Statistics and Probability Letters 78 (2008) 2957–2962

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Simple relations between principal stratification and direct and indirect effects

Tyler J. VanderWeele[*]

Department of Health Studies, University of Chicago, United States

ARTICLE INFO

Article history:
Received 3 November 2007
Received in revised form 28 February 2008
Accepted 12 May 2008
Available online 29 May 2008

ABSTRACT

The concepts of principal strata direct and indirect effects are compared with and contrasted to definitions of direct and indirect effects which make reference to interventions on intermediate variables. Certain simple relations hold for direct effects but not indirect effects.

1. Introduction

Statistical analyses concerning the role of an intermediate variable between a particular treatment and an outcome are important for understanding issues concerning mechanism. Several ways to conceptualize the mediatory role of an intermediate variable in the treatment–outcome relationship have been proposed in the causal inference literature (Joffe et al., 2007). One such approach considers what would happen to the treatment–outcome relationship under interventions on the intermediate variable (Robins and Greenland, 1992; Pearl, 2001). Another approach focuses on the treatment–outcome relation for strata defined by potential outcomes for the treatment–mediator relationship (Frangakis and Rubin, 2002; Rubin, 2004). In a couple of recent papers, Rubin (2004, 2005) has argued against the use of direct and indirect effects concepts which arise from interventions on the mediator. Rubin provides examples of cases in which there is no direct effect within principal strata but in which the intervention-oriented definitions of a direct effect lead to the conclusion that a direct effect is present. In this paper we clarify the relationships between the concepts of direct and indirect effects based on principal stratification and those based on interventions on the mediator. The remainder of this paper is organized as follows. Section 2 reviews the relevant concepts and clarifies what might be meant by a principal strata direct or indirect effect. Section 3 shows how the concepts of direct and indirect effects based on principal stratification and on interventions on the mediator are related. Section 4 provides some concluding discussion.

2. Concepts and definitions

We assume a potential outcomes framework Rubin (1974, 1978, 1990). We will let Ω denote the sample space of individuals in the population and we will use ω for a particular sample point. Let $A(\omega)$ denote the treatment received by individual ω. Let $Y(\omega)$ denote some post-treatment outcome for individual ω. Let $Z(\omega)$ be a post-treatment variable that will denote the intermediate variable. Let $Y_a(\omega)$ and $Z_a(\omega)$ denote respectively the counterfactual values (or potential outcomes) of Y and Z respectively for individual ω if treatment A were set, possibly contrary to fact, to the value a. Let $Y_{az}(\omega)$ denote the counterfactual value for Y if, possibly contrary to fact, A were set to a and Z were set to z. Note that we assume that the counterfactual values $Z_a(\omega)$, $Y_a(\omega)$ and $Y_{az}(\omega)$ for individual ω do not depend on the treatments received

[*] Corresponding address: Department of Health Studies, University of Chicago, 5841 South Maryland Avenue, MC 2007, Chicago, IL 60637, United States. Tel.: +1 773 834 2509; fax: +1 773 702 2453.
E-mail address: vanderweele@uchicago.edu.

0167-7152/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.05.029

by other individuals. Such assumptions are sometimes referred to as SUTVA, the stable unit treatment value assumption (Rubin, 1990). We also require the "consistency" assumption, i.e. that $Y_{A(\omega)}(\omega) = Y(\omega)$ so that the value of Y which would have been observed if A had been set to what it in fact was is equal to the value of Y which was in fact observed. Similarly, $Z_{A(\omega)}(\omega) = Z(\omega)$ and $Y_{A(\omega)Z(\omega)}(\omega) = Y(\omega)$.

