In Biometrics, 2002, 58, 21–29.

Principal Stratification in Causal Inference

Constantine E. Frangakis
Department of Biostatistics,
Johns Hopkins University
615 N. Wolfe St., Baltimore, MD
21205, U.S.A.
cfrangak@jhsph.edu

Donald B. Rubin
Department of Statistics,
Harvard University
1 Oxford St., Cambridge, MA
02138, U.S.A.
rubin@stat.harvard.edu

SUMMARY. Many scientific problems require that treatment comparisons be adjusted for post-treatment variables, but the estimands underlying standard methods are not causal effects. To address this defficiency, we propose a general framework for comparing treatments adjusting for post-treatment variables that yields "principal effects" based on "principal stratification". Principal stratification with respect to a post-treatment variable is a cross-classification of subjects defined by the joint potential values of that post-treatment variable under each of the treatments being compared. Principal effects are causal effects within a principal stratum. The key property of principal strata is that they are not affected by treatment assignment, and therefore, can be used just as any pre-treatment covariate, such as age category. As a result, the central property of our principal effects is that they are always causal effects, and do not suffer from the complications of standard post-treatment-adjusted estimands. We discuss briefly that such principal causal effects are the link between three recent applications with adjustment for post-treatment variables: (i) treatment noncompliance; (ii) missing outcomes (dropout) following treatment noncompliance; and (iii) "censoring by death". We then attack the problem of surrogate or biomarker endpoints, where we show, using principal causal effects, that all current definitions of surrogacy, even when perfectly true, do not generally have the desired interpretation as causal effects of treatment on outcome. We go on to formulate estimands based on principal stratification and principal causal effects, and show their superiority.

KEY WORDS: Biomarker; Causal inference; Censoring by death; Missing data; Noncompliance; Post-treatment variablee; Principal stratification; Quality of life; Rubin causal model; Surrogate.

1. Background.

Decisions in medicine, public health, and social policy depend critically on appropriate evaluation of competing treatments and policies. The extraction of information about such comparisons, which we can broadly view as causal inference, has been a growing area of statistical research in recent years. A statistical framework for causal inference that has received especially increasing attention is the one based on "potential outcomes", originally introduced by Neyman (1923) for randomized experiments and randomization-based inference, and generalized and extended by Rubin (1974, 1977, 1978) for nonrandomized studies and alternative forms of inference. Fundamentally, in this framework, often termed Rubin's causal model (Holland, 1986), a unit (e.g., a patient) is considered at a particular place and time; treatments are interventions each of which can be potentially applied to each unit; and potential outcomes are all the outcomes that would be observed when each of the treatments would be applied to each of the units. Then, a causal comparison between, say, two treatments is a comparison of the potential outcomes of the same group of units under the two treatment conditions.

A major difference between the potential outcomes and other frameworks for causal inference (e.g., simultaneous equations, Goldberger, 1972; Heckman, 1978) is that in the former, the definition of causal effects is separated from any probability models about the way in which units are assigned to treatments, namely the assignment mechanism (Rubin, 1978), and this separation is regarded broadly (though not uniformly; see, e.g., Dawid, 2000) as useful. This clarifying role of potential outcomes has been important in research, including, for example, the earlier works on the concept of ignorable assignment (def. Rubin, 1974, 1977, 1978); propensity scores (def. Rosenbaum and Rubin, 1983a); the concept of sequential ignorability and associated methods (Rubin, 1978; Robins, 1986), and others. More recently, methods are also becoming available to address treatment noncompliance using potential outcomes, starting mainly with work by Baker and Lindeman (1994), Imbens and Rubin (1994), Robins and

Greenland (1994), Angrist, Imbens, and Rubin (1996), and currently receiving even more attention (e.g., Frangakis and Rubin, 1999; Hirano et al., 2000), although an earlier related approach was discussed by Sommer and Zeger (1991).

In Sec. 2 we discuss the more general problem of how to formulate comparisons of treatments adjusting for a post-treatment outcome variable that is not the primary endpoint. We document that the current estimands called "net-treatment comparisons" are not causal effects, as noted by Rosenbaum (1984). We also discuss that other current estimands in this problem (e.g., Robins and Greenland, 1992) assume the post-treatment variable is controllable and thus are difficult to interpret when the post-treatment variable is not directly controlled.

In Sec. 3 we present a general framework for comparing treatments where the estimands are adjusted for post-treatment variables and yet are always causal effects: "principal effects" using "principal stratification". A principal stratification with respect to a post-treatment variable is a cross-classification of the units based on their joint potential values of that variable under each of the treatments being compared. Principal effects are comparisons of treatments within principal strata. The key property of a principal stratification is that it is not affected by treatment and, therefore, can be used as a pre-treatment covariate. Thus, the central property of our principal effects is that they are always causal effects. In Sec. 4, we discuss briefly that principal causal effects link three recent applications.

In Sec. 5 we discuss the problem of surrogate endpoints, and show, using principal effects, that all current definitions of surrogacy, even when true, do not generally define causal effects of treatment on outcome. We go on to formulate estimands based on principal stratification and principal effects, and show their superiority. Sec. 6 provides remarks and directions for further research; throughout, we focus on the fundamental issue of definition of estimands rather than methods of estimation.

