Discussion of "Principal stratification designs ..."

James Robins, Andrea Rotnitzky and Stijn Vansteelandt

We are grateful for the opportunity to discuss this paper. In this discussion, we i) question the plausibility of the authors' substantive assumptions, ii) discuss the authors' choice of scientific goals and their attainability, iii) comment on statistical issues and iv) describe a sensitivity analysis approach to the authors' problem.

i) Substantive assumptions. In §3 the authors show that under no interference and assumptions (I) $(A, S(0), S(1)) \perp\!\!\!\perp Z|X, C = 1$ where C is the binary indicator of the critical event, e.g. car accident, and (II) $S(1) = 1$ w.p.1, an application of Bayes' theorem implies identification of the joint distribution $f(A, S|C = 1, X)$ from the distributions $f(A|S = 1, C = 1, X)$ and $f(S|C = 1, X)$.

Assumption 2 stipulates a dichotomous treatment factor Z which is guaranteed to prevent death. In the authors' example, Z was transport time to hospital, a continuous variable that was dichotomized at 10 mins. As the authors recognize, treatments Z satisfying (II) rarely exist. For instance, a fraction of individuals injured in car accidents die almost immediately. For them, a "transport time to hospital" of less than 10 mins cannot prevent death. Yet, the empirical analysis of §4 reports that no deaths occurred in subjects with transport times of less than 10 mins. Possible explanations for the lack of deaths in the rapidly transported would include (i) ambulance paramedics appropriately transport victims found dead at the scene less quickly than injured survivors, (ii) the chosen cutpoint of 10 minutes was data driven, and (iii) the number of high risk rapidly transported subjects (i.e. 11) was sufficiently small that all survived by chance.

In §5 the authors replace assumption (II) with the monotonicity assumption that Z cannot cause death. However, it can be difficult to find variables Z that satisfy monotonicity. For instance, the authors suggest that thrombolytic drug therapy after stroke is a treatment

that never causes death. Yet, physicians are well aware that thrombolytic drugs can cause intracerebral hemmorraghe and death. Similarly, rapid transport to a hospital may cause death if, in their hurry, the paramedics fail to properly stabilize the patient. Indeed, it is a matter of debate whether fast transport is harmful or beneficial for accident victims[1].

We also question the validity of assumption (I). Subjects with limited pre-accident physical mobility (X) both have difficulty with activities of daily living (A) and are difficult to quickly extract from a wrecked automobile. We doubt one could measure physical mobility sufficiently well to insure $A \perp\!\!\!\perp Z | X, C = 1$ holds, for Z "transport time".

In conclusion, we regard neither the monotonicity assumption (much less the stronger assumption (II)) nor the ignorability assumption (I) as plausible in the authors' examples.

ii) Scientific goals. In §2.1 the authors list their scientific goals as estimation of (a) $f(A|S = 0, C = 1)$ and (b) $P(S = 0|A = a, C = 1)$ as a function of a.

Attainability of the Authors Goals. The authors show that $f(A|S = 0, C = 1)$ and $P(S = 0|A = a, C = 1)$ are identified under (I) and (II). In fact, a calculation using Bayes rule shows that they are identified under the weaker assumptions $A \perp\!\!\!\perp Z | X, C = 1$ and $\mathrm{P}(S = 1|Z = C = 1, X) = 1$. These assumptions do not require any reference to or assumptions about counterfactuals. Unfortunately, the arguments in i) above show that these weaker assumptions are also unrealistic.

In spite of our concerns about the monotonicity assumption, we now examine whether the authors' goals are attainable when this assumption holds. In §5.1, the authors state that the importance of their approach 'is essentially intact for addressing scientific goals', even when (II) is replaced by the weaker monotonicity assumption. We disagree because $f(A|S = 0, C = 1)$ and $P(S = 0|A = a, C = 1)$ are not identified, and thus not consistently estimable, when only (I) and monotonicity are imposed.

Perhaps the authors' claim is predicated on the fact that under (I) and monotonicity, the

principal stratum distributions $f(A|C=1, P=$ 'Z prevents death') and $f(A\,|C=1, P=$ 'always survive') are identified. However, the following example demonstrates that knowledge of these distributions does not suffice to address important scientific questions when $f(A|S=0, C=1)$, and thus $f(A|C=1)$, remain unidentified.

