The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 22

Reaction to Pearl's Critique of Principal Stratification

Arvid Sjolander, *Karolinska Institute*

Recommended Citation:

Sjolander, Arvid (2011) "Reaction to Pearl's Critique of Principal Stratification," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 22.

DOI: 10.2202/1557-4679.1324

Reaction to Pearl's Critique of Principal Stratification

Arvid Sjolander

Abstract

This Reader's Reaction contains some brief remarks regarding Pearl's concerns regarding the value of principal stratification.

KEYWORDS: causal inference, principal stratification, truncation

I thank the Editor for giving me the opportunity to comment on Judea's paper on Principal Stratification (PS). Judea argues that parameters derived through PS are often of little scientific interest. I am compelled to agree, to some extent, and would like to contribute with some further discussion of two specific parameters; the Principal Stratum Direct Effect (PSDE), and the Survivor Average Causal Effect (SACE) proposed in studies suffering from 'truncation by death'.

1 The PSDE

Several authors, me included (Sjölander et al, 2007), have advocated the use of PSDEs in scenarios where the aim is to estimate the direct effect of an exposure (X) on an outcome (Y), and the potential mediator (Z) is not manipulable. The argument goes as follows: if Z is not manipulable, then the counterfactual variable Y_{xz} is not well defined, which implies that the Controlled Direct Effect (CDE) and the Natural Direct Effect (NDE) are not well defined either. This problem is bypassed by the PSDE, which only requires Y_x and Z_x to be well defined. Thus, even though we may not have a specific interest in the PSDE *per se*, it may be the only choice we have, given that we are interested in direct effects.

I object to this argument. Specifically, I think that the reason for Y_{xz} being ill-defined also deprives the PSDE of its desired interpretation as a direct effect. Thus, when we focus on PSDEs we effectively throw the baby out with the bathwater; we solve the problem of ill defined counterfactuals by sacrificing the research question.

To make the point, consider the following example, adapted from Sjölander et al (2007). The argument that I will make is close in spirit to the argument made by Robins et al (2007). Let X , Z and Y stand for physical activity (PA) level during one year, body mass index (BMI) at the end of the year (t_0), and the binary indicator (0/1) for cardiovascular disease (CVD) during the following year, respectively. To simplify, suppose that X is randomized and binary (0/1). The CDE and the NDE require Y_{xz} to be well defined. Is it? I would say no. For instance, $Y_{x,35}$ is the CVD indicator, for a given subject with PA level equal to x and BMI equal to 35. But it is easy to imagine a wide range of body compositions (e.g. an extreme excess of body fat or an extreme excess of muscles) which all produce a BMI equal to 35, but which may have very different effects on Y . Thus, without further specification the empirical meaning of Y_{xz} is not clear.

It is instructive to formulate the problem in Directed Acyclic Graphs (DAGs) and Non Parametric Structural Equation Systems (NPSEs). The DAG in Figure 1 displays the scenario. In Figure 1, Z^* is intended to repre-

Figure 1: DAG illustrating the use of PSDEs.

sent a detailed description of the body composition at t_0 (e.g. height, muscle distribution, fat distribution etc). A sufficiently detailed description may prevent crucially different interpretations of Y_{xz^*} , and may render CDE and NDE well defined, when using Z^* instead of Z . From a NPSE perspective, it may be reasonable to consider Z^* as appearing in the true function determining Y , which motivates the arrow from Z to Y . Technically speaking, it may be reasonable to consider Z^* as an ‘atomic module’ or an isolated piece of a bigger machinery, which in principle can be subjected to arbitrary manipulations. On the contrary, Z does not appear in the function for Y . Indeed, in this example I don’t view Z as an atomic module in the underlying NPSE system at all, but rather as a logical coarsening of (i.e. a proxy for) Z^* , as indicated by the double arrow from Z^* to Z . Thus, in the NPSE underlying the DAG in Figure 1, Z can only be manipulated through Z^* , and Y_{xz} is not well defined.

Figure 1 clearly shows the problem with the PSDE: since Z is just a coarsening of Z^* , it cannot logically mediate anything. Thus, although $\text{PSDE}(z)$ is a proper causal effect, it cannot be interpreted as a *direct* effect. Yet, it is easy to show that almost any parametrization of the model would give a PSDE that is 1) not equal to the total effect, thus suggesting mediation through Z , and 2) not equal to the ‘true’ direct effect (i.e. the effect not mediated through Z^*), thus giving wrong conclusions about the degree of ‘true’ mediation.

I have not yet found any practical example in the PSDE literature which I feel cannot be represented by the generic DAG in Figure 1. Thus, the reason for Y_{xz} being ill defined is, in my experience, typically that the presumed mediator Z is not a mediator at all, but a coarse function of a ‘true’ mediator in the underlying NPSE. And under these circumstances, the PSDE cannot be interpreted as a direct effect. Conversely, whenever the PSDE can be interpreted as a direct effect, I would consider Y_{xz} , and thus also the CDE and the NDE, as being well defined.

