The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 30

Invited Commentary on Pearl and Principal Stratification

Ross Prentice, *Fred Hutchinson Cancer Research Center and University of Washington*

Recommended Citation:
Prentice, Ross (2011) "Invited Commentary on Pearl and Principal Stratification," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 30.
DOI: 10.2202/1557-4679.1359

Invited Commentary on Pearl and Principal Stratification

Ross Prentice

Abstract

Pearl (2011) posed the question of whether confinement of clinical trial analyses involving post-randomization variables to the principal stratum "framework" of Frangakis and Rubin (2002) unduly restricts the scientific questions that can be asked. Frangakis and Rubin illustrated their proposal through examples involving compliance, mediation, and surrogacy. Here the utility of the principal stratum framework, and the potential outcomes formulation from which it derives, are considered for these topics in the specific setting of the Women's Health Initiative randomized, placebo controlled trials of postmenopausal hormone therapy. It is argued that the essential issues related to study reliability and causal interpretation involve the avoidance of context-specific biases that are typically not closely related to whether or not treatment effects have a representation in terms of potential outcomes contrasts. Also, while the questions posed within principal strata may be of interest, some key questions in the hormone therapy setting would not be addressed if restricted to contrasts within principal strata.

KEYWORDS: causal inference, principal stratification, women's health initiative (WHI)

Author Notes: This work was supported by NIH grant CA53996.

1. INTRODUCTION

The editors have invited my comments on Pearl (2011), which questions whether requiring a principal stratification (PS) formulation for analyses having a causal intention unduly restricts the scientific questions that can be addressed, relative to post-randomization variables in clinical trials. The PS framework, proposed by Frankagis and Rubin (2002), aims to permit causal inference on such topics as adherence adjustment, treatment mediation, and surrogate outcomes, by using a potential outcomes formulation. As elaborated below, I share Pearl's concern, and I would like to again (Dawid,2000) raise the more basic question of whether it is reasonable to reserve causal terminology for estimation procedures where the 'estimand' can be expressed in terms of treatment contrasts between potential outcomes. In raising this question, I have no interest in engendering 'merciless philosophical discussion' (Cox, 2000), nor in trying to dissuade the use of potential outcomes formulations for those who find such to be helpful. Rather, I simply wish to express my view that whether or not an estimating function has a representation in terms of potential outcomes contrasts seems rather peripheral to the crucial issue of whether a study provides a fair comparison among treatment groups, and to the reliability and interpretation of study findings more generally To elaborate, I will consider the context of the Women's Health Initiative (WHI) clinical trials of postmenopausal hormone therapy (HT).

The WHI randomized, placebo-controlled trial of conjugated equine estrogens (E-alone, Premarin) enrolled 10,739 women who were post-hysterectomy, and the corresponding randomized, placebo-controlled trial of this same estrogenic formulation plus 2.5 mg/day medroxyprogesterone acetate (E+P, Prempro) enrolled 16,608 women with uterus. All women were postmenopausal and in the age range 50-79 when enrolled during 1993-1998, at 40 U.S. clinical centers. When the HT trials were initiated, these preparations were being used by about 8 million (E-alone) and 6 million (E+P) women in the U.S. alone, mostly for vasomotor symptom control, and/or for presumed chronic disease risk reduction benefits. The E+P trial was stopped early (Writing Group, 2002) when health risks were judged to exceed health benefits over a treatment period that averaged 5.6 years. Health risks included an early elevation in coronary heart disease (CHD) incidence, the primary trial outcome targeted for risk reduction, and more sustained elevations in stroke, venous thromboembolism, and breast cancer. The E-alone trial was also stopped early (Anderson et al, 2004) following an average 7.1-year treatment period, substantially because of a stroke elevation of similar magnitude to that observed for E+P, though overall health risks and benefits were rather balanced.