Example: Suppose Z is an anti-bird flu drug that is in limited supply and $C=1$ is contracting bird flu. Clearly, all else being equal, we should give the drugs to those most likely to be helped by the drug. Thus, we would like to know if $P(Z$ prevents death$|A=1, C=1) > P(Z$ prevents death$|A=0, C=1)$ for, then, we should give the drug to subjects with $A=1$ rather than $A=0$. By Bayes theorem, this inequality is $P(A=1|C=1, P=$ 'Z prevents death') $/P(A=0|C=1, P=$ 'Z prevents death') $> P(A=1|C=1\,)/P(A=0|C=1)$. When Z does not cause death but is not guaranteed to prevent death, we can identify the left hand side of the inequality but we cannot identify its right hand side and thus cannot determine whether the inequality is true.

Relevance of the authors' goals: The preceding example illustrates that knowledge of $f(A|S=0, C=1)$ can help address substantive questions. However, we argue that $P(S=0|A=a, C=1)$ is not relevant for predicting survival when Z is available. If, as the authors assume, data on a strong predictor Z are available, then clearly $P(S=0|A=a, C=1, Z=z)$ is a more relevant predictive distribution than $P(S=0|A=a, C=1)$. Indeed, if Z were a widely available non-toxic medical treatment that never caused death, it would be unethical to withhold Z and so $P(S=0|A=a, C=1, Z=1)$ would be the only predictive distribution of interest. Note that this implies that obtaining data on Z is a good idea irrespective of whether data on A are missing.

The authors state that knowledge of $P(S=0|A=a, C=1)$ [or, when data on X are collected, of $P(S=0|A=a, C=1, X)$] helps "medical research understand the pathways through which those inputs relate to critical events and death". The authors did not provide

any justification for this claim. Furthermore, they did not define what they meant by the term "pathway". To evaluate the authors' claim we first clarify the meaning of this term. Because our discussion applies even when no data are missing, we may assume A is always observed.

The term 'pathways' is generally used as shorthand for 'causal pathways'. Consider the query: does A have a causal effect on survival S through a pathway that does not involve the critical event C?. This query is often rephrased as whether A has a direct causal effect on survival not through C. The concept of direct effect has been formalized in three different ways. Let $S(a)$ and $C(a)$ denote a subject's counterfactual survival and critical event outcome when A is set to a, which we take to be well-defined. The subject's observed data S and C are $S = S(A)$ and $C = C(A)$ with A the observed treatment. Let the counterfactual $S(a, c)$ denote a subject's survival when A and C are set to a and c. When $S(a, c)$ is well defined, $S(a)$ equals $S(a, C(a))$. Suppose that, unlike earlier subsections, X is a variable that is causally unaffected by either A or C. The average *controlled direct effect* of A on S when C is set to c within levels of X is defined as $CDE(c) = E[S(1, c) - S(0, c)|X]$[2],[3]. The average *pure direct effect* of A on survival not through C given X is defined as $PDE = E[S(1, C(0)) - S(0, C(0))|X] = E[S(1, C(0)) - S(0)|X]$. This contrast measures the average effect of A on survival when C is set to its value $C(0)$ under $A = 0$[4],[5]. The *principal stratum average direct effect* of A on survival at level c given X is defined as $PSDE(c) = E[S(1) - S(0)|X, C(0) = C(1) = c]$[13].

The conditioning subset in $PSDE(c = 1)$ consists of those with covariate X who always suffer the critical event. Robins[2] §12.2 used this contrast to address the problem of censoring by competing causes of death, with $S = 1$ denoting death from a cause of interest (subsequent to a time t) and $C = 0$ denoting death from competing causes (before t). Subsequently, Robins[6], Robins and Greenland[7], Rubin[8],[9],[10], Little and Rubin[11] and Frankagis and

Rubin[12] also employed this contrast in addressing 'censoring by death'. Baker[12], Frankagis and Rubin[13], Rubin[14], Gilbert et. al.[15], Shepherd et. al.[16], Hudgens and Halloran[17], Matsuyama and Morita[18] used this contrast to address a number of other causal issues.

