Biometrika (2018), **105**, 4, *pp.* 979–986
Printed in Great Britain

doi: 10.1093/biomet/asy053
Advance Access publication 22 October 2018

Miscellanea

Principal ignorability in mediation analysis: through and beyond sequential ignorability

By LAURA FORASTIERE, ALESSANDRA MATTEI

Department of Statistics, Computer Science, Applications, University of Florence,
Viale Morgagni 59, Florence 50134, Italy
forastiere@disia.unifi.it alessandra.mattei@unifi.it

and PENG DING

Department of Statistics, University of California, 425 Evans Hall, Berkeley,
California 94720, U.S.A.
pengdingpku@berkeley.edu

Summary

In causal mediation analysis, the definitions of the natural direct and indirect effects involve potential outcomes that can never be observed, so-called a priori counterfactuals. This conceptual challenge translates into issues in identification, which requires strong and often unverifiable assumptions, including sequential ignorability. Alternatively, we can deal with post-treatment variables using the principal stratification framework, where causal effects are defined as comparisons of observable potential outcomes. We establish a novel bridge between mediation analysis and principal stratification, which helps to clarify and weaken the commonly used identifying assumptions for natural direct and indirect effects. Using principal stratification, we show how sequential ignorability extrapolates from observable potential outcomes to a priori counterfactuals, and propose alternative weaker principal ignorability-type assumptions. We illustrate the key concepts using a clinical trial.

Some key words: Causal inference; Identification; Potential outcome; Principal stratification.

1. Introduction

Mediation analyses decompose causal effects into channelled effects through some mediator that lies in the pathway between the treatment and the outcome, and unchannelled effects not through this mediator. We define channelled and unchannelled effects using the concepts of natural direct and indirect effects. The latter effects raise identifiability issues because they are defined as comparisons between potential outcomes of various types, on some of which data contain no or little information without strong assumptions. Inferences on these effects usually rest on sequential ignorability, which combines ignorability of treatment assignment given a set of pre-treatment covariates and ignorability of the mediator given the treatment and pre-treatment covariates (Robins & Greenland, 1992). Under sequential ignorability, natural direct and indirect effects can be identified from the data using the mediation formula (Pearl, 2001).

Sequential ignorability implies that, conditional on covariates, there is no unmeasured confounding of the treatment-mediator, treatment-outcome and mediator-outcome relationships. Therefore, these assumptions require that the mediator be, at least in principle, regarded as an additional treatment and could be potentially manipulated by an intervention. Sequential ignorability is not directly verifiable from the observed data and its plausibility is not always well understood.

We provide insight into sequential ignorability using the concepts of principal stratification (Frangakis & Rubin, 2002) and principal ignorability (Jo & Stuart, 2009; Ding & Lu, 2017) in the case of a binary mediator. We make the following contributions. First, we use principal ignorability to offer an alternative interpretation of sequential ignorability, which may seem more natural in some settings. Second, we use principal stratification to clarify the source of information on natural direct and indirect effects under sequential ignorability. Third, we elucidate the relationship between sequential and principal ignorability under an additional monotonicity assumption. Fourth, we propose a new set of assumptions to identify natural direct and indirect effects, and investigate their relationships with sequential ignorability.

2. Notation, framework and identifiability in mediation analysis

2·1. *Potential outcomes and causal effects*

For each individual i characterized by covariates X_i, let Z_i represent a binary treatment, with $Z_i = 1$ for those assigned to the active treatment and $Z_i = 0$ for those assigned to the control. Let $Y_i(z)$ and $M_i(z)$ be the potential outcomes for a primary endpoint, Y, and a binary post-treatment intermediate variable, M, that we would observe under treatment level z $(z = 0, 1)$ for unit i. In mediation analysis, M is referred to as a mediator.

For each unit i the observed data include covariates X_i, the treatment Z_i, and the observed values of the mediator and outcome, which can be defined, by consistency, as $M_i^{\text{obs}} = M_i(Z_i) = Z_iM_i(1)+(1-Z_i)M_i(0)$ and $Y_i^{\text{obs}} = Y_i(Z_i) = Z_iY_i(1) + (1 - Z_i)Y_i(0)$.

