Commentary

Ian Shrier*, Jay S. Kaufman, Robert W. Platt, and Russell J. Steele

Principal Stratification: A Broader Vision

***Corresponding author: Ian Shrier,** Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC H3T 1E2, Canada, E-mail: ian.shrier@mcgill.ca
Jay S. Kaufman: E-mail: jay.kaufman@mcgill.ca, **Robert W. Platt:** E-mail: robert.platt@mcgill.ca, Department of Epidemiology, Biostatistics and Occupational Health, Faculty of Medicine, McGill University, Montreal, QC, Canada
Russell J. Steele, Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada, E-mail: steele@math.mcgill.ca

1 Introduction

Recently, Pearl [1] challenged researchers to comment on whether the principal stratification framework (PS) is an objective or a tool. A series of commentaries and responses ensued [2–6] that have heightened interest in the approach originally proposed by several authors [7, 8], and later more formally defined by Frangakis and Rubin [9]. In this brief article, we address the specific issue of compliance in experimental studies.

The PS literature uses the taxonomy "compliers" to refer to participants who *would* follow the treatment assignment under all treatment arms (a baseline characteristic). However, the clinical literature uses "compliers" to refer to participants who *did* follow assigned treatment. To minimize confusion, we will use "adherence" or "adherers" [10–13] to refer to observed concordance between assigned and observed treatment, and "Baseline Compliers" to refer to those participants with baseline characteristics that would follow assigned treatment regardless of which treatment assignment they received.

Pearl [1] outlines the general principles of PS categorization succinctly:

> The population of units can be partitioned into a set of homogeneously responding classes, called "equivalence classes" ... such that all units in a given class respond in the same way to variations in X. [i.e. the assignment]

Pearl notes that this categorization, based on "response-type classification" (i.e. on the counterfactual outcomes for adherence under both treatment assignments) [7, 14], is considered advantageous, because it is more parsimonious compared to trying to determine equivalence classes based on unobserved baseline characteristics. We use causal diagrams to argue that standard compliance PS has limited value in estimating the adherence-based effects of clinical interventions in the absence of very strong assumptions that are unlikely to hold. Section 2 reviews the fundamental issues of non-adherence-based analyses, Section 3 uses causal diagrams to review weaknesses with the ways in which PS are currently used and Section 4 concludes by using concrete examples of the challenges raised and suggesting possible future directions that might lead to solutions.

2 Statement of the problem of non-adherence

In a randomized trial, the intention-to-treat (ITT) analysis provides an unbiased estimate of the causal effect of treatment assignment and, if there is 100% adherence, of the causal effect for receiving treatment.

In the context of non-adherence, Angrist et al. [14], Balke and Pearl [7], Imbens and Rubin [8] and others [9, 15] generally group participants into one of four strata according to their response-type. The potential outcomes for dichotomous exposure can be represented by the pair (X_0, X_1), where X_0 is the received treatment under assignment to control ($X_0 = 0$: received control; $X_0 = 1$: received active treatment) and X_1 is the received treatment under assignment to active treatment ($X_1 = 0$: received control; $X_1 = 1$: received active treatment). For now, we will assume that both groups in the study have access to and can receive either treatment. The four possible classes (adherence-based principal strata) defined by the pairs of values (X_0, X_1) are as follows:

1. "Always Takers" (AT): The participant will receive the active treatment regardless of assignment ($X_0 = X_1 = 1$).
2. "Baseline Compliers" (BC): The participant will receive the active treatment if prescribed active treatment and receive the control treatment if prescribed control ($X_0 = 0$; $X_1 = 1$).
3. "Never Takers" (NT): The participant will receive the control treatment regardless of assignment ($X_0 = X_1 = 0$).
4. "Defiers": The participant will receive the control treatment if prescribed active treatment or will receive the active treatment if prescribed control ($X_0 = 1$; $X_1 = 0$). In most (but not all) settings, it is typical to assume no defiers [7, 14].