The E+P trial results were a shock to practicing gynecologists and their patients, who had come to expect major benefits for CHD risk reduction, based on observational studies and intermediate outcome clinical trials. Shortly after the 2002 trial report about 70% of E+P users in the U.S. stopped these medications, as did about 40% of E-alone users, with corresponding changes worldwide that were accompanied by regulatory warnings and revised professional society position statements. The follow-up of trial women after their abrupt cessation of treatment revealed, for example, that the 2 to 3-fold breast cancer risk elevation, after about 5 years of E+P use, appeared to return to basal levels within a couple of years following cessation (Chlebowski et al, 2009). Under this scenario, the major change in clinical practice mentioned above can be projected to have led to about 15,000 fewer women being diagnosed with breast cancer in the U.S. each year since about 2004, a feature that is evident in national breast cancer incidence rate trends for postmenopausal women (e.g., Ravdin et al, 2007).

How does one answer scientific questions concerning the effect of E-alone or E+P on the risk of specific clinical outcomes, such as CHD, stroke, venous thromboembolism, breast cancer, colorectal cancer, fractures, dementia, or concerning overall health benefits versus risks; and how do available statistical formulations and procedures support the development of reliable and interpretable answers?

2. DATA ANALYSIS METHODS

2.1 ITT Analysis in Randomized Controlled Trials

Consider a clinical outcome Y, such as a possibly censored time to CHD diagnosis, and a randomization indicator Z (1-active, 0-placebo). Randomization of the treatment assignment has a most important role in studies of the effects of Z on Y. Randomization ensures the statistical independence between Z and all pre-randomization disease risk factors, whether or not such factors are even recognized as such. In short, randomization ensures a fair comparison between randomized groups, as is crucial to an argument of causation. There are well-established statistical methods (e.g., hazard ratio-based tests and estimators) for comparing randomization groups in terms of pertinent features of the corresponding distribution of Y.

These methods have nearly always been developed by modeling observable quantities (Z, Y), though it seems likely that any such method could be shown to arise also from a formulation in terms of the conceptual potential outcomes Y(0) and Y(1) that would obtain if the study subject were assigned Z=0 or Z=1 respectively. Even for the intention-to-treat (ITT) analysis in a randomized trial setting, however, some care may be needed to avoid ‘goats’ (Dawid, 2000) wherein an estimation procedure depends importantly on non-identifiable parameters that characterize the association between Y(0) and Y(1).

The use of semiparametric or nonparametric models for (Y(0), Y(1)) (e.g., Robins, 1999) can presumably avoid this problem for the basic trial comparison. Hence, whether one uses a 'classical' or a potential outcomes formulation for the ITT analysis seems primarily a matter of taste, and one that doesn't seem to be germane to the reliability and interpretation of trial results.

A causal interpretation of the ITT analysis does require non-differential ascertainment of the outcome (Y) between the randomized groups. This important requirement motivates various types of blindedness (or masking) that may be a part of a clinical trial protocol (e.g., blinding of study subjects, investigators, and outcomes adjudicators). To avoid bias, it may be crucial that study activities that may lead to the identification of an outcome (e.g., procuring of self-report of hospitalizations or health events from study participants) be identical for the treatment groups being compared. See Psaty and Prentice (2010) for a recent example where an open-label decision process as to which outcome packets to send for blinded adjudication may have compromised the validity of trial comparisons. Even if all forms of blindedness are in place, the treatment itself may affect the ability to diagnose disease, or the timing of disease diagnosis. For example, E+P tends to increase breast density, perhaps delaying malignant breast tumor diagnosis even under a protocol that includes regular mammographic screening and other efforts to ensure equal breast cancer ascertainment between randomized groups.

The challenge in working toward an argument of causation increases substantially when one moves away from the randomized controlled trial and its ITT analysis, whether in the direction of an observational study that may involve confounding, or in the direction of randomized trial analyses that condition in some manner on post-randomization intermediate outcome data.

2.2 Observational Study Analyses

The observational setting is quite central to population science research, since it is not practical to test many approaches to chronic disease prevention in randomized trials with clinical disease outcomes. Observational studies need to implement bias avoidance procedures to the greatest extent possible, to support an argument of causation. Of course, equal outcome ascertainment according to the primary 'exposure' (Z) under study is as important, say, in a prospective cohort study as in the randomized trial, and may be difficult to achieve unless there is a explicit outcome ascertainment protocol and a context for its enforcement. If, in addition, Z is difficult to measure, such as daily grams of fat consumption, or minutes per day of moderate physical activity, the reliability of exposure-disease associations may be compromised. These important nutritional and physical activity epidemiology research areas may require the acquisition of objective exposure data, and related careful statistical modeling (e.g., Neuhouser et al, 2008), for reliable inference.