2. Adjusting causal effects for post-treatment variables: goal and standard approaches.

Consider a group of units $i = 1, ..., n$ where each can be potentially assigned either a standard treatment $(z = 1)$ or a new treatment $(z = 2)$. (For more treatments, see Sec. 6). The objective is to measure an outcome Y (e.g. survival status) at a specific time after assignment of each unit. Let $Y_i(z)$ be the value of Y if unit i is assigned treatment z, for $z = 1, 2$. Then, a causal effect of assignment on the outcome Y is defined to be a comparison between the ordered sets of potential outcomes on a common set of units, e.g., a comparison between

$$\{Y_i(1) : i \in \text{set}_1\} \text{ and } \{Y_i(2) : i \in \text{set}_2\}, \tag{2.1}$$

given the groups of units, set_1 and set_2, being compared are identical (Neyman, 1923, Rubin, 1974, 1978). Examples include a comparison of the means of $Y_i(1)$ and $Y_i(2)$, or the median of $Y_i(2) - Y_i(1)$ for $i = 1, ..., n$. The potential outcomes and the causal effects are generally not all observable, even with random assignment, although such assignment simplifies estimation. When additional covariates are measured prior to the assignment, then comparisons in the subgroup of units with a given covariate value describe subgroup causal effects of the assignment. With no loss of generality and to avoid extra notation, we will subsequently assume we are already within cells defined by observed pre-treatment variables and ignore the issue of sampling of units from a population.

In many types of studies, after each unit i gets assigned treatment Z_i, a post-treatment variable S_i^{obs} is measured in addition to measuring the main outcome Y. For simplicity of notation, we assume the variable S_i^{obs} is binary (e.g., 1 for low, 2 for high), although our approach can be immediately extended to any format (e.g, see Sec. 6). Important types of studies where such post-treatment variables arise, include:

- clinical trials, where a post-treatment variable S^{obs} is a measure of subjects' compliance to the originally assigned treatment;

- studies with long follow-up, where whether or not the subject drops out is a post-treatment variable (missingness of outcome);
- studies where the outcome intended to be recorded can be "censored by death";
- studies comparing drugs for AIDS patients, where "surrogate" markers of progression, such as CD4 count and measures of viral load (Prentice, 1989; Lin, Fleming, and De Gruttola, 1997; Buyse et al., 2000), are post-treatment variables.

The first three are discussed briefly in Sec. 4, and the fourth is discussed at length in Sec. 5.

The variable S^{obs} generally encodes characteristics of the unit as well as of the treatment. For instance, in the example of clinical trials above, post-treatment noncompliance encodes information about efficacy – the effect of taking the treatment, as well as characteristics of compliance behavior of individual subjects. In such cases, an important study goal, and our objective, is to compare the effects of treatments on Y "after adjusting" for the post-treatment characteristics, in a way that the adjusted estimands are causal effects.

A standard method adjusts for the post-treatment variable using a comparison (e.g., difference in means) between the distributions

$$\text{pr}\{Y_i^{obs}|S_i^{obs}=s, Z_i=1\} \text{ and } \text{pr}\{Y_i^{obs}|S_i^{obs}=s, Z_i=2\}, \tag{2.2}$$

where $Y_i^{obs} = Y_i(Z_i)$, the observed outcome. Comparison (2.2) is called the "net treatment" effect of assignment Z adjusting for the post-treatment variable S^{obs} (Cochran, 1957; Rosenbaum, 1984), and compares outcomes under standard versus new treatment for subjects who *got* a common value s (e.g., s="high") of S^{obs}. For example, the current definitions for a surrogate endpoint by Prentice (1989), Freedman et al. (1992), Lin et al. (1997, eq. 2), Buyse and Molenbergh (1998), and Buyse et al. (2000) are all based on regressions in the sense of

(2.2). The key to understanding such adjustments is to recognize that S_i^{obs} is $S_i(Z_i)$, that is, the observed value of one of two potential values $S_i(1), S_i(2)$, depending on treatment assignment.

Assume for simplicity the condition that the treatment assignment Z_i is completely randomized, that is, $\text{pr}(Z_i = 1|S_i(1), S_i(2), Y_i(1), Y_i(2))$ is a common constant across subjects. Then the net treatment comparison (2.2) is equivalent to the comparison between

$$\text{pr}\{Y_i(1)|S_i(1) = s\} \text{ and } \text{pr}\{Y_i(2)|S_i(2) = s\}. \tag{2.3}$$

Comparison (2.3) is problematic if the treatment has any effect on the post-treatment variable (Rosenbaum, 1984), because the groups $\{i : S_i(1) = s\}$ (i.e., who get post-treatment value s under standard treatment) and $\{i : S_i(2) = s\}$ (i.e., who get post-treatment value s under new treatment) are not the same groups of subjects. Then, according to definition (2.1), the comparison (2.3) is not a causal effect. This concern is known to epidemiologists as post-treatment selection bias in estimating causal effects (e.g., see Rosenbaum, 1984; Robins and Greenland, 1992).

Potential values $S_i(1)$ and $S_i(2)$ were also used by Robins and Greenland (1992) (RG) but, like Rosenbaum (1984), RG did not use those values to define causal effects adjusted for the post-treatment variable. Instead, RG used a framework where both the treatment *and* the post-treatment variable are controllable, and defined a priori counterfactual values of outcomes Y that would have been observed under assignment to treatment z and if the post-treatment variable somehow were simultaneously forced to attain a value s. This framework with its a priori counterfactual estimands is not compatible with the studies we consider, which do not directly control the post-treatment variable. Specifically, most of the values of outcomes in this framework are not just unobserved-existent potential outcomes, but are nonexistent (a priori counterfactual) in the studies we consider. For example, consider a subject who, when assigned the standard treatment, yields a low value of the post-treatment CD4: for that subject,

the value of the outcome Y if assignment to standard treatment were to yield a high value of the post-treatment CD4 is nonexistent (i.e., a priori counterfactual) in the study (see also Sec. 5.2). Evidently, no existing approach has suitably addressed these limitations.

