The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 8

A Refreshing Account of Principal Stratification

Fabrizia Mealli, *University of Florence*
Alessandra Mattei, *University of Florence*

Recommended Citation:
Mealli, Fabrizia and Mattei, Alessandra (2012) "A Refreshing Account of Principal Stratification," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 8.
DOI: 10.1515/1557-4679.1380

A Refreshing Account of Principal Stratification

Fabrizia Mealli and Alessandra Mattei

Abstract

Pearl (2011) invites researchers to contribute to a discussion on the logic and utility of principal stratification in causal inference, raising some thought-provoking questions. In our commentary, we discuss the role of principal stratification in causal inference, describing why we view the principal stratification framework as useful for addressing causal inference problems where causal estimands are defined in terms of intermediate outcomes. We focus on mediation analysis and principal stratification analysis, showing that they generally involve different causal estimands and answer different questions. We argue that even when principal stratification may not answer the causal questions of primary interest, it can be a preliminary analysis of the data to assess the plausibility of identifying assumptions. We also discuss the use of principal stratification to address issues of surrogate outcomes. Our discussion stresses that a principal stratification analysis should account for all the principal strata and evaluate the distributions of potential outcomes in each of the principal strata. To this end, we view a Bayesian analysis particularly suited for drawing inference on principal strata membership and principal strata effects.

KEYWORDS: Causal Inference, Mediation Analysis, Principal Stratificafition, Principal Causal Effects, Surrogacy

Author Notes: We would like to thank a reviewer and the editor for comments. We would also like to thank Guido Imbens and Donald Rubin for insightful discussions, although they bear no responsibility for the views expressed in the paper.

1 Introduction

Our discussion on the role of principal stratification in causal inference queues up those of other authors (Baker et al., 2011, Egleston, 2011, Gilbert et al., 2011, Joffe, 2011, Prentice, 2011, Sjölander, 2011, VanderWeele, 2011), so we can benefit from their comments, by focussing on some issues, which we believe need to be further clarified, but neglecting some other aspects, which have already been discussed.

A principal stratification with respect to a post-treatment variable is a partition of units into latent classes defined by the joint potential values of that post-treatment variable under each of the treatments being compared. From this standpoint, some previous works (e.g., Robins (1986, 1998), Robins and Greenland (1989a,b, 1994), Baker and Lindeman (1994), Imbens and Angrist (1994), Angrist et al. (1996), Imbens and Rubin (1997a,b), Rubin (1998), Frangakis and Rubin (1999), Hirano et al. (2000), Heckman and Vytlacil (2001)), which temporally precede the formalization of the concept of principal stratification by Frangakis and Rubin (2002), can be viewed as examples of principal stratification. By definition, principal strata are not affected by treatment assignment, therefore a principal stratification can be used as any classification of units, to define meaningful causal estimands conditional on principal strata, to discover treatment effect heterogeneities, to state identifying assumptions as behavioral assumptions on the principal strata.

According to this definition, a principal stratification is a partition of units, therefore we essentially agree with the first of the four different interpretations of the term principal stratification proposed by Pearl (2011), where a principal stratification is defined 'as a partition of units by response type.' The concept of response type has been used more generally than the concept of principal stratification, as clarified in the next Sections. We view Pearl's interpretations 2 and 3 as possible applications of the framework of principal stratification rather than as an interpretation of the term principal stratification. Pearl's interpretation 4 highlights that a principal stratification analysis focuses on principal strata effects, but we do not view this approach as an 'intellectual restriction' (see the discussion in Sections 4 and 5).

A principal stratification is the core of the principal stratification framework, which can be defined as a general approach to formalize and address causal inference problems where causal estimands are defined in terms of intermediate outcomes, which are on the *causal pathway* between the treatment and the primary endpoint.

A principal stratification analysis focuses on the analysis of principal strata and principal strata effects: once a principal stratification has been defined, the objective is to draw inference on principal strata membership and principal strata ef-

fects (comparisons of treatments conditional on principal strata), or more generally on the distribution of potential outcomes within strata.

The role of principal stratification in drawing causal inference in the presence of noncompliance with treatment assignment and truncation by death has already been widely discussed, and we agree with comments by other authors, who recognize the importance of principal stratification in addressing these issues. Regarding the complier average causal effect (CACE), we would only like to further stress that CACE may be of intrinsic interest *per se* and cannot be generally interpreted as an approximation to the population average causal effect (ACE) unless additional assumptions are introduced, and we are in fact not aware of any paper, where CACE is interpreted as an approximation to the population ACE (see also the recent discussion in the economic literature, e.g., Imbens (2010)). We actually view it as a benefit of principal stratification, showing for which units the effect of treatment assignment can be interpreted as the effect of treatment received, and providing explicit assumptions for identification. We will briefly return on this issue in Section 4 because the role of the different assumptions is best seen (and generally accepted) in the noncompliance setting, while the same issues are still debated in mediation related analysis.

The role of principal stratification in dealing with issues of mediation and surrogate endpoints is still controversial and we believe that some further discussion on these topics may be useful.

