The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 28

Principal Stratification — Uses and Limitations

Tyler J. VanderWeele, *Harvard University*

Recommended Citation:
VanderWeele, Tyler J. (2011) "Principal Stratification — Uses and Limitations," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 28.
DOI: 10.2202/1557-4679.1329
Available at: http://www.bepress.com/ijb/vol7/iss1/28

©2011 Berkeley Electronic Press. All rights reserved.

Principal Stratification — Uses and Limitations

Tyler J. VanderWeele

Abstract

Pearl (2011) asked for the causal inference community to clarify the role of the principal stratification framework in the analysis of causal effects. Here, I argue that the notion of principal stratification has shed light on problems of non-compliance, censoring-by-death, and the analysis of post-infection outcomes; that it may be of use in considering problems of surrogacy but further development is needed; that it is of some use in assessing "direct effects"; but that it is not the appropriate tool for assessing "mediation." There is nothing within the principal stratification framework that corresponds to a measure of an "indirect" or "mediated" effect.

KEYWORDS: causal inference, mediation, non-compliance, potential outcomes, principal stratification, surrogates

Author Notes: The author thanks Michael Elliott, Linda Valeri, a reviewer and the editor for helpful comments. The research was supported by NIH grant HD060696.

Introduction

Principal stratification has become an increasingly popular approach to thinking about certain classes of causal effects. The notion of principal stratification is most closely associated with a paper of Frangakis and Rubin (2002). Although the idea of principal stratification had clear antecedents (Robins, 1986; Angrist et al., 1996), Frangakis and Rubin (2002) proposed that this approach to thinking about causal effects be used to address a broad class of related problems concerning noncompliance, censoring-by-death, and surrogate outcomes. In his commentary, Pearl (2011) has asked the causal inference community to reflect on and clarify the specific value of the "principal stratification framework." Pearl offered a four-fold classification of what he sees as the uses and misuses of the principal stratification framework: (i) partitioning of response types, (ii) defining effects that approximate those of interest (he lists non-compliance as an example), (iii) defining effects that are of genuine interest (he lists censorship by death as an example), and (iv) imposing an intellectual restriction e.g. by not allowing for counterfactuals defined by interventions on an intermediate (he lists surrogate outcomes and mediation as examples). In what follows I will offer my own thoughts on this issue and briefly survey the range of applications which have employed principal stratification ideas. As will be seen below, I believe a more nuanced evaluation is merited. The utility of the framework varies considerably across applications. Moreover, I would likely put non-compliance in the third, rather than the second, of Pearl's four categories. And perhaps more importantly, I think a sharp distinction should be drawn between surrogate outcomes and mediation; formally, the two applications look somewhat similar but the questions that are asked are in fact quite different and I believe that principal stratification holds more promise for the former than the latter. I will return to these points below.

We first review the notion of principal stratification itself. Stated briefly, if X denotes some binary treatment and S some post-treatment variable and if we let S_x denote the potential outcome (Rubin, 1974) for each individual that we would have observed had X, possibly contrary to fact, been x, then a principal stratum is simply a subgroup of individuals homogenous in their joint potential outcomes (S_0, S_1). If S is also binary then we have four principal strata: $(S_0 = 0, S_1 = 0)$ sometimes called "never-takers", $(S_0 = 0, S_1 = 1)$ sometimes called "compliers", $(S_0 = 1, S_1 = 0)$ sometimes called "defiers", and $(S_0 = 1, S_1 = 1)$ sometimes called "always takers." Suppose that, in addition, we have some other outcome Y and let Y_x denote the potential outcome for each individual that we would have observed had X, possibly contrary to fact, been x. The overall causal effect for the population is then given by $E[Y_1 - Y_0]$. However, we could also consider the causal effect of X on Y conditional on the principal stratum i.e. $E[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$. This

is what Frangakis and Rubin (2002) call a "principal causal effect." Conditioning on the principal stratum has certain advantages over conditioning on the observed posttreatment variable S. In general, if we condition on a post-treatment variable S, we will induce bias in analysis (see Shpitser et al., 2010, for exceptions). However, the principal stratum, (S_0, S_1), that an individual belongs to is essentially viewed as a pretreatment characteristic of an individual and thus we can, in principle, stratify on it as we could any other pretreatment variable. The difficulty is that we do not know who is in which principal stratum. As is discussed below, this creates problems for identifying quantities like $E[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$ from observed data and it also makes it difficult to know who the individuals are to which such estimates apply.