3. Principal Stratification and Principal Causal Effects.

Our proposal for adjustment for the post-treatment variable always generates causal effects because it always compares potential outcomes for a *common* set of people. Consider all the potential values of the post-treatment variable jointly, and construct the following partitions.

DEFINITION *(a) The basic principal stratification P_0 with respect to post-treatment variable S is the partition of units $i = 1, ..., n$ such that within any set of P_0, all units have the same vector $(S_i(1), S_i(2))$. (b) A principal stratification P with respect to post-treatment variable S is a partition of the units whose sets are unions of sets in the basic principal stratification P_0.*

An example of a principal stratification P is the partition of subjects into the set whose post-treatment variable is unaffected by treatment in this study (i.e., with $S_i(2) = S_i(1)$) and into the remaining subjects (i.e., with $S_i(2) \neq S_i(1)$). It is important to note that, generally, we cannot directly observe the principal stratum to which a subject belongs because we cannot directly observe both $S_i(1)$ and $S_i(2)$ for any i. For example, a subject with $S_i(1) = 2$ may belong to either stratum $\{i : S_i(1) = 2, S_i(2) = 1\}$ or stratum $\{i : S_i(1) = 2, S_i(2) = 2\}$. It is, nevertheless, also important at this stage to act as if we knew both $S(1)$ and $S(2)$ in order to determine which quantities are causal.

Generally, a principal stratification generates the following estimands.

DEFINITION *Let P be a principal stratification with respect to the post-treatment variable S, and let S_i^P indicate the stratum of P to which unit i belongs. Then, a principal effect with respect to that principal stratification is defined as a comparison of potential outcomes under*

standard versus new treatment within a principal stratum ς in P, that is, a comparison between the ordered sets

$$\{Y_i(1) : S_i^P = \varsigma\} \textit{ and } \{Y_i(2) : S_i^P = \varsigma\}. \tag{3.1}$$

The importance of principal effects draws from their conditioning on principal strata. Although the potential variable $S_i(1)$ generally differs from $S_i(2)$, the value of the ordered pair $(S_i(1), S_i(2))$ is, by definition, not affected by treatment, just like the pair (birthdate, gender). Therefore, we have

PROPERTY 1 *The stratum S_i^P, to which unit i belongs, is unaffected by treatment for any principal stratification P.*

And, by definition (2.1), we have:

PROPERTY 2 *Any principal effect, as defined in* (3.1)*, is a causal effect.*

Expressed in epidemiologists' terminology, if memberships S_i^P were known, stratification of the subjects by S_i^P would adjust for the personal characteristics reflected in the post-treatment variable without introducing treatment selection bias, for any principal stratification P.

The standard net-treatment comparisons (2.3) are functions of the basic principal causal effects and the corresponding distribution across these strata, $\text{pr}(S_i^{P_0} = \varsigma)$. Thus, if we have the basic principal causal effects and the counts of units in each of the basic principal stratum, we learn more, not less, about the problem than if we have only net-treatment comparisons. Moreover, because principal effects are causal effects, their estimation is critical for understanding the process by which treatments act on subjects, and in some situations also useful for more reliable generalization of results, as we shall see.

Setting principal causal effects to be the goal also helps focus the role of inference. Inference about the principal effects, for example, in P_0, requires prediction of the subjects' missing memberships to the principal strata, as determined by $S^{mis} = \{S_i(z) : \text{ all } i \,;\, z \neq Z_i\}$, as well

as prediction of the subjects' missing potential outcomes $Y^{mis} = \{Y_i(z) : \text{ all } i\,;\, z \neq Z_i\}$. Specifically, the observed data are $H^{obs} = (Y^{obs}, S^{obs}, Z)$ and the likelihood is

$$L(H^{obs}\,;\,\theta^S, \theta^Y) = \int\int \text{pr}\{Z \mid (Y(1), Y(2)),\ S^{P_0}\} \times \text{pr}(S^{P_0} \mid \theta^S) \times \text{pr}\{(Y(1), Y(2)) \mid\ S^{P_0}; \theta^Y\}\ dY^{mis}\, dS^{mis}, \qquad (3.2)$$

where θ^S and θ^Y denote parameters governing the proportions of basic principal strata, and the distribution of potential outcomes in these strata, respectively. In (3.2), omission of the unit subscript "i" means collection over all subjects in the data; integration over Y^{mis} operates on the decomposition $Y = (Y^{obs}, Y^{mis})$; and integration over S^{mis} operates on (S^{obs}, S^{mis}) that determine membership to the principal strata. (Note: in problems where a principal stratum implies that the outcome Y^{obs} itself is missing, e.g., as in those discussed in Sec. 4, the likelihood is a modification of (3.2).)

The likelihood function (3.2) can be used for estimation of principal causal effects as functions of θ^Y and θ^S, with either likelihood or Bayesian inference. With no additional assumptions, there is generally no unique maximum likelihood estimate of (θ^Y, θ^S). Nevertheless, we can often build plausible restrictions to capitalize on the scientific structure of each problem, for example, using covariates to predict principal strata, including information on dose-response curves within principal strata, or information on lag until and length of time for a treatment action based on pharmacokinetics. The framework can also host a combination of estimation with sensitivity analyses for the causal effects, for example in the sense of exploring ranges of unobserved quantities as done, in different contexts, in Rosenbaum and Rubin (1983b), Scharfstein, Rotnitzky, and Robins (1999), and Goetghebeur et al. (2000), and whose extreme application results in the use of bounds (e.g., Manski, 1990; Balke and Pearl, 1997).