Imbens and Rubin [8] posit that these strata should be considered pre-treatment variables, because the potential outcomes exist prior to receiving treatment, at least in the context of the current study.

Participants' observed behaviors give some indication as to their stratum memberships. The participants who take active treatment when assigned to active treatment (i.e. adherers in the active treatment group) belong to either the AT or BC principal strata, and those who take control treatment if assigned to the control group (i.e. adherers in the control group) belong to either the BC or NT principal strata. Under the no-defiers assumption, non-adherent participants assigned control who take active treatment are AT, and non-adherent participants assigned active treatment who take control are NT.

If the strata represent levels of a pre-treatment variable, they are not affected by treatment (assigned or received) even though one empirically classifies participants into a stratum, or the union of two strata, based on information obtained from treatment assignment and data observed after treatment. Imbens and Rubin [8] also explicitly state that the response-type groups define the principal strata and note that the reasons why one might have ended up in one of the response-type groups are not relevant in the classification schema. For example, a participant assigned to the active treatment arm who is accidentally given the wrong medication is considered an NT in the context of the study, even though the participant would have taken the active treatment if no error had been made [8]. In this sense, the monikers "never", "always" and "compliers" are potentially misleading if interpreted literally. They refer only to the two-element vector of potential outcomes in the current study, not necessarily to any more general statement about what would be expected to occur in other scenarios or even in future repetitions of the exact same study.

3 Causal diagrams for principal stratification

We begin with a simple causal diagram and elaborate gradually. The important nuance that we propose in this comment is to view the principal strata as a categorization for the combined effects of all pre-treatment common causes of adherence; the principal strata themselves are not caused by and do not cause any other variables. Note that this is different from the causal structure presented or described by others [7, 14, 16], where the PS lies in the causal pathway between unmeasured variables and adherence. This small difference in perspective leads to important implications for clinical usefulness. For these diagrams, we assume the most general case where those assigned to active treatment can refuse to take active treatment and those assigned to control can take active treatment.

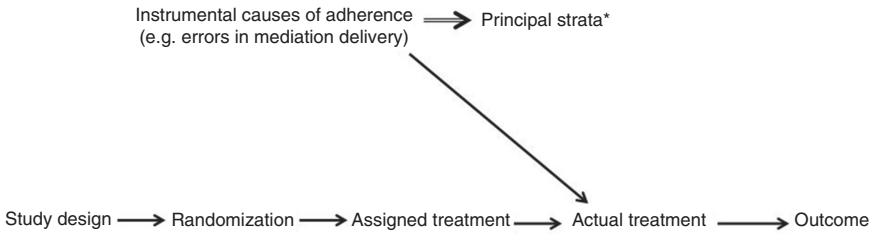

Figure 1 Causal diagram of a randomized trial with non-adherence. The principal strata are a classification (indicated by a double arrow as in Sjolander [17]) for the latent construct of the unmeasured causes of adherence. The asterisk indicates that there may be associated measurement error, since the principal strata are only partially identified by observed characteristics after the study is conducted. The unmeasured causes of adherence are not affected by the treatment or treatment assignment. They are called "instrumental", because they affect only the exposure (actual treatment) and not the outcome, except through the treatment

Figure 1 shows a causal diagram where the only reasons for adherence are unrelated to the outcome (e.g. some participants receive the wrong box of medication by mistake). The double arrow to principal strata indicates that the strata are simply a categorization and not an actual causal node [17]. In this scenario, actual treatment is the exposure of interest, and the latent variable "causes of adherence" is an instrumental variable for actual treatment in which it causes changes in actual treatment but has no effect on the outcome [14]. If this were the true causal diagram, an unbiased estimate of the average causal effect (ACE) of treatment receipt, as opposed to the ITT estimate for effect of treatment assignment, could be obtained by correctly conditioning on treatment received ("as-treated" or "per protocol" analyses) [18], and principal stratification would not be necessary.