Pearl (2001) gave definitions for controlled direct effects and natural direct and indirect effects based on interventions on the mediator or intermediate variable Z. Robins and Greenland (1992) provided related definitions. The controlled direct effect of treatment A on outcome Y comparing $A = a$ with $A = a^*$ and setting Z to z is defined by $CDE_{a,a^*}(z) = Y_{az} - Y_{a^*z}$ and measures the effect of A on Y not mediated through Z, i.e. the effect of A on Y after intervening to fix the mediator to some value z. The controlled direct effect for individual ω is denoted by $CDE_{a,a^*}(z)(\omega) = Y_{az}(\omega) - Y_{a^*z}(\omega)$. The average controlled direct effect for the population is denoted by $\mathbb{E}[CDE_{a,a^*}(z)] = \mathbb{E}[Y_{az} - Y_{a^*z}]$. The natural direct effect, $NDE_{a,a^*}(a') = Y_{aZ_{a'}} - Y_{a^*Z_{a'}}$, also measures the effect of A on Y not mediated through Z but, in contrast, considers the effect of A on Y intervening to fix the mediator to the value it would have taken if A had been set to a'. Corresponding to the natural direct effect is the concept of the natural indirect effect which measures the extent to which an intervention affects the outcome through the mediator. The natural indirect effect, $NIE_{a,a^*}(a') = Y_{a'Z_a} - Y_{a'Z_{a^*}}$, fixes A to a' and then measures the effect on the outcome Y of intervening to set the mediator to what it would have been if A were a in contrast to what it would have been if A were a^*. Note that for natural direct and indirect effects, Pearl (2001) considers only the case in which either $a' = a$ or $a' = a^*$ but the concepts of a natural direct and indirect effect extend more generally to those given above. The total effect $Y_a - Y_{a^*}$ can be decomposed into a natural direct effect and a natural indirect effect as follows: $Y_a - Y_{a^*} = Y_{aZ_a} - Y_{a^*Z_{a^*}} = Y_{aZ_a} - Y_{aZ_{a^*}} + Y_{aZ_{a^*}} - Y_{a^*Z_{a^*}} = NIE_{a,a^*}(a) + NDE_{a,a^*}(a^*)$. Robins and Greenland (1992), Pearl (2001), Robins (2003), Avin et al. (2006) and Peterson et al. (2006) all consider various identification strategies for controlled and natural direct and indirect effects. Didelez et al. (2006) and Geneletti (2007) apply similar concepts to a decision network without employing counterfactuals. Much of this work concerns identification strategies for causal graphical models. In this paper our concern is not with identification but with the conceptual relations between principal stratification and controlled and natural direct and indirect effects. We will thus assume knowledge on all potential outcomes.

We now consider definitions related to principal stratification (Frangakis and Rubin, 2002). Let $\mathcal{A}$ denote the support of the treatment variable A, then for individual ω define $Q(\omega) = \{Z_a(\omega)\}_{a\in\mathcal{A}}$. Frangakis and Rubin (2002) refer to the strata of Q as principal strata, i.e. a principal stratum is a group of individuals for whom for each a, $Z_a(\omega)$ takes the same value for individuals in that stratum. Frangakis and Rubin (2002) then define a principal effect as a comparison between the potential outcomes Y_a and Y_{a^*} within a particular stratum of Q. Barnard et al. (2003), Zhang and Rubin (2003), Shepherd et al. (2006), Cheng and Small (2006), Frangakis et al. (2007), Shepherd et al. (2007) and Imai (2008) consider various identification strategies, bounds and applications for the concepts of principal stratification and principal effects. Rubin (2004) provides some discussion relating principal stratification to direct and indirect effects. In this discussion, Rubin (2004) did not formally define principal strata direct and indirect effects. Rubin puts "direct effects" and "indirect effects" in quotations, presumably to indicate that he believes that these concepts have limited utility. His examples, however, seem to implicitly draw upon the following definitions which we now give formally. We will say A has no principal strata direct effect on Y, comparing $A = a$ to $A = a^*$, if whenever $Q = q$ is a principal stratum such that $Z_a = Z_{a^*}$ we have $P(Y_a - Y_{a^*} = 0|Q = q) = 1$; and we will say A has no principal strata indirect direct effect on Y, comparing $A = a$ to $A = a^*$, if $P(Y_a - Y_{a^*}|Q = q)$ is independent of q. For the definition of principal strata indirect effects one might alternatively say that A has no principal strata indirect effect on Y, comparing $A = a$ to $A = a^*$, if $P(Y_a - Y_{a^*}|Z_a = z, Z_{a^*} = z')$ is independent of z and z'. In the examples below, A is binary and so the two potential definitions coincide. In an attempt to follow the treatment in Rubin (2004), the definitions above are given negatively in terms of the absence of principal strata direct and indirect effects. If it is not the case that there is no principal strata direct effect we might propose defining the average principal strata direct effect for stratum q with $Z_a = Z_{a^*}$ as $\mathbb{E}[Y_a - Y_{a^*}|Q = q]$. We might also attempt proposing a definition for average principal strata indirect effects by considering contrasts of the form $\mathbb{E}[Y_a - Y_{a^*}|Q = q_1] - \mathbb{E}[Y_a - Y_{a^*}|Q = q_2]$; contrasts of this form however are not causal effects in the sense of being contrasts of potential outcomes of the same population under different interventions since the conditioning sets $Q = q_1$ and $Q = q_2$ are not the same. We will now formally relate these concepts of principal strata direct and indirect effects to controlled and natural direct and indirect effects.