4. Brief Review of Principal Effects in Three Examples.

We briefly review three examples of recently worked problems involving post-treatment variables, (i) treatment noncompliance; (ii) missing outcomes following treatment noncompliance; and (iii) "censoring by death".

An example of recent methods for addressing treatment noncompliance is Imbens and Rubin (1994, 1997) who reanalyzed a study on vitamin A by Sommer and Zeger (1991). In that study: (a) the controlled intervention was randomization of children to receive vitamin A or not and the outcome was mortality; (b) the uncontrolled post-treatment variable was the actual taking of vitamin A; and interest focused on (c) formulating and estimating the effect of taking versus not taking vitamin A (as opposed to the effect of being assigned or not assigned to take vitamin A). To address (c), Imbens and Rubin (1997) estimated the "complier average causal effect" (CACE), which is a causal effect of assignment on the subjects who would comply with treatment no matter the assignment ("compliers"). Therefore, this approach to adjusting for noncompliance is a special case of the framework of Sec. 3, where the compliers are a stratum in the principal stratification with respect to the post-treatment "compliance behavior". In that and related applications dealing with noncompliance, CACE is a special case of a principal effect. Thus, the following comparison of CACE to other estimands when faced with noncompliance shows the strengths of our framework.

First, CACE is, by Property 2, always a well defined causal effect. In contrast, a standard estimand to evaluate the actual taking of treatment compares the observed outcomes of subjects taking new treatment (vitamin A) to the observed outcomes of subjects taking control, within treatment assignment arm. That is, it compares $\text{pr}(Y^{obs}|S^{obs} = 2, Z = z)$ to $\text{pr}(Y^{obs}|S^{obs} = 1, Z = z)$ for $z = 0, 1$, which, in analogy to (2.2), is a "net-treatment" effect of the new treatment adjusted for assignment. The comparison of these estimands for $z = 0, 1$, also known as an "as-treated" estimand, is not a causal effect without the exchangeability of

prognosis for subjects who take and those who do not take new treatment within assignment arm. Quite generally, however, practitioners and regulatory agencies (e.g., US FDA) do not trust such exchangeabilty assumptions for uncontrolled compliance (e.g., The Coronary Drug Project Research Group, 1980; Zelen, 1990). Other estimands to address the actual taking of treatment are defined by comparing subjects' outcomes that, for a *fixed* level of the controllable assignment z, would have been observed under two scenarios: first, if all subjects (including noncompliers) would have somehow been forced to take the new treatment; second, if the same subjects would have somehow been forced to take the standard treatment. These estimands, therefore, involve outcome values that are a priori counterfactual (see also Frangakis, Rubin, and Zhou, 2001, rejoinder), that is, they do not exist as functions of the controllable factor (z) alone, and, therefore, their meaning as causal effects is not well defined.

Considerable growth of literature has followed or was proposed independently of Imbens and Rubin (1994, 1997) on methods to better address noncompliance (e.g., Baker and Lindeman, 1994; Robins and Greenland, 1994; Angrist et al., 1996, Goetghebeur and Molenberghs, 1996; Robins, 1998; Rubin, 1998). On the other hand, we are aware of no previous work that has linked such recent approaches for noncompliance to the more general class of problems with post-treatment variables.

An important such problem was recently reported by Barnard, Frangakis, Hill and Rubin (2001) in a large experiment to evaluate school choice programs, where (a) the randomized intervention was the offering of school vouchers to children of low income parents, and (b) uncontrolled post-treatment variables were both the actual use of vouchers, and the subsequent taking of tests to measure achievement. For such cases, Frangakis and Rubin (1997, 1999) had shown that in order to estimate even the "intention-to-treat" effect of randomized treatment on achievement ability (i.e, an *effect* that ignores compliance): (i) it is not appropriate to use "intention-to-treat" analyses (i.e., *analyses* that ignore compliance data); and (ii) the princi-

pal strata defined by both compliance and missingness of outcome must be used. Barnard et al. (2001) took into account these principal strata, and thereby proposed a more appropriate method of estimation of intention-to-treat effects as well as of other effects.

Another important such problem is discussed by Rubin (1998, Sec. 6; 2000), "censoring by death": subjects are assigned to treatments, the intended outome is quality of life at one year after assignment, and the post-treatment variable indicates death before the first year. Quality of life is "missing" for such cases, not because a non-null value exists and is unobserved, as often treated by standard approaches, but simply because a non-null value does not exist. Formulating causal effects of treatment on quality of life is subtle, first because such comparisons are restricted by the life of subjects, and second because life, as a post-treatment variable, can be affected by treatment. The outline described in Rubin (2000) to address this problem is another special case of principal stratification.

Other types of post-treatment censoring can also be addressed using principal stratification and effects. For example, Frangakis and Rubin (2001) use a related formulation to address design and estimation of survival curves using double sampling in the combined presence of administrative censoring and loss to follow-up (see also, Baker, Wax, and Patterson, 1993).