Then there is the classical problem of confounding: Important confounding factors may or may not be measured, or even recognized as such. Furthermore, there may be important confounders that are poorly measured (e.g., once again dietary or physical activity factors), and measured confounding factors need to be properly modeled and included in data analysis for bias avoidance.

One can attempt to avoid confounding biases through modeling and analysis of Y given (Z, X) for some set of disease risk factors X, or by stratifying study subjects on propensity scores (Rosenbaum and Rubin, 1983) that estimate exposure probabilities Z given X. The former targets association parameters conditionally on X, while the latter typically targets marginal association parameters. A recent commentary by Vansteelandt and Keiding (2012) describes the close connections between a marginal structural modeling approach with potential outcomes, and classical disease rate standardization in epidemiology.

The important issue for study interpretation is whether one can argue that comparisons of outcomes Y across categories of Z have been rendered ‘fair’ by virtue of controlling for X. Unfortunately, one is not in a position to judge whether this objective has been achieved in a particular application, though it certainly helps if there is a substantial knowledge base concerning disease risk factors and/or concerning the characteristics related to the exposure Z. Whether or not one formulates questions about the association between Y and Z on these observable variables, or on the potential outcomes (Y(0), Y(1)), does not seem central to the reliability and interpretation of the associations that emerge. Rather, crucial statistical issues, for a study to be persuasive as to Z causing Y, are the extent to which confounding has been controlled, the quality of the exposure/treatment data Z, and the extent to which outcome ascertainment is independent of (Z, X). Additional arguments concerning biological plausibility and the temporality of exposure and disease may also be needed.

Settings in which both randomized trial and prospective cohort data are available for a given treatment/exposure provide an opportunity to compare findings, and to compare associated biases. The WHI includes a prospective cohort study among 93,676 women drawn from the same catchment populations as the clinical trial, with much commonality in outcome and other data ascertainment, including a personal interview to ascertain hormone therapy histories at baseline followed by annual updates. We undertook a series of analyses of the individual-level data from the WHI HT trials and from this Observational Study (OS) cohort, toward understanding the apparently major differences between the HT trial and OS findings (e.g., Prentice et al, 2009a). For CHD, after controlling for a set of standard risk factors, and allowing the hazard ratio to vary as a function of duration of HT use, there was in fact reasonable agreement between the two data sources. Observational studies had collectively missed a rather dramatic increase in CHD incidence in the early months of HT use (especially for E+P), perhaps because of an undue focus on proportional hazards models, in conjunction with the typical inclusion of many women who were already some years into an episode of HT use at cohort enrollment. For breast cancer, the hazard ratio increased strongly with duration of use for E+P, but also was comparatively higher among women who started HT at or soon after the menopause. Trial and cohort analyses agreed well for both E+P and E-alone after taking account of these timing issues, along with conventional potential confounding factors and data on mammographic screening during cohort follow-up. For some other outcomes, however, including stroke, colorectal cancer, hip fracture, and total mortality, important differences between HT trial and OS cohort study results remained following these analytic maneuvers. These differences may reflect the need for confounding control efforts that are more comprehensive than is typical in epidemiologic practice, for these outcomes.

3. CLINICAL TRIALS AND POST-RANDOMIZATION VARIABLES

3.1 Adherence Adjustment

Suppose now that X is a post-randomization variate, or change from ‘baseline’ in some such variate, in a randomized controlled trial, and that one wishes to estimate aspects of the dependence of Y on Z, after controlling in some fashion for X. In proposing their principal stratification approach, Frankagis and Rubin (2002) list motivational special cases in which X involves measurements to assess adherence to the assigned treatment Z, intermediate variables that may help to understand pathways by which Z affects Y, and surrogate outcomes where X may be able to replace Y the evaluation of the effects of Z on Y.

The randomized treatment assignment provides an important advantage for these analyses as well, since it ensures a fair comparison between the distributions of (X, Y) between randomization groups, that is not confounded by pre-randomization factors. Hence, distributions, such as that for Y given (X, Z) that can be derived from that for (X, Y) given Z can also be fairly compared. However, related parameter estimates may not address some questions of scientific interest, necessitating additional assumptions and analytic methods.