The purpose of mediation analysis is to investigate the extent to which the mediator plays a role in the effect of the treatment on the outcome. To formalize causal effects that can answer such a question, Robins & Greenland (1992) and Pearl (2001) extended the above potential outcomes by introducing the double-indexed notation $Y_i(z, m)$, which denotes the potential outcome for unit i that would occur if the treatment were set to level z and if the mediator were manipulated to level m. Furthermore, we can define an additional potential outcome, $Y_i(z, M_{iz'})$, where the level of the mediator is determined by an intervention on the treatment. If $z' = z$, then $Y_i(z) = Y_i(z, M_{iz})$ under the composition assumption (VanderWeele, 2015). We use M_{iz} for $M_i(z)$ in the nested potential outcomes.

The average causal effect conditional on covariates at level $X_i = x$, $\textsc{ace}(x) = E\{Y_i(1) - Y_i(0) \mid x\}$, can be decomposed into the sum of a natural direct effect,

$$\textsc{nde}(z \mid x) = E\{Y_i(1, M_{iz}) - Y_i(0, M_{iz}) \mid x\} \quad (z = 0, 1),$$

and a natural indirect effect,

$$\textsc{nie}(z \mid x) = E\{Y_i(z, M_{i1}) - Y_i(z, M_{i0}) \mid x\} \quad (z = 0, 1),$$

as $\textsc{ace}(x) = \textsc{nde}(z \mid x) + \textsc{nie}(1 - z \mid x)$ (Robins & Greenland, 1992; Pearl, 2001). The natural direct effect $\textsc{nde}(z \mid x)$ is the average effect of the treatment when the mediator is kept at the level that would potentially be observed under treatment z, and the natural indirect effect $\textsc{nie}(z \mid x)$ is the average effect of a change in the mediator, achieved by a hypothetical intervention that sets the treatment to level z. All the effects are defined conditional on covariates.

Throughout the paper, we use a randomized clinical trial, the morphine study (Borracci et al., 2013), to convey the intuition behind the assumptions and illustrate how one can reason about their plausibility.

Example 1. Baccini et al. (2017) analysed the morphine study to assess the extent to which the effect of preoperative oral administration of morphine sulphate on postoperative pain intensity is mediated by postoperative self-administration of intravenous morphine sulphate by patients. A sample of patients undergoing an elective open colon-rectal abdominal surgery were randomly assigned to receive either oral morphine sulphate, $Z_i = 1$, or oral midazolam, $Z_i = 0$. The control is an active placebo with a sedative effect. For each patient, we observe gender and age. For patient i under treatment z, the potential outcome $Y_i(z)$ is the value of postoperative pain intensity, and $M_i(z)$ is a binary indicator equal to 1 or 0 depending on whether the patient self-administered a low or high level of morphine sulphate after surgery. Moreover, $Y_i(z,m)$ and $Y_i(z,M_{iz'})$ denote the values of postoperative pain intensity for patient i that would occur if his/her treatment was set to level z and if her/his postoperative morphine consumption was manipulated to levels m and $M_i(z')$, respectively.

2·2. *Identification issues and sequential ignorability*

Potential outcomes of the form $Y_i(z,M_{iz'})$, with $z \neq z'$, are referred to as cross-world counterfactuals (Robins & Greenland, 1992) or a priori counterfactuals (Frangakis & Rubin, 2002). They can never be observed in one experiment, because they result from hypothetically assigning each unit to two different treatments simultaneously (Mealli & Mattei, 2012; Forastiere et al., 2016). Although we can hypothesize their existence, a priori counterfactuals are conceptually different from potential outcomes of the form $Y_i(z)$, which are observable potential outcomes. The potential outcome $Y_i(z,M_{iz'})$ is observable only if either $z = z'$ or $M_i(z) = M_i(z')$, i.e., $Y_i(z) = Y_i(z,M_{iz}) = Y_i(z,M_{iz'})$, and is actually observed when the treatment received by unit i is $Z_i = z = z'$. Although ignorability of the treatment suffices to identify the marginal distributions of potential outcomes of the form $Y_i(z)$, and hence the average causal effect $\textsc{ace}(x)$, identification of the marginal distributions of a priori counterfactuals, and hence of natural direct and indirect effects, requires additional assumptions that would allow extrapolation to a priori counterfactuals based on the observed data.

