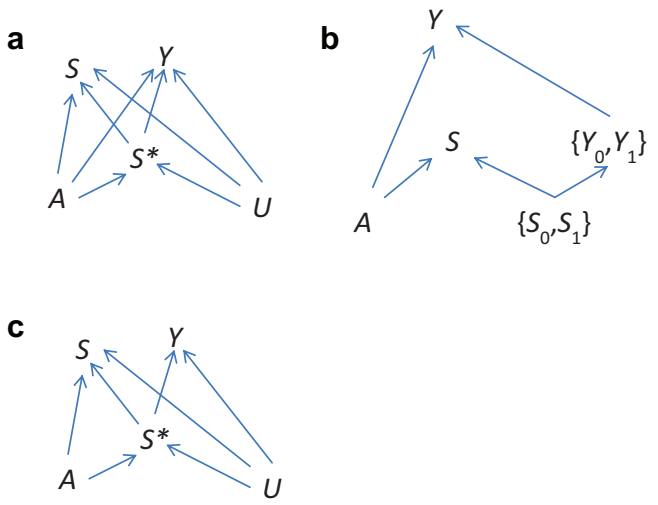

Figure 1. DAGs of proxy surrogates. (a) S is a proxy for the causal intermediate; (b) setup of putative principal surrogate; (c) S is a proxy for the causal intermediate which fully mediates effect of A on Y.

2. Proxy Surrogates

We use the term proxy surrogates to refer to variables which are not themselves on the causal pathway from treatment to outcome but nonetheless are proposed as surrogate endpoints. Most putative surrogates are selected to be related to the causal pathway from treatment to outcome, but may themselves not be on the pathway. While not logically necessary for a surrogate, being on or near the pathway provides a reason why one would expect the effect of the treatment on the surrogate to predict its effect on the clinical outcome. Consider, for example, hemoglobin A1c; it is thought itself not to directly affect outcomes in diabetes but to represent long-term average levels of blood sugar, high levels of which may affect clinical outcomes. Such settings are represented in Figure 1, where S^* represents the true causal intermediate (e.g., long-term course of blood glucose) and S is hemoglobin A1c. If the unmeasured S^* is eliminated from the graph, we recover the graph in Figure 3 in VW, which is also consistent with Figure 2 of VW.

Unfortunately, for proxy surrogates, even if the conditions in Propositions 3 and 4 of VW hold, they are not by themselves useful. To see this, consider a proxy surrogate for which Figure 3 of VW holds. We have that $E(Y|a,s,u) = E(Y|a,u)$ and so is nondecreasing in s, so satisfying part of condition a of Proposition 3. Suppose further that $E(Y|a,s,u)$ is nondecreasing in a for all u. Then $E(Y_a) = E(Y|a)$ is non-decreasing in a even if condition b of the proposition fails. Further, if $E(Y|a,s,u)$ is also increasing in a for some u with nonzero support, we can conclude that $E(Y_a) = E(Y|a)$ is increasing in a even without performing the surrogate experiment which would allow us to see the effect of A on S. Using such logic, one could conclude that a treatment is beneficial without ever performing a trial.

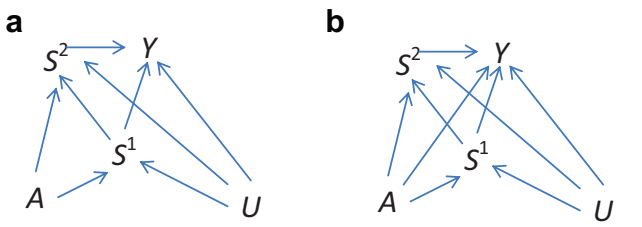

Figure 2. DAGs of time-varying surrogates. (a) $\{S_1, S_2\}$ fully mediates effect of A on Y. (b) $\{S_1, S_2\}$ partially mediates effect of A on Y

It would is thus valuable to consider criteria for consistent surrogates that are tailored to the structure of proxy surrogates. The reason to consider this would be that the structure of Propositions 3 and 4 involve only pretreatment unmeasured confounders U, whereas the structure of the proxy surrogate problem also involves post-treatment unmeasured variables S^*. It is thus plausible that restrictions on that structure would result in more useful operational criteria for identifying consistent surrogates.