For example, X may be a process that assesses treatment adherence (e.g., pill-taking) over time since randomization, while Y is a time-to-disease occurrence variate. An evaluation of the dependence of Y on Z among study subjects meeting a certain adherence criterion is operationally straightforward (assuming non-differential ascertainment of (X, Y) between randomized groups), but lack of adherence may arise for quite different reasons between the treatment groups. Furthermore, there may be a desire to compare the effects of Z on Y in the trial cohort in a 'counterfactual' setting in which full adherence was somehow achieved.

The HT trials provide a setting with rather complex adherence issues. Even though women in the community tended to use these preparations typically for just a few years, there was a desire to conduct a trial with a treatment period of sufficient duration for benefits and risks of a longer-term exposure to become apparent. The trial protocol included a complex program of dosage modification, or temporary stoppage, for women who were experiencing certain side effects,, or permanent stoppage in the face of certain safety issues. The blinded assignment was retained for both participant and provider to the extent practical. When the trials got underway, it soon became evident (to an unblinded data and safety monitoring group) that there was a high incidence of vaginal bleeding with E+P (e.g., 30-40% in first year of use), and that this issue continued for some years for a fraction of participating women. Similarly, there was a substantially elevated incidence of breast tenderness in the early years of use of either E-alone or E+P. On the other hand, there was the need to be attentive to sustained vaginal bleeding in the placebo group, as this may be suggestive of uterine cancer, resulting in the need to unblind the clinic gynecologists for some small fraction of women. Other reasons for non-adherence related to intercurrent illnesses that interfered with continued research project participation, and transportation issues related to clinic visits where pills were dispensed, among others. Pill-taking adherence was strongly time-dependent. By the time the E+P trial was stopped about 40-45% of participating women had failed to satisfy an 80% pill-taking criterion over the preceding year at at least one point in time, in both the treatment and placebo groups.

A principal stratification approach (see also Imbens and Rubin, 1997) to adherence-adjusted analyses would presumably aim to make a treatment comparison among women who would be adherent (e.g., ≥80% of pills taken for each year during the trial follow-up period), whether assigned to active treatment of placebo. Since reasons for non-adherence were rather different between treatment groups, and 40-45% of women in either group would fail to satisfy a treatment group-specific adherence criterion, such a 'complier group' assessment would presumably involve less than half of the study target population. Nevertheless, if a 'complier principal stratum' was identifiable, the resulting comparison would clearly be of interest. However, it would not address some questions of public health interest concerning the full adherence benefits and risks of HT in the broader target population in which the WHI trials were conducted. The lack of identifiability of this principal stratum, along with the need to 'integrate out' the missing outcomes for both the intermediate X and the main outcome Y, raises questions about the properties of any resulting inference, which if sheep-like (Dawid, 2000) may entail sensitivity assumptions that imply an imprecise inference, or if goat-like may entail influential assumptions on the joint distribution for potential outcomes (X(0), Y(0), X(1), Y(1)) that may complicate interpretation.