There are different sets of identifying assumptions for the natural direct and indirect effects (Pearl, 2001; Van Der Laan & Petersen, 2008; Hafeman & VanderWeele, 2011; Imai et al., 2010). Ten Have & Joffe (2012) provides a review. The difference between them is subtle and, broadly speaking, they all couple the ignorability of the treatment with the ignorability of the mediator conditional on covariates. Here we focus on the assumptions used by Imai et al. (2010):

Assumption 1 (Ignorability of the treatment). $\{Y_i(z,m), M_i(z')\} \perp\!\!\!\perp Z_i \mid X_i$ for all $z, z', m = 0, 1$;

Assumption 2 (Ignorability of the mediator). $Y_i(z,m) \perp\!\!\!\perp M_i(z') \mid (Z_i = z', X_i)$ for all $z, z', m = 0, 1$.

Imai et al. (2010) refer to Assumptions 1 and 2 together as sequential ignorability. Assumption 1 is the ignorability of the treatment, and Assumption 2 states that the mediator is ignorable given the observed treatment and covariates. Under Assumptions 1 and 2,

$$E\{Y_i(z,M_{iz'}) \mid x\} = \sum_{m=0,1} E(Y_i^{\text{obs}} \mid Z_i = z, M_i^{\text{obs}} = m, x) \times \text{pr}(M_i^{\text{obs}} = m \mid Z_i = z', x), \qquad (1)$$

which is referred to as the mediation formula (Pearl, 2001). We see from (1) that the average of the potential outcome $Y_i(z,M_{iz'})$ can be identified from the observed data by the conditional expectation of the observed outcomes given treatment level z and the observed mediator, averaged over the conditional distribution of the observed mediator given treatment level z'.

2·3. *Principal stratification*

Frangakis & Rubin (2002) introduced the principal stratification framework to deal with post-treatment variables. A principal stratification with respect to a post-treatment variable M is a partition of units into latent subpopulations, called principal strata, defined by the joint potential values of that post-treatment variable under each level of the treatment. Denote by $G_i = \{M_i(0), M_i(1)\}$ the principal strata membership. Given a binary mediator, $G_i \in \{00, 01, 10, 11\}$. In Example 1, we call $G_i = 00$ pain-intolerant patients, $G_i = 01$ compliant patients, $G_i = 10$ defiant patients, and $G_i = 11$ pain-tolerant patients.

A principal causal effect is a comparison between the potential outcomes within a particular principal stratum. We focus on average principal causal effects, defined as $\textsc{pce}(g \mid x) = E\{Y_i(1) - Y_i(0) \mid G_i = g, x\}$. The average causal effect is a weighted average of the principal causal effects, $\textsc{ace}(x) = \sum_g \textsc{pce}(g \mid x)\pi_{g|x}$, where the summation is over $g \in \{00, 01, 10, 11\}$ and $\pi_{g|x} = \text{pr}(G_i = g \mid x)$ is the conditional probability of the principal stratum g. Frangakis & Rubin (2002) call $\text{PCE}(11 \mid x)$ and $\text{PCE}(00 \mid x)$ dissociative effects, and $\text{PCE}(01 \mid x)$ and $\text{PCE}(10 \mid x)$ associative effects. The subgroups 00 and 11, for which the mediator is not affected by the treatment, provide information on the natural direct effect of the treatment. The subgroups 01 and 10, for which the mediator is affected by treatment, generally combine natural direct and indirect effects (Mealli & Mattei, 2012). See VanderWeele (2008) for more discussion.

The principal strata membership is in general unknown, as we cannot observe both potential values of the mediator in a single experiment. This inherent latent nature of principal strata jeopardizes the identification of principal causal effects.

3. Generalized strong principal ignorability and the mediation formula

Principal ignorability was introduced for the identification of principal causal effects (Jo & Stuart, 2009; Ding & Lu, 2017; Feller et al., 2017). Here, we generalize it for mediation analysis:

Assumption 3 (Generalized strong principal ignorability). $Y_i(z,m) \perp\!\!\!\perp G_i \mid X_i$ for all $z, m = 0, 1$.