2 The Potential Outcomes Approach

Consider a random sample of units, indexed by $i = 1, \ldots, n$. Following Pearl's notation as much as possible (Pearl, 2011), let X denote a binary treatment variable. Each unit i can be potentially assigned either a standard treatment ($X = 0$) or a new treatment ($X = 1$). The objective is to assess the causal effect of the $X = 1$ versus the $X = 0$ treatment on an outcome Y. Let Z stand for an additional post-treatment variable, which is on the *causal pathway* between the treatment and the main endpoint, Y: Z represents the actual treatment received in randomized experiments suffering from treatment noncompliance; the missing indicator in studies with missing outcome values; the survival indicator when outcomes are censored by death. When focus is on disentangling direct and indirect effects, Z stands for an intermediate variable, which may mediate the effect of the treatment on the primary outcome, in some way *channeling* part of the treatment effect. In problems of surrogate endpoints, the intermediate variable Z is a potential surrogate, that is, a variable that could be used in place of the primary endpoint, when measurement of the primary outcome is too expensive, inconvenient or unfeasible in a reasonable

time spell. Each of these situations, or combinations of them, can be viewed as special applications of principal stratification, which is a general framework that can be used to represent and tackle intrinsically different problems. While some principal stratification analyses may be mathematically equivalent, they can differ on fundamental issues of study design, on interpretation, on the specific (union of) principal strata of interest, and on the potential identifying structural and modeling assumptions.

We now briefly introduce the potential outcomes approach to causal inference, sometimes also referred to as the 'Rubin Causal Model' (RCM, Holland (1986)). For a more comprehensive account of the approach, readers can refer to, e.g., Rubin (1974, 2005). The RCM has two essential parts, where the concepts of potential outcomes and assignment mechanism have a leading role, and a third optional part, which involves extensions to include model-based inference. Therefore, although the concept of potential outcomes is basic in the RCM approach, it is just one of the several elements and concepts, which this framework consists of.

Let $Y_i(x)$ and $Z_i(x)$ denote the potential outcomes of Y and Z, respectively, if unit i were assigned treatment $X = x$, $x = 0, 1$[1]. The observed data include the assigned treatment level, X_i, and the observed values of the outcomes, which can be defined as $Z_i^{obs} = X_i Z_i(1) + (1 - X_i) Z_i(0)$ and $Y_i^{obs} = X_i Y_i(1) + (1 - X_i) Y_i(0)$. As a result, only one potential outcome can be observed for each unit once a treatment is applied. Therefore, in order to draw valid causal inferences, it is crucial to posit an assignment mechanism, which is the process that determines which units receives which treatment, hence which potential outcomes are observed.

The assignment mechanism is a well-defined mathematical concept, which describes, as a function of all observed covariates, and of all potential outcomes under study, the probability of any vector of assignments. Covariates are pre-treatment variables, which are not affected by the treatment. Formally, let $\mathbf{X}$ be the n-vector of treatment assignments, with ith element X_i, and let $\mathbf{Y}(x)$, and $\mathbf{Z}(x)$, $x = 0, 1$, denote the n-dimensional vectors of the potential outcomes with ith elements equal to $Y_i(x)$ and $Z_i(x)$, respectively. Finally, let $\mathbf{C}$ denote the $n \times K$ matrix of observed covariates, with ith row equal to $\mathbf{C}_i$. The assignment mechanism is defined as the conditional probability of each vector of assignments given the observed covariates and potential outcomes: $Pr(\mathbf{X} \mid \mathbf{C}, \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Z}(0), \mathbf{Z}(1))$.

The third optional part of the potential outcomes perspective involves a distribution on the quantities being conditioned on in the assignment mechanism, including the potential outcomes, thereby allowing model-based causal inference

[1]Note that the notation is adequate if we assume that the potential values $Z_i(x)$ and $Y_i(x)$ for individual i do not depend on the treatments received by other individuals and that there are no hidden versions of the treatment (Stable Unit Treatment Value Assumption: SUTVA; Rubin (1980)).

(Rubin, 2005). In model-based causal inference the potential outcomes are viewed as random variables, and any function of them, including causal estimands of interest, are also random variables. Here, we take this more general view, assuming that potential outcomes, on top of $\mathbf{X}$, are random variables. Therefore, we view the quantities associated with each sampled unit, X_i, $Y_i(0), Y_i(1), Z_i(0), Z_i(1)$, and the observed covariates, $\mathbf{C}_i$, as a joint draw from the population distribution, and we consider the observed values of these quantities to be realizations of random variables and the unobserved values to be unobserved random variables.

3 Principal Strata and Principal Causal Effects

The framework of principal stratification uses the potential outcomes of post-treatment intermediate variables to classify units into strata that are not affected by treatment assignment and therefore can be used just as any pre-treatment covariate. Formally, the (basic) principal stratification with respect to the post-treatment variable Z (with support $\mathscr{Z}$) is the partition of subjects into sets such that all subjects in the same set have the same vector $(Z_i(0); Z_i(1))$. A principal causal effect is a comparison between the potential outcomes $Y_i(0)$ and $Y_i(1)$ within a particular stratum (or union of principal strata). Henceforth we only focus for simplicity on average causal effects, therefore a principal causal effect (*PCE*) is formally defined as $PCE(z_0, z_1) = E[Y_i(1) - Y_i(0) \mid Z_i(0) = z_0, Z_i(1) = z_1]$.