Applications of the notion of principal stratification to develop methodology to address a wide range of problems now abound. This approach has been used in the context of non-compliance, censoring by death, and the related problem of post-infection outcomes, and more recently to issues of surrogate outcomes and "mediation." I will briefly discuss each of these with a special focus on the final topic, as this area of application has been somewhat more controversial.

Areas of Application

Non-compliance

The success of the principal stratification framework in addressing issues of non-compliance is now fairly clear. Numerous studies have fruitfully employed the idea (e.g. Angrist et al., 1996; Imbens and Rubin, 1997; Balke and Pearl, 1997; Cheng and Small, 2006; Cuzick et al., 2007; Little et al., 2009). The central insight was that in a randomized trial with non-compliance in which the group assigned the placebo had no access to treatment (i.e. no "defiers"), the instrumental variable (IV) estimator of the treatment effect would in fact only estimate the treatment effect for one principal stratum, that of the compliers. In other words, the IV estimate pertains only to the group for which treatment assignment actually changes treatment taken. In the notation above, if S denotes compliance status, then the IV estimator is consistent for $E[Y_1 - Y_0|S_0 = 0, S_1 = 1]$. This is a principal stratum effect and often now referred to as the local average treatment effect (LATE) or the complier average causal effect (CACE). Contrary to what is suggested by Pearl, the effect is not merely an approximation to the population average treatment effect, but is arguably of intrinsic interest as it is the effect of treatment for the only group that we can reasonably induce to take treatment (the group that would take treatment if they were assigned treatment). The use of principal stratification is by no means the only approach to handling issues of non-compliance within causal

DOI: 10.2202/1557-4679.1329

inference (Robins, 1994; Greevy et al., 2004; van der Laan et al., 2007) but that the ideas of principal stratification have been useful in this setting I think cannot be doubted. Many of the insights it has provided are applicable to the use of instrumental variables more generally (Angrist et al., 1996; Heckman and Vytlacil, 1999; Tan, 2006).

Censoring by death (and the analysis of post-infection outcomes)

A second application of the idea of principal stratification that has received considerable attention in the literature is the analysis of outcomes that have been censored or "truncated" by death. Consider a randomized trial comparing two drugs (X) and suppose we were interested in comparing quality of life outcomes (Y) at six months follow-up under these two drugs. If, however, some individuals die before the six month follow-up, their quality of life is not simply missing, it is undefined. We could attempt to simply compare outcomes amongst those who actually survived ($S = 1$). The trouble with this is that survival is a post-treatment variable and it may be affected by treatment; conditioning on it would essentially break randomization and could induce bias. Perhaps drug 1 was more likely to kill patients who are very sick at baseline than drug 2. A comparison of the quality of life outcomes between the two drugs amongst survivors would essentially be an unfair comparison because the sick patients are included in the average quality of life scores for drug 2 but they are not for drug 1 (because under drug 1, they die). An alternative comparison that would make sense in this setting is to compare the quality of life outcomes for the group that would have survived irrespective of which drug they were given. In the notation given above this is, $E[Y_1 - Y_0|S_0 = 1, S_1 = 1]$. This is, once again, a principal strata causal effect, sometimes referred to as the survivor average causal effect (SACE). In this context in which outcomes are effectively censored or truncated due to death, this is really the only comparison that is fair. The principal stratification approach is thus of considerable importance in addressing these questions as well and a number of papers have provided methods to try to assess this survivor average causal effect when outcomes are truncated due to death (Robins, 1986; Zhang and Rubin, 2003; Rubin, 2006; Frangakis et al., 2007; Imai, 2008; Egleston, 2009; Chiba and VanderWeele, 2011). A very closely related (essentially isomorphic) problem concerns the analysis of the effect of some treatment or vaccine on a post-infection outcome (e.g. HIV viral load) which is only defined for persons who are infected. In this context one would want to know the effect of treatment on the post-infection outcome within the principal stratum who would develop the infection irrespective of whether they were given treatment. Many of the important methodological contributions to analyzing these principal strata effects have been developed within this infectious disease context (Gilbert

