Biostatistics (2017) **00**, 00, *pp.* 1–13
doi:10.1093/biostatistics/kxx010

Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control

MICHELA BACCINI

Dipartimento di Statistica, Informatica, Applicazioni, University of Florence, Viale Morgagni, 59, 50134, Firenze, Italy

baccini@disia.unifi.it

ALESSANDRA MATTEI*

Dipartimento di Statistica, Informatica, Applicazioni, University of Florence, Viale Morgagni, 59, 50134, Firenze, Italy

mattei@disia.unifi.it

and

FABRIZIA MEALLI

Dipartimento di Statistica, Informatica, Applicazioni, University of Florence, Viale Morgagni, 59, 50134, Firenze, Italy

mealli@disia.unifi.it

Summary

We conduct principal stratification and mediation analysis to investigate to what extent the positive overall effect of treatment on postoperative pain control is mediated by postoperative self administration of intra-venous analgesia by patients in a prospective, randomized, double-blind study. Using the Bayesian approach for inference, we estimate both associative and dissociative principal strata effects arising in principal stratification, as well as natural effects from mediation analysis. We highlight that principal stratification and mediation analysis focus on different causal estimands, answer different causal questions, and involve different sets of structural assumptions.

Keywords: Bayesian inference; Causal inference; Mediation analysis; Principal stratification; Oral morphine; Premedication, Postoperative pain, Potential outcomes; Randomized Experiments.

*To whom correspondence should be addressed.

1. Introduction

Principal stratification analysis and casual mediation analysis are two ways to conceptualize the mediating role of an intermediate variable in the causal pathways between treatment and outcome. Nevertheless, they are often viewed as competing frameworks. One exception is VanderWeele (2008), who shows the relationships between them from a theoretical point of view. However, VanderWeele (2008) provides little insight on the substantive differences between principal stratification analysis and mediation analysis. Mealli and Mattei (2012) further investigate such relationships, although they do not provide empirical examples.

In this article, we aim to fill this gap, using a prospective, randomized, double-blind study concerning the effect of preoperative oral administration of morphine sulphate on postoperative pain relief as a motivating example. The study, in the sequel referred to as "the morphine study", involved adult patients who were undergoing an elective open colorectal abdominal surgery. Patients were randomized to receive, before surgery, the experimental treatment or an active placebo and the outcome of primary interest was postoperative pain intensity, measured using a visual analogue scale (VAS). According to the medical guidelines for pain control, after surgery patients received an Intra-Venous Patient Controlled Analgesia (IV-PCA) system programmed to give off fixed doses of morphine sulphate upon patient demand (see Borracci *and others*, 2013, for details).

The number of self-administered doses of morphine sulphate is a post-treatment intermediate variable lying on the causal pathway between the treatment (preoperative medication) and pain intensity. Then, the question is how to extricate, from one another, the channeled (indirect) effect mediated through postoperative self-administration of morphine sulphate, and the un-channeled (direct) effect. In this article, we use principal stratification and mediation analysis to get some information on the part of the effect of the treatment on the outcome that is channeled by the intermediate variable.

The role of principal stratification and mediation analysis in dealing with issues concerning causal mechanisms has often fired up heated discussions in the causal community. In this article, we aim to smooth these controversies over, by highlighting that mediation analysis and principal stratification analysis generally focus on different causal estimands, answer different questions and involve different sets of assumptions, which lead to the use of the information provided by the data in a substantially different way. Principal stratification focuses on causal effects for specific sub-populations named principal strata. Mediation analysis focuses on disentangling un-channeled and channeled effects, which are generally defined at the individual level and averaged over the whole population.

We use a parametric approach, specifying models for potential outcomes conditional on principal stratum membership in principal stratification analysis, and linear structural models in mediation analysis. We adopt a Bayesian approach for inference (e.g. Rubin, 1978). The Bayesian paradigm appears to be a natural and appealing approach to compare results from a principal stratification analysis and a mediation analysis, which involve different sets of assumptions with diverse inferential implications. Depending on the assumptions, causal estimands are fully or partially identified. Following Gustafson (2010), a target parameter is said to be *fully identified* if multiple values of the parameter "cannot correspond to the same distribution of observables", and to be *partially identified* if "it cannot be consistently estimated, but the possible set of values for the target which are consistent with the observed data law is smaller than the a priori set of possible values (at least for some such laws, if not all)". In Bayesian terms, this means that the support of the marginal posterior distribution of a partially identified parameter is asymptotically smaller than the corresponding prior support, but larger than a single value. In finite samples, posterior distributions of partially identified parameters usually have substantial region of flatness. This feature is called *weakly identifiability* (e.g., Hirano *and others*, 2000). The Bayesian approach does not require full identification: in Bayesian inference the posterior distribution of the parameters of interest is derived by updating a prior distribution to a posterior distribution via a likelihood, irrespective of whether the parameters are fully or partially identified. From a Bayesian perspective, inference is based on the posterior distributions of the causal estimands of interest, which are always proper if the prior distributions are proper.