Instead, a modeling approach involving observable quantities has been employed in HT trial adherence-adjusted analyses to date. These involve building a treatment group-specific adherence model for the probability of pill-taking adherence through each follow-up year, followed by time-dependent inverse adherence probability weighting in the Cox model partial likelihood score equation for hazard ratio parameter estimation (e.g., Chlebowski et al, 2009). This procedure was implemented by censoring the failure time response variable (Y) for a woman six months after she first becomes non-adherent according to the above pill-taking criterion. The six-month period was included in an attempt to avoid censoring the follow-up of women who adhered to study pills until a short time prior to diagnosis, during which time diagnosis-related procedures may have interfered with continued adherence. This approach allows women to contribute to the data analysis if adherent during the early years of trial follow-up, a feature that presumably could be incorporated into a PS analysis, though perhaps at the expense of abrogating the favorable 'cardinality' feature mentioned by Pearl (2011), since the number of principal strata may need to become large in an attempt to fully use available data. The inverse adherence probability analyses, using models for observable data only, include contributions from women who may not have been adherent if assigned to the other treatment group, and hence have a basic difference from PS 'complier' analyses. The reliability and interpretation of these 'classical' adherence-adjusted analyses hence depends on an assumption of a valid adherence probability model over the trial follow-up period, a situation analogous to confounding control modeling efforts in observational settings. The fact that the adherence probability model can depend on post-randomization data (e.g., vaginal bleeding, breast changes, or other symptoms experienced by women during hormone therapy trial follow-up) provides a framework for enhancing model completeness and validity. The inverse adherence probability weighting allows women having a certain symptom profile who remain adherent to 'represent' those having a similar profile who do not, thereby providing a simple attempt to estimate treatment effect parameters in the entire target population if all participating women had been willing and able to meet adherence criteria. This and a PS complier analysis each depends on assumptions that cannot be fully tested or justified, and the assumptions differ as befits their different aims. The PS analysis may lead to sensitivity analysis and imprecise inferences, which could be viewed as desirable if it prevents over-interpretation. The inverse adherence probability weighting approach can be expected to lead to inferences that became more strongly justified as additional variates are used to build adherence probability models. The two approaches can be viewed as complementary since they address different aspects of adherence-adjusted analyses.

3.2 Interacting and Mediating Variables

Although good clinical trial reporting practice would avoid undue data fragmentation into subset analyses, treatment effects can typically be expected to vary with some study subject characteristics, on any particular scale of treatment assessment. Statistical models for ratios (e.g., hazard ratios or odds ratios) may have distinct advantages in frequently leading to parsimonious models, compared to corresponding models for absolute risks. Analyses to identify interacting factors in a clinical trial context are much facilitated by randomization. For example, because of the orthogonality between Z and pre-treatment factors X, interaction analyses may be quite robust to classical measurement error in the ascertainment of X. Also, if Y represents the incidence of a fairly rare disease, analyses of treatment effects within subsets may be able to be based on efficient case-only analyses in which X needs to be assessed only for individuals who turn out to develop the study disease. See Prentice et al, 2009b, for an HT trial example where X is a high-dimensional genotype.

The situation is quite different, however, when it comes to mediation analyses that aim to identify the principal biological pathways whereby Z may affect Y. In this context X may, for example, involve short-term events that are influenced by treatment, or changes from baseline in blood-based biomarkers. Such intermediate variables may be strongly associated with Z, making it challenging to distinguish the role of X and Z in their association with Y. This is especially the case if X is subject to even a modest amount of measurement error since analytic procedures that ignore measurement error will tend to underweight the role of X, given that the randomization assignment Z is known without error.

A typical mediation analyses based on observable quantities would proceed by estimating a treatment effect in the absence of the potential mediator, and would examine the extent to which such effect is attenuated toward the null when X is added to the analytic model. For example, certain hazard ratio parameters may characterize the effect of Z on a disease incidence time Y, and to assess mediation, one compares such parameters without and with the inclusion of X in the hazard ratio model. A substantial effort of this type was mounted in the HT trials. This included a cardiovascular disease biomarker study that measured blood biomarkers at baseline and 1-year post-randomization, for women who developed cardiovascular disease beyond 1 year following randomization and 1-to-1matched controls. A substantial list of biomarkers was considered with a focus on markers of inflammation, coagulation/thrombosis, and lipids/lipoproteins. For example, many of these markers related strongly to the risk of CHD and stroke in the WHI study population and were affected by HT use, but none of those evaluated appeared to mediate the observed HT effects on these diseases (Kooperberg et al, 2007; Rossouw et al, 2008). For example, E+P and E-alone have apparently favorable effects on high-density (HDL) and low-density lipoprotein (LDL) cholesterol concentrations, and unfavorable effects on C-reactive protein (CRP), each of which is regarded as a risk factor for cardiovascular disease. Additional efforts have looked more agnostically in a high-dimensional fashion at changes to the plasma proteome following HT use, with a focus on changes that also related to CHD or stroke risk. Both E-alone and E+P were found to have a profound effect on plasma protein concentrations in multiple pathways relevant to observed clinical effects (Pitteri et al, 2010), and some of the affected proteins were found to be risk markers for CHD or stroke (Prentice et al, 2010), but to date, no important mediators on HT effects in these diseases have been identified.