et al., 2003; Hudgens et al., 2003; Hudgens and Halloran, 2006; Shepherd et al., 2006, 2007; Jemiai et al., 2007). With the analysis of problems concerning censoring by death or post-infection outcomes, the principal strata framework has once again given considerable insight.

It should be noted that whereas at least in some randomized trial non-compliance contexts, the principal stratum effect of interest is identified from the observed data, in more complex non-compliance settings the principal stratum effect is not identified and in the censoring-by-death setting the principal strata effect is again in general not identified. This lack of identification once again arises because we do not in general know which individuals are in which principal stratum. As a result most of the methodological approaches to the analysis of principal strata effect use either (i) bounds for the principal strata effects or (ii) sensitivity analysis techniques or (iii) take a Bayesian approach. With a sensitivity analysis approach one does not obtain a single point estimate but rather different estimates for each possible value of the sensitivity analysis parameters. With a Bayesian approach, because of lack of point identification, the length of posterior intervals will not shrink to 0 as the sample size increases to infinity and the posterior will depend on the prior even as the sample size tends to infinity (Richardson et al., 2011).

Surrogate Outcomes

Frangakis and Rubin (2002) had suggested the analysis of surrogate outcomes as an important application of the principal stratification framework. However, only more recently has there been further methodological development of the ideas they proposed (Gilbert and Hudgens, 2008; Wolfson and Gilbert, 2010; Li et al., 2010). The motivation for considering surrogate outcomes is that in certain randomized trials it may be very expensive or require considerable follow-up to assess the outcomes of interest. If measurements on a surrogate that is closely related to the outcome are easier to obtain then one might analyze the effect of the treatment on the surrogate rather than the effect of the treatment on the outcome. Frangakis and Rubin (2002) gave a definition of a "principal surrogate" that they argued was important in finding a good surrogate. With a binary outcome we would say that S is a principal surrogate for the effect of X on Y if for all s, $E[Y_1 - Y_0|S_0 = S_1 = s] = 0$ i.e. for the principal strata in which treatment does change the surrogate ($S_0 = s, S_1 = s$), the treatment should have no effect on the outcome. This property is sometimes referred to as one of "causal necessity."

Building on the work of Frangakis and Rubin (2002), Gilbert and Hudgens (2008) defined the "causal predictiveness surface" as $E[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$. This is simply the effect of treatment on the outcome in each of the principal strata. The idea of using this "causal predictiveness surface" to think about how the effect

DOI: 10.2202/1557-4679.1329

of the treatment on the surrogate relates to the effect of treatment on the outcome is theoretically appealing. Gilbert and Hudgens (2008) and Wolfson and Gilbert (2010) also extend this approach further to allow for the possibility that the outcome may occur for some individuals before the surrogate is measured. However, as noted above and as discussed by Gilbert and Hudgens (2008) and Wolfson and Gilbert (2010), identification of such principal strata effects is difficult. Gilbert and Hudgens (2008) and Wolfson and Gilbert (2010) show how some additional progress can be made if the surrogate takes the same value for all individuals under the control condition (i.e. $S_0 = c$), an assumption they refer to as a constant biomarker assumption. Li et al. (2010) take a Bayesian approach.

