The International Journal of Biostatistics Vol. 7: Issue 1, Article 20, Pages 1–13, ISSN (Online) 1557-4679,
DOI: 10.2202/1557-4679.1322, March 2011.

TECHNICAL REPORT
R-382
March 2011

Principal Stratification – A goal or a tool?

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

March 29, 2011

Abstract

Principal stratification has recently become a popular tool to address certain causal inference questions particularly in dealing with post-randomization factors in randomized trials. Here we analyze the conceptual basis for this framework and invite response to clarify the value of principal stratification in estimating causal effects of interest.

Keywords: causal inference, principal stratification, surrogate endpoints, direct effect, mediation

1 Introduction

The past few years have seen a substantial number of studies claiming to be using "The Principal Strata Approach" or "A Principal Strata Framework," a term coined by Frangakis and Rubin (2002). An examination of these studies reveals that they fall into four distinct categories, each subscribing to a different interpretation of "principal strata (PS)" and each making different assumptions and claims. The purpose of this note is to clarify this distinction and to identify areas of application where these interpretations may be useful.

2 Notation and preliminaries

We begin with the usual potential-outcome notation, $Y_x(u)$, which, for every unit u, defines a functional relationship

$$y = f(x,u) \tag{1}$$

between a treatment variable X and an outcome variable Y. Here y and x stand for specific values taken by Y and X, respectively, and u may stand either for the identity of a unit (e.g., a person's name) or, more functionally, for the set of unit-specific characteristics that are deemed relevant to the relation considered (Pearl, 2000, p. 94).

For any function f, the population of units can be partitioned into a set of homogeneously responding classes, called "equivalence classes" (Pearl, 2000, p. 264), such that all units in a given class respond in the same way to variations in X. For example, if X and Y are binary, then, for any given u, the relationship between X and Y must be one of four functions:

$$\begin{aligned} f_1 &: y = 0, & f_2 &: y = x, \\ f_3 &: y \neq x, & f_4 &: y = 1. \end{aligned} \tag{2}$$

Therefore, as u varies along its domain, regardless of how complex the variation, the only effect it can have is to switch the relationship between X and Y among these four functions. This partitions the domain of U into four regions, as shown in Fig. 1, where each region contains those points u that induce the same function.

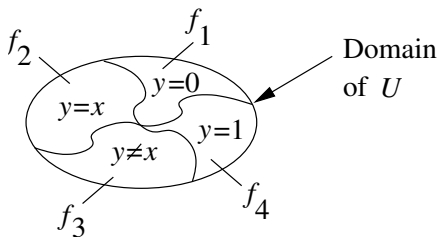

Figure 1: The canonical partition of U into four equivalence classes, each inducing a distinct functional mapping from X to Y for any given function $y = f(x,u)$.

If u is represented by a multi dimensional vector of unit-specific characteristics, we can regard the class membership of u as a lower dimension variable, R which, together with X, determines the value of Y. Pearl (1993) and Balke

and Pearl (1994a,b) called this variable "response variable" while Heckerman and Shachter (1995) named it a "mapping variable," (see Lauritzen, 2004).

The relation of this partition to the potential-outcome paradigm (Neyman, 1923, Rubin, 1974) is simple. Each equivalence class corresponds to an intersection of two potential outcomes (assuming binary variables), as shown in Table 1. The types described in Table 1 are often given problem-specific names, for

Response Type	Functional Description	Potential outcome Description
Type-1	$f(x,u) = 0$	$Y_0(u) = 0$ and $Y_1(u) = 0$
Type-2	$f(x,u) = x$	$Y_0(u) = 0$ and $Y_1(u) = 1$
Type-3	$f(x,u) = 1 - x$	$Y_0(u) = 1$ and $Y_1(u) = 0$
Type-4	$f(x,u) = 1$	$Y_0(u) = 1$ and $Y_1(u) = 1$

Table 1:

example, 1 - doomed, 2 - responders, 3 - hurt, 4 - always healthy (see Heckerman and Shachter (1995) or Imbens and Rubin (1997)).