3. Relations between principal stratification and direct and indirect effects

Rubin (2004) gives two examples, one in which principal strata direct effects are present but in which the correlation between the treatment variable and the outcome is zero within strata of the intermediate variable and another in which there is no principal strata direct effect but in which the correlation between the treatment variable and the outcome within strata of the intermediate variable is non-zero. In the context of Rubin's example, if there are no unmeasured confounding variables for the treatment–outcome or intermediate–outcome relationships then rules from causal graphical models (Pearl, 1995, 2000) imply that if the correlation between the treatment variable and the outcome within strata of the intermediate variable is non-zero then controlled direct effects must be present. Rubin's second example thus suffices to demonstrate that there can be controlled direct effects even if there are no principal strata direct effects. However, even if there are no unmeasured confounding variables for the treatment–outcome or intermediate–outcome relationships, a zero correlation between the treatment variable and the outcome within strata of the intermediate variable does not imply that there are no controlled direct effects; thus, Rubin's first example is an example where one cannot come to a conclusion about the absence of direct effects from the absence of an observed correlation. Note also that if there are unmeasured confounding variables of the treatment–outcome or intermediate–outcome relationships, conclusions about the presence of controlled direct effects cannot be drawn from data on the treatment, intermediate and outcome alone (Robins and Greenland, 1992; Cole and Hernán, 2002).

In this section we will consider in greater detail the relationship between principal strata direct and indirect effects and definitions of direct and indirect effects that arise in the context of interventions on the intermediate variable. We first show that if there is no natural direct effect for any individual then there is no principal strata direct effect.

Proposition 1. *If $NDE_{a,a^*}(a^*)(\omega) = 0$ for all ω then whenever $Q = q$ is a principal stratum such that $Z_a = Z_{a^*}$ we have $P(Y_a - Y_{a^*} = 0|Q = q) = 1$, i.e. A has no principal strata direct effect on Y comparing $A = a$ to $A = a^*$.*

Proof. Suppose $NDE_{a,a^*}(a^*) = 0$ and suppose that $Q = q$ is a principal stratum such that $Z_a = Z_{a^*}$; then conditional on $Q = q$

$$\begin{aligned} P(Y_a - Y_{a^*} = 0|Q = q) &= P(Y_{aZ_a} - Y_{aZ_{a^*}} + Y_{aZ_{a^*}} - Y_{a^*Z_{a^*}} = 0|Q = q) \\ &= P(Y_{aZ_a} - Y_{aZ_{a^*}} + NDE_{a,a^*}(a^*) = 0|Q = q) \\ &= P(Y_{aZ_a} - Y_{aZ_a} = 0|Q = q) \\ &= P(0 = 0|Q = q) = 1. \quad \square \end{aligned}$$

Similarly, it can be shown that if there are no controlled direct effects for any individual then there can be no principal strata direct effect.