The remainder of the article is organized as follows. In Section 2, we describe the morphine study and introduce the notation. In Section 3, we define the causal estimands of interest, clarifying the information they provide in the context of the morphine study. We present the structural assumptions in Section 4. In Section 5, we describe the Bayesian approach used for principal stratification analysis and mediation analysis, specifying our modeling assumptions. We present and discuss the results of the analyses in Section 6 and conclude the article in Section 7.

2. The morphine study

The morphine study, a double-blind randomized controlled trial conducted between October 2009 and June 2010 at the University Hospital of Florence in Italy, was designed to investigate the effects of preoperative oral administration of morphine sulphate on patients' postoperative pain. A random sample of $n = 60$ patients aged 18–80 who were undergoing an elective open colorectal abdominal surgery was enrolled in the study: 32 patients were randomly assigned to the treatment group, and 28 patients were randomly assigned to the control group. Before surgery, patients in the treatment group were administered oral morphine sulphate (Oramorph®, Molteni Farmaceutici, Italy), and patients in the control group received oral midazolam (Hypnovel®, Roche, Switzerland), a short-acting drug inducing sedation, which is here considered as an active placebo. After surgery all patients received a device for IV-PCA, programmed to deliver fixed doses of intravenous morphine sulphate upon patient demand, with a lock-out time of 5 minutes to avoid excess of sedation or overdose.

The outcome of primary interest was pain intensity measured using VAS scores at rest and for movement (that is, upon coughing). We refer to these outcomes as *static VAS* and *dynamic VAS*, respectively. VAS scores were measured using a line of 100 mm where the left extremity is no pain and the right one is extreme pain. Physicians consider a pain score not greater than 30 mm at rest, and not greater than 45 mm on movement as a satisfactory pain relief. For each patient, pain intensity at rest and for movement was measured at 4, 24, and 48 hours from the end of surgery. Here we focus on pain intensity at rest and for movement after 4 hours.

Our objective is to measure the causal effect of preoperative medication on pain relief, accounting for postoperative self-administration of morphine sulphate. Postoperative morphine consumption is a post-treatment intermediate variable, therefore it may be affected by the treatment, and, in turn, it may mediate the effect of the treatment on the primary outcome, in some way channeling part of the treatment effect.

2.1. *Notation and descriptive analyses*

We frame our discussion in the context of the potential outcome approach to causal inference (Rubin, 1974, 1978). Each patient who participates in the study can either be assigned to the oral morphine group, $z = 1$, or to the active placebo group, $z = 0$. Let Z denote the treatment variable. Under the standard Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980), for each patient there are two associated potential outcomes for each post-treatment variable. Formally, for each patient, indexed by i, $i = 1, \ldots, n = 60$, let $S_i(z)$ be the number of self-administered postoperative doses of morphine sulphate if the patient is exposed to treatment z, $z = 0, 1$. Analogously, let $Y_{i1}(z)$ and $Y_{i2}(z)$ define the potential outcomes for pain intensity at rest and for movement, respectively, if patient i is assigned treatment z, $z = 0, 1$.

For each patient i, we observe the treatment actually assigned, Z_i, and only one potential outcome for each post-treatment variable. Let $S_i^{obs} = S_i(Z_i)$ be the observed number of doses of morphine, and let $Y_{i1}^{obs} = Y_{i1}(Z_i)$ and $Y_{i2}^{obs} = Y_{i2}(Z_i)$ be the actual outcomes. Potential outcomes under the treatment status not assigned, $1 - Z_i$, are missing: $S_i^{mis} = S_i(1 - Z_i)$, $Y_{i1}^{mis} = Y_{i1}(1 - Z_i)$ and $Y_{i2}^{mis} = Y_{i2}(1 - Z_i)$. In the sequel, we drop the second subscript and use the generic simplified notation $Y_i(z)$ and Y_i^{obs}. For each patient we also observe two covariates, X_{i1}, gender, and X_{i2}, age in years. The vectors $\mathbf{Z}$, $\mathbf{S}^{obs}$, and $\mathbf{Y}^{obs}$ are