However, identification is not the only difficulty with a principal stratification approach to the analysis of surrogate outcomes. Chen et al. (2007) and Ju et al. (2010) note that a principal surrogate as defined by Frangakis and Rubin (2002) does not avert the so called "surrogate paradox." That is to say, a variable S may be a principal surrogate and the treatment may have a positive effect on the surrogate and the surrogate may have a positive effect on the outcome but it may still be the case the effect of the treatment on the outcome is negative! Chen et al. (2007) and Ju et al. (2010) discuss conditions beyond "principal surrogacy" that ensure that the surrogate paradox is avoided. The application of the principal stratification framework to the analysis of surrogates is theoretically appealing but, in my view, the jury is still out on how useful principal stratification ideas will in the end be in this context and a variety of other approaches to surrogate outcomes are also being pursued; see Joffe and Greene (2009) for a review.

Mediation

More recently there has been some interest in applying ideas of principal stratification to questions of mediation (Gallop et al., 2009; Elliott et al., 2010). Informally, we would generally say that the intermediate S mediates the effect of X on Y if X causes Y by changing S. As noted by Pearl (2011) in his commentary, Rubin (2004) considered the use of principal stratification framework for questions of mediation and essentially dismissed it. If we return to the original Frangakis and Rubin (2002) paper we see that they discuss two types of principal strata causal effect. They call an effect of treatment X on outcome Y within the principal strata in which X doesn't change S a "dissociative effect", i.e. $E[Y_1 - Y_0|S_0 = S_1 = s]$, and they call an effect of treatment X on outcome Y within the principal strata in which X does change S an "associative effect" i.e. $E[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$ when $s_0 \neq s_1$. Let us first examine the dissociative effect. The dissociative effect is the effect of treatment on outcome within the principal strata in which treatment doesn't change the intermediate. If, within these principal strata, the treat-

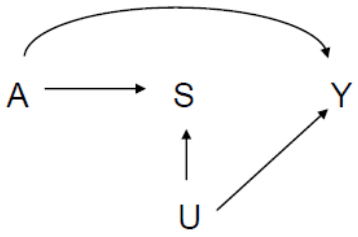

Figure 1: Example in which there may be "associative" principal strata effects without a mediated effect.

ment doesn't change the intermediate, then its effect cannot operate through the intermediate; it must be "direct." We might thus also call the dissociative effect a "principal strata direct effect" (PSDE). For a binary intermediate we would have two principal strata direct effects, $PSDE(0) = E[Y_1 - Y_0|S_0 = S_1 = 0]$ and $PSDE(1) = E[Y_1 - Y_0|S_0 = S_1 = 1]$. If one of these were non-zero then we would conclude that there were some pathway from treatment to outcome not through the intermediate (VanderWeele, 2008). This much seems relatively unproblematic. We still may have the same difficulties with the identification of these principal strata effects from observed data but sensitivity analysis techniques can be used to address principal strata direct effects (Sjölander et al., 2009; VanderWeele, 2010a; Chiba, 2010) and Bayesian methods have also been employed (Gallop et al., 2009; Elliott et al., 2010).

The principal stratification framework has a coherent notion of a direct effect. One might then be tempted to take the associative effects (the effect of treatment on the outcome when treatment does change the intermediate), i.e. $E[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$ with $s_0 \neq s_1$, as a measure of an indirect effect. As we will see, however, this does not work. The problem is that these "associative effects" are the overall effect of treatment within the relevant principal strata. Whatever we might call the "direct effect" and the "indirect effect", the associative effect will pick up both of them. We might in fact have very large associative effects with no "indirect effects" whatsoever. Consider the setting depicted in Figure 1 and suppose that S serves as a very good proxy for Y but has no effect on Y whatsoever. In this case, none of the effect would be mediated by S because S has no effect on Y; however, the associative effect $E[Y_1 - Y_0|S_0 = 0, S_1 = 1]$ might be large because whenever treatment changes S from 0 to 1, treatment is likely to have an effect on Y as well since S serves as a good proxy for Y.