Proposition 2. *If $CDE_{a,a^*}(z)(\omega) = 0$ for all z and for all ω then whenever $Q = q$ is a principal stratum such that $Z_a = Z_{a^*}$ we have $P(Y_a - Y_{a^*} = 0|Q = q) = 1$, i.e. A has no principal strata direct effect on Y.*

Proof. If $CDE_{a,a^*}(z) = 0$ for all z, then for any individual ω, $NDE_{a,a^*}(a^*)(\omega) = Y_{aZ_{a^*}(\omega)}(\omega) - Y_{a^*Z_{a^*}(\omega)}(\omega) = Y_{az'}(\omega) - Y_{a^*z'}(\omega) = CDE_{a,a^*}(z')(\omega) = 0$ where $z' = Z_{a^*}(\omega)$. Thus $NDE_{a,a^*}(a^*)(\omega) = 0$ for all ω and from this the conclusion follows from Proposition 1. □

We see then that if there are no controlled direct effects or no natural direct effects then there can be no principal strata direct effects. From the contrapositive of these results we have that if there are principal strata direct effects then there must be some individuals for whom there are controlled direct effects and natural direct effects. Thus in Rubin's (2004) first example in which there were principal strata direct effects but in which the correlation between the treatment variable and the outcome was zero within strata of the intermediate variable, Proposition 2 allows us to conclude that there are some individuals for whom there is a controlled direct effect even though the correlation between the treatment variable and the outcome is zero within strata of the intermediate variable. In the context of no confounding for the treatment–outcome and intermediate–outcome relationships, non-zero correlation between the treatment variable and the outcome within strata of the intermediate variable is a sufficient but not necessary condition for the presence of controlled direct effects. Rubin's second example shows that the converse of Proposition 2 does not hold, i.e. no principal strata direct effects does not imply the absence of controlled direct effects. Moreover, in Example 1 we show that the absence of principal strata direct effects does not imply the absence of natural direct effects and does not necessarily even imply that the average controlled or natural direct effect is zero.

Example 1. We show that there can be individual natural direct effects with no principal strata direct effects. Consider the potential outcomes given in Table 1. Note that the counterfactuals of the form Z_a and Y_{az} suffice to fix counterfactuals of the form Y_a because $Y_a = Y_{aZ_a}$. Suppose that one third of the subjects are in each principal stratum and that there is no individual for whom $Z_0 = 1$ and $Z_1 = 0$. In this example there are two principal strata for which $Z_0 = Z_1$, $Q = 1$ and $Q = 3$. For the principal strata $Q = 1$ we have that $P(Y_1 - Y_0 = 0|Q = 1) = P(Y_{1Z_1} - Y_{0Z_0} = 0|Q = 1) = P(Y_{10} - Y_{00} = 0|Q = 1) = P(0 - 0 = 0|Q = 1) = 1$; for the principal strata $Q = 3$ we have that $P(Y_1 - Y_0 = 0|Q = 3) = P(Y_{1Z_1} - Y_{0Z_0} = 0|Q = 3) = P(Y_{11} - Y_{01} = 0|Q = 3) = P(80 - 80 = 0|Q = 3) = 1$. Thus A has no principal strata direct effect on Y. However, for individuals in principal strata $Q = 2$, $NDE_{1,0}(0) = Y_{1Z_0} - Y_{0Z_0} = Y_{10} - Y_{00} = 55 - 25 = 30$ and the average natural direct effect for the population $\mathbb{E}[NDE_{1,0}(0)]$ is given by $\mathbb{E}[Y_{1Z_0} - Y_{0Z_0}] = \sum_q \mathbb{E}[Y_{1Z_0} - Y_{0Z_0}|Q = q]P(Q = q) = \frac{1}{3}\mathbb{E}[Y_{10} - Y_{00}|Q = 1] + \frac{1}{3}\mathbb{E}[Y_{10} - Y_{00}|Q = 2] + \frac{1}{3}\mathbb{E}[Y_{11} - Y_{01}|Q = 3] = 0 + \frac{1}{3}(55 - 25) + 0 = 10$. Also, for individuals in principal strata $Q = 2$, $NDE_{1,0}(1) = Y_{1Z_1} - Y_{0Z_1} = Y_{11} - Y_{01} = 60 - 45 = 15$ and the average natural direct effect for the population $\mathbb{E}[NDE_{1,0}(1)]$ is given by $\mathbb{E}[Y_{1Z_1} - Y_{0Z_1}] = \sum_q \mathbb{E}[Y_{1Z_1} - Y_{0Z_1}|Q = q]P(Q = q) = \frac{1}{3}\mathbb{E}[Y_{10} - Y_{00}|Q = 1] + \frac{1}{3}\mathbb{E}[Y_{11} - Y_{01}|Q = 2] + \frac{1}{3}\mathbb{E}[Y_{11} - Y_{01}|Q = 3] = 0 + \frac{1}{3}(60 - 45) + 0 = 5$. For controlled directed effects, for individuals in principal strata $Q = 1$, $CDE_{1,0}(0) = 0$ and $CDE_{1,0}(1) = 30 - 15 = 15$; for individuals in principal strata $Q = 2$, $CDE_{1,0}(0) = 55 - 25 = 30$ and $CDE_{1,0}(1) = 60 - 45 = 15$; for individuals in principal strata $Q = 3$, $CDE_{1,0}(0) = 60 - 30 = 30$ and $CDE_{1,0}(1) = 80 - 80 = 0$. Average controlled directed effects are given by $\mathbb{E}[CDE_{1,0}(0)] = \frac{1}{3}(0) + \frac{1}{3}(30) + \frac{1}{3}(30) = 20$ and $\mathbb{E}[CDE_{1,0}(1)] = \frac{1}{3}(15) + \frac{1}{3}(15) + \frac{1}{3}(0) = 10$. Clearly both natural and controlled direct effects can be present even when there are no principal strata direct effects; essentially this is because there can be direct effects in principal strata in which $Z_a \neq Z_{a^*}$.