5. Defining Surrogate Endpoints Using Principal Causal Effects.

5.1 The two goals of surrogate endpoints and previous approaches revisited.

Often in therapeutic trials, comparison of treatments for the outcome of primary importance, e.g., survival time, may require a long and practically infeasible follow-up. Nevertheless, if there exist variables measurable early in the follow-up and known to be linked to the effect of the treatments on survival, then such variables can arguably help understand the effect of treatment on the outcome. There is currently growing literature on such "surrogate" or "biomarker" endpoint variables (e.g., Prentice, 1989; Freedman et al., 1992; Lin et al., 1997; Buyse et al.,

2000). The most fundamental question is the definition of a surrogate endpoint so that it has an appropriate interpretation and can be used reliably for prediction.

To help fix ideas, consider a study where the treatments are standard $(z = 1)$ and new $(z = 2)$ therapy for AIDS patients. If patient i is assigned treatment z, let $Y_i(z)$ denote the outcome of survival time (the primary endpoint), and let $S_i(z)$ denote the measurement of CD4 count ("H"=high, "L"=low) at 2 months after treatment assignment. Also, to better present our arguments in a simple setting, we assume that: no patient dies before 2 months so that $S_i^{obs} = S_i(Z_i)$ is measured for all subjects, that treatment assignments $\{Z_i\}$ are completely randomized, and that Y_i^{obs} is measured for all subjects, thereby creating what we call a "validation" study.

In order to have an appropriate interpretation as a surrogate, the post-treatment variable S should possess two properties:

PROPERTY 3 Causal Necessity: S is necessary for the effect of treatment on the outcome Y in the sense that an effect of treatment on Y can occur only if an effect of treatment on S has occurred.

PROPERTY 4 Statistical Generalizability: S^{obs} should well predict Y^{obs} in an "application" study, where we do not wait for measurements Y^{obs}.

The property of causal necessity is important because it tells us if the treatment can act on the outcome without acting on the surrogate. This information is central for improving the focus of therapy or drug-development. The property of generalizability is important when it is not feasible to wait for the primary outcome in the application study.

In an early effort to satisfy these properties, Prentice (1989) defined S^{obs} to be a surrogate if it satisfies certain criteria, mainly that the observed outcome $Y_i^{obs}(= Y_i(Z_i))$ should be conditionally independent of the assigned treatment Z_i given the observed value S_i^{obs} of the

post-treatment variable in the validation study. (Prentice, 1989, used a hazard regression parameterization for multiple-time measurements on S^{obs}. For clarity, we discuss the single-time measurement case – the generalization is simple but notationally tedious.) When exact independence is not expected, related approaches have been proposed that compare results of the regression of the outcome on treatment before and after conditioning on the variable S^{obs}, as with comparison of parameter coefficients (Freedman, et al., 1992, Lin et al., 1997), and more recently with comparison of coefficients of determination (e.g., Buyse and Molenberghs, 1998; Buyse et al., 2000; Gail et al., 2000).

More generally, all these approaches are based on "net-treatment" comparisons (Sec. 2), where S^{obs} is considered a surrogate if S^{obs} is a good predictor (relative to treatment Z) of outcome Y^{obs} when conditioning on both S^{obs} and Z in the validation study. Thus, with respect to the way of "adjusting" for S^{obs}, we can collectively regard all such current definitions as variants generated from Prentice's main criterion for defining a statistical surrogate:

DEFINITION (Statistical Surrogate in a Randomized Experiment). *S is a* statistical *surrogate for a comparison of the effect of $z = 1$ vs. $z = 2$ on Y if, for all fixed s, that comparison of the distributions in (2.2) results in equality.*

It may appear that this definition of surrogate based on "net-treatment" comparisons is sufficient for both properties, causal necessity and generalizability. In contrast to standard beliefs (e.g, Prentice, 1989; "individual-level surrogacy" of Buyse et al., 2000), however, none of the approaches based on the definition of statistical surrogacy satisfies Property 3 of causal necessity. As we will show in the next section, in a study where the post-treatment S is a statistical surrogate, there will generally exist units with no causal effect of treatment on the statistical surrogate and who, nevertheless, experience causal effects of treatment on outcome. Conversely, in a study where there is no causal effect of treatment on outcome unless it occurs

together with a causal effect of treatment on the surrogate, S will generally not be a statistical surrogate.

We offer a new criterion for surrogacy using principal stratification and principal causal effects. We show that the new criterion does satisfy Property 3 of causal necessity. Moreover, in Sec. 5.3, we also discuss the role of principal stratification for better satisfying Property 4, statistical generalizability of the surrogate.

5.2 Definition of Principal Surrogate and Property of Causal Necessity.

Principal surrogate.

Consider the basic principal strata of the simple study example of Sec. 5.1:

- subjects whose CD4 count would be low and unaffected by treatment, $\{i : S_i(1) = S_i(2) = L\}$, whom we label for simplicity "sicker" patients;
- subjects whose CD4 count would be high and unaffected by the treatment $\{i : S_i(1) = S_i(2) = H\}$, and whom we label "healthier";
- subjects whose CD4 count under new treatment would be higher than under standard treatment, $\{i : S_i(1) = L$ and $S_i(2) = H\}$, and whom we label "normal";
- subjects whose CD4 count under new treatment would be lower than under standard treatment, $\{i : S_i(1) = H$ and $S_i(2) = L\}$, and whom we label "special";

We propose the following definition of a surrogate based on principal stratification.

DEFINITION *S is a* principal *surrogate for a comparison of the effect of $z = 1$ vs. $z = 2$ on Y if, for all fixed s, that comparison between the ordered sets*

$$\{Y_i(1) : S_i(1) = S_i(2) = s\} \text{ and } \{Y_i(2) : S_i(1) = S_i(2) = s\}, \tag{5.1}$$

results in equality.