Assumption 3 requires that the distribution of potential outcomes $Y_i(z,m)$ be the same across principal strata, conditional on covariates. Because the heterogeneity across principal strata can be interpreted as heterogeneity with respect to a latent variable (Forcina, 2006), Assumption 3 can also be seen as ruling out the presence of unmeasured confounding of the mediator-outcome relationship (Ding & Lu, 2017). In the following, we present results that help to clarify the relationship between Assumptions 2 and 3. While the former involves marginal independence between the potential outcomes and the two potential values of the mediator, the latter assumes joint independence. Therefore, Assumption 3 implies Assumption 2. Thus, there can be situations where principal strata are heterogeneous, i.e., Assumption 3 does not hold, but Assumption 2 holds. Even if the joint distribution of $M_i(0)$ and $M_i(1)$ depends on a latent variable also affecting the outcome, the marginal distribution of the two potential mediators might be free of unmeasured confounding. Then, the proposition below follows.

Proposition 1. *Under Assumptions 1 and 3, the mediation formula* (1) *holds.*

Proposition 1 implies that the average of a priori counterfactuals can be identified from the observed data in the same way, that is, by the mediation formula (1), under either Assumptions 1 and 2 or Assumptions 1 and 3. Although Assumption 3 is stronger than Assumption 2, in some cases the plausibility of Assumption 3 might be easier to justify, because it can help to think in terms of homogeneity across principal strata rather than in terms of no unmeasured confounding of the mediator-outcome relationship.

In Example 1, Assumption 2 requires that, at least in principle, we can conceive an intervention on postoperative morphine consumption, and assume that it is randomly assigned within each treatment group, conditional on covariates. Thus, Assumption 2 rules out unobserved confounders that causally affect both postoperative morphine consumption and pain intensity given the treatment and pre-treatment covariates. Although hypothetical interventions on postoperative morphine consumption might be conceivable, they might be unethical. Moreover, it might be difficult to argue that all relevant confounders of the relationship between postoperative morphine consumption and pain intensity have been observed, especially in the morphine study with only two covariates. It might be easier to envision the plausibility of Assumption 3, which requires that the potential outcomes for pain intensity that would occur if the treatment were set to level z and the postoperative morphine consumption were set to level m have the same distributions across pain-tolerant, pain-intolerant, compliant and defiant patients with the same value of the covariates.

4. Interpretation of the mediation formula: extrapolation across principal strata

We aim at clarifying the extrapolation of information on a priori counterfactuals performed by the mediation formula (1). In principle, the average potential outcome is a weighted average of the same potential outcome across principal strata, with weights given by principal strata proportions. The following proposition shows what part of the observed data and which type of units provide information on potential outcomes $Y_i(z, M_{iz'})$, which can be a priori counterfactuals for some units if $z \neq z'$.

Proposition 2. *Under Assumption 1, if either Assumption 2 or 3 holds, then*

$$
\begin{aligned}
&E\{Y_i(1, M_{i0}) \mid x\}\\
&=\left[E\{Y_i(1) \mid G_i = 00, x\}\frac{\pi_{00|x}}{\pi_{00|x}+\pi_{10|x}} + E\{Y_i(1) \mid G_i = 10, x\}\frac{\pi_{10|x}}{\pi_{00|x}+\pi_{10|x}}\right](\pi_{00|x}+\pi_{01|x})\\
&\quad+\left[E\{Y_i(1) \mid G_i = 11, x\}\frac{\pi_{11|x}}{\pi_{01|x}+\pi_{11|x}} + E\{Y_i(1) \mid G_i = 01, x\}\frac{\pi_{01|x}}{\pi_{01|x}+\pi_{11|x}}\right](\pi_{10|x}+\pi_{11|x}), \qquad (2)\\
&E\{Y_i(0, M_{i1}) \mid x\}\\
&=\left[E\{Y_i(0) \mid G_i = 11, x\}\frac{\pi_{11|x}}{\pi_{11|x}+\pi_{10|x}} + E\{Y_i(0) \mid G_i = 10, x\}\frac{\pi_{10|x}}{\pi_{11|x}+\pi_{10|x}}\right](\pi_{00|x}+\pi_{01|x})\\
&\quad+\left[E\{Y_i(0) \mid G_i = 00, x\}\frac{\pi_{00|x}}{\pi_{01|x}+\pi_{00|x}} + E\{Y_i(0) \mid G_i = 01, x\}\frac{\pi_{01|x}}{\pi_{01|x}+\pi_{00|x}}\right](\pi_{10|x}+\pi_{00|x}). \qquad (3)
\end{aligned}
$$

Each term of (2) and (3) is a product of a weighted average of an observable potential outcome, $Y_i(1)$ or $Y_i(0)$, and the sum of the proportion of two principal strata. This product reflects how information on observable potential outcomes for specific principal strata is used for potential outcomes of the type $Y_i(z, M_{iz'})$ for other principal strata.

