The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 35

Principal Stratification and Attribution Prohibition: Good Ideas Taken Too Far

Marshall Joffe, *University of Pennsylvania*

Recommended Citation:

Joffe, Marshall (2011) "Principal Stratification and Attribution Prohibition: Good Ideas Taken Too Far," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 35.

DOI: 10.2202/1557-4679.1367

Principal Stratification and Attribution Prohibition: Good Ideas Taken Too Far

Marshall Joffe

Abstract

Pearl's article provides a useful springboard for discussing further the benefits and drawbacks of principal stratification and the associated discomfort with attributing effects to post-treatment variables. The basic insights of the approach are important: pay close attention to modification of treatment effects by variables not observable before treatment decisions are made, and be careful in attributing effects to variables when counterfactuals are ill-defined. These insights have often been taken too far in many areas of application of the approach, including instrumental variables, censoring by death, and surrogate outcomes. A novel finding is that the usual principal stratification estimand in the setting of censoring by death is by itself of little practical value in estimating intervention effects.

KEYWORDS: principal stratification, causal inference

Author Notes: The author thanks Michael Elliott, Dylan Small, Sander Greenland, Judea Pearl, and Paul Rosenbaum for helpful comments. This research was supported by NIH grants DK090385 and DK090046.

1 Introduction

Judea Pearl (2011) has written a characteristically insightful and original review of principal stratification (PS). Nonetheless, in my view, his treatment of the topic does not mention several important strengths and weaknesses of the PS approach.

In the concluding section of his article, Pearl considers four interpretations of PS. In my view, the key two of these interpretations are 1) a partition of units into (principal) strata based on the intermediate response of those units to the primary intervention (treatment) of interest and characterization of causal effects within those strata; and 2) a prohibition against formal attribution of effects to the intermediate variable. The first of these is properly termed PS; we will use that term throughout to refer to this. We call the prohibition in 2) “no effect attribution” (NEA). PS and NEA need not go together; for example, Angrist, Imbens, and Rubin (1996) (as implicitly do Baker and Lindeman (1994)) use PS but not NEA in defining their estimands in the setting of noncompliance. Nonetheless, NEA has been a common factor in most treatments of PS since Imbens and Rubin (1997), and characterizes Frangakis and Rubin’s (2002) unifying paper on the topic.

This paper argues that these two key interpretations are based on important insights that have often been taken too far. To see this, it is instructive to consider how PS and NEA relate to scientific goals, including effect definitions, and to identification and estimation of those effects. Our treatment here will center on several examples in which principal stratification has been used, including noncompliance, other instrumental variables problems, censoring by death, surrogate outcomes, and the effects of partially manipulable variables. The structure here is thus similar to VanderWeele’s (2011) accompanying commentary. This essay will sometimes take issue with Pearl’s views and sometimes agree or expand on them.

The discussion of PS and alternative methods will sometimes present several approaches. We do this because of my sense that in many problems in biostatistics, epidemiology, and the social sciences, none of the currently available formal approaches fully answer all the questions that scientists and decisionmakers would like answered. We discuss this at length in the section on censoring by death, but this attitude applies to the other sections as well.

1.1 Notation

We adopt the potential outcomes framework (Neyman, 1923; Welch, 1937; Wilk, 1955; Copas, 1973; Rubin, 1974) and the notation of Pearl (2011). Thus, X denotes the primary treatment, Z the auxiliary post-treatment variable, Y the outcome of primary interest, U some unmeasured variable(s), and W other pretreatment

covariates. Denote by Z_x and Y_x the values Z and Y , respectively, would take if X were set to x , and by Y_{xz} the value Y would take if X were set to x and Z to z .

2 Instrumental variables

This section briefly discusses the application of PS and NEA to problems in instrumental variables, including noncompliance. We argue the PS and NEA have made useful contributions to IV analysis or randomized trials in settings where compliance is binary, but that these approaches can be severely limited when compliance behavior is more complex and when the proposed instrument is not under the control of the investigator.

2.1 Noncompliance

Randomized trials often suffer from noncompliance of subjects to their assigned treatment. The standard intent-to-treat estimand represents the effect of treatment assignment, not receipt. Clinicians and patients are usually and justifiably more interested in the effects of actually receiving treatment. Consider first binary compliance, where a subject either receives or does not receive treatment. Instrumental variables are often an attractive option here, with randomization as an instrument, since randomization should not be associated with measured or unmeasured characteristics predictive of the outcome, and it will often be plausible that randomization has no effect on outcome except that part mediated by treatment received. Standard approaches in econometrics took the IV estimand as a contrast of what would have happened had everyone been compliant to treatment versus control, and essentially ignored heterogeneity among subjects in treatment effects. In contrast, the principal stratification approach concentrates on the effect of treatment among the compliers (i.e., subjects for whom $Z_x = x$, where X is treatment assignment and Z is treatment received). For binary X and Z , monotonicity ($Z_1 \geq Z_0$ for all subjects) and an exclusion restriction ($Y_{xz} = Y_{x'z}$ for all x, x') are sufficient to identify the average causal effect for compliers, $E(Y_1 - Y_0 | Z_1 = 1, Z_0 = 0)$, without any assumption about potential effects of treatment in other principal strata (never- and always-takers). Under these often mild assumptions, the randomized experiment itself tells us little about treatment effects in those strata..