DOI: 10.2202/1557-4679.1329

Further insight into why associative effects cannot be used as measures of indirect effects can be gained by comparing them to what has been defined in the literature as natural direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001). These effects consider hypothetical interventions on the treatment X and the intermediate S so that it is possible to define the potential outcome Y_{xs}, the value of Y for each individual that would be observed if X were set to x and S were set to s. The average natural direct effect may then be defined as $E[Y_{1S_0} - Y_{0S_0}]$ i.e. the comparison of the outcome with versus without treatment in both scenarios setting the intermediate to the level it would have been without treatment. The average natural indirect effect can be defined as $E[Y_{1S_1} - Y_{1S_0}]$ i.e. a comparison of the outcomes under treatment when setting the intermediate to the level it would have been with versus without treatment. The natural indirect effect will thus only be non-zero if for some individual treatment changes the value of the intermediate and that change in the value of the intermediate changes the value of the outcome. With a non-zero natural indirect effect we have mediation: the treatment changes the outcome by changing the intermediate. These effects are also of interest because a total effect can be decomposed into a natural direct and indirect effect, $E[Y_1 - Y_0] = E[Y_{1S_1} - Y_{1S_0}] + E[Y_{1S_0} - Y_{0S_0}]$, even in models with interactions and non-linearities (Pearl, 2001). Methods to estimate these natural direct and indirect effects are now available (VanderWeele and Vansteelandt, 2009, 2010; Imai et al., 2010) as are sensitivity analysis techniques to address violations in the no-unmeasured-confounding assumptions required to identify these effects (VanderWeele, 2010a; Imai et al., 2010).

Let us now return to considering the associative effect. We can express the relationship between the associative effect and natural direct and indirect effects as follows:

$$\begin{aligned} E[Y_1 - Y_0|S_0 = 0, S_1 = 1] &= E[Y_{1S_1} - Y_{0S_0}|S_0 = 0, S_1 = 1] \\ &= E[(Y_{1S_1} - Y_{1S_0}) + (Y_{1S_0} - Y_{0S_0})|S_0 = 0, S_1 = 1] \\ &= E[(Y_{11} - Y_{10}) + (Y_{10} - Y_{00})|S_0 = 0, S_1 = 1] \end{aligned}$$

From the second line we see that the associative effect is the sum of the natural direct and indirect effects within the principal strata ($S_0 = 0, S_1 = 1$). From the final line we see that even if there were no effect of S on Y so that $(Y_{11} - Y_{10}) = 0$, we could still have a substantial associative effect if the direct of X, i.e. $Y_{10} - Y_{00}$, were non-zero. Again, associative effects do not correspond to indirect effects and they cannot be used to assess mediation. There is nothing within the principal stratification framework that corresponds to a measure of an indirect effect. For this reason, I think it would be best not to use the term "mediation analysis" (e.g. Gallop et al., 2009; Elliott et al., 2010) when estimating the principal stratification "associative" and "dissociative" effects in the context of intermediate S. These effects may be of

interest in their own right but one is not assessing mediation in these cases; one is not assessing whether treatment affects the outcome through the intermediate. As discussed above, one can use principal strata direct effects to assess whether there is a pathway from treatment to the outcome other than through the intermediate - and so "direct effects analysis" may still be an appropriate description; but one cannot assess, using principal strata effects, whether there is a pathway through the intermediate itself[1]. The potential outcomes framework and causal inference literature has clarified tremendously the causal effects that may be in view and that investigators might consider; I believe the causal inference literature is obscured when the label "mediation" is used when in fact mediation is not being assessed[2].