In Example 1, according to (2), a weighted average of the observable potential outcomes for pain intensity under oral morphine, $Y_i(1)$, for patients with $M_i(1) = 0$, who would self-administer a high level of morphine sulphate, i.e., pain-intolerant patients $G_i = 00$ and defiant patients $G_i = 10$, provides information on $Y_i(1, M_{i0})$ for patients with $M_i(0) = 0$, who would self-administer a high level of morphine sulphate under the placebo, i.e., compliant patients $G_i = 01$ and pain-intolerant patients $G_i = 00$. Moreover, the distributions of $Y_i(1)$ for patients with $M_i(1) = 1$, i.e., pain-tolerant patients $G_i = 11$ and compliant patients $G_i = 01$, are used to impute $Y_i(1, M_{i0})$ for patients with $M_i(0) = 1$, i.e., defiant patients $G_i = 10$ and pain-tolerant patients $G_i = 11$. A similar interpretation applies to (3).

Proposition 2 also provides valuable insights into the meaning of the natural indirect effects. Specifically, we have the following propositions, in which we use $\text{ACE}_M(x) = E\{M_i(1) - M_i(0) \mid x\}$ to denote the conditional average causal effect of the treatment on the mediation for notational simplicity.

Proposition 3. *Under Assumption 1, if either Assumption 2 or 3 holds, then*

$$\text{NIE}(1 \mid x) = \text{ACE}_M(x) \times [E\{Y_i(1) \mid G_i = 11 \text{ or } 01, x\} - E\{Y_i(1) \mid G_i = 00 \text{ or } 10, x\}], \qquad (4)$$

$$\text{NIE}(0 \mid x) = \text{ACE}_M(x) \times [E\{Y_i(0) \mid G_i = 11 \text{ or } 10, x\} - E\{Y_i(0) \mid G_i = 00 \text{ or } 01, x\}]. \qquad (5)$$

Proposition 3 decomposes the natural indirect effects into products of the average effect of the treatment on the mediator and a comparison of potential outcomes across different principal strata.

Under Assumptions 1 and 2, if we further introduce homogeneity assumptions of the potential outcome distributions across principal strata, then the second terms on the right-hand sides of (4) and (5) can be interpreted as the average causal effects of the mediator on the outcome.

Proposition 4. *Suppose Assumptions 1 and 2 hold. If $Y_i(1, m) \perp\!\!\!\perp G_i \mid X_i$, then*

$$\text{NIE}(1 \mid x) = \text{ACE}_M(x) \times E\{Y_i(1, 1) - Y_i(1, 0) \mid x\}. \qquad (6)$$

If $Y_i(0, m) \perp\!\!\!\perp G_i \mid X_i$, then

$$\text{NIE}(0 \mid x) = \text{ACE}_M(x) \times E\{Y_i(0, 1) - Y_i(0, 0) \mid x\}. \qquad (7)$$

The independence assumption $Y_i(z, m) \perp\!\!\!\perp G_i \mid X_i$ for a fixed value of z is implied by Assumption 3, so both (6) and (7) hold under Assumptions 1 and 3. Formulas (6) and (7) reflect the intuition of mediation: the treatment affects the mediator, and then the mediator affects the outcome given the treatment level $Z_i = z$ with either $z = 0$ or $z = 1$.

5. MONOTONICITY IN MEDIATION ANALYSIS

We now investigate the role of monotonicity in mediation analysis:

Assumption 4 (Monotonicity). $M_i(1) \geqslant M_i(0)$ for all i.

Assumption 4 rules out negative effects of the treatment on the mediator, but an alternative version of monotonicity, ruling out positive effects of the treatment on the mediator, could be considered. The plausibility of monotonicity in mediation analysis strongly depends on the substantive setting. In Example 1, monotonicity, ruling out the existence of defiant patients with $G_i = 10$, is likely plausible due to the pharmacological characteristics of the active placebo under control. See also Baccini et al. (2017).

When the treatment and the mediator are both binary, the following proposition holds under the monotonicity in Assumption 4.

