The International Journal of Biostatistics

Manuscript 1391

"Imagine a Can Opener"—The Magic of Principal Stratum Analysis

Philip Dawid, *University of Cambridge*
Vanessa Didelez, *University of Bristol*

“Imagine a Can Opener”—The Magic of Principal Stratum Analysis

Philip Dawid and Vanessa Didelez

Abstract

We extend Pearl's criticisms of principal stratification analysis as a method for interpreting and adjusting for intermediate variables in a causal analysis. We argue that this can be meaningful only in those rare cases that involve strong functional dependence, and even then may not be appropriate.

KEYWORDS: causal inference, principal stratification, instrumental variables

Author Notes: Vanessa Didelez's work was supported by grants from Leverhulme Trust RF-2011-320 and EPSRC

Pearl (2011) does a valuable service in opening up the methodology of principal stratification to detailed scrutiny. In our opinion this is overdue, and we welcome this opportunity to extend that scrutiny.

The principal stratum approach to adjusting for intermediate variables is to define one's way out of the problem. By a magical invocation of the mathematical construct of "potential outcomes", we conjure up a pre-treatment variable: the principal stratum. We can then interpret the intermediate variable as a partial observation on that. Since we think we know how to adjust for pre-treatment variables, the rest is mere technicality. This is a prime example of the pure mathematician's[1] approach to a practical problem: stranded on a desert island with no food supplies but a crate of tinned tuna, he imagines a can opener.

But what if there is no real can-opener—no real-world pre-treatment variable corresponding to the fictitious principal stratum? There is then no way of determining which principal stratum an individual belongs to. How can a principal stratum analysis then tell us anything relevant about the real world?

We are thus in general agreement with Pearl's position that—while there can be some special cases where a principal stratum analysis can make sense—in many cases it does not. In our opinion this latter situation will always apply, except in the rare case that the principal stratum variable can be identified with a real-world pre-treatment variable. And determining when such an exceptional case arises has to involve external subject matter considerations: it can not be purely a question of formal symbol manipulation. Similar distinctions have been made by Berzuini and Dawid (2010) in the context of defining mechanistic interaction.

In the sequel we elaborate on the above points with specific reference to instrumental variable analysis.

1 Determinism: Reality or formality?

Many problems of causal inference have been addressed in the framework of potential outcomes and counterfactuals. In past work, we have shown that many of these can be reformulated in terms of a "decision-theoretic" approach that does not rely on counterfactual constructs or reasoning (Dawid, 2000, 2002, 2007; Dawid and Didelez, 2010; Dawid, 2012), showing that even though many people find potential outcomes helpful they are often inessential. However, the idea of principal strata is inherently counterfactual and consequently we struggle to reformulate this approach in decision-theoretic terms.

[1] This old joke has also been told of economists, traffic consultants, ...

We begin with some general comments on non-parametric structural equations and potential outcomes, and then return to the specific case of principal strata.

Pearl's starting point is the functional equation[2]:

$$X = f(Z,U) \tag{1}$$

relating an exposure variable X to an input variable Z (both binary, for simplicity), and a unit-specific variable U, which has the status of a "pre-treatment" variable. More specifically, Pearl says that "U may stand either for the identity of a unit (*e.g.*, a person's name) or, more functionally, for the set of unit-specific characteristics that are deemed relevant to the relation considered." We will argue that these are two very different interpretations of U.

Pearl regards the functional relationship (1) as equivalent to a potential outcome formulation, and at a purely formal level this is indeed so. For if we start with (1), we can construct the pair of potential exposures (the mapping variable) as functions of U:

$$\mathbf{X} = (X_0, X_1) := (f(0,U), f(1,U)). \tag{2}$$

Conversely, if we start with a potential exposure pair $\mathbf{X}$, we can construct a functional model of the form of (1) on taking

$$U \equiv \mathbf{X} \tag{3}$$

$$f(z,(x_0,x_1)) := x_z. \tag{4}$$

In particular, we can use either the functional or the potential outcome representation to define principal strata.

However, this formal identity hides a fundamental difference of interpretation.

If we regard U as "a set of unit-specific characteristics that are deemed relevant to the relation considered", say age, sex, patient history *etc.*, this implies that it should be possible to measure (or at least to conceive of measuring) U. But then the functional form of (1) constitutes a very strong assumption of *determinism*: that, once we have measured these attributes (for a given unit), we will be able to predict, *without error*, exactly what value that unit's exposure X will take, in response to either input value for Z.[3] Such Laplacian determinism is out of favour as a general

[2]We have changed the symbols from Pearl's eqn. (1) to avoid a clash with the uses of X and Y below.