As I see it, in contexts in which pathways are of interest, the only advantage in considering the principal stratification framework rather than natural direct and indirect effects is that the latter requires conceiving of interventions on the intermediate and the former does not. The principal stratification framework only considers counterfactuals of the form Y_x and S_x; it does not require counterfactuals of the form Y_{xs}. In some settings, this will be an important advantage; the intermediate

[1] In contexts in which pathways are of interest, some papers (Gallop et al., 2009; Chiba, 2010; Elliott et al., 2010) have taken the usual principal strata labels "never-takers, compliers, defiers, always-takers" and adapted these to "never-mediators, compliant-mediators, defiant-mediators, always-mediators." For the same reasons as given above, I think, although the concepts are reasonable, the language should be changed. The use of the term "always-mediators" for the principal stratum $S_0 = 1, S_1 = 1$ is misleading in that it suggests that for this subgroup the effect of treatment is always mediated by the intermediate. This is not the case. In fact, in the subgroup constituted by this principal strata, the effect is never mediated for any individual since the value of the intermediate is the same ($S_0 = 1, S_1 = 1$) with or without treatment. I would suggest returning to the conventional: "never-takers, compliers, defiers, always-takers" or if adaption to the pathway setting is thought desirable, then perhaps the use of "never-intermediate, compliant-intermediate, defiant-intermediate, always-intermediate" would be appropriate. In any case, it seems strongly preferable to reserve the words "mediatior" and "mediation" for settings in which mediation is in fact in view.

[2] Elliott et al. (2010) recently proposed a measure of the "proportion mediated" within the principal stratification framework. They consider what the associative effect would be if it were entirely "unmediated" and also what it would be if there were no direct effect (i.e. entirely "mediated"). They define the "proportion mediated" as the ratio between: (i) the associative effect minus what it would be if it were entirely "unmediated" and (ii) what the associative effect would be if there were no direct effect minus what it would be if it were entirely "unmediated." However, the definition of "unmediated" (entirely direct) given by Elliott et al. (2010) is that the associative effect and the disassociative effect are equal. This is not what would generally be understood as entirely direct. In Figure 1, there is no effect of S on Y; there is no mediation; the effect is entirely direct. However, the associative and dissociative effects may differ. The effect of X on Y is entirely direct but this effect of X on Y may differ comparing those for whom X does or does not change S. The measure of Eilliott et al. (2010) may be of interest in its own right, but it is not assessing the "proportion mediated" through the variable S. Again, without allowing for counterfactuals defined by interventions on the intermediate S, it is not possible to formally define an effect corresponding to X changing S and that change in S changing Y i.e. to mediation.

DOI: 10.2202/1557-4679.1329

is not always under the control of the investigator; even hypothetical interventions are not always conceivable. However, whether one is willing to entertain counterfactuals of the form Y_{xs} that are needed for the definition of natural direct and indirect effects will depend on the context. In some cases such counterfactuals are quite reasonable. Calcium intake (X) might increase the risk of prostate cancer (Y) by decreasing vitamin D (S). Here hypothetical interventions on the intermediate (vitamin D) are as conceivable as those on the treatment itself (calcium) e.g. both could be changed by a supplement. In other cases, if the intermediate were beliefs say (as is sometimes the case in psychology experiments), hypothetical interventions on the intermediate are much less conceivable. There is no strict criterion for when we are or are not willing to entertain hypothetical interventions that give rise to counterfactuals and I think it is best to view the plausibility of hypothetical interventions and counterfactuals as a spectrum - some are more reasonable to entertain than others. Nor is this issue restricted to settings in which an intermediate is in view. Even in the analysis of overall causal effects within an observational study when we use potential outcomes notation we are entertaining counterfactual quantities and hypothetical interventions. These counterfactuals are ill-defined to the extent we have failed to specify the hypothetical intervention in view; they are arguably always partially ill-defined (Robins and Greenland, 2000). In settings in which multiple versions of treatment exist so that there are multiple ways of intervening on the exposure, ordinary estimates of causal effects can sometimes be interpreted as the effects of particular interventions in which the version of treatment is randomized (Hernán and VanderWeele, 2011). Future research extending this approach to settings in which there are multiple ways of changing the intermediate (i.e. multiple version of the mediator) may be promising in interpreting the estimates from methods for natural direct and indirect effects in settings in which hypothetical interventions on the intermediate are difficult to conceive.