In most cases, however, researchers believe that the baseline characteristics defining the PS also affect the outcome [19] as shown in Figure 2, which explicitly illustrates this assumption that there are additional baseline characteristics that are common causes of adherence to treatment and outcome. A trait such as susceptibility to "side effects" refers to baseline characteristics, such as an allergy to medication. The fact that we may only find out if a participant is allergic to medication after taking the treatment does not change the fact that this allergy is not caused by the treatment or the treatment assignment.

If Figure 2 were the true causal diagram, then the population ACE for treatment receipt could not be obtained unbiasedly by simply conditioning on treatment received, due to confounding by the common causes. Nonetheless, properly conditioning on all of the common causes of adherence and outcome could still recover the population ACE. In most settings, however, there are unmeasured common causes, making standard adjustment impossible and motivating the use of a compliance-stratum-specific average treatment effect estimate. Specifically, the ACE within the BC stratum (CACE) can be estimated even in the presence of

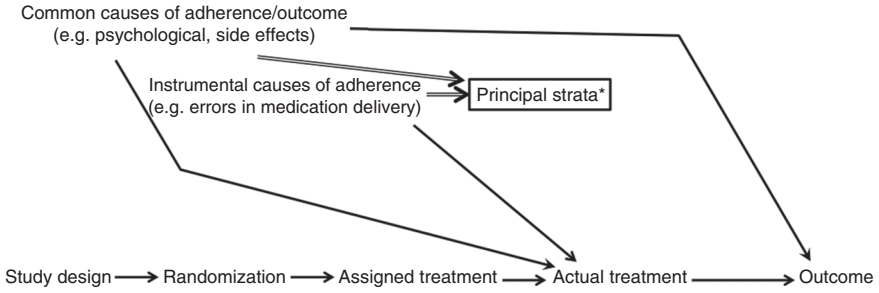

Figure 2 Causal diagram of a randomized trial with non-adherence. The principal strata* node again represents the classification of the latent causes of adherence in each context. In this diagram, causes of adherence are separated into instrumental causes, such as medication errors, and common causes of adherence and outcome, such as psychological traits

unmeasured common causes under well-known assumptions, which include monotonicity (i.e. no "defiers") and the exclusion restriction (i.e. no direct effect of assignment on outcome). Although the CACE may be considered deficient for population inference [20], it will be close to the population ACE to the extent that the conditioning on the PS approximates conditioning on the unmeasured common causes of adherence and outcome. For example, one could reasonably expect that those individuals whose "never taking" or "always taking" were due to instrumental mechanisms (i.e. those mechanisms not related to the outcome) would have characteristics similar to those found in the BC stratum and that the CACE provides an unbiased estimate of the causal effect of treatment receipt in these apparent "never takers" and "always takers". These instrumental causes of adherence effectively reduce the strength of the relationship between the common causes and the attained adherence.

In most studies, some causes of adherence may be measured, because they are known causes of the outcome, or they can be pursued as part of secondary objectives. In such studies, one could restrict the study sample to the subpopulation where adherence is mostly affected by common causes of adherence and outcome, thus assuring that PS would be more strongly associated with common causes of adherence and outcome. Figure 3 separates out the (instrumental) factors not affecting the outcome and the (common) factors that affect both adherence and outcome into unmeasured and measured groups.