Table 1. *Morphine study: Summary statistics*

	Mean			Mean
Outcome variable	All	$Z_i = 0$	$Z_i = 1$	difference
IV-PCA (S_i)	13.43	15.64	11.50	−4.14
Static VAS (Y_{i1})	36.08	45.36	27.97	−17.39
Dynamic VAS (Y_{i2})	55.08	66.61	45.00	−21.61

$n-$dimensional vectors with ith elements equal to Z_i, S_i^{obs}, and Y_i^{obs}, respectively. The $n \times 2$ matrix $\mathbf{X}$ has ith row equal to $\mathbf{X}_i' \equiv (X_{i1}, X_{i2})$.

From an epistemic perspective, data contain information on the marginal distributions of the intermediate and final potential outcomes, as well as some information on the distribution of the joint potential values of the intermediate outcome $(S_i(0), S_i(1))$, that is, on principal stratum membership, and on the conditional distribution of $Y_i(z)|S_i(0), S_i(1)$ (see Web Appendix A available at *Biostatistics* online).

In mediation analysis, potential outcomes for the primary endpoint are defined as functions of both the assignment variable and the intermediate variable. Specifically, under appropriate versions of SUTVA (see Web Appendix A available at *Biostatistics* online), mediation analysis assumes that there exist potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$: $Y_i(z, s)$ would be the value of the outcome Y if, possibly contrary to fact, the treatment were set to z and S were set to the value s, and $Y_i(z, S_i(z'))$ would be the value of Y if, possibly contrary to fact, the treatment were set to the level z and the mediator S were set to the value it would have taken if the treatment had been set to an alternative level, z'. For instance in the morphine study, potential outcomes of the type $Y_i(1, S_i(0))$ are the values of pain intensity under oral morphine, if the number of doses of morphine were somehow simultaneously forced to attain the value it would have taken under the active placebo. Note that $Y_i(z) = Y_i(z, S_i(z'))$ if $z = z'$.

The existence of potential outcomes of the form $Y_i(z, S_i(z'))$, $z \neq z'$, raises non-trivial ontological problems. Even if one is willing to conceive the intermediate variable as an additional treatment, the definition of a scientifically meaningful and feasible intervention corresponding to potential outcomes of the form $Y_i(z, S_i(z'))$, $z \neq z'$, is challenging and might be quite a stretch (Robins and Richardson, 2011; Imai *and others*, 2013). On the other hand, even if we are willing to hypothesize the existence of potential outcomes of the form $Y_i(z, S_i(z'))$, the intermediate variable is indeed a post-treatment variable, which can be potentially affected by treatment assignment. Therefore, some potential outcomes of the form $Y_i(z, S_i(z'))$ are a priori counterfactuals in the experiment (Frangakis and Rubin, 2002), because $Y_i(z, S_i(z'))$ cannot be observed for the type of units such that $S_i(z) \neq S_i(z')$ for $z \neq z'$. For such type of patients, a priori counterfactuals are not in the data, and in this specific experiment, they cannot be observed, not even on patients of the same type assigned to the opposite treatment. The lack of information in the data on some potential outcomes of the form $Y_i(z, S_i(z'))$ raises serious epistemic problems: The marginal distributions of $Y_i(0, S_i(1))$ and $Y_i(1, S_i(0))$ and their features (such as their means) are fully non-identified for subjects for whom $S_i(0) \neq S_i(1)$, even in randomized experiments.

Table 1 presents some summary statistics by treatment assignment, Z_i. There is some evidence that oral morphine reduces the number of doses of morphine and reduces pain intensity, both at rest and for movement.

3. Causal estimands

A causal effect of the treatment Z on an outcome Y is defined as a comparison of potential outcomes on a common set of units. Here, we focus on the average causal effect of oral morphine on pain intensity, defined as the expected difference between potential outcomes for the study population: $ACE = \mathbb{E}[Y_i(1) - Y_i(0)]$. However, this causal estimand does not account for postoperative morphine consumption, S.

Principal stratification analysis may provide useful information by looking at the joint value of the mediating variable under treatment and under control, $(S_i(0), S_i(1))$, which is essentially a characteristic of a subject, describing how an individual reacts to the treatment. The framework of principal stratification focuses on causal effects for specific subpopulations defined on the basis of $(S_i(0), S_i(1))$, named principal strata. It may provide useful insights, and has the advantage to avoid a priori counterfactuals, because it uses only potential outcomes of the form $Y_i(z)$ and $S_i(z)$.