3 Applications

The idea of characterizing units by their response function, rather than their baseline features has several advantages, stemming primarily from the parsimony achieved by the former. Whereas each unit may have thousands of features, standing in unknown relationships to X and Y, the number of functions that those features can induce is limited by the cardinality of X and Y, and each such function defines the response of Y unequivocally.

Robins and Greenland were the first to capitalize on this advantage and have used classification by response type as a cornerstone in many of their works, including confounding (1986) attribution (1988, 1989a,b) and effect decomposition (1992).

Pearl (1993) and Balke and Pearl (1994a,b) formulated response types as variables in a graph, and used the low dimensionality (16) of two response variables to derive tight bounds on treatment effects under condition of noncompliance (Balke and Pearl, 1997).

Chickering and Pearl (1997) as well as Imbens and Rubin (1997) used the parsimony of response type classification in a Bayesian framework, to obtain pos-

terior distributions of causal effects in noncompliance settings. It is obviously easier to assign meaningful priors to a 16-dimensional polytope than to a space of the many features that characterize each unit (see Pearl, 2009a, Ch. 8).

Baker and Lindeman (1994) and Imbens and Angrist (1994) introduced a new element into this analysis. Realizing that the population averaged treatment effect (ATE) is not identifiable in experiments marred by noncompliance, they have shifted attention to a specific response type (i.e., compliers) for which the causal effect was identifiable, and presented the latter as an *approximation* for ATE. This came to be known as LATE (Local Average Treatment Effect) and has spawned a rich literature with many variants (Angrist, Imbens, and Rubin, 1996, Heckman and Vytlacil, 2001, Heckman, 2005) all focusing on a specific stratum or a subset of strata for which the causal effect could be identified under various combinations of assumptions and designs. However, most authors in this category do not state explicitly whether their focus on a specific stratum is motivated by mathematical convenience, mathematical necessity (to achieve identification) or a genuine interest in the stratum under analysis.

Though membership in response-type classes is generally not identifiable and is vulnerable to unpredictable changes,[1] such membership may occasionally be at the center of a research question. For example, the effect of treatment on subjects who *would have survived regardless of treatment* may become the center of interest in the context of censorship by death (Robins, 1986). Likewise, survival in cancer *cases caused by* hormone replacement therapy need be distinguished from survival in cancer *cases caused by* other factors (Sjölander, Humphreys, and Vansteelandt, 2010). In such applications, expressions of the form

$$P(Y_x = y | Z_x = z, Z_{x'} = z') \tag{3}$$

emerge organically as the appropriate research questions, where Z is some post-treatment variable, and the condition $(Z_x = z, Z_{x'} = z')$ specifies the response-type stratum of interest.

4 The Frangakis-Rubin Framework

Thus far, we discussed principal strata as a classification of units into equivalence classes that arises organically from the logic of counterfactuals and the

[1] For example, those who comply in the study may or may not comply under field conditions, where incentives to receive treatment are different.

inference challenges posed by the study. A different perspective was proposed in the paper of Frangakis and Rubin (2002) who attached the label "principal strata" to this classification. Frangakis and Rubin viewed the presence of a stratum $(Z_x = z, Z_{x'} = z')$ behind the conditioning bar (Eq. (3)), as a unifying conceptual principle, deserving of the title "framework," because it seems to be correcting for variations in Z without the bias produced by standard adjustment for post-treatment variables. In their words: "We are aware of no previous work that has linked such recent approaches for noncompliance to the more general class of problems with post-treatment variables." The approach that subsequently emerged from this perspective, and came to be known as the "principal strata framework" presumes that most if not all problems involving post-treatment variables can, and should be framed in terms of strata-specific effects.

We have seen in Section 3, however, that the class of problems involving post-treatment variables is not monolithic. In some of those problems (e.g., noncompliance), post-treatment variables serve as useful information sources for bounding or approximating population-wide questions, while in others, they define the research question itself. More importantly, while some of those problems can be solved by conditioning on principal strata, others cannot. In those latter cases, constraining solutions to be conditioned on strata, as in (3), may have unintended and grossly undesirable consequences, as we shall see in the sequel.