PROPOSITION 5. *Under Assumptions 1 and 4, Assumptions 2 and 3 are equivalent.*

Proposition 5 implies that, under ignorability of treatment assignment and monotonicity, sequential ignorability and strong principal ignorability are equivalent, so we can use the mediation formula in (1) to identify and estimate natural direct and indirect effects invoking either Assumption 2 or Assumption 3, whichever is easier to justify in a specific case study. In Example 1, Assumption 1 holds by design and Assumption 4 is very plausible. Therefore, we can identify the natural direct and indirect effects using (1), if we can provide convincing arguments on the plausibility of either Assumption 2, i.e., no unmeasured confounding between morphine consumption and pain intensity, or Assumption 3, i.e., homogeneity of the distributions of the potential outcomes across pain-tolerant, pain-intolerant, and compliant patients.

6. IDENTIFICATION UNDER GENERALIZED WEAK PRINCIPAL IGNORABILITY

Here we propose a set of alternative assumptions for identification of natural direct and indirect effects, involving generalizations of weak principal ignorability assumptions (Jo & Stuart, 2009; Ding & Lu, 2017; Feller et al., 2017) to potential outcomes of the form $Y_i(z, m)$:

Assumption 5. $Y_i(1, 1) \perp\!\!\!\perp M_i(0) \mid \{M_i(1) = 1, X_i\}$;

Assumption 6. $Y_i(1, 0) \perp\!\!\!\perp M_i(1) \mid \{M_i(0) = 0, X_i\}$.

Assumption 5 is a generalized weak principal ignorability of $Y_i(1, 1)$ across strata $G_i = 11$ and $G_i = 01$, and Assumption 6 is a generalized weak principal ignorability of $Y_i(1, 0)$ across strata $G_i = 00$ and $G_i = 01$. Assumptions 5 and 6 together are weaker than Assumption 3, because the independence in Assumptions 5 and 6 refers to specific potential outcomes and is conditional on specific values of $M_i(0)$ and $M_i(1)$.

In general, we cannot rank sequential ignorability and Assumptions 5 and 6. However, when the treatment and the mediator are both binary, relying on Proposition 5 we have the following result.

PROPOSITION 6. *Under Assumptions 1 and 4, Assumption 2 implies Assumptions 5 and 6.*

Proposition 6 implies that the set of Assumptions 1, 4, 5 and 6 is weaker than the set of Assumptions 1, 4 and 2 or 3, and thus may be more plausible. Therefore, it might be valuable to investigate whether we can identify natural direct and indirect effects under Assumptions 1, 4, 5 and 6.

Assumptions 5 and 6 involve homogeneity of two different potential outcomes, $Y_i(1, 1)$ and $Y_i(1, 0)$, across two different sets of principal strata. In particular, Assumption 5 states that the distribution of

$Y_i(1,1)$ is the same for strata $G_i = 11$ and $G_i = 01$, i.e., pain-tolerant and compliant patients for whom $Y_i(1,1) = Y_i(1, M_{i1}) = Y_i(1)$. Assumption 5 implies that we can use the observed data to estimate the distribution of $Y_i(1,1)$ for the two principal strata that are mixed together in the observed set with $Z_i = 1$ and $M_i^{\text{obs}} = 1$, i.e., patients who are treated with preoperative oral morphine and who self-administer a low level of morphine sulphate after surgery.

The second homogeneity in Assumption 6 refers to the potential outcome $Y_i(1,0)$ across strata $G_i = 00$ and $G_i = 01$, i.e., pain-intolerant and compliant patients for whom $Y_i(1,0) = Y_i(1, M_{i0})$. This homogeneity has a slightly different flavour, because it allows for identifying the a priori counterfactual for compliant patients $G_i = 01$ using information on $Y_i(1,0)$ for pain-intolerant patients $G_i = 00$. Under Assumptions 1 and 4, we can estimate the distribution of $Y_i(1,0)$ for $G_i = 00$ using information on the observed outcome for units with $Z_i = 1$ and $M_i^{\text{obs}} = 0$, i.e., patients who are treated with preoperative oral morphine and who self-administer a high level of morphine sulphate after surgery.

We formalize these arguments in the following proposition.