That is, causal effects of treatment on outcome Y may only exist when causal effects of treatment on the post-treatment variable S exist. Thus our criterion based on principal stratification immediately satisfies Property 3 of causal necessity of the previous section.

Although definition (5.1) does not involve an assumption about the assignment model for Z_i, under randomization, criterion (5.1) implies that the same comparison applied to

$$\text{pr}\{Y_i^{obs} | S_i(1) = S_i(2) = s, Z_i = 1\} \text{ and } \text{pr}\{Y_i^{obs} | S_i(1) = S_i(2) = s, Z_i = 2\}. \qquad (5.2)$$

also results in equality. The following result then asserts that Property 3 is not shared by a statistical surrogate.

RESULT 1. *In a randomized experiment, and with respect to any comparison, we have that: (a) If the post-treatment variable S is a principal surrogate, then it is not, generally, a statistical surrogate. (b) If the post-treatment variable S is a statistical surrogate, then it is not, generally, a principal surrogate.*

To understand better the implications of Result 1, we offer a proof by discussing the two examples of Figure 1 for the comparison of averages (to show the result, in the figures we need only consider scenarios with no "special" subjects). First consider Fig. 1(a). The subgroups of patients who experience no causal effect of treatment on the CD4 counts ("sicker" and "healthier") experience no causal effect of treatment on survival. Therefore, by criterion (5.2), CD4 count is a principal surrogate in this study.

However, when $s = L$, the subgroup $\{i : S_i^{obs} = L, Z_i = 1\}$ of subjects in the left side conditioning of (2.2) is the mixture of "sicker" and "normal" patients under standard treatment, whereas the subgroup $\{i : S_i^{obs} = L, Z_i = 2\}$ is, in fact, a different group of subjects – the "sicker" patients only – under new treatment. Using the numbers of Fig. 1(a), the left side of (2.2) has mean 20 months, whereas the right side of (2.2) has mean 10 months. It follows that

CD4 is not a statistical surrogate. Therefore, although the standard interpretation would be that the new treatment decreases survival whenever it cannot change a low value of the surrogate, that conclusion is incorrect, as the principal surrogacy of S clearly indicates.

Consider now Fig. 1(b). For the "sicker patients", the new treatment has no causal effect on their CD4 count, but does have a 10 month causal effect on increasing survival (comparing sicker patients' survival under new vs. standard treatment) . Similarly, a 10-month increase in survival holds for the "healthier" patients in the study. Therefore, CD4 count is not a principal surrogate, that is, there can be an effect of treatment on survival when there is no effect of treatment on the surrogate. Using the criterion of statistical surrogacy, however, we obtain that, for $s = L$, both the left and right sides of (2.2) have mean 20 months, and that, for $s = H$, both the left and right sides of (2.2) have mean 50 months, so CD4 is, by definition, a statistical surrogate for the average comparison. Therefore, although, here, the standard interpretation would be that treatment does not change survival without changing the surrogate, this conclusion is incorrect. The discrepancy indicated in Result 1 occurs more generally because a statistical surrogate does not generally involve causal effects.

Associative and Dissociative effects.

We also propose, more generally than assessing principal surrogacy, to evaluate the effects of treatment on outcome that are associative and dissociative with effects on the post-treatment variable in the validation study. An effect on outcome that is dissociative with an effect on surrogate is defined as a comparison between the ordered sets

$$\{Y_i(1) : S_i(1) = S_i(2)\} \text{ and } \{Y_i(2) : S_i(1) = S_i(2)\}, \tag{5.3}$$

that were equated in (5.1). An effect on outcome that is associative with an effect on surrogate is defined as a comparison between the ordered sets

$$\{Y_i(1) : S_i(1) \neq S_i(2)\} \text{ and } \{Y_i(2) : S_i(1) \neq S_i(2)\}. \tag{5.4}$$

Both (5.3) and (5.4) can, in principle, be further stratified on basic principal strata.

Both the associative effect (5.4) and the dissociative effect (5.3) are causal effects, by Property 2 of Sec. 3. If the dissociative effect is large (small), then we are to conclude that there is large (small) causal effect of treatment on outcome for subjects for whom treatment *does not* affect CD4. Similarly, if the associative effect is large (small), then we are to conclude that there is large (small) causal effect of treatment on outcome for subjects for whom treatment *does* affect CD4. A comparison between (5.4) and (5.3) then measures the degree to which a causal effect of treatment on outcome occurs together with a causal effect of treatment on the surrogate. For example, if this association is high, it can indicate that developing a drug to target biophysiological characteristics of the surrogate may be a good way to target the clinical endpoint Y. It is important to note that causal interpretation of the latter association is not automatic, in contrast to (5.4) and (5.3), and should be examined experimentally in a new (perhaps laboratory) study where an intervention manipulating a factor *in addition* to z would be applied, e.g., to increase CD4. For that new study, the potential outcomes would be regarded as functions, not of the uncontrolled post-treatment CD4, but of the new factorial interventions used to change it.

Finally, we emphasize that the approach we present is applicable to continuous post-treatment variables as well, where analogous comparisons are formulated as the conditional distributions of the causal effect of treatment on outcome given principal strata of the post-treatment variable, which differ from the "individual level" comparisons of Buyse et al. (2000, Sec. 4.2) (the latter still being net-treatment comparisons).