According to this definition, a principal stratification can be interpreted 'as a partition of units by response type', where a 'response type' is defined by the joint potential values of the intermediate post-treatment outcome. In the presence of a post-treatment intermediate variable, the term 'response type' is often used to refer to a cross-classification of units defined by the joint potential values of both the intermediate outcome and the primary outcome (e.g., Robins and Greenland (1992), Pearl (1993), Balke and Pearl (1994a,b), Heckerman and Shachter (1995), Balke and Pearl (1997), Chickering and Pearl (1997), Cai et al. (2008)). This partition of units is generally finer than the one that would be used in the principal stratification framework, where units are usually classified into strata defined by the joint potential values of the intermediate post-treatment outcome only, without involving the potential values of the primary outcome. We argue that both these cross-classifications of units are conceptually well-sound and may be reasonable, although one or the other partition can be more attractive in some settings.

For instance, when focus is on causal estimands that depend on the joint distribution of potential outcomes, the partition of units defined by the joint potential values of both the intermediate outcome and the primary outcome might be useful. On the other hand, when causal estimands of interest are defined as

comparisons of marginal distributions of potential outcomes within principal strata, cross-classifying units into strata defined by the joint potential values of both the intermediate and the primary post-treatment outcome can be cumbersome and might introduce severe complications and practical difficulties, although it is conceptually straightforward. Limiting the analysis to principal causal effects, even if it may be viewed as a restriction by someone, has led to extended methods to address causal questions involving categorical or continuous intermediate variables (e.g., Jin and Rubin (2008), Schwartz et al. (2011)); multivariate intermediate variables (e.g. Mattei and Mealli (2007), Frumento et al. (2012)); intermediate variables with categorical, continuous or multivariate treatment variables.

A cross-classification of units defined by the joint potential values of both the intermediate outcome and the primary outcome sometimes involves potential outcomes of the form $Y_i(x,z)$, which would be the value of Y if, possibly contrary to fact, X were set to x and Z were set to z. The potential outcomes $Y_i(x, Z_i(x) = z)$, which are generally avoided in a principal stratification analysis, are a priori counterfactuals for units who exhibit a value of the intermediate outcome Z under treatment x not equal to z, because in one specific experiment, they can be never observed for such type of units (Rubin, 2004)[2]. The role of potential outcomes of the form $Y_i(x,z)$ in the potential outcomes approach is controversial and some discussion on it can help understanding.

In the RCM, an intermediate variable is initially viewed as a post-treatment variable, which can be potentially affected by treatment assignment; therefore a principal stratification with respect to that variable is defined and principal causal effects are generally the causal estimands of initial interest. Given a principal stratification with respect to Z, we can still hypothesize that there exist potential outcomes of the form $Y_i(x,z)$, but some of these potential outcomes are 'a priori counterfactuals': for units with $Z_i(x) \neq z$, $Y_i(x,z)$ is not in the data, and in a specific experiment or study it is 'a priori counterfactual' because it cannot be observed, not even on units of the same type assigned the opposite treatment. Therefore, although we can hypothesize the existence of 'a priori counterfactuals', they are conceptually different from potential outcomes of the form $Y_i(x)$, which are observable potential outcomes: we observe $Y_i(1)$ for some units under treatment and $Y_i(0)$ for some other units under control. For instance, and to repeat the concept again, when both the treatment, X, and the intermediate variable, Z, are binary, four potential outcomes could be defined for each unit, i: $Y_i(0,0)$, $Y_i(1,0)$, $Y_i(0,1)$ and $Y_i(1,1)$. However,

[2]This notation includes potential outcomes of the form $Y_i(x, Z_i(1-x))$, which would be the value of Y if X were set to x and Z were set to the value that it would attain under treatment $X = 1-x$. Without an assignment mechanism on the intermediate variable Z, it is not clear to us if this notation is fully consistent with SUTVA.

two out of four possible potential outcomes, $Y_i(0,z_0)$ and $Y_i(1,z_1)$, with $z_0 \neq Z_i(0)$ and $z_1 \neq Z_i(1)$, become 'a priori' counterfactuals, because they cannot be observed for any subset of units in a specific study or experiment.

Causal estimands involving potential outcomes of the form $Y_i(x,z)$ require that both the treatment and the post-treatment variable Z can somehow be manipulated. When an hypothetical intervention on the intermediate variable is conceivable and Z can be regarded as an additional treatment, there are no 'a priori counterfactuals'. In such a case potential outcomes have to be defined as a function of a multivariate treatment variable, (X,Z), and a compound assignment mechanism should be specified. In other words, all values $Y_i(x,z)$, $x = 0,1$, $z \in \mathscr{Z}$, are potentially observable, although only one will ultimately be realized and therefore possibly observed: the potential outcome corresponding to the treatment actually assigned. For instance, as before, if both treatments are binary, there are four potential outcomes for each unit i: $Y_i(0,0)$, $Y_i(1,0)$, $Y_i(0,1)$, $Y_i(1,1)$, and none would be 'a priori counterfactual.' This distinction will be further discussed in Section 4.