Causal mediation analysis focuses on disentangling un-channeled and channeled effects, the definition of which usually involves potential outcomes defined as a function of both the treatment and the mediator, such as $Y_i(z, S_i(z'))$.

3.1. *Principal stratification and principal strata effects*

In the morphine study principal strata are defined by $(S_i(0), S_i(1))$, the joint potential values of the number of doses of morphine under the oral morphine treatment and under the active placebo. The intermediate variable takes on several values, thus the principal stratification leads to classify units into several principal strata. Here, we prefer to focus on a simplified setting with a binary intermediate variable, considering a binary variable equal to 1 if a patient uses a number of morphine doses greater than a pre-fixed cut-off point s^*, and 0 otherwise. Formally, let $S_i^* \equiv \mathbb{I}\{S_i > s^*\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. It should be noted that principal stratification *per se* does not require the intermediate variable to be binary or categorical. Nevertheless, multi-valued or continuous intermediate variables introduce serious challenges to principal stratification analysis. The methods proposed in the literature face these challenges using advanced statistical methodologies, including sophisticated models with a complex structure, which may make inference demanding, especially in the presence of reduced sample sizes, as in the morphine study (e.g., Schwartz *and others*, 2011). Therefore, we prefer to avoid introducing heavy model structures, sacrificing some generality and information for gaining clarity. Note that the dichotomization of the intermediate variable implies focusing on channeled and un-channeled effects of a binary mediator, which are different from (and not directly comparable with) channeled and un-channeled effects of the continuous mediator.

Principal stratification with respect to the binary intermediate variable S^* partitions patients into four latent groups: (*i*) patients who would self-administer a low number of doses of morphine irrespective of their treatment assignment: $00 = \{i : S_i^*(0) = 0, S_i^*(1) = 0\}$, whom we label as "pain-tolerant patients"; (*ii*) patients who would self-administer a high number of doses of morphine under the active placebo, but would self-administer a low number of doses of morphine under oral morphine: $10 = \{i : S_i^*(0) = 1, S_i^*(1) = 0\}$, whom we label as "normal patients"; (*iii*) patients who would self-administer a high number of doses of morphine irrespective of their treatment assignment: $11 = \{i : S_i^*(0) = 1, S_i^*(1) = 1\}$, whom we label as "pain-intolerant patients"; and (*iv*) patients who would self-administer a low number of doses of morphine under the active placebo, but would self-administer a high number of doses of morphine under oral morphine: $01 = \{i : S_i^*(0) = 0, S_i^*(1) = 1\}$, whom we label as "special patients."

A principal causal effect (PCE) is a comparison between the potential outcomes for the outcome Y, within a particular principal stratum (or union of principal strata). Here, we focus on average PCEs:

$$PCE(s_0, s_1) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid S_i^*(0) = s_0, S_i^*(1) = s_1\right]. \qquad (3.1)$$

If one is seeking information on causal mechanisms, it is sensible to start looking at the effects of treatment on the outcome that are associative and dissociative with the effects of treatment on the mediating variable. Associative PCEs are causal effects within principal strata where the mediating variable is affected by the treatment, that is, causal effects for normal and special patients: $PCE(s_0, s_1)$ with $s_0 \neq s_1$. Dissociative PCEs are causal effects within principal strata where the mediating variable is unaffected by the treatment, that is, causal effects for pain-tolerant and pain-intolerant patients: $PCE(s, s)$, $s = 0, 1$.

The average total effect is the weighted average of PCEs across units belonging to different principal strata:

$$ACE = \sum_{s_0,s_1} PCE(s_0, s_1)\pi_{s_0,s_1} = \sum_{s_0=s_1=s} PCE(s, s)\pi_{s,s} + \sum_{s_0\neq s_1} PCE(s_0, s_1)\, \pi_{s_0,s_1},$$

where π_{s_0,s_1} is the proportion of units belonging to $\{i : S_i^*(0) = s_0, S_i^*(1) = s_1\}$.