Table 1
Individual controlled and natural direct effects with no principal strata direct effects

Principal stratum Q	Z_0	Z_1	Y_{00}	Y_{01}	Y_{10}	Y_{11}
1	0	0	0	15	0	30
2	0	1	25	45	55	60
3	1	1	30	80	60	80

Table 2
Natural indirect effects with no principal strata indirect effects

Principal stratum Q	Z_0	Z_1	Y_{00}	Y_{01}	Y_{10}	Y_{11}
1	0	0	0	0	20	20
2	0	1	40	40	55	60
3	1	1	80	80	100	100

Table 3
Principal strata indirect effects with no natural indirect effects

Principal stratum Q	Z_0	Z_1	Y_{00}	Y_{01}	Y_{10}	Y_{11}
1	0	0	0	0	20	20
2	0	1	40	40	40	40
3	1	1	80	80	100	100

It is thus clear that no individual controlled or natural direct effects is a stronger condition than no principal strata direct effects. In his examples, Rubin (2004) does not give counterfactuals of the form Y_{az} but only of the form Y_a and Z_a. Knowing the individual potential outcomes Y_a and Z_a for all a for each individual will not in general suffice to determine controlled and natural direct and indirect effects because counterfactuals of the form Y_{az} and $Y_{aZ_{a^*}}$ are not identified from information on the individual potential outcomes Y_a and Z_a when $z \neq Z_a$ or when $Z_{a^*} \neq Z_a$ respectively. Whether it is reasonable to consider counterfactual variables of the form Y_{az} will depend on whether an intervention on the intermediate variable is conceivable; this is discussed further below. Propositions 1 and 2, however, allow us to conclude that when it is proper to use counterfactuals of the form Y_{az} then, if there are principal strata direct effects, there must be individuals for whom there are controlled and natural direct effects regardless of what the values of the counterfactual variables Y_{az} which are not determined by Y_a and Z_a may be.

For indirect effects, there are no relations similar to Propositions 1 and 2. Examples can be constructed in which there are no natural indirect effects for any individual but for which there are principal strata indirect effects. Examples can also be constructed in which there are no principal strata indirect effects but in which natural indirect effects are present. Examples 2 and 3 below illustrate these cases respectively.