A somewhat earlier approach to IV analysis based on structural nested models (Robins, 1989, 1994; Robins and Tsiatis, 1991; Robins, Blevins, Ritter, and Wolfson, 1992; Robins, Rotnitzky, and Scharfstein, 2000) also in principle allows one to estimate the effect of treatment on a subgroup of the population; in one parameter

causal models corresponding to the simple PS approach. In this approach, the effect of treatment may vary with observed variables, including (in this setting) treatment received. Interactions between treatment effects and level of treatment received are termed “current treatment interactions.” Here, a simple 1 parameter structural mean model might be written $E(Y|X,Z) - E(Y_{00}|X,Z) = Z\psi$, where ψ is a possibly unknown parameter. When subjects randomized to control cannot receive treatment (i.e., $Z_0=0$ for all I), ψ equals the CACE. When they can, this model assumes that the effect of treatment among subjects randomized to treatment and receiving treatment is the same as the effect among subjects randomized to control but receiving treatment. In many settings, the homogeneity of effects assumption will be less plausible than the monotonicity assumption justifying the PS approach to noncompliance. A two parameter structural model $E(Y|X,Z) - E(Y_{00}|X,Z) = X\psi_1 + ZX\psi_2$ which relaxes this homogeneity of effects assumption is not identified solely on the basis of randomization and these modeling assumptions. A standard G-estimator in this model is consistent for the same quantity as the standard IV estimator.

The attention to the possibly different effects of treatment in different subgroups of the data and to which effects are identified, as developed both in the PS and the structural nested models approaches, is salutary. In either approach, it should be clear that the direct policy relevance of the IV estimates of the subgroup-specific effects is limited. This is because the subgroups for whom treatment effects are estimated are, in general, not identified prior to treatment decisions, and further may change in conditions outside the trial. The latter may occur either because inducements to use treatment outside the trial may be different than in the trial; this, in turn could be due to differences between the conduct of the trial and standard medical practice, or due to knowledge after a successful trial that one is receiving an active drug rather than a placebo and that the drug has been determined to be efficacious (Robins, 1989). For this reason, the view that the CACE is generally of primary interest is too strong.

Beginning at least with Imbens and Rubin (1997) and continuing with Frangakis and Rubin (2002), there has been a shift from doubly to singly-indexed potential outcomes in PS-based IV analyses, with attention focused on the effect of treatment assignment in each of the principal strata; this shift is due to a reluctance to attribute causal effects to variables (treatment received) not under the control of the investigator. A justification for this shift is that, under the conditions of the randomized experiment, it is randomization rather than treatment received that is under the control of the investigator, and so the effects of assignment (within subgroups) are what is really being estimated by the experiment. This reasonable argument comes with a price: it removes causal mechanistic ideas from the formal notation for effects and enjoins formal attribution of effects to treatment received or

adherence. As we shall see, this choice has fewer benefits in the context of many other types of IV analyses.

Probably in part because the mathematics works nicely for binary X and binary Z in the noncompliance setting, there has been a tendency to dichotomize Z in the applied literature on principal stratification even when treatment received may take many values. Let Z^* denote a dichotomized version of Z . In general, even if the exclusion restriction holds for the original Z , it will not hold for a binary version (Robins and Greenland, 2000; VanderWeele, 2011). The directed acyclic graph (DAG) in figure 1 is useful for illustrating this¹. The graphical criterion for instrumental variables (Pearl, 2009) include the criterion that X and Y be d-separated in the graph in which arrows pointing out of the putative instrument are removed. In figure 1, this is true for Z but not Z^* . The sort of mechanistic thinking embodied in the graphical approach enables one to see this immediately. In contrast, the prohibition against attributing effects decried by Pearl (2011) might forbid or at least discourage graphical representations like figure 1, and so lead to missing the fact that X is an instrument for the effect of Z but not Z^* .

Figure 1

2.2 Other IV settings

Instrumental variables were not originally devised for randomized trials with noncompliance. Here, the PS approach to IV can have additional drawbacks. First, the IV may not itself be the subject of intervention. Consider the case of Mendelian randomization, in which some genetic component is viewed as a putative instrument for the effect of some endogenous variable Z affected by some genetic factors related to the measured gene (e.g., a single nucleotide polymorphism (SNP)) (Davey Smith

¹Sjolander (2011) uses a similar DAG to illustrate the problems with dichotomization (or, more generally, coarsening) in characterizing principal stratification direct effects.

and Ebrahim, 2003; Shinohara *et al.*, 2009). The SNP may be a marker for other genes on the same chromosome that actually affect the levels of the exposure Z . Figure 2 provides an illustrative DAG, where X denotes the collection of genetic factors which may affect exposure but are otherwise unrelated to outcome, and X^* denotes the marker SNP of interest. X^* satisfies the graphical criteria for an IV (Pearl, 2009) but is not itself the subject of the intervention; here, the actual intervention is taken to be the process of meiosis and fertilization, which is viewed as a sort of natural experiment. The measured SNP may be a proxy for this intervention.

Figure 2

Another popular candidate for an instrument is distance to a source of care or treatment. One may view distance as variable subject to external manipulation. Typically, there is a degree of vagueness in this hypothetical manipulation. If distance is measured in miles, the effect moving a subject or a treatment center to a particular location with a given distance may depend on other aspects of the move, including driving time between the locations or availability of public transportation. There is thus a degree of vagueness to the counterfactual quantities involved in manipulating the instrument and a failure of the consistency (Cole and Frangakis, 2009; VanderWeele, 2009) or stable unit treatment value (SUTVA) (Rubin, 1980) assumptions.

If one takes the view of Robins and Greenland (2000) and of VanderWeele's (2011) accompanying commentary, there is always some vagueness to counterfactuals, and there can be a range of degrees of discomfort with defining effects in the presence of such vagueness. It will not always be the case that the hypothetical intervention on the instrument is better-defined than the intervention on the treatment or exposure of interest. Because of this vagueness and because the usual goal of IV analysis is to estimate the effect of a treatment or exposure and not the instrument, the NEA-associated redefinition of an instrument as a manipulable variable whose effects are potentially of interest (see Tan (2010) for an example outside the PS framework) is not justifiable as a general or overall formulation for IVs.