Dissociative PCEs naturally provide information on the existence of an un-channeled causal effect of the treatment on the primary outcome for the sub-population of patients for whom treatment does not affect the intermediate variable in this study (Rubin, 2004). If dissociative PCEs are all zero, then there is no evidence on the un-channeled (direct) effect of the treatment (Rubin, 2004; Mattei and Mealli, 2011). This does not mean that there is no direct effect of the treatment, because associative effects generally combine un-channeled and channeled effects (e.g., VanderWeele, 2008). Although no PCE can be directly interpreted as a channeled effect, under the assumption that un-channeled effects are homogeneous across strata, we can use associative and dissociative effects to get some clues on the causal mechanisms by which the treatment affects the outcome. Specifically, if associative effects are large in magnitude relative to dissociative effects, it is reasonable to believe that the mediator channels a part of the treatment effect on the outcome. On the other hand, associative effects that are similar in magnitude relative to the dissociative effects suggest that the treatment affects the outcome mainly through other causal pathways rather than through the mediator of interest (e.g., Zigler *and others*, 2012; Mealli and Mattei, 2012; Forastiere *and others*, 2016).

3.2. *Natural direct and indirect effects*

We conduct mediation analysis focusing on the average natural direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001), which are defined as follows:

$$NDE(z) = \mathbb{E}\left[Y_i(1, S_i(z)) - Y_i(0, S_i(z))\right] \qquad NIE(z) = \mathbb{E}\left[Y_i(z, S_i(1)) - Y_i(z, S_i(0))\right], \tag{3.2}$$

for $z = 0, 1$. $NDE(z)$ measures the effect of the treatment on the outcome Y when the mediator is fixed to the value it would have taken if Z had been set to z, that is, it measures the effect of oral morphine on pain intensity not mediated through the number of doses of morphine. $NIE(z)$ measures the effect on the outcome Y of intervening to set the mediator to what it would have been if Z were set to $z = 1$ in contrast to what it would have been if Z were set to $z = 0$, that is, it measures the extent to which oral morphine affects pain intensity, through the number of doses of morphine. Natural direct effects measure the part of the effect of oral morphine on pain intensity that is not due to a change in the number of doses of morphine, while natural indirect effects measure the effect on pain intensity of a change in the number of doses of morphine, which is due to oral morphine (see Imai *and others*, 2013, and discussion for an insightful debate on natural effects).

Natural effects provide a decomposition of the average total causal effect into the sum of a natural direct effect and a natural indirect effect: $ACE = NDE(z) + NIE(1 - z)$, $z = 0, 1$.

Table 2. *Principal stratification and observed data*

Principal Stratification			Observed Data		
Stratum	$\mathbb{I}\{S_i(0) > s^*\}$	$\mathbb{I}\{S_i(1) > s^*\}$	Z_i	$\mathbb{I}\{S_i^{obs} > s^*\}$	Stratum
00	0	0	0	0	$00 \cup 01$
01	0	1	0	1	$10 \cup 11$
10	1	0	1	0	$10 \cup 00$
11	1	1	1	1	$01 \cup 11$

4. Structural assumptions

Randomization implies

Assumption 4.1 *Ignorability of treatment assignment*. For each $i = 1, \ldots, n$,

$$Pr\,(Z_i \mid S_i(0), S_i(1), Y_i(0), Y_i(1), \mathbf{X}_i) = Pr\,(Z_i)\,.$$

Under Assumption 4.1, we can easily identify the total average causal effect, *ACE*, but here focus is on PCEs and natural direct and indirect effects. In this section we review the assumptions that are usually invoked in principal stratification analysis and mediation analysis, highlighting their different nature.

4.1. *Structural assumptions in principal stratification analysis*

Randomization guarantees that principal strata have the same distribution in both treatment arms, and implies that the treatment is independent of potential outcomes given the principal stratum: $Pr\,(Z_i \mid S_i(0), S_i(1), Y_i(0), Y_i(1), \mathbf{X}_i) = Pr\,(Z_i \mid S_i(0), S_i(1))$, so that treated and control units can be compared conditional on a principal stratum. This is also true if principal strata are defined dichotomizing the intermediate variable, S.

Unfortunately we cannot, in general, observe the principal stratum to which a subject belongs, because we cannot directly observe both $S_i(0)$ and $S_i(1)$ for any subject. In the morphine study, each observed group, defined by the treatment actually received and the observed level of postoperative morphine consumption, is a mixture of two principal strata (see Table 2).

The latent nature of principal strata makes the identification of PCEs a challenging task. Under randomization, without additional structural assumptions, principal strata proportions and PCEs are only partially identified. In principle, we can avoid introducing structural assumptions using a fully Bayesian approach for inference, which does not need full identification. Bayesian inference can proceed in the usual way, and posterior distributions of partially and fully identified parameters can still be compared (see Introduction for further discussion).