An alternative structure for proxy surrogates would involve principal surrogacy or generalizations thereof (Frangakis and Rubin, 2002; Gilbert and Hudgens, 2008). Here the structure of the problem would replace U in Figure 3 of VW by the potential surrogates $\{S_0, S_1\}$ and possibly outcomes $\{Y_0, Y_1\}$ (Figure 1b). Unlike the assumptions underlying Propositions 3 and 4, the assumptions used to rule out the surrogate paradox (i.e., 1-sided average causal sufficiency ($S_1 > S_0$ implies $P(Y_1 = 1|S_1 = s_1, S_0 = s_0) > P(Y_0 = 1|S_1 = s_1, S_0 = s_0)$) and causal necessity) in this paradigm require data from the surrogate experiment to estimate the effect of treatment on the surrogate, and so this might be a fruitful paradigm for using external knowledge for proposing variables as surrogates.

For surrogates that do not fully mediate the effect of treatment (e.g., Figure 2 of VW) or that are proxies for intermediate variables (Figure 1a,c), the condition of causal necessity will generally not hold. Thus, the conditions of Gilbert and Hudgens (2008) outlined in the last paragraph for avoiding the surrogate paradox are not directly relevant. It is easy to show that 1-sided average causal sufficiency together with monotonicity ($S_1 \geq S_0$) is sufficient to rule out the surrogate paradox, even for nonbinary Y; monotonicity, while not fully testable, has testable implications (i.e., that $E(S|A = 1) \geq E(S|A = 0)$).

The original work by Prentice on surrogates involved a setting in which the putative surrogate is not a scalar but is measured repeatedly over time and the outcome a failure-time outcome. Most of the subsequent literature has considered a scalar surrogate measured once and an outcome measured at a fixed follow-up time. The simplification allows more straightforward presentation of certain conceptual issues but also obscures some important aspects of the surrogacy problem.

This is especially true for the causal effects paradigm (Joffe and Greene, 2009). To see this, consider the directed acyclic graph in Figure 2a. Here, the generic putative surrogate S fully mediates the effect of the intervention on the outcome. However, it is only the entire history of S which fully mediates the effect of treatment A; the individual variable S^t at any given time only mediates a part of the treatment effect. Additionally, in typical settings, S^t is measured only intermittently, and so the entire measured S^t process does not fully mediate the effect of treatment. Thus, an individual S^t and

even the entire measured S^t process may be viewed as, at best, partial proxy surrogates for the effect of the treatment of interest.

In the situation in Figure 2a, some specific S^t (e.g., S^1) might mediate only a portion of the effect of A on Y; nonetheless, the effect of A on that S^t might predict the effect of A on Y well if the effect of A on S^t were similar for different times t. Further, the surrogate paradox might be avoided if the nature of the effects of A on S^t and of S^t on Y were common for different times t. It would thus be useful to generalize the conditions of Propositions 3 and 4 to the time-varying surrogate considered in Figure 2a,b (which allows indirect effects of the treatment), or even time-varying proxy surrogates (i.e., $\{S^t\}$ does not mediate effect of A on Y but is proxy for time-varying $\{S^{t*}\}$ which at least partially mediates effect of A on Y). In this setting, the criterion of causal necessity for a perfect principal surrogate would not be satisfied by a single S^t.

3. Discussion

Since Prentice's original formalization, there have been multiple attempts at formalization of criteria for good surrogate endpoints. Several of these have included criteria for perfect surrogates, which typically involve standards unattainable in typical applications, including conditional independence (Prentice, 1989), complete mediation (Chen, Geng, and Jia, 2007), and causal necessity (Frangakis and Rubin, 2002), and a recent manuscript on transportability by Pearl and Barenboim (2011). Attempts to relax these standards have been made, sometimes later (see, e.g., Freedman, Graubard, and Schatzkin, 1992; Gilbert and Hudgens, 2008), and sometimes simultaneously (Frangakis and Rubin, 2002). VW provides an important contribution in this line of thought.

The technical development in VW regard criteria that a putative surrogate should satisfy to avoid the surrogate paradox. As such, these involve criteria that should apply in a hypothetical trial to be performed testing the efficacy of a new agent. Like many previous attempts at criteria for surrogate outcomes, these criteria can, at best, be partially validated for a treatment only after a trial of the treatment has been performed. They are potentially more useful in two ways: (1) using data from earlier trials to examine whether a particular variable seemed reasonable as a surrogate for a previous combination of treatment, outcome, and population, thus making it more plausible for a future combination, and (2) considering whether the presumed causal relationships among variables encoded in graphs would make a particular candidate variable a reasonable surrogate. Because these criteria do not allow prior validation of a potential surrogate for a new setting, VW considers the meta-analytic approach (Daniels and Hughes, 1997; Burzykowski, Molenberghs, and Buyse, 2006), which deals with transportability of results from several studies to new settings, more appropriate for this purpose. It would be valuable to also see whether the approach of VW might apply to modifying the graph-based criteria of Pearl and Barenboim (2011) for transportability to allow for imperfect surrogates and yet avoid the surrogate paradox.