3 Censoring by death

“Censoring by death” has become shorthand for the problem of characterizing the effect of a treatment or exposure X on an outcome Y when death (Z ; 1=alive, 0=dead) precludes the development or observation of that outcome. Pearl (2011) and VanderWeele (2011) accept the notion advanced by Frangakis and Rubin (2002) and others that the combined PS/NEA approach defines a target quantity, the survivor average causal effect (SACE), of great interest². This section will provide simple numerical examples to demonstrate that the SACE, while an interesting quantity, does not fully capture the notion of the effect of treatment on the censored outcome in the presence of censoring by death and is in general inadequate for decision making or regulatory purposes. We then discuss other quantities of interest, and consider important differences between censoring by death and the problem of post-infection outcomes, which has sometimes been treated as the same problem.

3.1 Problems with principal stratification

A continuous outcome Y may be well-defined only for subjects who live. Frangakis and Rubin (2002) argue that the causal effect of treatment on Y is only defined in the principal stratum in which subjects live whether or not they are treated (i.e., $Z_0=Z_1=1$) and so Y is well-defined under both treatment levels. Thus, they argue, causal effects on Y , as contrasts of the potential outcomes, are well defined only for this principal stratum.

It is productive to consider the contrast of the poorly-defined or missing potential outcome for subjects who die and the well-defined potential outcome for a subjects who do not to be a causal contrast. Let $\cdot$ denote the value of Y when a subject is dead and so Y cannot be measured. When the proportions of the population for whom treatment causes and prevents death are the same and so treatment has no net effect on mortality, this results in a simple measure of the net effect of treatment on Y adjusting for survival, as described below in this section. Section 3.2.3 considers a generalization to settings where treatment affects marginal survival distributions. This redefinition of the effects of X on Y allows us to avoid the apparent paradox in which treatment has no net effect on Z and no effect on Y (under Frangakis and Rubin’s definition), yet affects the distribution of Y .

The data in tables 1 and 3 illustrate these points. In these examples, treatment X has no effect overall on mortality, but causes and prevents death in equal numbers

²Robins (1986) first proposed the PS estimand for this setting, where he dismisses it because of nontransitivity of treatment comparisons.

of subjects. In table 1, treatment has no effect on Y in the “immune to death” principal stratum (i.e., $E(Y_1 - Y_0 | Z_0 = Z_1 = 1) = 0$). We would be tempted to conclude that treatment is neither beneficial nor harmful, since it has no effect on Y in the immune principal stratum and has no overall effect on mortality.

Table 1

Z_0	Z_1	N	$E(Y_0 Z_0, Z_1)$	$E(Y_1 Z_0, Z_1)$
0	0	250	.	.
0	1	250	.	1
1	0	250	0	.
1	1	250	1	1

Even though treatment does not effect Y in the immune principal stratum, it does affect the distribution of Y in the population (i.e., $pr(Y_0 = y) \neq pr(Y_1 = y)$, where $Y_x = \cdot$ if $Z_x = 0$), and further, the joint distribution of the potential outcomes is different under the two levels of treatment (i.e., $pr(Y_0 = y, Z_0 = z) \neq pr(Y_1 = y, Z_1 = z)$). Thus, the average outcome conditional on surviving differs under both levels of treatment (i.e., $E(Y_0 | Z_0 = 1) = 0.5 \neq E(Y_1 | Z_1 = 1) = 1$); this is represented in table 2. One should conclude from these data that treatment is, overall, beneficial (if higher levels of Y are good); this is supported by a decision-theoretic approach to the problem (Joffe, Small, and Hsu, 2007), in which the utility under a treatment is a function only of the marginal (with respect to X) distribution of variables under that treatment and not a function of the joint (with respect to X) distribution of potential outcomes under different treatments. A sole concentration on the immune principal stratum (this focus is nearly explicit in Rubin (2006)) would lead to an incorrect conclusion, even about the marginal effects of treatment on Y .

Table 2

Z_x	$x=0$		$x=1$	
	N	$E(Y_x Z_x)$	N	$E(Y_x Z_x)$
0	500	.	500	.
1	500	0.5	500	1

Consideration of all principal strata simultaneously is here interesting and informative. We would correctly note that treatment has no effect on Y in the immune principal stratum, but that in the “die only if treated” stratum (i.e., $Z_1=1, Z_0=0$), the level of Y is higher than the “die only if untreated” stratum ($Z_1=0, Z_0=1$). A difficulty here is that membership in a principal stratum is not directly observed, and both individual-level membership and proportions in a stratum are not identified without further assumptions. Since these strata have the same size, it would be nice to say that, overall, treatment increases Y even after adjusting appropriately for mortality. We might do this by allowing the missing values for Y to cancel here, since the strata are of equal size. It is not immediately apparent how to do this when treatment affects the overall probability of death and the two strata are not of equal size.

Table 3

Z_0	Z_1	N	$E(Y_0 Z_0, Z_1)$	$E(Y_1 Z_0, Z_1)$
0	0	250	.	.
0	1	250	.	1
1	0	250	0	.
1	1	250	1	0

In table 3, treatment reduces the level of Y in the immune principal stratum but has no effect on the marginal distribution of Y or the joint distribution of Y and Z . If lower levels of Y are undesirable, focus solely on the distribution of Y in the immune principal stratum (with or without also considering the effect of X on Z) would lead to the incorrect conclusion that treatment is harmful, since it is harmful in the immune principal stratum and has no overall effect on mortality. Nonetheless, treatment has no overall effect on the distribution of Y in the population (i.e., $pr(Y_0=y)=pr(Y_1=y)$, where $Y_x=\cdot$ if $Z_x=0$), and further, the joint distribution of the potential outcomes is the same under both levels of treatment (i.e., $pr(Y_0=y, Z_0=z)=pr(Y_1=y, Z_1=z)$), and so the average outcome conditional on surviving is the same under both levels of treatment (i.e., $pr(Y_0=y|Z_0=1)=pr(Y_1=y|Z_1=1)=0.5$). One should conclude from these data that treatment is, overall, neither beneficial or harmful. A sole concentration on the immune principal stratum would lead to an incorrect conclusion, even about the marginal effects of treatment on Y .