Recall that principal stratum membership for the sampled subjects is partially unknown and, therefore, not directly applicable at the individual patient level when making recommendations for individual patients or entire populations. For these reasons, if all common causes could be measured and properly controlled for, many clinicians would prefer an estimate of the ACE, as this represents the ACE of patients actually receiving treatment over the entire patient population, rather than the CACE, applicable only to the unknown subpopulation of BC. Thus, a primary motivation for using a PS-based model is to replace the

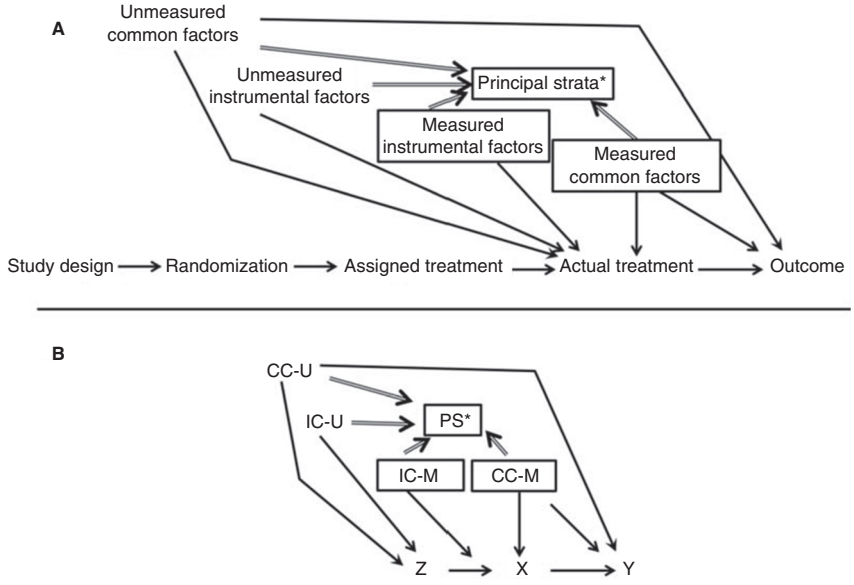

Figure 3 Panel A illustrates a causal diagram of a randomized trial with non-adherence. Factors affecting adherence (both instrumental causes and common causes of adherence and outcome) are now separated into unmeasured factors that cannot be conditioned on and measured factors that can be conditioned on (indicated by the surrounding box). The principal strata represent a classification of the latent measured and unmeasured adherence factors that affect the outcome, and the measured and unmeasured instrumental factors that only affect adherence. In a well-conducted randomized trial, the most important prognostic factors affecting outcome would presumably be measured, and some may be considered as the Measured Common Factors affecting the principal strata classification of participants. Panel B is an equivalent diagram, but the words have been replaced with symbols Z (randomization), X (treatment), Y (outcome), CC-U (unmeasured common causes of X and Y), CC-M (measured common causes of X and Y), IC-U (unmeasured instrumental cause of adherence), IC-M (measured instrumental cause of adherence) and PS (principal strata)

as-treated and per protocol estimators as approximations to the ACE, not to elevate the CACE to be the primary causal effect of interest. Additionally, generalization of the CACE to populations outside the study may be invalid if the effect of instrumental causes were different in different populations, even if all common causes (the actual confounding bias structure) remained the same.

Although uncertainty could exist as to whether a cause of adherence is instrumental or not, generalizability might still be improved compared to current PS methods, even if a weak common cause is considered instrumental. Still, more general limitations of the PS method remain and are described below.

4 Challenges and possible approaches to a solution

The CACE is necessarily limited to the context of the individual observed study, because this effect depends critically on the influence of the instrumental causes. Any change in context will lead to a different CACE, therefore limiting the clinical usefulness of the PS approach. As a concrete example, when controls do not have access to active treatment in the study, the current PS method (and any instrumental variable estimator) states that there are no ATs [8, 21]. However, once the active treatment is approved, ATs will very likely exist, and the CACE will be different due to the change in the membership of the BC stratum. Note that all else held constant, the ACE should not change, since granting access to treatment should not generally change the causal effect of receiving the treatment in the population of ATs.