PROPOSITION 7. *If Assumptions 1, 4, 5 and 6 hold, then*

$$E\{Y_i(1, M_{i0}) \mid x\} = \sum_{m=0,1} E(Y_i^{\text{obs}} \mid Z_i = 1, M_i^{\text{obs}} = m, x) \times \text{pr}(M_i^{\text{obs}} = m \mid Z_i = 0, x),$$

$$\begin{aligned}\textsc{nde}(0 \mid x) = \sum_{m=0,1} & E(Y_i^{\text{obs}} \mid Z_i = 1, M_i^{\text{obs}} = m, x) \times \text{pr}(M_i^{\text{obs}} = m \mid Z_i = 0, x) \\ & - E(Y_i^{\text{obs}} \mid Z_i = 0, x),\end{aligned}$$

$$\begin{aligned}\textsc{nie}(1 \mid x) = & \left\{E(Y_i^{\text{obs}} \mid Z_i = 1, M_i^{\text{obs}} = 1, x) - E(Y_i^{\text{obs}} \mid Z_i = 1, M_i^{\text{obs}} = 0, x)\right\} \\ & \times \left\{E(M_i^{\text{obs}} \mid Z_i = 1, x) - E(M_i^{\text{obs}} \mid Z_i = 0, x)\right\}.\end{aligned}$$

In the Supplementary Material, we give analogous results for $\textsc{nde}(1 \mid x)$ and $\textsc{nie}(0 \mid x)$.

7. DISCUSSION

Generalized strong principal ignorability in Assumption 3 implies ignorability of the mediator in Assumption 2. Proposition 5, however, shows that under monotonicity, the two assumptions are equivalent with a binary mediator. This allows us to derive alternative and weaker assumptions to identify natural direct and indirect effects, namely the weak principal ignorability in Assumptions 5 and 6. Unfortunately, monotonicity, ignorability of the mediator and weak principal ignorability assumptions are not directly testable from the observed data, and they may be implausible in some contexts. Therefore, it is valuable to think about what we can learn from the data about the causal estimands of interest when some of the underlying critical assumptions cannot be invoked.

ACKNOWLEDGEMENT

Peng Ding was partially supported by the U.S. Institute of Education Sciences and National Science Foundation. The authors thank the editor, associate editor and a reviewer for helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes all the proofs.

REFERENCES

BACCINI, M., MATTEI, A. & MEALLI, F. (2017). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. *Biostatistics* **18**, 605–17.

BORRACCI, T., CAPPELLINI, I., CAMPIGLIA, L. PICCIAFUOCHI, F., BERTI, J., CONSALES, G. & DE GAUDIO, A. (2013). Preoperative medication with oral morphine sulphate and postoperative pain. *Minerva Anestesiologica* **79**, 525–33.

Ding, P. & Lu, J. (2017). Principal stratification analysis using principal scores. *J. R. Statist. Soc.* B **79**, 757–77.

Feller, A., Mealli, F. & Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *J. Educ. Behav. Statist.* **42**, 726–58.

Forastiere, L., Mealli, F. & VanderWeele, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *J. Am. Statist. Assoc.* **111**, 510–25.

Forcina, A. (2006). Causal effects in the presence of non compliance: A latent variable interpretation. *Metron*, **64**, 275–301.

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58**, 21–9.

Hafeman, D. M. & VanderWeele, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* **22**, 753–64.

Imai, K., Keele, L. & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25**, 51–71.

Jo, B. & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statist. Med.* **28**, 2857–75.

Mealli, F. & Mattei, A. (2012). A refreshing account of principal stratification. *Int. J. Biostatist.* **8**, 246–54.

Pearl, J. (2001). Direct and indirect effects. In *Proc. 17th Conf. Uncert. Artif. Intel.*, ed. J. S. Breese and D. Koller. San Francisco: Morgan Kaufman, pp. 411–20.

Robins, J. M. & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–55.

TenHave, T. R. & Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Statist. Meth. Med. Res.* **21**, 77–107.

Van Der Laan, M. & Petersen, M. (2008). Estimation of direct causal effects. *Int. J. Biostatist.* **4**, 23.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statist. Prob. Lett.* **78**, 2957–62.

VanderWeele, T. J. (2015). *Explanation in Causal Inference. Methods for Mediation and Interaction*. New York: Oxford University Press.

[*Received on* 24 *July* 2017. *Editorial decision on* 15 *June* 2018]