Pearl (2011) suggests that a principal stratification analysis has several advantages, 'stemming primarily from the parsimony achieved by' characterizing units by their principal stratum membership, rather than their baseline features (denoted by u). Although the parsimony of principal strata classification is undoubtedly useful, we argue that another advantage of principal stratification is its role in dealing with nonignorability issues.

Assume for simplicity that the treatment assignment mechanism is unconfounded, a special case of ignorable treatment assignment mechanism (Rubin, 1978). Unconfoundedness of the treatment, which usually holds by design in randomized experiments, amounts to assuming that within cells defined by the values of observed pre-treatment variables, C_i, the treatment is assigned independently of the relevant post-treatment variables: $Y_i(0), Y_i(1), Z_i(0), Z_i(1) \perp X_i \mid C_i$, for all i. However, unconfoundedness of the treatment does not in general hold conditional on Z_i^{obs}. Therefore net comparisons of treated and control units conditional on Z_i^{obs} generally lack causal interpretation, because these two groups of units are obtained by conditioning on different variables ($Z_i(0)$ and $Z_i(1)$, respectively), or, in other words, on different subsets of the baseline features, u, for units under treatment and under control. On the other hand, unconfoundedness of the treatment implies that $Y_i(0), Y_i(1) \perp X_i \mid Z_i(0), Z_i(1), C_i$ so that potential outcomes are independent of the treatment given the principal strata, and treated and control units can be compared conditional on a principal stratum, which includes points u characterizing units that have the same vector $(Z_i(0), Z_i(1))$. Therefore principal stratification represents the coarsest choice of groups of units (i.e., subpopulations, types of units, or subsets of

the support of u, named equivalence classes in Pearl (2011)), conditional on which we still have ignorability of the treatment, X_i.

4 Principal Stratification and Mediation Analysis

The framework of principal stratification focuses on local causal effects, that is, causal effects for specific subpopulations (principal strata). Despite the local nature of principal strata effects, we view the concept of principal stratification as a useful principle for addressing the topic of direct and indirect causal effects. Principal stratification makes it clear that evidence on the direct effect of the treatment on the primary outcome is provided by principal strata where the intermediate variable is unaffected by the treatment, i.e., $Z_i(0) = Z_i(1)$. Formally, the (average) Principal Strata Direct Effect (*PSDE*) of X on Y at level z, $z \in \mathscr{X}$, is simply the principal causal effect for the stratum where $Z_i(0) = Z_i(1) = z$, i.e., $PSDE(z) = E[Y_i(1) - Y_i(0) \mid Z_i(0) = Z_i(1) = z]$. A principal strata direct effect can also be named 'dissociative effect', because it measures an effect on the outcome that is dissociative with an effect on the intermediate variable. Therefore, only in strata where the intermediate variable is unaffected by the treatment (strata comprising units with $Z_i(0) = Z_i(1)$) can we hope to learn something about the direct effect of the treatment, even if these strata may not be interesting strata *per se*.

Causal mediation analysis focuses on disentangling direct and indirect effects, which are generally defined at the individual level and averaged over the whole population. Formally, Robins and Greenland (1992) and Pearl (2001) give the following definitions of (average) natural direct an indirect effects: $NDE(x) = E[Y_i(1, Z_i(x)) - Y_i(0, Z_i(x))]$ and $NIE(x) = E[Y_i(x, Z_i(1)) - Y_i(x, Z_i(0))]$, $x = 0, 1$. These effects provide a decomposition of the average total causal effect (*ACE*) into the sum of a natural direct effect and a natural indirect effect: $ACE = NDE(x) + NIE(1-x)$. Conversely, principal stratification does not in general allow one to decompose the total effect into overall direct and indirect effects, unless additional assumptions are made, but the average total effect of the treatment X on the outcome Y (*ACE*) is the weighted average of *PCE*s across units belonging to different principal strata:

$$\begin{aligned} ACE &= E[Y_i(1) - Y_i(0)] \\ &= \sum_{(z_0,z_1)} PCE(z_0,z_1)\pi_{z_0,z_1} = \sum_{z_0=z_1=z} PSDE(z)\pi_z + \sum_{z_0 \neq z_1} PCE(z_0,z_1)\pi_{z_0,z_1}, \end{aligned}$$

where π_{z_0,z_1} is the proportion of subjects belonging to principal stratum $\{i : Z_i(0) = z_0, Z_i(1) = z_1\}$, and $\pi_z = \pi_{z,z}$.