Bellamy et al. [22] posit that the results from the original study where access is limited to active treatment are generalizable to the new context: "For example, consider a completed trial, in which we found beneficial treatment effects in compliers (CACE). Under the principal stratification approach, this finding can be disseminated to the public by policy makers on the assumption that compliers exist in the population." In addition, Vanderweele [6] argues:

> Contrary to what is suggested by Pearl, the effect [compliance average causal effect (CACE) measured by principal stratification or instrumental variable analysis when the control group has no access to treatment] is not merely an approximation to the population average treatment effect, but is arguably of intrinsic interest as it is the effect of treatment for the only group that we can reasonably induce to take treatment (the group that would take treatment if they were assigned treatment).

We disagree with this view and provide an example where a change in context affects the way in which subjects are classified into strata and, therefore, the interpretation of the CACE.

In one study, participants with on-going symptoms were randomized to an advanced ankle rehabilitation program or to no further rehabilitation (no access to the advanced rehabilitation program) and followed to determine reinjury risk [23, 24]. Although the classical compliance PS approach would assume that there are no ATs in this study, five participants in the control group sought additional rehabilitation outside the protocol. Therefore, given that these BC (by definition) sought additional treatment when the active treatment was not available to them, it is irrational to believe that they would be BC in a context once the treatment was available to them. As it happens, none of these participants was reinjured. Under the plausible assumption that these participants had baseline characteristics that included health-seeking behaviors that affected both adherence and outcome, the PS risk calculated for the BC in the control group for the study does not represent the risk for BC in the context of interest (because it included participants who would be AT in the real world context and who had a lower risk of injury). Thus, the CACE in the study context will not necessarily be equal to the CACE in a context where subjects have access to the advanced rehabilitation program.

Alternatively, one could try to estimate the risk in the PS that would occur in the context of clinical interest (in this case, when treatment is available to everyone). Adherence to assigned treatment is rarely all-or-none [25]. Therefore, one might be able to use "partial adherers" as a group of participants who are more likely to become "always takers" in a context where treatment becomes accessible to the control

group, creating two new principal substrata from the classical AT stratum. Using risk calculations for this substratum, with appropriate weighting according to proportions within each group, might provide some bounds for the causal effect within the clinical context of interest. Challenges to this approach include making explicit all of the necessary underlying assumptions, as well as estimation with potentially small numbers of participants and appropriate coverage for confidence intervals.

Changes to context may not only affect the PS assigned to individual participants but could also affect the strength of the relationship between causes and their effects (either causes of adherence or causes of outcome). Currently, the PS literature does not detail any methodology for converting the effects calculated from a context not representative of reality, into meaningful causal effects in the context of clinical interest. However, if one focused on the causes themselves as covariates, then the challenge appears to be the same as with any question of external generalizability from a randomized trial. If instrumental causes of adherence are excluded from the PS categorization, it might be possible to consider the PS as a latent variable within a causal diagram. Pearl and Bareinboim [26] have represented non-generalizability in terms of causal selection diagrams and illustrated how causal effects from one study in one context can be adjusted using observational data from a target context. In principle, this approach appears promising for PS as well, but further study is required to determine limitations and explicitly document the assumptions. Some particular issues that need to be considered include:

1. Defining the effects when the values for a subject's pre-treatment characteristics change with context even if causal relationships of these characteristics with respect to adherence remain the same, or do not remain the same;
2. The extent to which a participant's potential outcomes for compliance differ from those of the current single study;
3. The implications of a situation in which receiving the active treatment is the same across contexts for each subject.

5 Conclusions

We agree with Pearl [1] that the principal strata based on response-types represent an objective to categorize participants into clinically meaningful groups. However, failure to distinguish instrumental causes of adherence from common causes of adherence and outcome render the categorization problematic. However, we also believe that the underlying fundamentals of the principal stratification approach [8, 9] may still be applicable. If future work illustrates that one can restructure the PS to separate instrumental causes from common causes of adherences/outcome and that the results can be made generalizable using other existing analytical frameworks, then PS may become a clinically useful tool. Until then, researchers interested in causal effects within strata (or combinations of strata) should match their study design to the clinical context in which the active treatment will be used.