Acknowledgements

This work was supported by NIH grant RC4-MH092722. The content is the responsibility of the author alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

Burzykowski, T., Molenberghs, G., and Buyse, M. (2006), *The Evaluation of Surrogate Endpoints*. New York: Springer.

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

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

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

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

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

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

Ju, C. and Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *Journal of the Royal Statistical Society Series B* **72**, 129–142.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd edition, Cambridge University Press, Cambridge.

Pearl, J. and Barenboim, E. (2011). Transportability across studies: a formal approach," in UCLA Department of Computer Science.

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

VanderWeele, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics*, in press.

Judea Pearl

University of California, Los Angeles, Computer Science Department
Los Angeles, California 90095-1596, U.S.A.
email: judea@cs.ucla.edu

I commend Professor VanderWeele for providing a lucid description of the "surrogate paradox" and, through it, a comprehensive discussion of the current state of thinking about surrogate endpoints, their function in experimental studies,

and the various approaches devised to give them formal underpinnings.

The first question that came to mind in reading Vander-Weele's paper was: can we explain the phenomenon in simple terms, divorced from the technical vocabulary that was devised to formulate notions such as "indirect effect," "principled strata," "proportion-mediated," and perhaps others? My second question was: If we take the negation of the "surrogate paradox" as a criterion for "good" surrogate, why cannot we create a new, formal definition of "surrogacy" that (1) will automatically avoid the paradox and (2) will settle, once for all, the disputes (among theoreticians) as to what "approach" is best for defining surrogates (Joffe and Green, 2009, pp. 530–538; Pearl, 2011).

In thinking about these two questions, I came across a simple way of explaining how the paradox comes about and, indirectly, why the requirement of avoiding the paradox could not, in itself, constitute a satisfactory definition of surrogacy.

As with other paradoxes of causal inference (e.g., Simpson's paradox, Berkson's paradox, suppression effect, and reverse regression) a good starting point is linear models, where the emergence of "paradoxical" phenomena can be examined under the powerful "microscope" of path analysis and elementary linear regression (Pearl, 2013a). If a paradox emerges in linear models, we can be sure that its origin does not rest with effect heterogeneity or idiosyncratic non-linearities, but with the age-old confusion between regression and causation (Pearl, 2013b).

Indeed, starting with the simple linear model of Figure 1(a), we can write the effects of A on S and on Y, as well as the correlation between S and Y in terms of the structural parameters α, β, γ, and δ.[1]

$$E(Y_1 - Y_0) = \alpha + \beta\gamma, \tag{1}$$

$$E(Y|S = 1, A = a) - E(Y|S = 0, A = a) = \beta + \delta, \tag{2}$$

$$E(S_1 - S_0) = \gamma. \tag{3}$$

The surrogate paradox will be exhibited when the effect of treatment A on the surrogate S (3) is positive, S and Y are positively correlated (2), but the effect of A on Y is negative (1), that is, when the structural parameters satisfy:

$$\alpha + \beta\gamma < 0, \tag{4}$$

$$\beta + \delta > 0, \tag{5}$$

$$\gamma > 0. \tag{6}$$

Clearly, for any $\gamma > 0$ and any β, one can find α sufficiently negative and δ sufficiently positive so as to satisfy (4) and (5). Moreover, even for the unconfounded case, $\delta = 0$, shown in Figure 1(b), the three inequality can be satisfied with $\beta > 0$ and α sufficiently negative, namely, $\alpha < -\beta\gamma$.

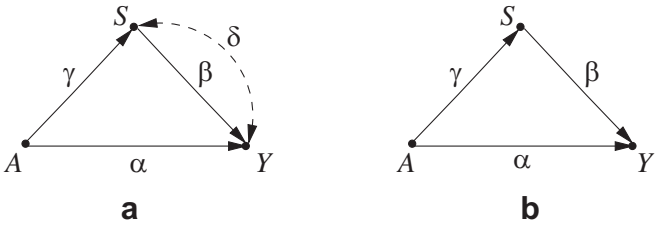

Figure 1. Path diagram in which S acts as a surrogate for the effect of A on Y, demonstrating the "surrogate paradox" under both confounded (a) and unconfounded (b) models.