Table 4

Z_0	Z_1	N	$E(Y_0 Z_0, Z_1)$	$E(Y_1 Z_0, Z_1)$
0	0	500	.	.
0	1	0	.	.
1	0	0	.	.
1	1	500	0.5	1

An assumption often made in the PS literature is monotonicity of the potential auxiliary variable (i.e., $Z_1 \geq Z_0$ with probability 1 or $Z_1 \leq Z_0$ with probability 1). Consider the data in table 2, which could represent observed data in a randomized trial (changing x to X). The combination of monotonicity and the data in table 2 leads to the data in table 4. Under the (incorrect) monotonicity assumption, the effect of treatment in the immune principal stratum leads to the same contrast as derived above; i.e., $E(Y_1|Z_0=Z_1=1) - E(Y_0|Z_0=Z_1=1) = E(Y_1|Z_1=1) - E(Y_0|Z_0=1) = 1 - 0.5 = 0.5$. That incorrect assumptions can lead to an appropriate estimand, while correct assumptions can lead to unreasonable conclusions is curious but not accidental in this setting. This will be explored further below.

3.2 Alternatives to principal stratification

The above examples suggest that sole focus on the PS estimand can be misleading. While our example above is suggestive of problems with the PS approach, it involves a situation in which treatment has no effect on the marginal distribution of mortality. The motivation for the PS approach derives from settings in which treatment does affect mortality and so standard comparisons conditional on observed survival are biased; one may view our examples above as a special case of settings in which treatment may have arbitrary effects on the survival distribution. We consider several alternative approaches: death blocking, marginal joint distributions, and marginal adjusted conditional contrasts. Finally, we discuss and compare the roles of each of the different estimands.

3.2.1 Death blocking

One approach to this situation is to suppose that there are latent outcomes that would have been measured if each subject received his/her treatment X but death were

prevented. This approach is implicit in some approaches to survival analysis in the presence of competing risks and is explicit in Robins (1986; chapter 12). These approaches rely on the assumption that, “at least conceptually, deaths ... could be eliminated in a manner that does not affect past or future covariate status or ...” the (latent) outcome Y_x . The combination of the implausibility of this assumption with the vagueness of the associated counterfactuals (How is death to be eliminated? Would all means of blocking death lead to the same outcomes?) make this option unattractive (e.g., Kalbfleisch and Prentice, 2002); the NEA view seems attractive here.

We might view this as the controlled direct effect of treatment controlling for death (by setting it to 0). A less drastic intervention might involve equalizing death between different levels of treatment (Robins and Greenland, 1989); this might be formulated in terms of natural or pure direct effects (Robins and Greenland, 1992; Pearl, 2001) or in terms of stochastic interventions on death (Geneletti, 2007); such approaches would still be unattractive because of the vagueness of the counterfactuals or hypothetical interventions.

3.2.2 Marginal joint distributions

One can consider the effects of treatment on the joint distribution of the outcomes Z and Y . This would lead to comparisons of $pr(Z_0, Y_0)$ and $pr(Z_1, Y_1)$. Such contrasts are adequate for making decisions and comparing utilities of various treatments. Let $U(x, Z_x, Y_x)$ denote the utility function for a particular decision; in general, it may be a function of all three arguments. In a decision-theoretic approach (Joffe, Small, and Hsu, 2007), one would seek to choose the value x of treatment which maximizes the expected utility $E\{U(x, Z_x, Y_x)\} = \sum_{x,z} U(x, z, y) pr(Z_x=z, Y_x=y)$ (here shown for discrete

Y and Z . Because different decisionmakers will have different utility functions, one would thus want to have estimates of the marginal joint distributions $pr(Z_x, Y_x)$ and not just of $E\{U(x, Z_x, Y_x)\}$. In the setting of censoring by death, death is an important outcome that one would generally want to consider in making decisions (Rosenbaum, 2006; Joffe, Small, and Hsu, 2007). These contrasts of joint distributions, however, fail to isolate the effects of X on Y and so are unsatisfying for explanatory purposes.

One may factorize the marginal joint distributions as $pr(Z_x, Y_x) = pr(Z_x)pr(Y_x|Z_x)$. While contrasts of $pr(Z_1)$ and $pr(Z_0)$ are standard aggregate causal effects, contrasts of $pr(Y_1|Z_1=1)$ with $pr(Y_0|Z_0=1)$ are not standard causal effects, because they involve contrasts between different sets of subjects (Rosenbaum, 1984; Robins, 1986; Frangakis and Rubin, 2002). In particular, when

X affects the marginal distribution of Z , these contrasts will be unsatisfying for explanatory purposes, because subjects with $Z_1=1$ may systematically differ from subjects with $Z_1=0$ with respect to Y_1 or Y_0 .