It seems that the crucial barrier to progress, in this continuing work, is that of biomarker process modeling and measurement error. Observed biomarker values may be subject to non-trivial technical measurement error and, importantly, the biomarker functions that relate to cardiovascular disease risk may involve, say, average values over a period of some years. Also, the underlying 'latent' biomarker changes of interest may correlate very strongly with treatment assignment making it difficult to distinguish Z from the underlying biomarker changes, X, in relation to Y, especially on the basis of a limited set of measurements on the biomarker process (e.g., from blood specimens at baseline and 1-year only in the HT trials). Hence, it may not be surprising that the measured biomarker change does not appear to mediate in this type of context (see Prentice and Zhao, 2011, for further discussion). These biomarker measurement or modeling issues have not been carefully addressed in the statistical literature, and this lacking could substantially reduce the reliability and usefulness of mediation analyses efforts to date, in contexts such as the WHI HT trials. Of course, biased mediation analyses may prevent identification of the important biological pathways, perhaps leading to additional unnecessary research while deterring the development of related treatments that may have an improved health benefit versus risk profile.

One could consider a PS approach to studying the extent to which, say, changes in HDL or LDL cholesterol or CRP help to explain observed E-alone or E+P effects on CHD or stroke. Doing so would presumably involve strata defined by categories of change in some or all such variables under both the active treatment and under placebo. For example, one could entertain a stratum formed by women whose measured HDL would increase by more than 10 mg/dl and measured CRP would increase by more than 50% if assigned to active HT, and for whom measured HDL would not change by as much as 10 mg/dl and measured CRP would not change by as much as 20% if assigned to placebo. Clearly such strata would need to be rather narrowly defined if the within-stratum treatment effect is to be viewed as controlled for the potential biomarker changes, and some means of combining treatment contrast information across principal strata may be needed for an inference of useful precision. Such a PS approach would not acknowledge the biomarker modeling and measurement issues discussed in the preceding paragraph. It seems that the PS approach would require considerable further development before its utility for mediation analysis, in a context such as the HT trials, could be demonstrated.

3.3 Surrogate Outcomes

There has been a longstanding interest among clinical investigators in identifying an intermediate outcome variable X that becomes evident earlier or with greater frequency than does the ‘true’ outcome Y, for the purpose of evaluating the effects of a treatment Z, and that could substitute for Y for treatment effect evaluation. Since treatments often have multiple effects, some intended and some not, a putative surrogate for a particular Y may be unlikely to serve as such for other clinical outcomes. Also, mechanisms and pathways can vary substantially among treatments, even if they are in the same general class, so the applicability of an X, that may seem to satisfy surrogacy criteria in a given trial, to a future trial with a somewhat different treatment, or even the same treatment in a different population, needs to be carefully considered and justified.

To provide concreteness to a discussion of surrogate outcomes and criteria, I proposed (Prentice, 1989) that the term ‘surrogate outcome’ for a specific outcome Y and treatment Z be reserved for outcomes X for which the null hypothesis of no effect of Z on X was equivalent to that of no effect of Z on Y. I then translated this definition into mathematical conditions on the joint distribution of (X, Y) given Z, in the typical setting where Y is a time to response variates and X is a stochastic process recorded over trial follow-up. In present notation, for X independent of Z to imply Y independent of Z, it is sufficient that

(i) Y given (X, Z) be independent of Z.

To see this, using f generically for distribution and E for expectation, one can write

$$\begin{aligned} f(Y|Z) &= E_{X|Z}\ \{f(X,Y|Z\} \\ &= E_{X|Z}\ \{f(Y|X,Z)\ f(X|Z)\}, \end{aligned}$$

so f(X|Z) independent of Z and f(Y|X,Z) independent of Z, implies f(Y|Z) independent of Z. Hence, if Y depends on Z then X must also depend on Z under (i). This seems to be the intent of a property that Frangakis and Rubin (2002) label 'causal necessity'. However, they argue that (i) is insufficient for causal necessity, evidently because it doesn't have an expression in general in terms of a potential outcome contrast. They then label variables X satisfying (i) as 'statistical surrogates', whereas those satisfying a corresponding criterion to (i) with X replaced by potential outcomes X(0) and X(1) are labeled as 'principal surrogates'. The randomized trial assignment provides (assuming non-differential outcome ascertainment) a causal basis for a test of null hypothesis concerning the effect of Z on X, so that if null hypotheses for Z on X and Z on Y coincide, one induces a causal basis for the null hypothesis concerning the effects of Z on Y. While one can debate whether the equivalence of null hypothesis is the desired surrogate outcome definition, it is not clear what alternative definition Frangakis and Rubin (2002) have in mind that leads to their principal surrogate criterion.