We conclude that the surrogate paradox may occur in very common models; it does not require confounding, nor interaction or heterogeneity. It requires only that the direct effect of A on Y be sufficiently negative for the paradox to surface. This of course is an unlikely situation in practice. A treatment that has such a negative direct effect on outcome would rarely be a candidate for surrogacy analysis. In practice, the paradox is more likely to take place under confounding conditions ($\delta > 0$) where even a positive α and a negative β will permit it to surface.

We now address the question of why we cannot pose the avoidance of the surrogate paradox, namely, the positivity of all quantities on the left hand side of Eqs. (1)–(3) as a formal definition of a "good" surrogate. Indeed, unlike Simpson's paradox, which stems from a misinterpretation of statistical data (Pearl, 2009, Ch. 6), negating the surrogate paradox expresses precisely what we expect a "good" surrogate to do. It is expected to provide a good prediction of outcome, once it is found to be positively affected by the treatment. Why, then, have researchers labored to define "good" surrogates using fancy formalisms such as "indirect effect," "principal strata,[2]" or "proportion-mediated" (Joffe and Green, 2009) instead of constraining Eqs. (1)–(3) with the proper inequalities?

The reason, I believe, is that definitions are expected to be formulated in terms of the knowledge available to the investigator at the time of the study, and this knowledge consists of qualitative understanding of the model's structure prior to seeing the data, or quantitative assessments of the parameters after examining the data. Eqs. (4)–(6) show that structural knowledge is not sufficient to protect us from the paradox. The paradox may surface even when $\alpha = 0$ (strong surrogacy) or $\beta = 0$. About the only structural condition to prevent the paradox is $\alpha = \delta = 0$, which amounts to perfect mediation (Prentice, 1989). As to quantitative protection from the paradox, the confounding model of Figure 1 does not permit the identification of α, β, and δ, or, in the nonparametric case, of direct and indirect effects.

Another important consideration is robustness. Pearl and Bareinboim (2011) argued that good prediction of the effect of A on Y should not be the sole criterion for judging surrogacy, but must be accompanied with a requirement of robustness.

[1]We assume a randomized trial, hence, A and S are not confounded nor are A and Y. δ stands for the covariance of the "disturbances" affecting S and Y.

[2]The choice of "principal strata" to define surrogacy is particularly inadequate, for these strata are empty in the case of continuous S (Pearl, 2011).

Let us imagine two studies. In the first, we measure the effects of A on both S and Y and confirm that S is a good surrogate, that is, knowing the effect of treatment on S allows prediction of the effect of treatment on the outcome. Once S is proclaimed a "surrogate," it invites efforts to find direct means of controlling S. For example, if cholesterol level (S) is found to be a predictor of heart disease in a long run (Y), drug manufacturers would rush to offer cholesterol-reducing substances for public consumption. As a result, both the prior $P(S = s)$ and the treatment-dependent probability $P(S = s|A = a)$ would undergo a change. For S to be a good surrogate, we should be able to re-assess the effect of the treatment $E(Y_1 - Y_0)$ in a new population in which the effect of treatment on S has changed, and in which access to Y is no longer available. Instead, we have an experiment to assess the new value of $E(S_1 - S_0)$. Pearl and Bareinboim (2011) have shown that, if we assume that the disparity between the two populations lies only in the difference in $E(S_1 - S_0)$ (the surrogate's susceptibility to treatment) the effect of treatment on the outcome under the new conditions can still be estimated from the two studies, provided S and Y are not confounded.

Acknowledgements

I thank the editor for the opportunity to comment on this important paper. This research was supported in parts by Grants from NSF IIS-1249822 and ONR N00014-13-1-0153 and N00014-10-1-0933.

References

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

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd edition. New York: Cambridge University Press.