If $PSDE(z) = 0$, for each $z \in \mathscr{Z}$, then there is no evidence on the direct effect of the treatment after controlling for the mediator, because the causal effect of treatment on the outcome exists only in the presence of a causal effect of treatment on the intermediate variable. This does not mean that there is no natural direct effect of the treatment: The *PCE*s for units belonging to principal strata where the post-treatment variable is affected by treatment (also named 'associative effects') generally combine natural direct and indirect effects. Formally, we can easily show that $NDE(x) = \sum_{z_0=z_1=z} PSDE(z)\pi_z + \sum_{z_0 \neq z_1} E[Y_i(1, Z_i(x)) - Y_i(0, Z_i(x)) \mid Z_i(0) = z_0, Z_i(1) = z_1)]\pi_{z_0,z_1}$, $x = 0, 1$. Therefore, even if $PSDE(z) = 0$, for each $z \in \mathscr{Z}$, $NDE(x)$ can be non-zero (VanderWeele, 2008).

The assumptions that allow us to identify principal strata effects and natural direct and indirect effects are of a different nature and a careful evaluation of their plausibility is crucial.

To clarify the role of the different assumptions, we first turn back to randomized studies suffering from treatment noncompliance, where these issues are best seen and, to some extent, accepted. In a randomized study, let X be the initial binary treatment assignment (or instrument); $Z(x)$ represents the actual binary treatment received under assignment x, $x = 0, 1$ (the intermediate variable), $Y(x)$, $x = 0, 1$, are the two potential outcomes if units are assigned treatment or control. A principal stratification approach cross-classifies units into four groups based on their compliance behavior: compliers (if $Z_i(x) = x$ for $x = 0, 1$), never-takers (if $Z_i(x) = 0$ for $x = 0, 1$), always-takers (if $Z_i(x) = 1$ for $x = 0, 1$) and defiers (if $Z_i(x) = 1 - x$ for $x = 0, 1$). Assuming monotonicity, thus ruling out the existence of defiers, compliers are the only group where we can learn something about the effect of treatment received, as never-takers (always-takers) are never (always) observed taking the treatment in this experiment. The average causal effect for this subpopulation, the compliers average causal effect, $CACE = E[Y_i(1) - Y_i(0) | Z_i(0) = 0, Z_i(1) = 1]$, is an associative effect, which essentially combines direct and indirect effects of X. In order to interpret *CACE* as the causal effect of the receipt of the treatment, i.e., as an indirect effect *only* of the assignment through the treatment received, an additional assumption is required, which rules out direct effects of the assignment for compliers, for whom the treatment assignment and the treatment received are completely confounded. This exclusion restriction for compliers is an assumption of a different nature from the exclusion restriction for noncompliers, because it is about the interpretation of CACE, not about issues concerning identifying or estimating it (e.g., Mealli and Rubin (2002)). Note that the exclusion restriction for compliers is routinely made, often implicitly, also in randomized experiments with full compliance, where the desire to make this assumption more plausible underlies the widespread practice of blinding and double blinding experiments. In order to iden-

tify CACE, the following assumptions are sufficient: unconfoundedness of treatment assignment (which holds by design in randomized studies), monotonicity of compliance and compound exclusion restriction for never-takers and always-takers. However, these assumptions are not sufficient to identify the average effect of the treatment for the full population (ACE): If ACE is the causal estimand of interest, additional assumptions are required, which allow one to carry out extrapolation of causal effects of the treatment received, Z, for non-compliers. Similarly, we argue that assumptions of a different nature are required depending on if the focus is on principal causal effects, which are local treatment effects (like CACE), or natural direct and indirect effects, which are population causal effects (like ACE). See, for example, Ten Have and Joffe (2012) for a recent review of the alternative sets of assumptions needed to identify and estimate different direct and indirect effects.

In principal stratification analysis, challenges in identifying principal strata effects stem primarily from the fact that we cannot, in general, observe the principal stratum to which a subject belongs, because we cannot directly observe both $Z_i(0)$ and $Z_i(1)$. The observed groups defined by the treatment, X_i, and the observed value of the intermediate outcome, Z_i^{obs}, generally comprise mixtures of principal strata, therefore assumptions that allow us to untie these mixtures of principal strata are required to identify *PCE*s. Unfortunately, outside the noncompliance/instrumental variables setting, assumptions such as the exclusion restrictions cannot be invoked (exclusion restrictions would rule out a priori the direct effects that are being sought!). Depending on the substantive empirical setting, other assumptions can be introduced, which however generally lead only to partial identification of *PCE*s (e.g., Zhang and Rubin (2003), Lee (2009), Imai (2008), Mattei and Mealli (2011)), unless coupled with distributional assumptions (e.g. Hirano et al. (2000), Mattei and Mealli (2007), Jin and Rubin (2008), Zhang et al. (2009), Schwartz et al. (2011)). Those may be critical, and that is the reason why we are advocating the use of Bayesian methods that allow one to also conduct sensitivity analysis to model specification (e.g., Mattei and Mealli (2007), Jin and Rubin (2008)). Although these additional assumptions may be arguable in some settings, they generally do not involve comparisons of units belonging to different strata, because they aim at identifying local causal effects rather than overall average direct and indirect effects.