4.1 The principal strata direct effect (PSDE)

A typical example where definitions based on principal stratification turned out inadequate is the problem of *mediation*, which Rubin (2004, 2005) tried to solve using an estimand called "Principal Stratification Direct Effect" (PSDE). In mediation analysis, we seek to measure the extent to which the effect of X on Y is mediated by a third variable, Z, called "mediator." Knowing the direct (unmediated) effect permits us to assess how effective future interventions would be which aim to modify, weaken or enhance specific subprocesses along various pathways from X to Y. For example, knowing the extent to which sex discrimination affects hiring practices would help us assess whether efforts to eliminate educational disparities have the potential of reducing earning disparity between men and women.

Whereas causal notions of "direct effect" (Robins and Greenland, 1992, Pearl, 2001) measure the effects that would be transmitted in the population with all mediating paths (hypothetically) deactivated, the PSDE is defined as the effects transmitted in those units only for whom mediating paths *happened to be deactivated* in the study. This seemingly mild difference in definition leads to un-

intended results that stand contrary to common usage of direct effects (Robins, Rotnitzky, and Vansteelandt, 2007, Robins, Richardson, and Spirtes, 2009, VanderWeele, 2008), for it excludes from the analysis all individuals who are both directly and indirectly affected by the causal variable X (Pearl, 2009b). In linear models, as a striking example, a direct effect will be flatly undefined, unless the $X \rightarrow Z$ coefficient is zero. In some other cases, the direct effect of the treatment will be deemed to be nil if a small subpopulation exists for which treatment has no effect on both the outcome and the mediator (Pearl, 2011). These definitional inadequacies point to a fundamental clash between the Principal Strata Framework and the very notion of mediation.

Indeed, taking a "principal strata" perspective, Rubin found the concept of mediation to be "ill-defined." In his words: "The general theme here is that the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful to clear statistical thinking in real, as opposed to artificial problems" (Rubin, 2004). Conversely, attempts to define and understand mediation using the notion of "principal-strata direct effect" have concluded that "it is not always clear that knowing about the presence of principal stratification effects will be of particular use" (VanderWeele, 2008). It is now becoming widely recognized that the natural direct and indirect effects formulated in Robins and Greenland (1992) and Pearl (2001) are of greater interest, both for the purposes of making treatment decisions and for the purposes of explanation and identifying causal mechanisms (Joffe, Small, and Hsu, 2007, Albert and Nelson, 2011, Mortensen, Diderichsen, Smith, and Andersen, 2009, Imai, Keele, and Yamamoto, 2010, Robins et al., 2007, 2009, Pearl, 2009a, Petersen, Sinisi, and van der Laan, 2006, Hafeman and Schwartz, 2009, Kaufman, 2010, Sjölander, 2009).

This limitation of PSDE stems not from the notion of "principal-strata" per se, which is merely a benign classification of units into homogeneously responding classes. Rather, the limitation stems from adhering to an orthodox philosophy which prohibits one from regarding a mediator as a *cause* unless it is manipulable. This prohibition prevents us from defining the direct effect as it is commonly used in decision making and scientific discourse – an effect transmitted with all mediating paths "deactivated" (Pearl, 2001, Avin, Shpitser, and Pearl, 2005, Albert and Nelson, 2011), and forces us to use statistical conditionalization (on strata) instead. Path deactivation requires counterfactual constructs in which the mediator acts as an antecedent, written Y_z, regardless of whether it is physically manipulable. After all, if we aim to uncover causal mechanisms, why should nature's pathways depend on whether we have the technology to manipulate one variable or another. The whole philosophy of extending the potential outcome analysis from

experimental to observational studies (Rubin, 1974) rests on substituting physical manipulations with reasonable assumptions about how treatment variables are naturally chosen by the so called "treatment assignment mechanism." Mediating variables are equally deserving of such substitution.