Pearl, J. (2011). Principal stratification a goal or a tool? *The International Journal of Biostatistics* 7, Article 20, DOI: 10.2202/1557-4679.1322. Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r382.pdf>

Pearl, J. (2013a). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* **1**(1): 155–170, DOI: 10.1515/jci-2013-0003. Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r409.pdf>

Pearl, J. (2013b). Trygve Haavelmo and the emergence of causal calculus. Tech. Rep. R-391, <http://ftp.cs.ucla.edu/pub/stat_ser/r391.pdf>, University of California Los Angeles, Computer Science Department, CA. Forthcoming, *Econometric Theory*, special issue on Haavelmo Centennial.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*. Menlo Park, CA. Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r372a.pdf>

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

Rejoinder

Tyler J. VanderWeele

I thank Professors Pearl, Joffe, and Elliott and colleagues for their comments and insights. It seems to me that one of central methodological question concerning surrogates is, or at least should be, under what conditions the research community should allow a surrogate to be used as the primary outcome in a randomized trial. I will respond to the comments of Pearl, Joffe, and Elliott et al. with this central question in view. In my article (VanderWeele, 2013), I laid out criteria for ruling out settings in which (i) the treatment had a positive effect on the surrogate, (ii) the surrogate and the outcome were strongly positively correlated but (iii) the effect of the surrogate on the outcome was negative, a setting described as the "surrogate paradox." When the sign of the effect of the treatment on the surrogate is the same as that of the treatment on the outcome, the surrogate is said to be "consistent" (the surrogate paradox is absent). In settings manifesting the surrogate paradox we would not want to use the surrogate as the primary outcome in a randomized trial because it could give us the wrong conclusions about the outcome of ultimate interest. The three criteria laid out in my article were (A) if there is a direct effect of the treatment on the outcome not through the surrogate then it is positive, (B) the positive correlation between the surrogate and the outcome indicates an actual positive effect of the surrogate on the outcome and is not simply due to confounding, and (C) if the surrogate is not binary then the effect of treatment on the surrogate is not simply positive on average but rather the treatment increases the whole distribution of the surrogate. This last criterion rules out settings in which the treatment positively changes the surrogate for different individuals than for whom the surrogate positively changes the outcome resulting in lack of transitivity and effect reversal. As pointed out in my article and by Pearl and Joffe, these criteria are sufficient to rule out the surrogate paradox but not necessary. As also pointed

out in my article and by Pearl and Joffe, these criteria cannot in general be evaluated empirically and must be decided upon on a priori substantive grounds. I will argue below that these criteria can be useful in addressing the fundamental question of under what conditions a surrogate should be allowed to be used as the primary outcome in a randomized trial.

Concepts Versus Decisions and Decision-Making Contexts

In their commentaries, both Elliott et al. and Joffe turn to notions of principal stratification (Frangakis and Rubin, 2002; Gilbert and Hudgens, 2008). Within this framework we hope that for individuals for whom the treatment does not change the surrogate, the treatment has no effect on the outcome ("causal necessity"); and for individuals for whom the treatment changes the surrogate, the treatment has an effect on the outcome ("causal sufficiency"). These notions seem to capture well the meaning of surrogacy. Unfortunately, because we do not know who is in which stratum, we cannot identify effects within such principal strata. As discussed by Joffe and Elliott et al., by using Bayesian approaches and sensitivity analysis, we can, to a certain extent, bound these effects. Although these concepts are theoretically appealing and arguably capture what is meant by surrogacy, what still seems unclear is how they are useful in decision-making in the context of surrogate outcomes.

This brings us to the question of what are the relevant decision-making contexts for surrogate outcomes. Suppose we have one or more trials with data on the treatment and the surrogate and the outcome and, on some grounds, we believe we have identified a good surrogate. As indicated by Joffe, one context of interest with surrogates might be evaluating the same treatment in a new population using a new trial with only the surrogate as the outcome; another context of interest might be evaluating a different treatment in the same population using a trial with only the surrogate as the outcome. We would then be interested in when it is sufficient to only use the surrogate alone in these new trials? As suggested by Joffe, one could imagine two approaches to try to settle such questions. First, one might use the data from one or more of the existing trials with data on the treatment, surrogate, and outcome to *empirically* assess whether the surrogate would have in some sense been a good surrogate in the existing trial(s) and if so, then hope that it would likewise be a good surrogate for a new drug or in a new population. Second, it might be possible to articulate criteria (as was done in my article), that could be assessed on a priori substantive grounds, and that would ensure that the direction of the effect of the treatment on the surrogate would match that of the treatment on the outcome. One could attempt to assess whether these criteria were likely to be satisfied with a new drug or in a new population.