Causal mediation analysis focuses on causal estimands defined using potential outcomes of the form $Y_i(x,z)$, which are not observed in a specific experiment for units in some principal strata. In order to identify and estimate natural direct and indirect effects assumptions are required, which generally involve estimating causal effects for units for which the data contains no or little information. These assumptions generally require to specify an assignment mechanism for the medi-

ating variable, Z, thereby requiring that Z could be, at least in principle, regarded as an additional treatment and could be at least potentially controlled by external interventions. If we are willing to entertain hypothetical interventions on the intermediate variable Z, assumptions on the compound assignment mechanism for the multivariate treatment variable, (X,Z) should be contemplated. We agree with Pearl (2011) that these assumptions may be reasonable in some studies, but there are also studies where hypothetical interventions on the intermediate variable are not conceivable. Therefore, as in the analysis of observational studies, an important preliminary step in mediation analysis is to evaluate very carefully the possibility and plausibility to conceptualize interventions on the mediating variable. If this preliminary step is successful, that is, if an intervention on the mediating variable is, at least in principle, conceivable, we can try to substitute physical manipulations with reasonable assumptions concerning the rules used to assign the values of the intermediate variable, positing an assignment mechanism on the mediating variable (Mealli and Rubin, 2003, Jin and Rubin, 2008).

For instance, sequential ignorability assumptions are often made in mediation analysis (e.g., Robins (1999), Jo (2008), Imai et al. (2010), Ten Have and Joffe (2012)). These assumptions imply unconfoundedness of the mediator (conditional on some observed confounders), which allows one to extrapolate information on potential outcomes of the form $Y_i(x,Z_i(1-x))$ from the observed data. In noncompliance settings, a similar assumption would require unconfoundedness of the treatment received, which amounts to assuming that within cells defined by the values of pre-treatment variables, the treatment received is randomly assigned. This assumption guarantees that conditional on the covariates, comparing individuals by the actual treatment received leads to valid inference on causal effects. Therefore, unconfoundedness of the treatment received implies that we can compare treated and untreated units with the same value of the covariates, also if they belong to different principal compliance strata. In other words, this assumption allows us to use extrapolation across principal strata to draw inference about potential outcomes for units on which the data contains no or little information (such as, the potential outcomes $Y_i(x,Z_i(x)=1)$ for never-takers and the potential outcomes $Y_i(x,Z_i(x)=0)$ for always-takers). In noncompliance/instrumental variable settings, this is specifically the assumption we want to avoid, because it is believed that the treatment received is plausibly confounded. The nature of these extrapolations may be less credible than the inferences for a particular subpopulation, such as the compliers (see, e.g., Imbens (2010), pages 414–416) or the subpopulations of units for which the intermediate variable is unaffected by the treatment.

If one is seeking natural direct and indirect effects, we suggest to preliminary conduct a principal stratification analysis, looking at the distribution of covari-

ates and outcomes within each principal stratum (which is possible using likelihood or Bayesian analysis, e.g., Frumento et al. (2012)), and to decide whether mixing information across principal strata is reasonable, instead of doing it a priori under no confounding assumptions. In fact, although a principal stratification analysis may not answer the primary question of interest of a mediation analysis, it can turn out to be useful also to clearly understand the different nature of the assumptions leading to identify and estimate natural direct and indirect effects and principal strata direct effects. To be more specific: The observed data, X_i, Z_i^{obs} and Y_i^{obs}, contain information on the potential outcome $Y_i(x, Z_i(1-x))$ only for the subpopulation of units for which the intermediate variable is unaffected by the treatment. For this type of units $Z_i(0) = Z_i(1)$, which implies that $Y_i(x, Z_i(1-x)) = Y_i(x, Z_i(x))$ and, hence, $Y_i(x, Z_i(1-x))$ is observed for units receiving treatment x. As a result, the natural direct effect for this type of units corresponds to the weighted average of the principal strata direct effects $PSDE(z)$ with weights π_z, $z \in \mathscr{Z}$. The observed data are uninformative regarding the natural direct effects for other subpopulations of units, for which the treatment affects the intermediate variable, as associative effects combine direct and indirect effects. However, we argue that a principal stratification analysis that involves all the principal strata, including strata where $Z_i(0) \neq Z_i(1)$, may provide some insights on the mediated effects by comparing associative and dissociative effects, at least under specific assumptions. For instance, if units belonging to different principal strata had similar distributions of the covariates (that is, similar observed characteristics) and/or similar outcome levels under one of the treatment levels, we could reasonably assume that subpopulations where the treatment affects the intermediate outcome are characterized by the same direct effects as subpopulations where the mediating variable is unaffected by the treatment. Under this assumption indirect effects within principal strata can be derived by difference between associative effects and principal strata direct effects, and the overall (natural) indirect effect defined as weighted average of these indirect effects with weights the strata proportions. To be clear, suppose that the intermediate variable Z is binary taking on values 0 and 1 and focus on the natural direct effect of X on Y intervening to fix the mediator Z to the value it would have taken if X had been set to the control level $X = 0$, $NDE(0)$. Because the intermediate variable is binary, the basic principal stratification partitions units into four latent groups: $00 = \{i : Z_i(0) = 0, Z_i(1) = 0\}$; $01 = \{i : Z_i(0) = 0, Z_i(1) = 1\}$; $10 = \{i : Z_i(0) = 1, Z_i(1) = 0\}$; and $11 = \{i : Z_i(0) = 1, Z_i(1) = 1\}$. If the mean potential outcome under control for units belonging to principal stratum 01 is estimated to be approximately equal to the mean potential outcome under control for units belonging to principal stratum 00: $E[Y_i(0)|Z_i(0) = 0, Z_i(1) = 1] = E[Y_i(0)|Z_i(0) = 0, Z_i(1) = 0]$, then it may be reasonable to assume that $E[Y_i(1, Z_i(0))|Z_i(0) = 0, Z_i(1) = 1] =$