4.2 Principal surrogacy

A second area where a PS-restricted definition falls short of expectation is "surrogate endpoint" (Frangakis and Rubin, 2002). At its core, the problem concerns a randomized trial where one seeks "endpoint surrogate," namely, a variable Z that would allow good predictability of outcome for both treatment and control (Ellenberg and Hamilton, 1989). More precisely, "knowing the effect of treatment on the surrogate allows prediction of the effect of treatment on the more clinically relevant outcome" (Joffe and Green, 2009).

To meet this requirement, Frangakis and Rubin offered a definition called "principal surrogacy" which requires that *causal effects of X on Y may exist if and only if causal effects of X on Z exist* (see Joffe and Green (2009)). Again, their definition rests solely on what transpires in the study, where data is available on both the surrogate (Z) and the endpoint (Y), and does not require good predictions under the new conditions created when data on the surrogate alone are available.[2]

Pearl and Bareinboim (2011) present examples where a post-treatment variable Z passes the "principal surrogacy" test and yet it is useless as a predictor under the new conditions. Conversely, Z may be a perfect surrogate (i.e., a robust predictor of effects) and fail the "principal surrogacy" test. In short, a fundamental disconnect exists between the notion of "surrogate endpoint," which requires robustness against future interventions affecting Z and the class of definitions that the principal strata framework can articulate, given its resistance to conceptualizing such interventions.[3]

[2]Note that if conditions remain unaltered, the surrogacy problem is solved trivially by correlation. Therefore, no formal definition of surrogacy is complete without making change in conditions an integral part of the definition.

[3]The resistance, as in the case of mediation, arises from the prohibition on writing expressions containing the term Y_z, in which Z acts as an counterfactual antecedent.

5 Conclusions

The term "principal strata (PS)" is currently used to connote four different interpretations

1. PS as a partition of units by response type,
2. PS as an approximation to research questions concerning population averages (e.g., bounds and LATE analysis under noncompliance),
3. PS as a genuine focus of a research question (e.g., censorship by death),
4. PS as an intellectual restriction that confines its analysis to the assessment of strata-specific effects (see Addendum, p. 13).

My purpose in writing this note is to invite investigators using the PS framework to clarify, to their readers as well as to themselves, what role they envision this framework to play in their analysis, and how it captures the problem they truly care about.[4]

I have no reservation regarding interpretations 1-3, though a clear distinction between the three would be a healthy addition to the principal stratification literature. I have strong reservation though regarding the 4th; frameworks should serve, not alter research questions.

The popularization of Frangakis and Rubin "Principal Strata Framework" has had a net positive effect on causal research; it attracted researchers to the language of counterfactuals and familiarized them with its derivational power. At the same time, it has created the impression that conditioning on strata somehow bestows legitimacy on one's results and thus exonerates one from specifying research questions and carefully examining whether conditioning on strata properly answers those questions. It has further encouraged researchers to abandon policy-relevant questions in favor of a framework that restricts those questions from being asked, let alone answered.

I hope by bringing these observations up for discussion, a greater clarity will emerge as to the goals, tools and soundness of causal inference methodologies.

[4] I purposely refrain from discussing the issue of identification, namely, the assumptions needed for estimating principal strata effects in observational studies. Such issues tend to conflate the more fundamental problems of definition and interpretation which should take priority in methodological discussions. Joffe and Green (2009) compare identification conditions for both principal surrogacy and mediation-based surrogacy.

Addendum – Questions, Answers, Discussions

Question to Author

You state that you are concerned that individuals might be using principal stratification 'as an intellectual restriction that confines its analysis to the assessment of strata-specific effects.' Can you provide any examples in the literature where you felt that researchers might be using principal stratification in this manner?

Author's Reply

In retrospect, I feel that way about ALL PS papers that I have read, with the exception of ONE - Sjölander et al. (2010) on Hormone Replacement Therapy, which explicitly justifies why one would be interested in stratum-specific effects.