Example 2. We show that there can be natural indirect effects with no principal strata indirect effects. Consider the potential outcomes given in Table 2. Again suppose that one third of the subjects are in each principal stratum and that there is no individual for whom $Z_0 = 1$ and $Z_1 = 0$. Let $1(A)$ denote the indicator function for A. We see that:

$$P(Y_1 - Y_0 = v|Q = 1) = P(Y_{1Z_1} - Y_{0Z_0} = v|Q = 1) = P(Y_{10} - Y_{00} = v|Q = 1) = 1(v = 20 - 0) = 1(v = 20)$$
$$P(Y_1 - Y_0 = v|Q = 2) = P(Y_{1Z_1} - Y_{0Z_0} = v|Q = 2) = P(Y_{11} - Y_{00} = v|Q = 2) = 1(v = 60 - 40) = 1(v = 20)$$
$$P(Y_1 - Y_0 = v|Q = 3) = P(Y_{1Z_1} - Y_{0Z_0} = v|Q = 3) = P(Y_{11} - Y_{01} = v|Q = 3) = 1(v = 100 - 80) = 1(v = 20).$$

Thus $P(Y_a - Y_{a^*}|Q = q)$ is independent of q and so there are no principal strata indirect effects. However, the indirect effect $Y_{1Z_1} - Y_{1Z_0}$ is non-zero in principal stratum 2 since in this principal stratum $Y_{1Z_1} - Y_{1Z_0} = Y_{11} - Y_{10} = 60 - 55 = 5$. The average natural indirect effect $\mathbb{E}[Y_{1Z_1} - Y_{1Z_0}]$ is $\mathbb{E}[Y_{1Z_1} - Y_{1Z_0}] = \sum_q \mathbb{E}[Y_{1Z_1} - Y_{1Z_0}|Q = q]P(Q = q) = \frac{1}{3}\mathbb{E}[Y_{10} - Y_{10}|Q = 1] + \frac{1}{3}\mathbb{E}[Y_{11} - Y_{10}|Q = 2] + \frac{1}{3}\mathbb{E}[Y_{11} - Y_{11}|Q = 3] = 0 + \frac{1}{3}(60 - 55) + 0 = \frac{5}{3}$. Natural indirect effects may be present with no principal strata indirect effects because the condition for the latter imposes restrictions on $P(Y_{1Z_1} - Y_{0Z_0}|Q = q)$ which does not restrict natural indirect effects of the form $Y_{1Z_1} - Y_{1Z_0}$ and $Y_{0Z_1} - Y_{0Z_0}$.

Example 3. We show that even if there are no natural indirect effects for any individual, the condition for no principal strata indirect effects may not hold. Consider the potential outcomes given in Table 3. Again suppose that one third of the subjects are in each principal stratum and that there is no individual for whom $Z_0 = 1$ and $Z_1 = 0$.

$$P(Y_1 - Y_0 = v|Q = 1) = P(Y_{1Z_1} - Y_{0Z_0} = v|Q = 1) = P(Y_{10} - Y_{00} = v|Q = 1) = 1(v = 20 - 0) = 1(v = 20)$$
$$P(Y_1 - Y_0 = v|Q = 2) = P(Y_{1Z_1} - Y_{0Z_0} = v|Q = 2) = P(Y_{11} - Y_{00} = v|Q = 2) = 1(v = 40 - 40) = 1(v = 0)$$
$$P(Y_1 - Y_0 = v|Q = 3) = P(Y_{1Z_1} - Y_{0Z_0} = v|Q = 3) = P(Y_{11} - Y_{01} = v|Q = 3) = 1(v = 100 - 80) = 1(v = 20).$$

Thus $P(Y_a - Y_{a^*}|Q = q)$ is not independent of q and so in this example the condition for no principal strata indirect effects does not hold. However, the natural indirect effect $Y_{1Z_1} - Y_{1Z_0}$ is zero for all individuals because in principal stratum 1, $Y_{1Z_1} - Y_{1Z_0} = Y_{10} - Y_{10} = 0$ and in principal stratum 2, $Y_{1Z_1} - Y_{1Z_0} = Y_{11} - Y_{10} = 40 - 40 = 0$ and in principal stratum