The contrasts $CDE(c)$ and PDE are well-defined only when $S(a,c)$ is well defined. In contrast $PSDE(c)$ is well defined whenever $S(a)$ and $C(a)$ are well defined. How do we decide whether a counterfactual is well-defined? This has been a hotly debated issue in philosophy. The following example, due to Quine[19], effectively ended counterfactual analysis among philosophers until the late 60's. "If Bizet and Verdi had been of the same nationality, they both would have been French." Quine argued that, since Bizet was French and Verdi Italian, by symmetry considerations, this counterfactual was neither true nor false and thus was ill-defined. David Lewis[20] later rejoined that, even though some counterfactuals may be ill-defined and all are somewhat vague, many are useful. Robins and Greenland[7] agreed but went further. They argued that counterfactuals are "vague" to the degree to which one fails to make precise the hypothetical interventions.

Following REFS [7] and [12], we believe that for subjects with $C = 0$, the intervention corresponding to setting C to 1 is ill-defined because (i) $C = 1$ only encodes the occurrence of an accident, failing, for example, to distinguish high speed head-on collisions from rear-enders at moderate speed and (ii) there is no basis for choosing among them as the intervention. As a consequence $S(a,c)$ is ill-defined. Thus among the three direct effects, only $PSDE(c)$ is well-defined. Unfortunately, the following somewhat humourous example demonstrates that knowledge of $PSDE(c)$ may add little to our understanding of the pathways by which A relates to critical events and death. Like the authors, we restrict attention to $PSDE(c = 1)$.

Example: Suppose a psychiatrist hypothesizes that, conditional on pre-accident health and the seriousness of the crash injury, hypervigilant, controlling individuals ($A = 1$) have higher post-accident in-hospital mortality ($S = 0$) than distractible laid-back individuals

$(A = 0)$. His theory is that the loss of personal control during hospitalization causes controlling individuals to have serious life threatening arrhythmias. Suppose intervention on A is well defined. For example, there might exist drugs that can change a person from state $A = 1$ to $A = 0$ (Prozac and Valium) and vice-versa (amphetamines). Suppose, conditional on X, $PSDE\,(c = 1)$ is very negative. Does this refute a skeptic who believes the psychiatrist's hypothesis is false? It does not because $PSDE\,(c = 1)$ would also be negative under the following scenario. The psychiatrist's hypothesis is false. However, hypervigilant individuals avoid most potential accidents. Those they cannot avoid are usually serious head-on collisions with speeding cars that cross the centerline, leaving no time to react. In contrast, distractible, laid-back individuals have frequent, less serious collisions, because they are neither in a hurry nor do they look where they are going. Thus, individuals in the stratum 'always an accident' will tend to have serious accidents and thus a high in-hospital mortality rate when $A = 1$, but less serious accidents when $A = 0$. Thus, a negative $PSDE\,(c = 1)$ may arise because $A = 1$ increases mortality over $A = 0$; (i*) by directly causing increased in-hospital mortality, as hypothesized by the psychiatrist or, (ii*) solely by preventing minor accidents, as in the last scenario. We conclude that negative values of $PSDE\,(c = 1)$ fail to indicate the presence of direct effects of A not through its effect on accidents.

The difficulties with $PSDE\,(c = 1)$ are due to the fact that the event $C = 1$ lumps together the occurrence of accidents of varying severity. Thus, the natural solution is to replace C with a multivariate variable C^* that records relevant details of an accident including the type and seriousness of the injuries sustained. Then a non-zero c^*-specific principal stratum contrast $PSDE\,(c^*)$ could still be explained by pathway (i*) but no longer by (ii*), thus surmounting the difficulties of $PSDE\,(c = 1)$.Unfortunately, replacing C with C^* creates a major problem for the principal stratum approach: there is no subject with $C^*\,(0) = C^*\,(1)$ if, as is likely, A has an effect on at least one component of every subject's C^*. In that case, the

event $C^*(0) = C^*(1) = c^*$ has probability zero for all c^*, rendering the principal stratum approach useless. Even if there were subjects with $C^*(0) = C^*(1)$, their numbers would likely be few. Consequently, the principal stratum approach would only apply to a small subset of the population. Robins and Rotnitzky[21]catalogue analogous difficulties in substantively important examples. We believe these difficulties are sufficiently problematic to suggest that the principal stratum approach to direct effects is, at times, of little scientific value.