To substantiate the basis of my perception I cite the lead articles by Rubin (2004, 2005) where the PSDE is motivated and introduced. What if not an "intellectual restriction" could spawn a definition of "Direct Effect" that excludes from the analysis all units for whom X has both direct and indirect effect on Y?

According to Rubin (2005), the problem was originally posed by Fisher (1935) and Cochran (1957) in the context of agricultural experiments. Forgiving their mistaken solutions for a moment (they had no graphs for guidance), we find that both Fisher and Cochran were very clear about what their research questions were: To estimate the effect of X on Y after allowance is made for the variations in Y DIRECTLY DUE TO variations in Z itself.

The phrase "due to variations in Z" clearly identifies Z as a secondary CAUSE of Y, for which allowance needed to be made.

What, if not an "intellectual restriction" could compel us to change the research question from its original intent and replace it with another, in which Z is NOT treated as a secondary cause of Y, but only as a variable affected by X. This is, in essence, a restriction against writing down the counterfactual Y_{xz}.

Frangakis and Rubin (2002) state explicitly that they refrain from using "a priori counterfactuals," namely Y_z. In their words: "This [Robins and Greenland's] framework with its a priori counterfactual estimands, is not compatible with the studies we consider, which do not directly control the post-treatment variable (p. 23)." This resolution to avoid counterfactuals that are not directly controlled in the study is a harsh and unjustifiable "intellectual restriction," especially when the problem statement calls for an estimand involving Y_{xz}, and especially when refraining from considering Y_{xz} leads to unintended conclusions (e.g., that the direct

effect of a grandfather's income on a child education can only be defined in those families where that income did not affect the father's education.)

One should note that in the agricultural experiments considered by Fisher and Cochran, the post-treatment variables (e.g., plant-density or eelworms) were not controlled either, yet this did not prevent Fisher and Cochran from asking a reasonable, policy relevant question: To what extent do these post-treatment variables affect the outcome?

But your question highlights an important observation: most PS-authors do not view PS as a restriction. True; they actually view it as a liberating intellectual license; a license to assess quantities with a halo of legitimacy, without telling us why these quantities are the ones that the author cares about, or how relevant they are for policy questions or scientific understanding.

This is the power of the word "framework"; working within a "framework" assures an investigator that someone would surely find a reason to appreciate the quantity estimated, as long as it fits the mold of the "framework."

But how do we alert researchers to the possibility that they might be solving the wrong problem? One way is to present them with weird conclusions that emerge from their method, and ask them: "Do you really stand behind such conclusions?" This is what I tried to do with "PS direct effects," and "principal surrogacy." I hope the examples illuminate what the PS framework computes.

References

Albert, J. M. and S. Nelson (2011): "Generalized causal mediation analysis," *Biometrics*, DOI: 10.1111/j.1541–0420.2010.01547.x.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of causal effects using instrumental variables (with comments)," *Journal of the American Statistical Association*, 91, 444–472.

Avin, C., I. Shpitser, and J. Pearl (2005): "Identifiability of path-specific effects," in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, Edinburgh, UK: Morgan-Kaufmann Publishers, 357–363.

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

Balke, A. and J. Pearl (1994a): "Probabilistic evaluation of counterfactual queries," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume I, Menlo Park, CA: MIT Press, 230–237.

Balke, A. and J. Pearl (1994b): "Counterfactual probabilities: Computational methods, bounds, and applications," in R. L. de Mantaras and D. Poole, eds., *Uncertainty in Artificial Intelligence 10*, San Mateo, CA: Morgan Kaufmann, 46–54.

Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1172–1176.

Chickering, D. and J. Pearl (1997): "A clinician's tool for analyzing noncompliance," *Computing Science and Statistics*, 29, 424–431.

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

Ellenberg, S. and J. Hamilton (1989): "Surrogate endpoints in clinical trials: Cancer," *Statistics in Medicine*, 405–413.