3, $Y_{1Z_1} - Y_{1Z_0} = Y_{11} - Y_{11} = 0$. Furthermore the natural indirect effect $Y_{0Z_1} - Y_{0Z_0}$ is also zero for all individuals because in principal stratum 1, $Y_{0Z_1} - Y_{0Z_0} = Y_{00} - Y_{00} = 0$ and in principal stratum 2, $Y_{0Z_1} - Y_{0Z_0} = Y_{01} - Y_{00} = 40 - 40 = 0$ and in principal stratum 3, $Y_{0Z_1} - Y_{0Z_0} = Y_{01} - Y_{01} = 0$. Note that there may be no natural indirect effects even if the condition for no principal strata indirect effects does not hold because the condition that there are no natural indirect effects imposes restrictions on $Y_{1Z_1} - Y_{1_{Z_0}}$ and $Y_{0Z_1} - Y_{0Z_0}$ which does not place restrictions on $P(Y_{1Z_1} - Y_{0Z_0} \mid Q = q)$.

4. Discussion

Rubin (2004) presented a hypothetical example in which there were no principal strata direct effects but in which the treatment variable was correlated with the outcome variable within strata of the intermediate variable, thereby arguing that graphical causal models were misleading for the assessment of direct and indirect effects. The discussion above indicates that there is no tension between Rubin's example and the definitions of direct and indirect effects that arise from graphical causal models and interventions on intermediate variables - it is possible to have no principal strata direct effects but for controlled and natural direct effects to still be present. We have discussed above how and when the concepts of principal strata direct and indirect effects and controlled and natural direct and indirect effects are related.

In practice, whether a researcher chooses to examine principal strata direct and indirect effects or controlled and natural direct and indirect effects will depend on (i) what variables in the study the treatment, the intermediate and the outcome actually represent, (ii) the scientific goals of the study and (iii) the plausibility of the assumptions required for estimation. Principal strata direct and indirect effects have the advantage that the concepts are defined irrespective of whether an intervention on the intermediate variable is conceivable. The definitions of controlled and natural direct and indirect effects require counterfactual variables of the form Y_{az} and thus require that some at least hypothetical intervention on the intermediate Z is conceivable. Rubin (2005) gives an agricultural example in which the units are plots of land, the treatment variable is fertilizer type, the intermediate is the number of plants established in each plot and the outcome is the yield in each plot. If the intermediate (the number of plants established) and the outcome (the total crop yield) are measured concurrently, one could, if desired, attempt to estimate principal effects and then examine whether e.g. principal direct effects are present. However, in this context it would not make sense to examine controlled and natural direct and indirect effects. If, however, the number of plants established were measured considerably prior to the measurement of total crop yield and if one could conceivably intervene on the number of plants established by transplanting plants from one plot to another then one could, in this context, define and estimate controlled and natural direct and indirect effects.

A second consideration in choosing whether to examine principal strata direct and indirect effects or controlled and natural direct and indirect effects concerns the plausibility of the assumptions required for estimation. The estimation of principal effects and controlled and natural direct and indirect effects require different substantive assumptions, different sensitivity analysis techniques and different approaches for obtaining bounds. These matters are beyond the scope of this paper but much work has already been done on characterizing the assumptions need for the estimation of or bounds on principal effects (Zhang and Rubin, 2003; Shepherd et al., 2006; Cheng and Small, 2006; Frangakis et al., 2007; Shepherd et al., 2007; Imai, 2008) and controlled and natural direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001; Robins, 2003; Kaufman et al., 2005; Avin et al., 2006; Peterson et al., 2006). Certain studies may more plausibly satisfy the assumptions required for inference about principal direct effects; other studies may more plausibly satisfy the assumptions required for inference about controlled and natural direct and indirect effects.