3.2.3 Marginal adjusted conditional contrasts

In this section, we consider another quantity that attempts to capture the net effect of treatment on Y adjusting for censoring by death. Suppose that we observe not only whether but also when a subject fails T . Under treatment level x , we may observe T_x , Y_x . For observations made at a fixed follow-up time m , $Y_x = \cdot$ if $T_x < m$. Let $T_{0,x} \equiv T_0(T_x, x)$, where $T_0(t, x) \equiv S_0^{-1}\{S_x(t)\}$, where $S_x(t) \equiv \text{pr}(T_x \geq t)$, and $S_x^{-1}(t) \equiv q: \text{pr}(T_x \geq t) = q$. $T_0(t, x)$ involves a mapping of survival distributions under treatment at level x to treatment level 0. For illustration, consider a simple accelerated failure-time model, in which withholding treatment lengthens the distribution of lifetimes by a factor $\exp(\beta)$. We then have $T_0(t, x) = t \exp(x\beta)$, and $T_{0,x} = T_x \exp(x\beta)$. $T_{0,x}$ is a variable whose distribution is not a function treatment. We can then factorize the marginal joint distribution of T_x, Y_x as $\text{pr}(T_x, Y_x) = \text{pr}(T_{0,x}, Y_x) = \text{pr}(T_{0,x}) \text{pr}(Y_x | T_{0,x})$. Contrasts of $\text{pr}\{Y_1 | T_{0,1} = t\}$ with $\text{pr}\{Y_0 | T_{0,0} = t\}$ are a sort of marginal (with respect to the joint distribution of potential outcomes under different treatments) contrast of the potential outcomes Y_x adjusted for treatment-adjusted survival, a variable whose distribution is not affected by treatment. These are numerical contrasts of levels of the outcome for $T^* \equiv \min(T_{0,0}, T_{0,1}) \geq m$, $T_{0,1} \geq m$. We can also marginalize contrasts over $T^* \geq m$. When rank preservation for T holds (i.e., $T_{0,x} = T_0$ with probability 1), this contrast is a contrast of $\text{pr}\{Y_1 | T_1 = t, T_0 = T_0(t, 1)\}$ with $\text{pr}\{Y_0 | T_1 = t, T_0 = T_0(t, 1)\}$; i.e., a PS type of estimand. Rank-preservation is stronger than and implies monotonicity. Rank-preservation is usually an implausible assumption.

We view these contrasts as appropriate control for a survival variable whose distribution is not affected by treatment. The contrast is adjusted for the effect of treatment on survival. The joint distribution may be parametrized in part using joint structural nested failure-time models for the effect of treatment on survival and structural nested distribution (or mean) models for the conditional effects of treatment on outcome Y (see Robins, 2008, who considers joint models with outcomes Y meaningful after failure. Greene (2011, personal communication)

proposed a version marginalizing over $T^* > m$, which may most closely approximate an overall assessment of the effect of X on Y controlling for death..

In the data in tables 1 and 3, the marginal adjusted conditional contrasts are simply contrasts of $pr(Y_1|T_1 > m) = pr(Y_1|Z_1 = 1)$ with $pr(Y_0|T_0 > m) = pr(Y_0|Z_0 = 1)$, since treatment has no effect on the distribution of Z . These contrasts appropriately find that, in table 1 (see table 2), treatment has no effect, whereas in table 3 it does not.

Contrasts between $pr(Y_1|T_{0,1} = t)$ and $pr(Y_0|T_{0,0} = t)$ are not formal causal contrasts in the sense of comparisons of individual outcomes among a common group of subjects. Nonetheless, they may be viewed in a broader sense as causal inasmuch as they are a component of the effect of treatment on the joint distribution of T and Y . The factorization provided above best isolates the effect of X on Y from its effect on T , or adjusts the effect of X on Y for its effect on T . More expanded treatment of these sorts of joint models will be provided elsewhere.

3.3 Post-infection outcomes

The problem of post-infection outcomes can productively be viewed as similar to the problem of censoring by death (Gilbert, Bosch, and Hudgens, 2003). The purpose of some vaccines in development for HIV is to reduce the level of some adverse outcome (e.g., viral load) after infection instead of just reducing the rate of infection. The outcome of interest Y is considered meaningful only if a subject is not infected; thus, if $Z=1$ denotes absence of infection, Y is considered available only if $Z=0$. For binary Z , the problem is isomorphic with censoring by death if one switches the coding of Z .

This isomorphism is broken when infection is considered a failure-time outcome. Now, if T denotes the time of infection, Y is available if $T \geq m$. Joint structural nested models would need to be reformulated to reflect this change.

3.4 Roles of different estimands

Based on the foregoing considerations, we consider the roles of different estimands in the setting of censoring by death. We view many of the different quantities as complementary and so are loathe to identify a single estimand as being of sole or primary interest³.

³In reaction to an earlier version of the manuscript, Pearl was frustrated by my failure to select a single quantity of greatest interest. This section is a partial response.

Of primary interest for decision problems is the joint distribution of the failure-time outcome and other outcomes that would be seen under a given treatment and the contrasts of these joint distributions under different treatments; i.e., we would like to estimate and contrast $pr(Z_0, Y_0)$ and $pr(Z_1, Y_1)$. The decisionmaker can then supply a utility to each combination of treatment and potential outcomes and choose a treatment to maximize the expected utility. In populations and conditions identical to those from which extant data are derived, we can often (e.g., in randomized trials) simply estimate $pr(Z_x, Y_x)$ from data on the joint distribution of $pr(X, Z, Y)$.

Often we are interested in decisions under different conditions than those obtaining in a particular study. Following Pearl and Barenboim (2011), denote the marginal joint distributions of the potential outcomes under the new conditions by $pr^*(Z_x, Y_x)$. Here decisions should ideally be made based on $pr^*(Z_x, Y_x)$ and a decisionmaker's utility function. Identifying $pr^*(Z_x, Y_x)$ under other than the precise conditions of the study from which data are collected can be challenging.