5.3 Principal Stratification and Property of Statistical Generalizability.

We now examine the use of principal stratification to predict outcomes in a randomized application study. Here, distinguish the distributions of principal strata and of outcomes given

principal strata between a validation and an application study, respectively:

$$\text{pr}^{\text{V}}\{S(1), S(2)\}, \quad \text{pr}^{\text{V}}\{Y^{obs} \mid S(1), S(2), Z\}, \tag{5.5}$$

$$\text{pr}^{\text{A}}\{S(1), S(2)\}, \quad \text{pr}^{\text{A}}\{Y^{obs} \mid S(1), S(2), Z\}, \tag{5.6}$$

and assume all distributions are available except $\text{pr}^{\text{A}}\{Y^{obs} \mid S(1), S(2), Z\}$.

Before the outcomes Y_i^{obs} in the application study are known, they could be predicted by their predictive distribution, denoted by $\text{pr}^{\text{A}}(Y^{obs} \mid S^{obs}, Z)$. Because the distributions (5.6) determine the distributions of all observable data in that study, we have (under randomization):

$$\text{pr}^{\text{A}}(Y^{obs} \mid S^{obs}, Z) = \frac{\int \text{pr}^{\text{A}}\{Y^{obs} \mid S(1), S(2), Z\} \, \text{pr}^{\text{A}}\{S(1), S(2)\} \, dS^{mis}}{\int \text{pr}^{\text{A}}\{S(1), S(2)\} \, dS^{mis}}. \tag{5.7}$$

Without waiting for any outcome Y^{obs}, however, the correct predictive distribution is not available because $\text{pr}^{\text{A}}\{Y^{obs} \mid S(1), S(2), Z\}$ is not available. To address this, the standard approach predicts the outcomes Y^{obs} in the application study using the predictive distribution from the validation study, $\text{pr}^{\text{V}}(Y^{obs} \mid S^{obs}, Z)$, effectively replacing in (5.7) both distributions of (5.6) with those of (5.5). But the application study can differ from the validation study in *either* the distribution of principal strata or the potential outcomes given principal strata, in which case the validation predictive distribution will be incorrect for the application study. This may help to explain empirical evidence that regressions $\text{pr}^{\text{V}}(Y^{obs} \mid S^{obs}, Z)$ in one validation study can be quite different in another study with the same type of treatment, outcome, and surrogate (e.g., Fleming and DeMets, 1996).

Consider, alternatively, replacing only the outcome component in the right side of (5.7) with that of the validation study, to obtain the synthetic predictive distribution defined as,

$$\text{pr}^{\text{SYN}}(Y^{obs} \mid S^{obs}, Z) = \frac{\int \text{pr}^{\text{V}}\{Y^{obs} \mid S(1), S(2), Z\} \, \text{pr}^{\text{A}}\{S(1), S(2)\} \, dS^{mis}}{\int \text{pr}^{\text{A}}\{S(1), S(2)\} \, dS^{mis}}. \tag{5.8}$$

By any measure, it is more likely that "the left side of (5.6) equals the left side of (5.5)" than it is that "both the right side and the left side of (5.6) equal, respectively, those in (5.5)" . Therefore, using the synthetic predictive distribution (5.8) should be a more plausible approximation to the correct predictive distribution in the application study, than the predictive distribution from the validation study.

6. Remarks and Extensions.

For comparing treatment effects on outcomes adjusting for post-treatment variables, we focused on estimands before estimation, by formulating principal causal effects. We compared our estimands with existing estimands, and separated this discussion from issues of their estimation, which can only be relevant when the estimands are relevant. We discuss the estimation of principal effects in subsequent papers specifically for each open application.

As discussed in Sec. 3, membership of subjects to the principal strata is not generally fully observed, and so estimation must involve techniques for incomplete (missing) data. Moreover, because with no restrictions there is generally a range of parameter values that maximize the likelihood, it is important to couple our framework with plausible additional assumptions specific to each context. Such explicit restrictions (e.g., "latent ignorability" of outcome missingness or the "compound exclusion restriction", Frangakis and Rubin, 1999), can be scientifically more plausible than the implicit assumptions of standard approaches and can also lead to increased precision of estimated principal causal effects. It is, therefore, a distinct advantage that our framework formalizes why and what types of assumptions are needed, and how to incorporate them to make inference in these problems.

Although we concentrated on examples with two treatments and binary post-treatment variable, the framework is immediately applicable to post-treatment variables that are multivariate, (e.g., as in the experiment on school choice, Sec. 4) or time-dependent, or continuous (end of

Sec. 5.2), and to multiple treatments, say $z = 1, ..., k$. In the latter case, the basic principal strata with respect to S are subgroups of subjects with the same vector $(S_i(1), ..., S_i(k))$. Then, as in (3.1), principal causal effects are comparisons of the potential outcomes among strata that are unions of the basic principal strata.

In summary, continued use of the current frameworks in problems with post-treatment variables (e.g., surrogate endpoints) in principle makes incorrect attributions of effects of treatments. As Buyse (2000) noted recently about the comparison of our framework to the existing ones for surrogate endpoints: "Until now, we had always thought that the roles of biology and statistics did not mix in these complex problems. But principal causal effects set the framework for allowing biological assumptions in statistical methods and vice versa." We hope that this paper provokes the development and dissemination of more principled frameworks.