Conclusion

I have offered here a very brief survey of what, as I see it, are the uses and limitations of the principal stratification framework. I believe that principal stratification has shed considerable light on non-compliance, the analysis of instrumental variables, settings in which the outcome is censored by death, and settings in which a post-infection outcome is in view. I believe that the framework holds promise for the analysis of surrogate outcomes but it remains to be seen how useful this will be in practice. I believe the framework can also be of some use for thinking about whether there may be a pathway from treatment to outcome other than through a particular intermediate. However, it is not of use in mediation analysis itself, conceived of as assessing whether there is an effect of the treatment on the outcome that

operates through the intermediate. The framework cannot be used to assess indirect effects. I have focused here on some of the more common applications of principal stratification ideas. The framework may, however, shed light on other areas as well. Ideas from principal stratification can be used in settings with interference between units to formalize the notion of an "infectiousness effect" in infectious disease epidemiology (VanderWeele and Tchetgen Tchetgen, 2011). Work on causal interactions (VanderWeele, 2010b) could essentially be construed as detecting certain principal strata defined by two cross-classified exposures. Further applications may emerge as the ideas are applied to other areas. While Pearl's four-fold assessment of the principal stratification framework is thought-provoking, the variety of potential applications of principal stratification ideas does not always neatly fit into one of his four categories.

In this survey I have tried to highlight the potential uses of principal stratification. However, mentioning at least a couple of further caveats is in order. First, as noted above, even after we have used statistical and sensitivity analysis techniques to assess principal strata effects, the principal strata themselves in general remain unidentified. We do not know who is in which stratum and this makes the framework somewhat more difficult to use in informing policy questions. Second, in describing the various applications, I considered the setting of a binary intermediate. Although the framework is not, in principle, restricted to the setting of a binary intermediate (Frangakis and Rubin, 2002), the analysis becomes much less tractable with intermediates with more categories. If the intermediate is continuous there may be no more than one individual in any of the principal strata. Dichotomization of a continuous or ordinal intermediate, as is often done, can give rise to misleading inferences (Robins et al., 2007). Identification difficulties are also compounded when the intermediate has more than two levels. The notions of principal stratification in practice thus seem most useful in settings in which the intermediate is in fact binary: all-or-nothing compliance, censoring-by-death versus survival, and the analysis of post-infection outcomes all fit the bill quite well.

DOI: 10.2202/1557-4679.1329

References

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

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

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

Cheng, J. and Small, D.S. (2006). Bounds on causal effects in three-arm trials with non-compliance. Journal of the Royal Statistical Society, Series B, 68:815-836.

Chiba, Y. (2010). Bias analysis for the principal stratum direct effect in the presence of confounded intermediate variables. Journal of Biometrics and Biostatistics, 1:101.

Chiba, Y. and VanderWeele, T.J. (2011). A simple sensitivity analysis technique for principal strata effects when the outcome has been truncated due to death. American Journal of Epidemiology, in press.

Cuzick, J., Sasieni, P., Myles, J., and Tyrer, J. (2007). Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. Journal of the Royal Statistical Society, Series B, 69:565–588.

Egleston, B., Sharfstein, D.O, MacKenzie, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. Biometrics, 65:497-504.

Elliott, M.R., Raghunathan, T.E. and Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes, Biostatistics, 11:353-372.

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

Frangakis, C.E., Rubin, D.B., An, M.W. and MacKenzie, E. (2007). Principal stratification designs to estimate input data missing due to death (with discussion). Biometrics, 63:641-662.

Gallop, R., Small, D.S., Lin, J.Y., Elliott, M.R., Joffe, M. Ten Have, T.R. (2009). Mediation analysis with principal stratification. Statistics in Medicine, 28:1108-1130.