[3]A further assumption implicit in (1) is that the identical functional relationship holds, between the identical variables, no matter how Z is assigned—be it by the free choice of an experimenter or by the stochastic whim of Nature. Similarly, a potential exposure X_z is implicitly supposed to have a single unique value, no matter how Z may be assigned. Such implicit conditions of "stability across regimes" are needed to justify causal inference from observational data in these frameworks.

scientific or philosophical principle, and it seems odd to make it the cornerstone of a general theory of causality—especially since there exist non-deterministic alternatives, such as the decision-theoretic approach mentioned above. (An additional philosophical objection is that, if we take Laplacian determinism seriously, no scope remains for the exercise of free will, *e.g.* in choosing how to intervene in a system—and causal questions degenerate into meaninglessness.)

On the other hand, if instead we regard U as a label identifying the unit itself, then the potential outcome approach takes for granted the existence—indeed co-existence—of the various potential exposures (X_z) for each unit; but it is not necessarily assumed that there is any measurement on the unit we could make in the real world that would reveal the values of all these variables. In that case we can not construct a functional model (1) *in which U can be given a real-world interpretation* (certainly (3) will not so serve), and so the functional model (4) is valid only in a purely formal sense.

In our view, it will only be in very special circumstances that we can take the deterministic functional model (1) seriously, with U representing quantities that are (at least in principle) measurable (Berzuini and Dawid, 2010). In such a case, we could also take seriously the real-world existence of $\mathbf{X}$. However, in most applications this would overstretch credulity. Formally indistinguishable expressions of principal stratum concepts and arguments, while they may sometimes be meaningful in the former case, are always meaningless and misleading in the latter.

Several of the other commentators on Pearl's paper (Sjölander, 2011; Joffe, 2011; Prentice, 2011) have expressed the concern that there are situations where it is difficult or impossible meaningfully to assume the existence of the potential response of one variable to the setting of another. Their arguments and examples have force, but they do not clearly identify when such an assumption might be meaningful. In our view this is only in the very untypical case of functional dependence on real-world variables (and, in contrast to certain other views, is in no way related to whether or not the putative causal variable is manipulable).

2 Instrumental variables

The DAG of Figure 1 describes the general "instrumental variable" set-up (Bowden and Turkington, 1984; Imbens and Angrist, 1994; Angrist et al., 1996; Greenland, 2000; Pearl, 2009; Hernán and Robins, 2006).

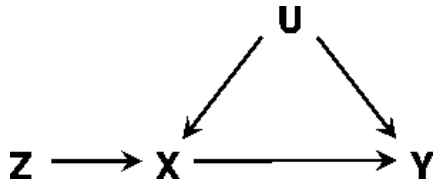

Figure 1: Instrumental variable

Here X is the treatment variable, Y the response variable, Z the instrumental variable, and U a set of unobserved "confounders". For simplicity we assume both Z and X are binary. This DAG is nothing more nor less than a graphic way of displaying the following conditional independence properties (Dawid, 1979), assumed to hold in the observational setting:

$$U \perp\!\!\!\perp Z \tag{5}$$

$$Y \perp\!\!\!\perp Z \,|\, (X, U). \tag{6}$$

In particular, the arrows serve a purely incidental function and should not immediately be given causal interpretations. To support causal interpretation and inference, the DAG needs further elaboration to connect the observational behaviour of the system with its behaviour under real or hypothesised interventions: see Dawid (2003); Didelez and Sheehan (2007b); Dawid (2010). We shall here take this as done, but not dwell on it.

Under (5) and (6) and suitable additional model assumptions (*e.g.*, linearity—see Didelez et al. 2010 for an overview and summary), it is possible to identify the average causal effect (ACE) of X on Y from observational data on (Z,X,Y). Without these additional assumptions ACE is typically unidentifiable, although it may be possible to set bounds on it (Robins, 1989; Manski, 1990; Balke and Pearl, 1994; Dawid, 2003).

The principal stratum formulation of this problem (Imbens and Angrist, 1994; Angrist et al., 1996; Imbens and Rubin, 1997) applies to the case that Z is externally set or randomized[4], and introduces the potential exposures (X_0,X_1) of X to the setting of Z, and the potential responses (Y_0,Y_1) of Y to the setting of X. In the context of Figure 1, and as discussed in § 1 above, this is formally equivalent to assuming we have functional dependencies $X = f(Z,U)$ and $Y = g(X,U)$: then $X_z := f(z,U)$, $Y_x := g(x,U)$. In that case we can even replace U by (X_0,X_1,Y_0,Y_1). But again, we regard this formal equivalence as of little interest outside the very restrictive case that the initial variable U is, in principle, measurable in the real world—in which case so are X_0, X_1, Y_0, Y_1. And only in the rare cases that we can

[4]See § 2.2, point 2 for why this is already a restrictive assumption

accept this could it be of interest to focus attention on the *local average treatment effect (LATE)*:

$$\mathrm{E}(Y_1 - Y_0 \,|\, X_0 = 0, X_1 = 1) \tag{7}$$

the causal effect of X on Y within a specific principal stratum (those individuals for whom $X_0 = 0, X_1 = 1$)—which in this very special case describes a condition that can in principle be verified before treatment, because it is fully determined by the measurable pre-treatment variable U.