$E[Y_i(1,Z_i(0))|Z_i(0)=0,Z_i(1)=0]$. Analogously, if the mean potential outcome under control for units belonging to principal stratum 10 is estimated to be approximately equal to the mean potential outcome under control for units belonging to principal stratum 11: $E[Y_i(0)|Z_i(0)=1,Z_i(1)=0]=E[Y_i(0)|Z_i(0)=1,Z_i(1)=1]$, then it may be reasonable to assume that $E[Y_i(1,Z_i(0))|Z_i(0)=1,Z_i(1)=0]=E[Y_i(1,Z_i(0))|Z_i(0)=1,Z_i(1)=1]$. Under these assumptions, units belonging to principal strata 01 and 10 are characterized by the same natural direct effect as units belonging to principal stratum 00 and 11, respectively: $NDE_{01}(0)=NDE_{00}(0)$, and $NDE_{10}(0)=NDE_{11}(0)$. Therefore,

$$\begin{aligned} NDE(0) &= NDE_{00}(0)\pi_{00}+NDE_{11}(0)\pi_{11}+NDE_{10}(0)\pi_{01}+NDE_{10}(0)\pi_{10} \\ &= NDE_{00}(0)(\pi_{00}+\pi_{01})+NDE_{11}(0)(\pi_{11}+\pi_{10}) \\ &= PSDE(0)(\pi_{00}+\pi_{01})+PSDE(1)(\pi_{11}+\pi_{10}). \end{aligned}$$

and the natural indirect effect, $NIE(1)$, can be derived by difference: $NIE(1)=ACE-NDE(0)=ACE-PSDE(0)(\pi_{00}+\pi_{01})-PSDE(1)(\pi_{11}+\pi_{10})$. A similar reasoning could be applied to derive the natural direct effect of X on Y intervening to fix the mediator Z to the value it would have taken if X had been set to the treatment level $X=1$, $NDE(1)$.

5 Surrogate Endpoints

In studies where measurement of the primary outcome, Y, may be too expensive or unfeasible in a practical time spell, surrogate variables are often used to evaluate the effects of the treatment on Y.

Pearl (2011) shows 'strong reservation' regarding the use of principal stratification for addressing surrogate problems, interpreting the use of principal stratification in this setting as 'an intellectual restriction that confines its analysis to the assessment of strata-specific effects.' Although a principal stratification analysis focuses on principal strata causal effects (strata-specific effects), we argue that it should not be used only to assess the 'principal surrogacy' condition, which only involves dissociative effects: an intermediate variable Z is a principal surrogate if all the dissociative effects are zero. In order to gain insights on how the effect of treatment on the surrogate relates to the effect of treatment on the outcome, a *full* principal stratification analysis, which aims at evaluating the effect of the treatment in each of the principal strata, should be conducted. In addition to providing information on the principal surrogacy condition, a *full* principal stratification may warn against potential paradoxes, such as the 'surrogate paradox' (Chen et al., 2007). Specifically, a principal surrogate may lead to a surrogate paradox if the treatment

has a positive (negative) effect on the surrogate, which in turn has a positive (negative) effect on the primary outcome, but the treatment has a negative (positive) effect on the primary outcome. In this situation, some associative effects should have opposite signs, so investigating the effect of the treatment in each of the principal strata may provide some insights on the transportability of the effect of the treatment on the surrogate to the effect of the treatment on the outcome, which we agree with Pearl (2011) is the primary purpose of a surrogate.

As said before, a pitfall of a full principal stratification analysis is that identification of principal strata effects may create challenges, because the groups which these effects refer to are only partially observed. However, advanced statistical methodologies, such as flexible Bayesian models (Schwartz et al., 2011), have been recently developed to face identification and estimation issues.

6 Conclusion

Our discussion aims at clarifying the role of principal stratification in causal inference and explaining why it can be useful to address causal problems involving post-treatment variables. Principal stratification does not always answer the causal question of primary interest, but it often provides useful insights, compelling to explicitly delineate the critical assumptions needed for a causal interpretation of the estimands of interest, and allowing for a clear assessment of the consequences of violation of these assumptions.