References

1. Pearl J. Principal stratification – a goal or a tool? Int J Biostat 2012;7:Article 20.
2. Baker SG, Lindeman KS, Kramer BS. Clarifying the role of principal stratification in the paired availability design. Int J Biostat 2012;7:Article 25.
3. Egleston BL. Response to pearl's comments on principal stratification. Int J Biostat 2012;7:24.
4. Gilbert PB, Hudgens MG, Wolfson J. Commentary on "principal stratification – a goal or a tool?" by Judea Pearl. Int J Biostat 2011;7:Article 36.
5. Joffe M. Principal stratification and attribution prohibition: good ideas taken too far. Int J Biostat 2011;7:Article 35.

6. Vanderweele TJ. Principal stratification – uses and limitations. Int J Biostat 2011;7:Article 28. Available at: http://bayes.cs.ucla.edu/jp_home.html
7. Balke A, Pearl J. Nonparametric bounds on causal effects from partial compliance data (R-199). UCLA Cognitive Systems Laboratory,1994: 1–25.
8. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. Ann Stat 1997;25:305–27.
9. Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002;58:21–29.
10. Kimmel SE, Troxel AB, Loewenstein G, et al. Randomized trial of lottery-based incentives to improve warfarin adherence. Am Heart J. 2012;164(2):268–274.
11. Jakicic JM, Wing RR, Butler BA, Robertson RJ. Prescribing exercise in multiple short bouts versus one continuous bout: effects on adherence, cardiorespiratory fitness, and weight loss in overweight women. Int J Obes Relat Metab Disord 1995;19:893–901.
12. Lerner BH. From careless consumptives to recalcitrant patients: the historical construction of noncompliance. Soc Sci Med 1997;45:1423–31.
13. Van Gool CH, Penninx BW, Kempen GI, Rejeski WJ, Miller GD, Van Eijk JT, et al. Effects of exercise adherence on physical function among overweight older adults with knee osteoarthritis. Arthritis Rheum 2005;53:24–32.
14. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;91:444–55.
15. Egleston BL, Cropsey KL, Lazev AB, Heckman CJ. A tutorial on principal stratification-based sensitivity analysis: application to smoking cessation studies. Clin Trials 2010;7:286–98.
16. Pearl J. Imperfect experiments: bounding effects and counterfactuals. Causality: models, reasoning and inference, 2nd ed. Cambridge, New York: Cambridge University Press, 2009:259–81.
17. Sjolander A. Reaction to pearl's critique of principal stratification. Int J Biostat 2011;7:Article 22.
18. Pearl J. Causality: models, reasoning and inference. Cambridge: University of Cambridge, 2000.
19. Dimatteo MR, Haskard KB, Williams SL. Health beliefs, disease severity, and patient adherence: a meta-analysis. Med Care 2007;45:521–8.
20. Robins JM, Greenland S. Identification of causal effects using instrumental variables: comment. J Am Stat Assoc 1996;91:456–8.
21. Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika 1999;86:365–79.
22. Bellamy SL, Lin JY, Ten Have TR. An introduction to causal modeling in clinical trials. Clin Trials 2007;4:58–73.
23. Hupperets MD, Verhagen EA, Van Mechelen W. Effect of unsupervised home based proprioceptive training on recurrences of ankle sprain: randomised controlled trial. Br Med J 2009;339:b2684.
24. Verhagen EA, Hupperets MD, Finch CF, Van Mechelen W. The impact of adherence on sports injury prevention effect estimates in randomised controlled trials: looking beyond the CONSORT statement. J Sci Med Sport 2011;14:287–92.
25. Vanderweele TJ. Mediation analysis with multiple versions of the mediator. Epidemiology 2012;23:454–63.
26. Pearl J, Bareinboim E. Transportability across studies: a formal approach (R-372-A). Los Angeles, CA: University of California, 2011:1–19.