2 The SACE

In its simplest form, the problem of ‘truncation by death’ involves three observed variables. X is a treatment of interest, which is randomized at baseline, and Y is an outcome of interest, which is realized at the end of follow-up. Z is a binary (0/1) indicator of survival until end of follow-up. If $Z = 1$ (survival), then Y is realized and can be observed, if $Z = 0$ (death), then we define $Y = *$ (not realized). To simplify, suppose that X is binary (0/1). The scenario can be depicted by the DAG in Figure 2.¹ Clearly, a

Figure 2: DAG illustrating the problem of ‘truncation by death’.

¹Arguably, the counterfactual Y_{xz} is not well defined when Z is the crude indicator of survival. Thus, I have refrained from drawing an arrow from Z to Y . Arguing as for the PSDE we may view Z as a logical coarsening of a complex variable Z^* , sufficiently detailed to deserve the status of an ‘atomic module’ in the underlying NPSE system. The deterministic relation $Z = 0 \Rightarrow Y = *$ should then be viewed as shorthand for $Z_{z^*} = 0 \Rightarrow Y_{z^*} = *$.

comparison of the distributions $\Pr(Y|X = 1, Z = 1)$ and $\Pr(Y|X = 0, Z = 1)$ does not produce a causal effect of X on Y ; by restricting the analysis to the observed survivors (conditioning on $Z = 1$) we open the path $X \rightarrow Z \leftarrow U \rightarrow Y$.

The SACE restricts attention to those who would survive under *both* treatment levels - the always survivors. It is defined as some comparison between the distributions $\Pr(Y|X = 1, Z_1 = Z_0 = 1)$ and $\Pr(Y|X = 0, Z_1 = Z_0 = 1)$. The SACE is a proper causal effect, since it compares the same group of people under the two levels of a randomized treatment. But is it scientifically interesting? Generally, I think any causal effect deserves attention if 1) it can be used to improve decision making (e.g. about which treatment level to recommend in clinical practice), or 2) it generates important insights about the problem at hand (e.g. about etiology or biological mechanisms). I doubt whether the SACE meets any of these criteria.

Consider the first criterion. Suppose, rather unrealistically, that clinicians can somehow identify the always survivors as they arrive to the clinic. Then, obviously, the SACE can be used in clinical practice to determine the optimal treatment level for this group of patients. In reality however, the group of always survivors is typically non-identifiable, and the best we can do without further assumptions is to derive bounds for the population proportion. It thus becomes less clear whether the SACE is of any practical/clinical use. An additional complication is that the SACE itself is typically non-identifiable, and can at most be bounded without further assumptions. Thus, a clinician who asks for guidance may in practice be left with bounds for a causal effect, which applies to a non-identifiable group of patients.

When the always survivors cannot be identified, then, for any given patient the treatment may have an effect on survival Z as well as on the outcome Y , which clearly must be taken into account.² A natural approach is to consider Z and Y as a joint outcome, and compare the distributions $\Pr(Z, Y|X = 0)$ and $\Pr(Z, Y|X = 1)$. This comparison is not hampered by any identification problems, it can be given a causal interpretation, and it may be easier to interpret than the SACE.³

²If X has no effect on Z for any patient, then the path $X \rightarrow Z \leftarrow U \rightarrow Y$ no longer exists, and the analysis restricted to the observed survivors ($Z = 1$) produces a causal effect. Whether X affects Z or not can easily be tested by comparing $\Pr(Z = 1|X = 1)$ and $\Pr(Z = 1|X = 0)$.

³A parallel can be drawn to more standard survival analysis; comparing $\Pr(Y|X = 1, Z = 1)$ and $\Pr(Y|X = 0, Z = 1)$ is like comparing hazards, whereas comparing

What about the second criterion? Does the SACE contribute with any important insights about the problem depicted in Figure 2? I fail to see this. Although the SACE is mathematically flawless, its definition is rather abstract and the connection to any biological mechanism is unclear.⁴ I admit though, that this is to some extent an issue of subject matter expertise (and to a large extent a matter of opinion), and I'm happy to be persuaded the opposite by the other discussants.

References

- Hernan M. (2010). The hazard of hazard ratios. *Epidemiology* 21(1): 13-15.
- Robins JM, Rotnitzky A, Vansteelandt S. (2007). Discussion of Principal stratification designs to estimate input data missing due to death. by Frankakis CE, Rubin DB, An M, MacKenzie E. *Biometrics* 63:650-654.
- Rothman KJ, Greenland S. (1998). *Modern Epidemiology*, 2 ed. Lippincott Williams & Wilkins: Philadelphia.
- Sjölander A, Humphreys K, Vansteelandt S, Bellocco R, 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(2): 514-520.

$\Pr(Z, Y|X = 0)$ and $\Pr(Z, Y|X = 1)$ is like comparing survival curves. Hernan (2010) noted that hazard ratios cannot in general be given a causal interpretation, and recommended a comparison of survival curves for causal inference.

⁴A parallel can be drawn to the concept of 'biological interaction', which was formally defined by Rothman and Greenland (1998) (RG). RG's definition relies on a classification of each subject into one of 16 possible response types with respect to binary exposures A and B, i.e. a PS. They proved that the presence of principal strata for which A and B interacts 'biologically' can be inferred by testing for additive interactions in a statistical model. RG clearly stated that their definition of biological interaction 'does not depend on any specific mechanistic model for the disease process'. Nevertheless, it is in my experience not uncommon that reviewers and senior epidemiological researchers dismiss the use of standard multiplicative models, claiming that additive models are more informative about biological mechanisms.