Fisher, R. (1935): *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Frangakis, C. and D. Rubin (2002): "Principal stratification in causal inference," *Biometrics*, 1, 21–29.

Greenland, S. and J. Robins (1986): "Identifiability, exchangeability, and epidemiological confounding," *International Journal of Epidemiology*, 15, 413–419.

Greenland, S. and J. Robins (1988): "Conceptual problems in the definition and interpretation of attributable fractions," *American Journal of Epidemiology*, 128, 1185–1197.

Hafeman, D. and S. Schwartz (2009): "Opening the black box: A motivation for the assessment of mediation," *International Journal of Epidemiology*, 3, 838–845.

Heckerman, D. and R. Shachter (1995): "Decision-theoretic foundations for causal reasoning," *Journal of Artificial Intelligence Research*, 3, 405–430.

Heckman, J. (2005): "The scientific model of causality," *Sociological Methodology*, 35, 1–97.

Heckman, J. and E. Vytlacil (2001): "Policy-relevant treatment effects," *The American Economic Review*, 91, 107–111, papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Association.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, inference, and sensitivity analysis for causal mediation effects," *Statistical Science*, 25, 51–71.

Imbens, G. and J. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

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

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

Joffe, M., D. Small, and C.-Y. Hsu (2007): "Defining and estimating intervention effects for groups that will develop an auxiliary outcome," *Statistical Science*, 22, 74–97.

Kaufman, J. (2010): "Invited commentary: Decomposing with a lot of supposing," *American Journal of Epidemiology*, 172, 1349–1351.

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

Mortensen, L., F. Diderichsen, G. Smith, and A. Andersen (2009): "The social gradient in birthweight at term: Quantification of the mediating role of maternal smoking and body mass index," *Human Reproduction*, 24, 2629–2635.

Neyman, J. (1923): "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Statistical Science*, 5, 465–480.

Pearl, J. (1993): "Aspects of graphical models connected with causality," in *Proceedings of the 49th Session of the International Statistical Institute*, Tome LV, Book 1, Florence, Italy, 391–401.

Pearl, J. (2000): *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press, 2nd edition, 2009.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 411–420.

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

Pearl, J. (2009b): "Causal inference in statistics: An overview," *Statistics Surveys*, 3, 96–146, <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

Pearl, J. (2011): "The mediation formula: A guide to the assessment of causal pathways in nonlinear models," Technical Report R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles, to appear in C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), *Statistical Causality*. Forthcoming, 2011.

Pearl, J. and E. Bareinboim (2011): "Transportability across studies: A formal approach," Technical Report R-372, <http://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

Petersen, M., S. Sinisi, and M. van der Laan (2006): "Estimation of direct causal effects," *Epidemiology*, 17, 276–284.

Robins, J. (1986): "A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect," *Mathematical Modeling*, 7, 1393–1512.

Robins, J. and S. Greenland (1989a): "Estimability and estimation of excess and etiologic fractions," *Statistics in Medicine*, 8, 845–859.

Robins, J. and S. Greenland (1989b): "The probability of causation under a stochastic model for individual risk," *Biometrics*, 45, 1125–1138.

Robins, J. and S. Greenland (1992): "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, 3, 143–155.

Robins, J., T. Richardson, and P. Spirtes (2009): "On identification and inference for direct effects," Technical report, Harvard University, MA.

Robins, J., A. Rotnitzky, and S. Vansteelandt (2007): "Discussion of principal stratification designs to estimate input data missing due to death," *Biometrics*, 63, 650–654.

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

Rubin, D. (2004): "Direct and indirect causal effects via potential outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Rubin, D. (2005): "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 100, 322–331.

Sjölander, A. (2009): "Bounds on natural direct effects in the presence of confounded intermediate variables," *Statistics in Medicine*, 28, 558–571.

Sjölander, A., K. Humphreys, and S. Vansteelandt (2010): "A principal stratification approach to assess the differences in prognosis between cancers caused by hormone replacement therapy and by other factors," *The International Journal of Biostatistics*, 6, article 20.

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