It is the applied context that must dictate whether the determinism assumptions underlying this formulation are apppriate, and so whether the concept of local average treatment effect is meaningful.

2.1 Incomplete compliance

In the scenario of *incomplete compliance*, Z denotes treatment assignment—it being understood that a subject may not comply with that assignment. The variable U incorporates attributes of both the individual subject and the surrounding context that might influence both his compliance behaviour and his outcome. The LATE now becomes the *complier causal effect*.

We could only determine which principal stratum a subject belongs to by knowing in advance how he would react to either treatment assignment—requiring deterministic dependence of X on (Z,U). However in general such reactions are stochastic and unpredictable, and the principal stratum is not meaningful (a criticism also raised by Sjölander, 2011).

We explore the possibility that a principal stratum *might* be meaningful by two different stories.

2.1.1 Homo Economicus

An Economist might treat each subject as an ideally rational agent, with a coherent preference pattern, who can be guaranteed to choose the most preferred available option—*e.g.*, by acting so as to maximise expected utility. If U specifies the subject's preference pattern, as well as enough detail to determine (together with the other variables in the problem) a well-defined consequence, then, knowing U, we can in principle predict exactly how the subject will react to being assigned either treatment: the technique of *control variables* (Heckman and Robb, 1985) involves building models of such behaviour. In such a case X will be completely determined by (Z,U) as in (1), and since U represents real-world, not just formally defined, quantities, we can construct meaningful principal strata. Given the right information (which may be no mean task), we could identify, in advance, which stratum

any individual belongs to; and it might well be of scientific interest to study (say) the causal effect of treatment on those who are thus classified as compliers.

2.1.2 Homo Psychologicus

An Experimental Psychologist objects that her subjects do not behave as if they had a fully coherent preference pattern; and even when they announce in advance just how they would react in various hypothetical circumstances, they often react differently when those circumstances are realised. But if we can not take the Economist's story seriously, then we have to regard the relationship between X, Z and U as incorporating an irreducibly stochastic element. Lacking now the fully functional dependence of (1), we no longer have real-world access to potential exposures or principal strata, and it does not even make sense to ask which principal stratum an individual belongs to—still less to estimate a causal effect for such a non-existent group of individuals.

2.2 Mendelian randomization

Mendelian randomization (Katan, 1986; Davey Smith and Ebrahim, 2003; Didelez and Sheehan, 2007a) is an important application of instrumental variable analysis. Here Z refers to a gene[5] that is associated with the phenotype X, a putative risk factor for disease outcome Y. We here list two reasons, over and above our general philosophical objections, why one should *not* use principal stratum analysis in this context (for examples where this has been done, see Shinohara et al., 2012; von Hinke Kessler Scholder et al., 2010a,b).

1. **The complier causal effect is not of interest**
   Let us, for the sake of argument, even suppose "biological determinism", in the sense that, in Nature, the phenotype X is fully determined by the gene Z and some background biological variable U, such as possession of some other gene. A complier is now some one whose phenotype X must be the "same" (in a suitably defined sense) as her gene Z—a property which is determined by U.

   However, in any one individual Nature will have determined just one value of the gene Z—while we ourselves may be interested in (say) the impact of population level policy interventions, but have no interest in modifying a specific individual's phenotype X by manipulating her gene Z, even were we able to do so. So even when there is no problem in defining the stratum

[5] We use this term loosely to include genotype, SNP, *etc.*

of compliers, it is simply not of interest (Joffe, 2011). This contrasts with the case of partial compliance, where compliers—if they exist—could be of direct interest, since in real life we will be able to encourage, but not force, patients to take a drug (see also the commentary by VanderWeele, 2011).