A final consideration concerns scientific goals. In the context of mediation, it is not always clear that knowing about the presence of principal strata direct and indirect effects will be of particular use. In fact, Joffe et al. (2007) have recently argued that within the context of mediation, the estimation of average controlled and natural direct and indirect effects is of greater interest than the estimation of average effects within principal strata, both for the purposes of making treatment decisions and for the purposes of explanation and identifying causal mechanisms. The utility of estimating average effects within principal strata for the purposes of making treatment decisions is limited because it will often be difficult to identify to which principal strata an individual belongs. In contrast, estimates of average controlled and natural direct and indirect effects concern the whole population (or alternatively a subset defined by observed covariates). Furthermore, unlike the principal stratification approach, the estimation of natural direct and indirect effects also allows for the decomposition of the total effect into direct and indirect components and thus for the estimation of the relative contribution of different causal pathways by which the treatment may be operating. A number of the applications of principal stratification, such as “truncation-by-death” (Frangakis and Rubin, 2002; Zhang and Rubin, 2003; Imai, 2008; Frangakis et al., 2007), do not concern questions of mediation. However, in the context of mediation, when interventions on the intermediate variable are conceivable, and when the identification conditions hold, the estimation of controlled and natural direct and indirect effects are arguably more informative than examining the presence of principal strata direct and indirect effects.

References

Avin, C., Shpitser, I., Pearl, J., 2006. Identifiability of path-specific effects. In: Proceedings of the International Joint Conferences on Artificial Intelligence, pp. 357–363.

Barnard, J., Frangakis, C.E., Hill, J.L., Rubin, D.B., 2003. A principal stratification approach to broken randomized experiments: A case study of School Choice vouchers in New York City with discussion. J. Amer. Statist. Assoc. 98, 299–323.

Cheng, J., Small, D.S., 2006. Bounds on causal effects in three-arm trials with non-compliance. J. Roy. Statist. Soc., Ser. B 68, 815–836.

Cole, S.R., Hernán, M.A., 2002. Fallibility in estimating direct effects. Int. J. Epidemiol. 31, 163–165.

Didelez, V., Dawid, A.P., Geneletti, S., 2006. Direct and indirect effects of sequential treatments. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, pp. 138–146.

Frangakis, C.E., Rubin, D.B., 2002. Principal stratification in causal inference. Biometrics 58, 21–29.

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

Geneletti, S., 2007. Identifying direct and indirect effects in a non-counterfactual framework. J. Roy. Statist. Soc., Ser. B 69, 199–216.

Imai, K., 2008. Sharp bounds on causal effects in randomized experiments with "truncation-by-death". Statist. Probab. Lett. 78, 144–149.

Joffe, M., Small, D., Hsu, C.-Y., 2007. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. Statist. Sci. 22, 74–97.

Kaufman, S., Kaufman, J.S., MacLehose, R.F., Greenland, S., Poole, C., 2005. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. Stat. Med. 24, 1683–1702.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82, 669–688.

Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.

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

Peterson, M.L., Sinisi, S.E., van der Laan, M.J., 2006. Estimation of direct causal effects. Epidemiol. 17, 276–284.

Robins, J.M., Greenland, S., 1992. Identifiability and exchangeability for direct and indirect effects. Epidemiol. 3, 143–155.

Robins, J.M., 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P., Hjort, N.L., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford University Press, New York, pp. 70–81.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701.

Rubin, D.B., 1978. Bayesian inference for causal effects: The role of randomization. Ann. Statist. 6, 34–58.

Rubin, D.B., 1990. Formal modes of statistical inference for causal effects. J. Statist. Plan. Infer. 25, 279–292.

Rubin, D.B., 2004. Direct and indirect effects via potential outcomes. Scand. J. Statist. 31, 161–170.

Rubin, D.B., 2005. Causal inference using potential outcomes: Design, modeling, decisions. J. Amer. Statist. Assoc. 100, 322–331.

Shepherd, B.E., Gilbert, P.B., Jemiai, Y., 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., Lumley, T., 2007. Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. J. Amer. Statist. Assoc. 102, 573–582.

Zhang, J.L., Rubin, D.B., 2003. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". J. Ed. Behav. Statist. 28, 353–368.