Gilbert, P. B., Bosch, R., 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.

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

Greevy, R., Silber, J. H., Cnaan, A., Rosenbaum, P. R. (2004). Randomization inference with imperfect compliance in the ACE-Inhibitor After Anthracycline Randomized Trial. Journal of the American Statistical Association, 99:7-15.

Heckman, J.J., and Vytlacil, E.J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences. 96:4730–4734.

Hernán, M.A. and VanderWeele, T.J. (2011). Compound treatments and transportability of causal inference. Epidemiology, in press.

Hudgens, M.G. and Halloran, M.E. (2006). Causal vaccine effects on binary post-infection outcomes. Journal of the American Statistical Association, 101:51-64.

Hudgens, M.G., Hoering, A., and Self, S.G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. Statistics in Medicine, 22:2281-2298.

Imai, K. (2008). Sharp bounds on causal effects in randomized experiments with "truncation-by-death." Statistics and Probability Letters, 78:144-149.

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

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

Jemiai, Y., Rotnitzsky, A., Shepherd, B.E. and Gilbert, P.B. (2007). Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. Journal of the Royal Statistical Society, Series B, 69:879-902.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman Data. Journal of the American Statistical Association, 103:101-111.

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

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

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

Little, R.J., Long, Q. and Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. Biometrics, 65:640-649.

DOI: 10.2202/1557-4679.1329

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. San Francisco: Morgan Kaufmann, 411-420.

Pearl, J. (2011). Principal stratification - a goal or a tool? International Journal of Biostatistics: 7(1), Article 20.

Richardson, T.S., Evans, R.J. and Robins, J.M. (2011). Transparent parametrizations of models for potential outcomes. In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.), Bayesian Statistics 9, 569–610. Oxford: Oxford University Press. (ISBN 9780199694587)

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

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics, 23:2379-2412.

Robins, J.M. and Greenland S. (2000). Comment on "Causal Inference Without Counterfactuals" by A.P. Dawid. Journal of the American Statistical Association - Theory and Methods, 95:477-482.

Robins, J.M., Rotnitzky, A., and Vansteelandt, S. (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 non-randomized studies, Journal of Educational Psychology, 66:688-701.

Rubin, D.B. (2004). Direct and indirect effects via potential outcomes. Scandinavian Journal of Statistics, 31:161-170.

Rubin, D.B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death (with discussion). Statistical Science, 21:299-321.

Shepherd, B.E., Gilbert, P.B., Jemiai, Y. and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. Biometrics 62:332-342.

Shepherd, B.E., Gilbert, P.B. and Lumley, T. (2007). Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. Journal of the American Statistical Association, 102:573-582.

Shpitser, I., VanderWeele, T.J. and Robins, J.M. (2010). On the validity of covariate adjustment for estimating causal effects. Proceedings of the 26th Conference on Uncertainty and Artificial Intelligence, 527-536, AUAI Press in Corvallis, WA.

Sjölander, A., Humphreys, K., Vansteelandt, S., Bellocco, R. and Palmgren, J. (2009). Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. Biometrics, 65:514-520.

van der Laan, M.J., Hubbard, A., Jewell, N.P. (2007), Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome. Journal of the Royal Statistical Society, Series B, 69:443-482.

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

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

VanderWeele, T.J. (2010b). Epistatic interactions. Statistical Applications in Genetics and Molecular Biology, 9, Article 1:1-22.

VanderWeele, T.J. and Tchetgen Tchetgen, E.J. (2011). Bounding the infectiousness effect in vaccine trials. Epidemiology, in press.

VanderWeele T.J., Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface - Special Issue on Mental Health and Social Behavioral Science, 2:457-468.

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

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

Zhang, J.L. and Rubin, D.B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death." Journal of Educational and Behavioral Statistics, 28:353-368.

DOI: 10.2202/1557-4679.1329