2. **The complier causal effect may not be well-defined**
The definition of the complier causal effect relies on Z being a *causal* (functional) gene for X, in the sense that if we (or Nature) were to intervene in the individual's genome to change Z (and nothing else), then that would result in a change of the value or distribution of X, at least for some individuals. However, it is commonly the case that the relationship between the genetic instrument Z and the phenotype X is not itself causal, but rather Z is in linkage disequilibrium with the true functional gene for X; then this premise is violated. In other approaches to instrumental variable analysis, it is generally enough to require a robust *association*—but not necessarily "causal" in any sense—between the instrument Z and the exposure X (Hernán and Robins, 2006; Didelez and Sheehan, 2007a, § 4.3, Fig. 3). However, this is not enough to define a complier causal effect, which does require that Z be causal for X. If we were to use an instrument Z that is associated with, but not necessarily causal for, X, then it is not clear what the principal stratum method is estimating (Joffe, 2011, §2.2). Formally, if Z is not a causal gene for X then the potential outomes X_1 and X_0 must be equal, because intervening to set Z to different values (without changing anything else) has no effect on X. In such a case there are no compliers.

3 Conclusions

We have focused our comments on instrumental variable analysis, which is just one of the problem areas where principal strata are used. Other problem areas discussed by Pearl and his commentators include direct/ indirect/ mediated effect analyses, surrogate outcomes, and truncation by death. Our reservations apply equally to these other applications. In particular, in the last case we consider that it is typically not helpful to focus on the effect of treatment in the principal stratum of "those who survive in any case", since individuals can not be so identified from information knowable at the time of the treatment decision. The decision-theoretic alternative suggested by Sjölander (2011) and Joffe (2011) (see also Dawid, 2000, §5 of Rejoinder) supplies a more directly meaningful approach.

References

Angrist, J., Imbens, G., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455.

Balke, A. A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In de Mantaras, R. L. and Poole, D., editors, *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 46–54.

Berzuini, C. and Dawid, A. P. (2010). Deep determinism and the assessment of mechanistic interaction between categorical and continuous variables. `arXiv:1012.2340`.

Bowden, R. J. and Turkington, D. A. (1984). *Instrumental Variables*. Cambridge University Press.

Davey Smith, G. and Ebrahim, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32:1–22.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B*, 41:1–31.

Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, 95:407–448.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189. Corrigenda, *ibid*., 437.

Dawid, A. P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with Discussion). In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 45–81. Oxford University Press.

Dawid, A. P. (2007). Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London. 94 pp. `http://www.ucl.ac.uk/statistics/research/pdfs/rr279.pdf`.

Dawid, A. P. (2010). Beware of the DAG! In Guyon, I., Janzing, D., and Schölkopf, B., editors, *Proceedings of the NIPS 2008 Workshop on Causality*, volume 6 of *Journal of Machine Learning Research Workshop and Conference Proceedings*, pages 59–86. `http://tinyurl.com/33va7tm`.

Dawid, A. P. (2012). The decision-theoretic approach to causal inference. In Berzuini, C., Dawid, A. P., and Bernardinelli, L., editors, *Causal Inference: Statistical Perspectives and Applications*, pages 25–42. John Wiley and Sons Ltd., Chichester.

Dawid, A. P. and Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistical Surveys*, 4:184–231.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25:22–40.

Didelez, V. and Sheehan, N. A. (2007a). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330.

Didelez, V. and Sheehan, N. A. (2007b). Mendelian randomisation: Why epidemiology needs a formal language for causality. In Russo, F. and Williamson, J., editors, *Causality and Probability in the Sciences*, volume 5 of *Texts In Philosophy*, pages 263–292. College Publications, London.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29:722–729.

Heckman, J. and Robb, R. (1985). Alternative methods for estimating the impact of interventions. In Heckman, J. and Singer, B., editors, *Longitudinal Analysis of Labor Market Data*, pages 156–245. New York: Cambridge University Press.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17:360–372.

Imbens, G. W. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62:467–476.

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

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

Katan, M. B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *The Lancet*, 327(8479):507–508.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge. Second Edition.

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

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

Robins, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In Sechrest, L., Freeman, H., and Mulley, A., editors, *Health Service Research Methodology: A Focus on AIDS*, pages 113–159. NCSHR, U.S. Public Health Service.

Shinohara, R. T., Frangakis, C. E., Platz, E., and Tsilidis, K. (2012). Designs combining instrumental variables with case-control: Estimating principal strata causal effects. *International Journal of Biostatistics*, 8(1). Article 2.

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

VanderWeele, T. J. (2011). Principal stratification—uses and limitations. *International Journal of Biostatistics*, 7(1). Article 28.

von Hinke Kessler Scholder, S., Smith, G. D., Lawlor, D. A., Propper, C., and Windmeijer, F. (2010a). Child height, health and human capital: Evidence using genetic markers. CMPO Working Paper 10/245, University of Bristol.

von Hinke Kessler Scholder, S., Smith, G. D., Lawlor, D. A., Propper, C., and Windmeijer, F. (2010b). Genetic markers as instrumental variables: An application to child fat mass and academic achievement. CMPO Working Paper 10/229, University of Bristol.