Counterfactuals regained. As we record more details in C^*, the intervention that sets C^* to c^* and the counterfactual $S(a, c^*)$ becomes less and less vague. Consequently, $CDE(c^*)$ and $PDE^* = E[S(1, C^*(0)) - S(0, C^*(0))|X]$ will often be reasonably well-defined. In our opinion, these are the contrasts that best serve to distinguish among different pathways. For example, they distinguish pathway (i*) from (ii*) above: PDE^* or $CDE(c^*)$ equal to 0 for all c^* is consistent with (ii*) but not with (i*), while non-zero values of PDE^* or $CDE(c^*)$ can be explained by (i*) but not by (ii*). Of course, even $S(a, c^*)$ is somewhat vague. The only counterfactuals free of vagueness are the treatment-assignment potential outcomes of a randomized experiment, but they are often uninformative about pathways. Because PDE^* only requires $S(a, c^*)$ to be defined for $a = 1$ and $c^* = C^*(0)$, there exist studies in which PDE^* may be regarded as well-defined even when $CDE(c^*)$ is not for some c^*[22].

We end this section by noting that none of the three contrasts $CDE(c^*)$, PDE^* and $PSDE(c)$ are identifiable from knowledge of $P(S = 0|A = a, C = 1, X)$, $f(A|C = 1, X)$ and $f(X|C = 1)$ without additional strong assumptions that were not either assumed or considered by the authors. We conclude that, even had the authors succeeded in their goal of learning these distributions, this success would not have helped 'understand the pathways through which inputs relate to critical events and death'.

iii) Statistical issues. In §6 the authors discuss similarities between the problem treated in §5 and the problem of treatment noncompliance in randomized trials. We now

show that these problems are statistically not merely similar but isomorphic. As a consequence, i) some of the material in §5 simply reproves previously published results and, ii) doubly robust semiparametric methods already exist[23] that address the modelling issues of §5.2 and do not require that Z be dichotomous.

Assumptions of §5.2 are exactly the same as the monotonicity, exclusion and randomization assumptions considered in the non-compliance literature, upon appropriate identification of the authors' variables with those in a non-compliance model. Specifically, identify X with a pre-randomization variable, Z with randomized arm and $S(z)$ with the actual treatment received when $Z = z$. Then $S = S(Z)$. In the authors' problem, A is a variable that is uninfluenced by Z and would be recorded, if, possibly contrary to fact, the person survived. Thus, we can regard A as the potential outcome $A(s = 1, z) = A(s = 1)$ for any z. This identity is the exclusion restriction. Further, assumption (I) is the assumption that Z is randomized and the assumption that Z never causes death is the monotonicity assumption. Under these assumptions, Abadie[24] has shown that $E[A|S(1) > S(0), X] = \frac{\pi(X,1)\eta(X,1)-\pi(X,0)\eta(X,0)}{\pi(X,1)\ -\pi(X,0)}$ where $\eta(x,z) = E[A|S = 1, X = x, Z = z]$ and $\pi(x,z) = P[S = 1|X = x, Z = z]$. The right hand side is precisely the right hand side of the last displayed equation in §5.1 in the case of no $X's$. Tan[23] showed how to obtain doubly-robust estimators of $E[A|S(z) > S(z')] = \frac{E[\pi(X,1)\eta(X,1)-\pi(X,0)\eta(X,0)]}{E[\pi(X,1)\ -\pi(X,0)]}$ when $z > z'$, with high dimensional X and Z possibly non-binary, even continuous, that are consistent if either a working model for $f_Z[z|X = x]$ is correct or working models for both $\pi(x,z)$ and $\eta(x,z)$ are correct.

iv) A sensitivity analysis. Because we wish not to impose assumptions (I) and (II), the distributions $f(A|S = 0, C = 1)$ and $P(S = 0|A = a, C = 1)$ of interest are not identified. Instead we suggest a sensitive analysis motivated by the observations that a) $f(A|S = 0, C = 1)$ and $P(S = 0|A = a, C = 1)$ would be identified were $P(A = 1|C = 1)$ identified and (ii) with $W \equiv (X, Z)$, $P(A = 1|C = 1)$ is identifed under the non-parametric