Making decisions is not the sole object of scientific inference, and sometimes not even the primary object. Understanding of the processes involved in generating the data can be a primary goal; this has been true in a wide variety of scientific disciplines, including astronomy, meteorology, and evolutionary biology. Such understanding may be enhanced by appropriate factorization of the marginal joint distributions of potential outcomes $pr(Z_x, Y_x)$ and by consideration of the joint distributions of the potential outcomes (e.g., $pr(Z_0, Y_0, Z_1, Y_1)$) used in PS.

Table 5

Z_0	Z_1	N	$E(Y_0 Z_0, Z_1)$	$E(Y_1 Z_0, Z_1)$
0	0	500	.	.
0	1	0	.	.
1	0	0	.	.
1	1	500	0.5	0.5

To see this, consider the data in table 5. These data are indistinguishable in observable data from the data in table 3. Nonetheless, the story told by the two tables is rather different. In table 5, treatment has no effect on Z and no average effect on Y in the always survivor stratum; in table 3, treatment affects Z , and Y , but the positive and negative effects on both balance each other. The different tables would

generate different expectations of what would happen in populations with different survival experiences. Although even experimental data cannot distinguish between the two settings, external theory or knowledge is sometimes available which allows us to say something about the joint distribution.

Similarly, when treatment affects survival, comparisons of $pr(Y_1|T_{0,1}>m)$ and $pr(Y_0|T_{0,0}>m)$ are likely more informative than contrasts of $pr(Y_1|T_1>m)$ with $pr(Y_0|T_0>m)$. Differences in the latter quantities may be due to the effect of treatment on Y or the net effect of X on T , which is associated with Y . In contrast, the marginal adjusted conditional contrast of $pr(Y_1|T_{0,1}>m)$ and $pr(Y_0|T_{0,0}>m)$ has sought to adjust for the average effect of treatment on survival and so is closer to a standard causal contrast. Pearl (2009) has argued for the primacy of causal over observational knowledge, in part because causal knowledge is based on a more fundamental understanding of causal processes and more readily generalizes to new situations; these considerations apply here as well.

Similarly, qualitative aspects of causal effects are often more fundamental building blocks of our knowledge than their precise quantification, which may change from one setting to the next. Thus, the statement that smoking causes lung cancer is based on a fundamental understanding of biological process and generalizes more broadly than statements about the precise value of the average causal effect of smoking, which will be dependent on a host of measured and unmeasured factors. In such circumstances, we can gain a useful understanding of causal processes without being able to choose definitively an optimal treatment. A similar phenomenon arises in DAGs, where knowledge of the causal structure is more fundamental and prior to precise quantification.

Pearl (2009) has argued for the complementary nature of multiple conceptualizations of causality, including the graphical, nonparametric structural equations, and counterfactual. Here too, multiple approaches and estimands can provide complementary information which can enhance scientific understanding and may aid prediction⁴.

⁴Despite our overall negative assessment of this approach, the idea of death blocking is sometimes necessary. In particular, when we are interested in what would happen to a population similar to the study population after some new medical technology is introduced that prevents or delays death in some subjects, death blocking (i.e., intervention on T) represents these ideas more closely than conditioning (i.e., principal stratification or marginal adjusted conditional contrasts).

4 Surrogate outcomes

PS and NEA have made important contributions to understanding of surrogate outcomes and the related issue of characterizing the effects of partially manipulable variables. Nonetheless, these approaches, taken by themselves, can be somewhat limiting. The following sections explore these limitations.

4.1 Causal paradigms for surrogacy

Randomized trials of new treatments can be expensive and time-consuming to run, especially when the outcomes of interest are failure-time outcomes (e.g., mortality, cancer incidence, cardiac events) that are rare (i.e., occur in a small proportion of a study population over the course of follow-up). There is thus much interest in using surrogates for the outcomes of clinical significance which would allow shorter follow-up and fewer subjects. Ideally, a surrogate outcome is a variable for which knowledge of the effect of treatment on the surrogate would allow prediction of the effect of the treatment on the clinical outcome (Joffe and Greene, 2009).

Framed this way, surrogacy ultimately involves causal questions. Nonetheless, the earliest formal approaches to surrogacy (Prentice, 1989; Daniels and Hughes, 1997) did not formally invoke causal ideas. Although Prentice (1989) did not use formal causal terminology, it appears that he took a good surrogate to be a variable which completely mediates the effect of the treatment on the clinical outcome. The approach may be viewed as an example of the causal effects paradigm, where knowledge of the effect of X on Z and of Z on Y (as well as the direct effect of X on Y controlling for Z) allows prediction of the effect of X on Y . Unfortunately, his criteria (in particular, conditional independence of X and Y given Z) can be shown to be inappropriate (Joffe and Greene, 2009) based on ideas of direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001).

Frangakis and Rubin (2002) were, to my knowledge, the first to apply formal causal ideas to the problem of surrogacy. They showed that Prentice's criteria are not appropriate. Additionally, they adopt the notion that a surrogate need not itself be subject to direct manipulation and have its own causal effect. There are at least two justifications for this. First, predicting the effect of X on Y based on its effects on Z does not necessarily require Z to be a fully manipulable variable with its own effects. Second, many putative surrogates (e.g., hemoglobin A1c in diabetes, blood pressure in hypertension) are not directly manipulable variables. Frangakis and Rubin's reluctance to attribute effects to Z is consistent in principle with the nature of the surrogate outcomes problem.