ACKNOWLEDGEMENT

We thank the Editor, the Associate Editor, two anonymous reviewers, and Stuart Baker, Marc Buyse, Steve Goodman, Jennifer Hill, Sue Marcus, Susan Murphy, Dan Scharfstein and Scott Zeger for constructive comments, and the H.-C. Yang Memorial Fund, the U.S. National Institute of Child Health and Human Development (R01 HD38209), and the National Science foundation for partial support.

REFERENCES

Angrist, J., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, **91**, 444-472.

Baker, S. G. and Lindeman, K. S. (1994). The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statistics in Medicine* **13**, 2269–2278.

Baker, S. G., Wax, Y., and Patterson, B. H. (1993). Regression analysis of grouped survival data: informative censoring and double sampling. *Biometrics*, **49**, 379–389.

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

Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. (2001). School Choice in NY City: A Bayesian Analysis of an Imperfect Randomized Experiment. Forthcoming in *Case Studies in Bayesian Statistics* (with discussion), C. Gatsonis et al. (eds.). New York: Springer-Verlag.

Buyse, M. (2000). Rejoinder to Discussion by C. E. Frangakis on "Validation of Surrogate Endpoints" by M. Buyse. Presentation at the Biostatistics Grand Rounds Seminar, The Johns Hopkins University.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014-1029.

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

Cochran, W. G. (1957). Analysis of covariance: its nature and uses. *Biometrics* **13**, 261–281.

Dawid, A. P. (2000). Causal inference wihout countefactuals (with discussion). *Journal of the American Statistical Association* **95**, 407–448.

Fleming, T. R. and DeMets D. L.(1996). Surrogate end points in clinical trials: are we being misled ? *Annals of Internal Medicine.* **125**, 605-613.

Frangakis, CE, and Rubin, DB (1997). A new approach to the idiosyncratic problem of drug-noncompliance with subsequent loss to follow-up. In: *American Statistical Association, Proc. Biopharm. Sec.*, pp. 206-211.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 365–379.

Frangakis, C. E., and Rubin, D. B. (2001). Addressing an idiosyncrasy in estimating survival curves using double-sampling in the presence of self-selected right censoring. *Biometrics* (with discussion), **57**, 333–353.

Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2001). Clustered encouragement design with individual noncompliance: Bayesian inference and application to Advance Directive Forms. Forthcoming in

Biostatistics (with discussion).

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

Gail, M., Pfeiffer, R., Houwelingen, H., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 3, 231–246.

Goetghebeur, E. and Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91**, 928–934.

Goetghebeur, E. Kenward, M., Molenberghs, G., and Vansteelandt, S. (2000). Inferential tools for sensitivity analysis and noncompliance in clinical trials. Paper presented at the Annual Meeting of the American Statistical Association, Indianapolis, IN.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica.* **40**, 979–1001.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrika* **46**, 931-959.

Hirano, K., Imbens, G., Rubin, D. B., and Zhou, X.-H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1**, 69–88.

Holland, P. (1986). Statistics and causal inference. *J. Am. Statist. Assoc.* **81**, 945-70.

Imbens, G. W. and Rubin, D. B. (1994). Causal inference with instrumental variables. Discussion paper # 1676. Cambridge, MA: Harvard Institute of Economic Research.

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

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

Manski, C. F. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers & Proceedings* **80**, 319–23.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on prin-

ciples, Section 9. Translated in *Statistical Science*, **5**, 465–480, 1990.

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

Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. Mathematical Modelling, 7, 1393-1512.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statist. Med.* **17**, 269–302.

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

Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association* **89**, 737–479.

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

Rosenbaum, P., and Rubin, D. B. (1983a). The Central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society* B **45**, 212–218.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* **2**, 1-26.

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

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with noncompliance (with discussion). *Statistics in Medicine* **17**, 371–389.

Rubin, D. B. (2000). Comment on “Causal inference without counterfactuals”, by AP Dawid, *Journal of the American Statistical Association* **95**, 435–437.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models (with discussion). *Journal of the American Statistical Association*, 94, 1096–1146.

Sommer, A. and Zeger, S. (1991). On estimating efficacy from clinical trials. *Statist. Med.* **10**, 45–52.

The Coronary Drug Project Research Group. (1980). Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine* **303**, 1038–1041.

Zelen, M. (1990). Discussion of presidential address 'Biostatistical collaboration in medical research' by J. H. Ellenberg. *Biometrics* **46**, 28–29.

principal stratum of subject i [1]	Full data: Post−treatment variable − CD4 $S_i(1)$	$S_i(2)$	Full data: Potential outcome survival (average) $Y_i(1)$	$Y_i(2)$	Observed data from randomized study: (S_i^{obs}, Y_i^{obs}) (average) given assignment $Z_i=1$	$Z_i=2$
(a) Case where post−treatment S is a principal surrogate but not a statistical surrogate						
sicker :	L	L	10	10	(2) (L, 20)	(L, 10)
normal :	L	H	30	50		(3) (H, 50)
healthier :	H	H	50	50	(H, 50)	
(b) Case where post−treatment S is a statistical surrogate but not a principal surrogate						
sicker :	L	L	10	20	(2) (L, 20)	(L, 20)
normal :	L	H	30	40		(4) (H, 50)
healthier :	H	H	50	60	(H, 50)	

(1) We set equal proportions for each principal stratum, for simplicity of demonstration.

(2) (1/2)10+ (1/2)30.

(3) (1/2)50+ (1/2)50.

(4) (1/2)40+ (1/2)60.

Figure 1. Distinction between statistical and principal surrogates. Dashed boxes represent missing information, solid boxes represent observed information.