A difficulty with the first approach is that, as discussed in my article, the empirical approaches that have been considered most often to date (Joffe and Greene, 2009) generally do not suffice to preclude the surrogate paradox. The *empirical* criteria for "good surrogates" used in the proportion-explained approach, the indirect effects approach, and the principal stratification approach do not, without further assumptions, ensure the surrogate paradox is absent. Fortunately, a recent set of empirical criteria have been put forward by Wu, He, and Geng (2011) that do suffice to ensure consistent surrogates. Wu et al. (2011) show that if one of $E(Y|A=1,s)$ or $E(Y|A=0,s)$ is non-decreasing in s, and if $E(Y|A=1,s) \geq E(Y|A=0,s)$ for all s, then the surrogate paradox is avoided in the sense that if A has a positive distributional effect on S, that is, $P(S>s|a)$ is non-decreasing in a, then $E(Y|A=1)-E(Y|A=0) \geq 0$. These empirical criteria have the advantage over the empirical approaches previously proposed in that they ensure consistent surrogates. However, when considering a new drug or a new population, one would still have to hope that the empirical criteria that ensured surrogate consistency in the original population with the original drug also held with the new population or the new drug. As suggested by Joffe, it is arguably this move to a new drug or to a new population in which a priori criteria to assess whether a surrogate is consistent (judged on substantive grounds and background knowledge, rather than empirically) are of most use. The criteria articulated in my article could be used to assist with such decision-making in the context of a new population or a new drug. For instance, before proceeding with a trial that uses only the surrogate as the outcome, one would want to know, for a new drug, whether the drug might have a negative direct effect (e.g., side effects) not through the surrogate; whether the surrogate really does have a positive effect (not just a positive association with the outcome); and whether the drug might change the surrogate for different individuals than for whom the surrogate changes the outcome. As pointed out by Pearl, the criteria in my article are sufficient but not necessary, and confounding that exaggerates the surrogate-outcome relationship may, for instance, be offset by a positive direct effect. It may be possible to further refine my criteria. With neither the a priori criteria, nor with empirical criteria assessed in the original population and with the original drug, can we be certain that a surrogate will be consistent and suitable for use with a new population and a new drug without making assumptions. But together, empirical and a priori approaches can perhaps be of some use in informing such decision-making.

If none of the major existing *empirical* approaches considered by Joffe and Greene (2009) can ensure surrogate consistency, we might ask whether these empirical approaches are at all useful for other purposes. Note first that, as suggested by Pearl, the consistency of a surrogate is in some sense a relatively minimal requirement for a surrogate. Generally, if we think we have a good surrogate we hope not only to get the direction of the effect of the treatment on the outcome right from the trial of the treatment on the surrogate, but we also hope that if the effect of the treatment on the surrogate is large, then the effect of the treatment on the outcome should be large as well. The meta-analytic approach (Burzykowski, Molenberghs, and Buyse, 2005) can help assess this latter objective concerning effect sizes; the principal stratification approach might also be useful in assessing this if sufficiently narrow bounds on the principal stratum effects can be achieved. Second, instead of deciding whether to allow a trial in a new population or with a new drug with only the surrogate as the outcome, a distinct decision-making context

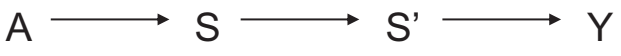

Figure 1. Two surrogates, S and S', both with 100 % proportion explained but with S' a better predictor for Y than S.

might also arise concerning surrogacy when the original population and the original drug are in view. We might, for example, want to assess whether a surrogate is "good" in a prognostic sense, that is, whether once we know treatment status A and surrogate value S, we can do a good job predicting the outcome Y. This might be of interest if a patient is trying to decide how to spend the last weeks or months or years of her life and this decision will depend in part on how much longer she is expected to live. This, I believe, is quite a distinct decision-making context and the proportion-explained approach (Freedman, Graubard, and Schatzkin, 1992) perhaps best captures, of the approaches considered by Joffe and Greene (2009), a measure of the importance of a surrogate in this decision-making context, but one could presumably do even better with more sophisticated predictive modeling approaches. Moreover, two surrogates (e.g., S and S' in Figure 1) might each have a proportion explained of 100% but S' might be much better than S for predictive/prognostic purposes.