The PS approach to surrogacy may be viewed as an example of the causal association paradigm (Joffe and Greene, 2009). Let θ_j denote the effect of X on Z in subgroup j , and ψ_j denote the effect of X on Y in that subgroup. In this paradigm, a good surrogate is one in which θ_j predicts ψ_j well, and in which $\theta_j \approx 0$ implies $\psi_j \approx 0$. In the PS approach, j indexes principal strata. One could also use j to index subgroups defined by pretreatment covariates W . An important advantage of subgroups defined by pretreatment covariates W over those defined by principal strata is that the effects θ_j and ψ_j are identified solely on the basis of randomization, whereas in the PS approach, θ_j and ψ_j are not identified, as j indexes unidentified subgroups. Nonetheless, observed covariates W may do a poor job of predicting effects, leading to little variation in θ_j and to Z being a poor surrogate. In contrast, when j indexes principal strata, the variability of θ_j will be larger. In the context of vaccine trials, Follman (2006) has proposed useful approaches for obtaining good proxies for the PS effects θ_j through additional testing and data collection. In some cases, these proxies are observed pretreatment covariates W .

4.2 Transportability

Pearl and Barenboim (2011), cited in Pearl (2011), make an important contribution to the surrogate outcomes literature by framing the problem of surrogacy as one of transportability of effects to a new situation. They note that the problem of surrogate outcomes typically involves inference for the effect of a new treatment (or conditions) S on Y based on the effect of S on Z , where data are not (yet) available the joint distribution of S , Z , and Y . Thus, even if Z is a surrogate for the effect of X on Y (which may be evaluated using existing data from earlier trials), it may not be a surrogate for the effect of S on Y . This appropriate criticism applies both to Frangakis and Rubin's criteria of principal surrogacy and to most other extant methods (Joffe and Greene, 2009) based on both the causal association and causal effects paradigms.

Pearl and Barenboim's solution is to frame the problem within a setting in which we know the structure of the causal relations among important variables (as represented by a graph), where we can estimate relationships of X , Z , and Y based on data from earlier data, and the effect of S on Z based on a new trial. In the examples they provide in which Z is a surrogate, it fully mediates the effect of S on Y . Further, in these cases, its effect is nonparametrically identified based on the graph. These requirements are very difficult to satisfy and so make it extremely hard to identify putative surrogates. Further, no quantitative measures of the potential usefulness of

an imperfect surrogate are provided (in contrast to the PS approach, which proposes such measures as associative or dissociative proportion (Frangakis and Rubin, 2002), and the causal effects paradigm, which proposes such measures as the proportion of effect explained (Freedman, Graubard, and Schatzkin, 1992; Joffe and Greene, 2009). Their solution may be of benefit in identifying putative surrogates for the effect of a particular new treatment from extant data, but may be less useful in evaluating empirically whether these are good surrogates for future untested treatments.

The meta-analytic approach to surrogate outcomes provides an alternative approach which deals, albeit imperfectly, with transportability. This approach, like others in the causal association paradigm (including PS), looks at the association between effects θ_j and ψ_j ; in the meta-analytic approach j indexes separate studies, possibly of different treatments. If one finds a strong association between effects θ_j of treatment on the putative surrogate and its effects ψ_j on the clinical outcome, this provides some evidence that θ_j will predict ψ_j for the next agent. Formally, this supposes that a new study is randomly sampled from the same hypothetical population of potential studies giving rise to earlier estimates of the joint distribution of θ_j and ψ_j .

4.3 Effects of partially manipulable variables

Characterizing the effects of partially manipulable variables is a problem related to surrogacy. Here, investigators, individuals, and clinicians may profess to be interested in the effect of a variable that is not directly or completely manipulable given current technology. Examples include the effect of blood pressure, blood sugar, or inflammation on clinically meaningful outcomes. For scientific understanding, we would ideally like to attribute effects to these endogenous variables. However, the consistency assumption or condition required for defining effects may be violated, and so the relevant counterfactuals may be ill-defined (see van der Laan, et al. (2005) and Hernan and VanderWeele (2011) for discussion of related topics).

Sometimes, one may be satisfied with answering a practical question: is it worthwhile to try to affect the partially manipulable variable with a possibly unspecified intervention (e.g., reduce blood pressure, sugar, or inflammation)? Such questions are easily framed within the causal association paradigm; will a reduction in inflammation as measured by, say, C-reactive protein, be associated with reductions in mortality, heart disease, etc.? Here, answering the practical question does not necessarily require attributing causal effect to directly manipulating C-reactive protein. This idea is implicit in some of the work under the PS/NEA approaches (e.g., Shinohara *et al.*, 2009). The causal effects and causal association

paradigms may be viewed as complementary. The causal associations paradigm is black box in nature and provides little understanding (VanderWeele, 2011); prediction of effects may be enhanced by considering mechanistic approaches.

5 Conclusions

PS and the associated NEA approach have spawned a large literature on both methods and applications and have sparked controversy. The defining characteristics of these linked approaches are concentration on effects in subgroups which are often not identifiable, and a reluctance to attribute effects to variables not directly under the control of the investigators.

PS itself has had a salutary effect by bringing to the fore possible heterogeneity in effect across subgroups characterized by response to the main treatment. However, its sometimes exclusive concentration on effects in certain subgroups can be harmful, as seen in the problem of censoring by death. Further, its focus on subgroups not identifiable without special assumptions even in well-conducted randomized trials is somewhat problematic in itself and has often led to inappropriate simplifications (e.g., dichotomization).

The reluctance to consider all variables as having effects is salutary. In many cases, however, proponents of the PS/NEA approach have taken this tendency too far, leading to impoverishment of explanatory power as well as to concentration on quantities that are not of practical value in characterizing the effects of an intervention.