Equivalence of the null hypotheses concerning effects on X and Y also requires an implication in the other direction, which may be more demanding. That is, a dependence of a short-term variable X on Z may exist, without a corresponding dependence of Y on Z. Conditions beyond (i) that ensure that Y independent of Z implies X independent of Z (Prentice, 1989) are (ii) Y depends on X and (iii) any dependence of Y given (X, Z) on Z not average out over the distribution of X given Z. Condition (ii) is described as 'statistical generalizability' by Frangakis and Rubin (2002). Conditions (i)-(iii) combine to constitute my proposed surrogate outcome definition.

In a recent paper Li et al (2011) discuss the application of the PS framework to surrogacy issues. They note that 'assumption free analysis is impossible…because of the large 'missing data' problem implied by the unobserved potential outcomes', and that additional assumptions are needed for a useful inference. One such assumption, so-called monotonicity, which assumes that it is impossible for anyone to benefit more from the control than the active treatment, would typically undermine the need for and ethics of a randomized trial.

The major challenge in applying the 'equivalence of null hypotheses' definition of surrogate outcomes in any particular trial context is that of justifying that condition (i) holds in the study context, even approximately. For this stringent condition to be plausible, it may be necessary to require X to be high-dimensional and to involve measurements over much of the trial follow-up period (Prentice, 2009). An attempt to apply principal strata in that context would presumably require a large number of strata based on X(0) and X(1).

In the WHI HT trial, in spite of well-characterized cohorts with many biomarker substudies, there are presently no known credible surrogate outcomes for any of the major clinical outcomes.

4. CONCLUSION

In biomedical research settings, such as the WHI HT trials, investigators are often presenting arguments toward the causality of a treatment of exposure Z in relation to an outcome Y. The persuasiveness of the argument depends on the ability of the study design, conduct, and analysis to avoid recognized sources of potential bias. Such biases may involve differential outcome ascertainment as a function of Z; systematic or random biases in measuring Z itself; and the comprehensiveness or lack thereof of measurement and modeling of pre-randomization or post-randomization measurements X, that may need to be controlled to address specific questions. The question of whether or not estimating functions for treatment contrasts have a representation as contrasts between potential outcomes (Y(0), Y(1)) and (X(0), X(1)) seems peripheral to these crucial bias avoidance issues, and it seems unwise to label analytic procedures as causal or not based on whether such a representation is available. In particular, the principal stratification framework seems to provide a too narrow context for data analytic procedures that aim to support causation arguments in clinical trials when post-randomization variables are involved.

In response to Reader Reaction Pearl (2011) indicates that restriction to the PS framework unduly confines the questions that can be addressed in all PS papers that he has read, with the exception of Sjolander et al (2010) on hormone therapy. This paper examines prognosis following a breast cancer diagnosis and defines principal strata according to the non-identifiable feature of whether or not the tumor was caused by hormone therapy. They conclude, from their observational data sources, that women whose tumors are caused by HT have comparable better prognosis. In contrast, data from the WHI E+P trial (Chlebowski et al, 2010), with its randomized HT assignment and carefully controlled clinical context, indicate that women using E+P not only have a higher incidence of invasive breast tumors, but such tumors are of higher stage of diagnosis. Though the data are not extensive, breast cancer mortality is evidently also increased and survival since diagnosis reduced in the active treatment versus the placebo group. While the basis for this observational versus clinical trial discrepancy needs to be investigated, it illustrates the major challenge in obtaining reliable information, and in establishing causation, for this type of treatment effect evaluation. The use of certain types of data analysis methods alone is by no means sufficient to merit the ‘causal’ label.