Most of existing studies, either using principal stratification or criticizing this framework in favor of alternative approaches, focus only on some principal causal effects ignoring the other ones. For instance, studies aiming at evaluating surrogate endpoints or direct effects of the treatment usually focus on causal effects for subpopulations of units where the potential surrogate outcome/mediating variable is unaffected by the treatment (principal strata direct effects), and neglect information on principal strata where $Z_i(0) \neq Z_i(1)$. We argue that a principal stratification analysis should involve all the principal strata, studying the characteristic of each principal stratum, and evaluating the distributions of potential outcomes in each principal stratum. Conducting a *full* principal stratification analysis may be a challenging task, due to the fact that principal strata membership is only partially observed. To address this issue, our preference would go for the Bayesian paradigm, which appears to be particularly appropriate for dealing with problems of causal inference. With a Bayesian principal stratification analysis, we can transparently specify causal models, explicitly define and separate structural behavioral assumptions and model assumptions, and clearly define priors on parameters. A principal

stratification analysis that does not neglect any strata may provide substantial information on the causal problem at hand, by discovering heterogeneities in the treatment effects across principal strata, providing insights even on indirect effects of the treatment, and warning against potential paradoxes, such as the 'surrogate paradox' (Chen et al., 2007). In our view, principal stratification is a principal framework for addressing causal inference problems in the presence of post-treatment variables.

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): Identification of Causal Effects Using Instrumental Variables (with discussion). *Journal of the American Statistical Association* **91**, 444–472.

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

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.

Baker, S. G., K. S. Lindeman, and B. S. Kramer (2011): Clarifying the role of principal stratification in the paired availability design. *The International Journal of Biostatistics* **7**(1), Article 25.

Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* **64**, 695–701.

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

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

Egleston, B. L. (2011) Response to Pearl's comments on principal stratification. *The International Journal of Biostatistics* **7**(1), Article 24.

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

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

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2012): Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistcial Association*. Forthcoming.

Gilbert, P. B., M.G. Hudgens, and J. Wolfson (2011): Commentary on 'Principal stratification - a goal or a tool?' by Judea Pearl. *The International Journal of Biostatistics* **7**(1), Article 36.

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

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.

Hirano, K., G. W. Imbens, D. B. Rubin, and X-H. Zhou (2000): Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 1–20.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–970.

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

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

Imai, K., D. Tingley, and T. Yamamoto (2012): Identification, inference, and sensitivity analysis for causal mediation effects. *Journal of the Royal Statistical Society (Series A)*, Forthcoming.

Imbens G. W. (2010): Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature, American Economic Association* **48**, 399–423.

Imbens G. W., and Angrist, J. D. (1994): Identification and Estimation of Local Average Treatment Effects. *Econometrica* **62**, 467–475.

Imbens G. W., and D. B. Rubin (1997a): Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies* **64**, 555–574.

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

Jin, H. L., and D. B. Rubin (2008): Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association* **103**, 101–111.

Jo, B. (2008): Causal inference in randomized experiments with mediational processes. *Psychological Methods* **13**, 314–336.

Joffe, M. (2011): Principal stratification and attribution prohibition: Good Ideas taken too far. *The International Journal of Biostatistics* **7**(1), Article 35.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* **76**, 1071–1102.

Mattei, A., and F. Mealli (2007): Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63**, 437–446.

Mattei, A., and F. Mealli (2011): Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society (Series B)* **73**, 729–752.

Mealli, F. and D. B. Rubin (2002): Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services & Outcomes Research Methodology*, **3**, 225–232.

Mealli, F., and D..B. Rubin (2003): Assumptions allowing the estimation of direct causal effects. Commentary on 'Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status' by P. Adams, M. D. Hurd, D. McFadden, A. Merrill, T. Ribeiro. *Journal of Econometrics* **112**, 79–87.

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. (2001): Direct and indirect effects. In *Proceeding 17th Conference on Uncertainty in Artificial Intelligence* (eds. J. S. Breese & D. Koller), 411–420. Morgan Kaufman, S. Francisco, CA.

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

Prentice, R. (2011): Invited commentary on Pearl and principal stratification. *The International Journal of Biostatistics* **7**(1), Article 30.

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

Robins, J. M., (1999): Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment and Clinical Trials* (M. E. Halloran and D. A. Berry, eds.) 95–134. Springer, New York.

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

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

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

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

Rubin, D. B. (1974): Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 68–701

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

Rubin, D. B. (1980): Comment on 'Randomization analysis of experimental Data: The Fisher randomization test' by D. Basu. *Journal of the American Statistical Association* **75**, 591–593.

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

Rubin, D. B. (2004): Direct and indirect causal effects via potential outcomes (with discussion and reply). *Scandinavian Journal of Statistics* **31**, 161–170; 196–198.

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

Schwartz, S. L., F. Li, and F. Mealli (2011): A Bayesian semiparametric approach to intermediate variables in Causal inference. *Journal of the American Statistcial Association*, **106**, 1331–1344.

Sjölander, A. (2011): Reaction to Pearl's critique of principal stratification. *The International Journal of Biostatistics* **7**(1), Article 22.

Ten Have, T. R., and M. M. Joffe (2012): A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research* **21**, 77–107.

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

VanderWeele, T. L. (2011): Principal stratification − Uses and limitations. *The International Journal of Biostatistics* **7**, Article 28.

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

Zhang, J., D. B. Rubin, and F. Mealli (2009): Likelihood-based analysis of causal effects via principal stratification: New approach to evaluating job-training program. *Journal of the American Statistical Association*, **104**, 166–176.