just-identified non-ignorable model for $\pi(W,A) \equiv P(S=0|W,A,C=1)$ that specifies $\pi(W,A) = \{1+\exp\{-[h(W)+Q]\}\}^{-1}$ where $h(W)$ is an unknown function and $Q = q(A,W)$ is a user-specified selection bias function. However, because Q itself is not identified, we later vary it in a sensitivity analysis. Since W is high-dimensional, we also specify flexible parametric models $B(\eta) = b(W;\eta)$ and $h(W;\alpha)$ for $b(W) \equiv E[A\exp(Q)|C=S=1,W]/E[\exp(Q)|C=S=1,W]$ and $h(W)$. We compute the estimators $(\widehat{\alpha},\widehat{\eta})$ given in Scharfstein et al.[25] and $\widehat{P}(A=1|C=1)$ as the sample average over $C=1$ of $\left[S\{1-\pi(W,A;\widehat{\alpha})\}^{-1}\{A-B(\widehat{\eta})\}+B(\widehat{\eta})\right]$. $\widehat{P}(A=1|C=1)$ is a doubly robust estimator of $P(A=1|C=1)$. That is, with $q(A,W)$ known, the estimator is consistent and asymptotically normal if either model $h(W;\alpha)$ or model $B(\eta)$ is correct. Final substantive conclusions depend on the set of functions $q(A,W)$ considered scientifically plausible[26]. Robins et al.[27] showed this sensitivity analysis can be used as input for a full Bayesian analysis.

References

[1] Lerner, E.B., Billittier, A.J., Dorn, J.M. and Wu, Y.W.B. (2003). Is total out-of-hospital time a significant predictor of trauma patient mortality? *Academic Emergency Medicine* **10**, 949-954.

[2] Robins, J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393-1512.

[3] Robins, J.M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease* (40, Supplement) **2**, 139s-161s.

[4] Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143-155.

[5] Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* San Francisco, CA: Morgan Kaufmann, pp. 411-420.

[6] Robins, J.M. (1995). An analytic method for randomized trials with informative censoring: Part I. *Lifetime Data Analysis,* 1:241-254.

[7] Robins, J.M. and Greenland, S. (2000). Causal inference without counterfactuals - Comment. *Journal of the American Statistical Association* **95**, 431-435.

[8] Rubin, D.B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine* **17**, 371-385.

[9] Rubin, D.B. (2000). Causal inference without counterfactuals - Comment. *Journal of the American Statistical Association* **95**, 435-438.

[10] Rubin, D.B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with 'censoring' due to death. *Statistical Science* **21**, 299-309.

[11] Little, R. and Rubin, D. (1999). Discussion of Adjusting for non-ignorable drop-out using semiparametric non-response models by Scharsftein, Rotnitzky and Robin. *Journal of the American Statistical Association*, 94:1121-1146.

[12] Baker S. (2000). Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance. *Journal of the American Statistical Association*; 95:43-50.

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

[14] Rubin, D.B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**, 161-170.

[15] Gilbert P., Bosch R., Hudgens M. (2003) Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials, *Biometrics*, 59 (3), 531-41.

[16] Shepherd, B., Gilbert, P., Jemiai, Y. and 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

[17] Hudgens, M. and Halloran, B. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the Royal Statistical Society*, 101:51-64.

[18] Matsuyama, Y. and Morita, S. (2006). Estimation of the average causal effect among subgroups defined by post-treatment variables. *Clinical Trials*, 2, vol. 3: pp. 1 - 9.

[19] Quine, W. V. (1950). *Methods of Logic*. New York: Holt, Reinhardt and Winston

[20] Lewis, D. (1973). Causation, *Journal of Philosophy*, 70, pp.556-67.9-688.

[21] Robins, J.M. and Rotnitzky, A. (2007). On direct effects, surrogate markers and principal stratification. Unpublished technical report. Department of Biostatistics. Harvard School of Public Health.

[22] Petersen, M.L., Sinisi, S.E. and van der Laan, M.J. (2006). Estimation of direct causal effects. *Epidemiology* **17**, 276-284.

[23] Tan, Z.Q. (2006). Regression and weighting methods for causal inference using instrumental variables *Journal of the American Statistical Association* **101**, 1607-1618.

[24] Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* **113**, 231-263.

[25] Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Rejoinder to Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*94:1121-1146.

[26] Robins JM. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies.". by Paul R. Rosenbaum. *Statistical Science*, 17(3):286-327.

[27] Robins, J.M., Rotnitzky, A. and Scharfstein, D. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In: *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Halloran, M.E. and Berry, D., eds. IMA Volume 116, NY: Springer-Verlag, pp. 1-92.