Different Surrogates, Different Approaches: A Structural Classification of Surrogates

Joffe pointed out that the criteria in my article were of most use in settings in which the surrogate is on the pathway from the treatment to the outcome, and that different criteria may be more useful in cases of what he called "proxy surrogates." More generally, we might distinguish three types of surrogates: those that are on the pathway from treatment to outcome as in Figure 2a; those that are related to variables on the pathway from treatment to outcome as in Figures 2b, c; and those that are not on the pathway and unrelated to variables on the pathway, as in Figure 2d. We might refer to these three classes as mediator surrogates (Figure 2a), proxy-mediator surrogates (Figures 2b, c), and prognostic surrogates (Figure 2d), respectively. More formally, under faithfulness (Pearl, 2009), we might say that a variable S is a surrogate for the effect of A on Y if there is a directed path from A to S and there is an unblocked path from S to Y not through A. A surrogate S is a mediator surrogate if it is on a directed path from A to Y. A surrogate S is a proxy-mediator surrogate if S is not on a directed path from A to Y but there is an unblocked path from S, not through A, to some variable M that is on a directed path from A to Y. A surrogate S is a prognostic surrogate if S is not on a directed path from A to Y and there is no unblocked path from S, not through A, to a variable M that is on a directed path from A to Y.

As argued by Joffe, the criteria in my article are of primary use for mediator surrogates. Joffe suggests it may be useful to have alternative criteria for surrogate consistency for proxy-mediator surrogates (Figures 2b, c). In fact, the

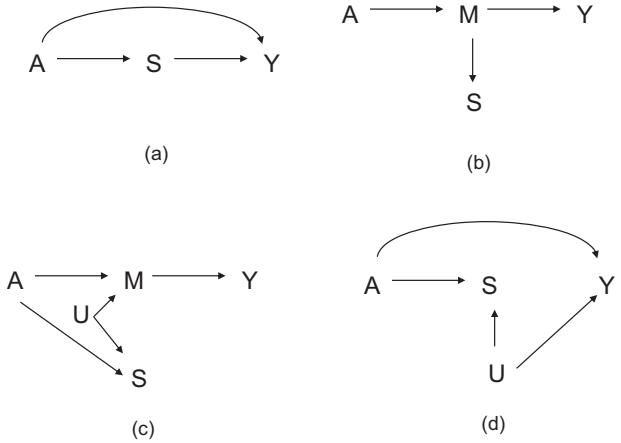

Figure 2. Different types of surrogates: mediator surrogates (a), proxy-mediator surrogates (b,c), and prognostic surrogates (d).

criteria in my article could potentially be extended to proxy-mediator surrogates. If we have a proxy mediator surrogate S, then the surrogate paradox is avoided if there is some variable M on the pathway from A to Y such that the criteria in my article (criteria A–C above) hold with respect to M and if a positive distributional effect of A on S ($P(S > s|a)$ is non-decreasing in a) implies a positive distributional of A on M ($P(M > m|a)$ is non-decreasing in a). If M is binary, this latter condition can be reduced to a positive average effect of A on S implies a positive average effect of A on M. However, a priori criteria would still need to be articulated for this latter condition (almost the reverse of the problem of surrogacy).

This brings us to prognostic surrogates (Figure 2d), that is, surrogates unrelated to variables on the pathway from treatment to outcome. Such surrogates may still be of interest if, for example, they were to denote a side effect that occurs only when the drug is going to be effective (e.g., an individual's hair turning blue if and only if the drug will be effective). In such cases, the surrogate may predict the outcome very well even though it does not cause the outcome. Here, although the criteria in my article are still valid, and may be useful in diagnosing the potential for the surrogate paradox, the criteria are less useful in establishing that a surrogate is consistent because, as suggested in a related context by Joffe, in Figure 2d the criteria reduce to simply assessing whether the treatment itself improves the outcome; we would not need data from the trial of the treatment on the surrogate at all. As suggested by Joffe and discussed in more detail by Elliott et al., a priori criteria based on ideas of principal stratification, may be of more use here. Many of the a priori criteria noted by Joffe and Elliott et al. make strong individual-level monotonicity assumptions and assumptions about causal necessity, and so require more stringent conditions than those in my article. However, Elliott et al. do also provide one set of criteria, based on principal strata probabilities, that do not require the monotonicity assumption. More research along these lines, particularly in the case in which average causal necessity does not hold, would be of interest. It

would also be of interest to assess whether and to what extent we have real examples of surrogates that are prognostic surrogates with a structure such as that of Figure 2d, rather than Figures 2a–c.

The discussion here also raises the issue considered by Joffe of the extent to which a surrogate must be a mediator or be related to a mediator. The final setting of prognostic surrogates in Figure 2d suggests that a good surrogate need not be a mediator or even related to a mediator. It was in part for this reason that I argued in my article that approaches to surrogacy based on assessing mediation and indirect effects may not be adequate. A good surrogate need not be a mediator. In settings which the surrogate is in fact a mediator, assessing the extent of mediation may be useful in trying to refine treatment to better change or target the mediator, but again mediation is not a logically necessary criterion for surrogacy itself.