REFERENCES

Chlebowski RT, Anderson GL, Gass M, et al. Estrogen plus progestin and breast cancer incidence and mortality in postmenopausal women. JAMA 2010;304(15):1684-1692.

Chlebowski RT, Kuller LH, Prentice RL, et al; WHI Investigators. Breast cancer after use of estrogen plus progestin in postmenopausal women. N Engl J Med 2009;360(6):573-587.

Cox DR. Comment on 'Causal Inference Without Counterfactuals'. J Am Stat Assoc 2000;95(450):424-425.

Dawid AP. Causal inference without counterfactuals. J Am Statist Assoc 2000;95(450):407-424.

Frangakis CE and Rubin DB. Principal stratification in causal inference. Biometrics
2002;58:21-29.

Imbens GW and Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. Ann Statist 1997;25(1):305-327.

Kooperberg C, Cushman M, Hsia J, et al. Can biomarkers identify women at increased stroke risk? The Women's Health Initiative Hormone Trials. PLoS Clin Trials 2007;2(6):e28.

Li Y, Taylor JMG, Elliott MR. Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. Biostatistics 2011;12(3):478-492.

Neuhouser ML, Tinker L, Shaw PA, et al. Use of recovery biomarkers to calibrate nutrient consumption self-reports in the Women's Health Initiative. Am J Epidemiol 2008;167(10):1247-1259.

Pearl, Judea (2011) "Principal Stratification — a Goal or a Tool?," The International Journal of Biostatistics: Vol. 7: Iss. 1, Article 20.

Pitteri SJ, Hanash SM, Aragaki A, et al. Postmenopausal estrogen and progestin effects on the serum proteome. Genome Medicine 2009;1(12):121.1 – 121.14.

Prentice RL. Surrogate endpoints in clinical trials: discussion, definition and operational criteria. Stat Med 1989;8:431-440.

Prentice RL. Surrogate and mediating endpoints: current status and future directions (editorial). J Natl Cancer Inst 2009;101:216-217.

Prentice RL and Zhao S. On the use of biomarkers to elucidate clinical trial results: examples from the Women's Health Initiative. Invited paper for Proceedings of the 4th Seattle Biostatistics Symposium: Clinical Trials, 2011.

Prentice RL, Manson JE, Langer RD, et al. Benefits and risks of postmenopausal hormone therapy when it is initiated soon after menopause. Am J Epidemiol 2009a;170(1):12-23.

Prentice RL, Huang Y, Hinds DA, et al. Variation in the FGFR2 gene and the effects of postmenopausal hormone therapy on invasive breast cancer. Cancer Epidemiol Biomarker Prev 2009b;18(11):3079-3085.

Prentice RL, Paczesny SJ, Aragaki A, et al. Novel proteins associated with risk for coronary heart disease or stroke among postmenopausal women identified by in-depth plasma proteome profiling. Genome Med 2010;2(7):48-60.

Psaty BM and Prentice RL. Minimizing bias in randomized trials. The importance of blinding. JAMA 2010;304(7):793-794.

Ravdin PM, Cronin KA, Howlader N, et al. The decrease in breast-cancer incidence in 2003 in the United States. N Engl J Med 2007;356(16):1670-1674.

Robins, JM. "Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference," in Statistical Models in Epidemiology: The Environment and Clinical Trials, eds. M. E. Halloran and D. Berry, New York: Springer-Verlag,1999, pp. 95-134.

Rosenbaum P and Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.

Rossouw JE, Cushman M, Greenland P, et al. Inflammatory, lipid, thrombotic, and genetic markers of coronary heart disease risk in the Women's Health Initiative trials of hormone therapy. Arch Intern Med 2008;168(20):2245-2253.

Sjolander A, Vansteelandt S, Humphreys K. A principal stratification approach to assess the differences in prognosis between cancers caused by hormone replacement therapy and by other factors. Int J Biostat 2010;6(1):Article 20.

Vansteelandt S and Keiding N. Invited commentary: G-computation—lost in translation? Am J Epidemiol 2011;173(7):739-742. Epub 2011 Mar 16.

Writing Group for the Women's Health Initiative (Rossouw JE, Anderson G, Prentice RL, et al). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. JAMA 2002;288:321-333.