Reference List

- 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.
- Cole, S. R., and Frangakis, C. E. (2009), "The consistency statement in causal inference: a definition or an assumption?," *Epidemiology*, 20, 3-5.
- Copas, J. B. (1973), "Randomization models for the matched and unmatched 2×2 tables," *Biometrika*, 60, 467-476.
- Daniels, M. J., and Hughes, M. D. (1997), "Meta-analysis for the evaluation of potential surrogate markers," *Statistics in Medicine*, 16, 1965-1982.

- Davey Smith, G., and Ebrahim, S. (2003), "Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?," *International Journal of Epidemiology*, 32, 1-22.
- Freedman, L., Graubard, B., and Schatzkin, A. (1992), "Statistical validation of intermediate endpoints for chronic disease," *Statistics in Medicine*, 11, 167-178.
- Follman, D. (2006), "Augmented Designs to Assess Immune Response in Vaccine Trials," *Biometrics*, 62, 1161-1169.
- Frangakis, C. E., and Rubin, D. B. (2002), "Principal stratification in causal inference," *Biometrics*, 58, 21-29.
- Geneletti, S. (2007), "Identifying direct and indirect effects in a non-counterfactual framework," *Journal of the Royal Statistical Society, Series B*, 69, 199-215.
- Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003), "Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials," *Biometrics*, 59, 531-541.
- Hernan, M. A., and Vanderweele, T. J. (2011), "Compound Treatments and Transportability of Causal Inference," *Epidemiology*, 22, 368-377.
- Imbens, G. W., and Rubin, D. B. (1997), "Bayesian inference for causal effects in randomized experiments with noncompliance," *Annals of Statistics*, 25, 305-327.
- Joffe, M. M., Small, D., and Hsu, C. Y. (2007), "Defining and estimating intervention effects for groups that will develop an auxiliary outcome," *Statistical Science*, 22, 74-97.
- Joffe, M. M., and Greene, T. (2009), "Related causal frameworks for surrogate outcomes," *Biometrics*, 65, 530-538.
- Kalbfleisch, J. D., and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data* 2nd ed.), Hoboken, NJ: Wiley-Interscience
- Neyman, J. (1990), "On the application of probability theory to agricultural experiments. Essay on principles. Translated by D.M. Dabrowska and edited by T. P. Speed," *Statistical Science*, 5, 465-472.

- Pearl, J. (2001), "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann.
- Pearl, J. (2009), *Causality: models, reasoning, and inference*, Second ed., Cambridge University Press.
- Pearl, J. (2011), "Principal Stratification - a Goal or a Tool?," *International Journal of Biostatistics*, 1, 20.
- Pearl, J., and Bareinboim, E. (2011), "Transportability across studies: a formal approach," in UCLA Department of Computer Science Technical Reports, R-372.
- Robins, J. (1986), "A new approach to causal inference in mortality studies with a sustained exposure period- application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393-1512.
- Robins, J. M. (1989), "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies," in *Health Service Research Methodology: A Focus on AIDS*, ed. L. A. Sechrest, NCHSR, U.S. Public Health Service.
- Robins, J. M. (1994), "Correcting for non-compliance in randomized trials using structural nested mean models," *Communications in Statistics-Theory and Methods*, 23, 2379-2412.
- Robins, J. M. (2008), "Causal models for estimating the effects of weight gain on mortality," *International Journal of Obesity*, 32, S15-S41.
- Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992), "G-estimation of the effect of prophylaxis therapy for pneumocystic carinii pneumonia on the survival of AIDS patients," *Epidemiology*, 3, 319-336.
- Robins, J. M., and Greenland, S. (1989), "Estimability and estimation of excess and etiologic fractions," *Statistics in Medicine*, 8, 845-859.
- Robins, J., and Greenland, S. (1992), "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, 3, 143-155.

- Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical Models in Epidemiology*, eds. E. Halloran and D. Berry, New York: Springer-Verlag.
- Robins, J. M., and Tsiatis, A. A. (1991), "Correcting for non-compliance in randomized trials using rank preserving structural failure time models," *Communications. in. Statistics,. Theory. and. Methods.*, 20, 2609-2631.
- Rosenbaum, P. R. (1984), "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society*, 147, 656-666.
- Rosenbaum, P. R. (2006), "Comment: the place of death in the quality of life," *Statistical Science*, 21, 313-316.
- Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D. B. (1980), "Comment on "Randomization analysis of experimental data: The Fisher randomization test," by D. Basu," *Journal of the American Statistical Association*, 75, 591-593.
- Rubin, D. B. (2006), "Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with "Censoring" Due to Death," *Statistical Science*, 21, 299-309.
- Shinohara, R. T., Frangakis, C. E., Platz, E., and Tsilidis, K. (2009), "Estimating effects by combining instrumental variables with case-control designs: the role of principal stratification," in Johns Hopkins University, Department of Biostatistics Working Papers, paper 198, Berkeley Electronic Press.
- Sjolander, A. (2011), "Reaction to Pearl's critique of principal stratification," *International Journal of Biostatistics*, 7, 22.
- Tan, Z. (2010), "Marginal and Nested Structural Models Using Instrumental Variables," *Journal of the American Statistical Association*, 105, 157-169.
- van der Laan, M. J., Haight, T. J., and Tager, I. B. (2005), "van der Laan et al. respond to "Hypothetical interventions to define causal effects"," *American Journal of Epidemiology*, 162, 621-622.

- Vanderweele, T. J. (2009), "Concerning the consistency assumption in causal inference," *Epidemiology*, 20, 880-883.
- Vanderweele, T. J. (2011), "Principal stratification - uses and limitations," *International Journal of Biostatistics*, 7, 28.
- Welch, B. L. (1937), "On the z-test in randomized blocks and Latin squares," *Biometrika*, 29, 21-52.
- Wilk, M. (1955), "The randomization analysis of a generalized randomized block design," *Biometrika*, 42, 70-79.