In summary, with the concepts, approaches and criteria currently available, I believe the notions from principal stratification perhaps best capture what we mean by a surrogate. However, in establishing surrogate consistency on a priori grounds, which is useful in deciding whether to allow a trial using just the surrogate as an outcome, I believe my criteria are most useful for mediator surrogates and proxy-mediator surrogates; and I believe the principal strata criteria in Elliott et al. may be the most useful to date for prognostic surrogates. In establishing surrogate consistency on empirical grounds, I believe the criteria of Wu et al. (2011) are most useful. In establishing whether the magnitude of the effect size in a trial of the treatment on the surrogate would correspond to the effect size in a trial of the treatment on the outcome, I believe the meta-analytic approach (Burzykowski et al., 2005) is most useful. In refining treatments to better target mechanisms and increase effect sizes, I believe the indirect effects approach is most useful. And in using surrogates for the purposes of predicting outcomes with the same drug and same population I believe the proportion-explained approach, or preferably better predictive modeling approaches, are most useful. How we approach surrogacy should depend on the nature of our goals and the type of surrogate.

Community Consensus and Practical Considerations: The Sociology of Surrogacy

I began this discussion with the central question of under what conditions the research community should allow a surrogate to be used as the primary outcome in a trial in which some other outcome is ultimately of interest. In actual practice, research communities do, with time, settle upon various surrogates which are considered acceptable to use. The HIV treatment literature often takes CD4 count as a surrogate for mortality. Joffe mentions the use of hemoglobin A1c as a surrogate for diabetes outcomes. While the process by which consensus within a research community is established is not always clear, I suspect, in many cases at least, it involves a combination of empirical evidence and substantive understanding which are thought to establish that the surrogate is either an important mediator or related to an important mediator, and that if there is a direct effect of the treatment on the outcome not through the mediator then either it is not too large or it is in the correct direction. I believe these are the practical circumstances under which a community often comes to consensus that a particular surrogate is acceptable to use. Often the surrogates established by communal consensus, for example, CD4 count in HIV treatment, work reasonably well. Sometimes, however, as was seen with the drugs to treat ventricular arrhythmia (Moore, 1995), the surrogates have disasterous consequences. One way to interpret my criteria is that they provide an additional set of questions and guidelines to consider in the process of establishing community consensus concerning the use of a surrogate. The questions my criteria pose are: (A) Might the treatment have a negative direct effect not through the surrogate? (B) Might observed associations between the surrogate and the outcome be due to confounding rather than causation? (C) Might the treatment affect the surrogate for a very different group of individuals than for whom the surrogate affects the outcome? If the answer to all three questions is "no," researchers can more confidently proceed with using a particular surrogate with less concern that the direction of the effect, in trials with the surrogate as the primary outcome, is wrong. If the answer to one or more of these questions is "yes," this does not necessarily mean the surrogate is bad, but merely that the research community should proceed much more cautiously. The surrogate paradox may be present.

I would like to conclude by taking up the issue of surrogate robustness raised by Pearl. Pearl argued that the usefulness of a surrogate should not depend on whether the prevalence of the surrogate changes when efforts are made to alter the surrogate distribution to prevent disease. While I think it would in general be good to know something about the "robustness" of a surrogate, so defined, I do not think such robustness should be a requirement for a good surrogate. Stated more generally, I think whether a surrogate is useful will depend heavily on time and context. For HIV patients, CD4 count may at present be a good surrogate for mortality. If, due to technological change and medical advance, we are eventually able to keep patients alive with much lower CD4 counts, then it may later no longer be a good surrogate for mortality. What occurs in the future does not, however, mitigate its usefulness at the present time. Surrogates, when used carefully and properly, should assist in the reduction or eradication of disease. As this is accomplished, the surrogate's usefulness may become diminished. The utility of a surrogate ultimately depends on context.

References

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

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

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

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

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

Moore, T. (1995). *Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster*. New York: Simon and Schuster.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

VanderWeele, T. J. (2013). Surrogate outcomes and consistent surrogates. *Biometrics*, in press.

Wu, A., He, P., and Geng, Z. (2011). Sufficient conditions for concluding surrogacy based on observed data. *Statistics in Medicine* **30**, 2422–2434.