In our study, Bayesian inference is conducted under an additional assumption:

Assumption 4.2 *Monotonicity of morphine consumption*. For each $i = 1, \ldots, n$, $S_i^*(1) \leq S_i^*(0)$.

Assumption 4.2 rules out the presence of special patients (01 principal stratum). Although this assumption is not necessary for Bayesian inference, it helps sharpen inference, because under it we can identify principal strata proportions. Assumption 4.2 is not directly verifiable, although it has some testable implications (e.g., Mattei and Mealli, 2011) that we have verified. We also discussed it with physicians and experts, who found it substantially plausible due to the pharmacological characteristics of the active placebo.

4.2. *Structural assumptions in mediation analysis*

The definition of natural direct and indirect effects involves potential outcomes of the form $Y_i(z, S_i(z'))$, $z, z' \in \{0, 1\}$, where $Y_i(z, S_i(z)) = Y_i(z)$, for $z = 0, 1$, and thus we need to incorporate them in the assumption about ignorability of treatment assignment:

Assumption 4.3 *Ignorability of treatment assignment in the presence of potential outcomes of the form $Y_i(z, s)$. $Pr\ (Z_i \mid S_i(0), S_i(1), \{Y_i(0, s), Y_i(1, s); s \in \mathcal{S}\}, \mathbf{X}_i) = Pr\ (Z_i)$* for each $i = 1, \ldots, n$, where $\mathcal{S}$ is the support of S.

Moreover, due to the fact that potential outcomes of the form $Y_i(z, S_i(z'))$ are not observed in this specific experiment, in order to infer natural direct and indirect effects we need to introduce additional structural assumptions that allow us to extrapolate information on $Y_i(z, S_i(z'))$ from the observed data, thus mixing information across principal strata (Mealli and Mattei, 2012). To face this issue, mediation analysis usually invokes assumptions that posit an assignment mechanism for the mediating variable, thereby implying that the mediating variable could be, at least in principle, regarded as an additional treatment. Here, we make the following assumption:

Assumption 4.4 *Sequential ignorability* (Imai *and others*, 2010). For each $i = 1, \ldots, n$,

$$Pr(S_i(z) | Y(z', s), Z_i = z, \mathbf{X}_i) = Pr(S_i(z) | Z_i = z, \mathbf{X}_i) \quad \text{for each } s \in \mathcal{S} \text{ and } z', z \in \{0, 1\}.$$

Assumption 4.4 is not verifiable. It is violated if there exist unobserved variables that are correlated with both the mediator and the outcome even after conditioning on the observed treatment and the observed pretreatment covariates. The plausibility of this assumption rests heavily on the amount of information contained in the covariates, $\mathbf{X}$. Although the presence of a high number of pretreatment variables does not guarantee that Assumption 4.4 is satisfied, the presence of unobserved variables that may make Assumptions 4.4 untenable is less likely if $\mathbf{X}$ is rich enough. In the morphine study, we only have information on two covariates, gender and age, so Assumption 4.4 might be arguable, and results from mediation analysis might not be defensible.

Under Assumptions 4.3 and 4.4, natural effects are point identified (e.g., Imai *and others*, 2010). From a statistical perspective, full identification is valuable, making inference on natural effects relatively straightforward. Nevertheless, it is worth noting that without Assumption 4.4 data contain no information on potential outcomes of the form $Y_i(z, S_i(z'))$, $z \neq z'$. Under ignorability of treatment assignment only, natural effects are partially identified, but without additional assumptions, the worst case bounds (the values at the upper and lower end of the support of the outcome) are required to construct the identification regions. From a Bayesian perspective, the support of the marginal posterior distributions of $\mathbb{E}[Y_i(z, S_i(z'))]$, $z \neq z'$, is the same as the corresponding prior support, and posterior distributions are identical to the prior distributions if parameters are a priori independent (see Gustafson, 2010, and Web Appendix A available at *Biostatistics* online for further details). Recently, alternative causal estimands, named "randomized interventional analogues of natural direct and indirect effects" have been introduced (e.g., VanderWeele *and others*, 2014), which can be identified from the data under milder assumptions, even when Assumption 4.4 fails to hold.

5. Bayesian inference

Bayesian inference is conducted conditional on pretreatment covariates. Covariates do not enter the treatment assignment mechanism (by design), but they enter the sequential ignorability assumption (Assumption 4.4) in mediation analysis. In principal stratification analysis, conditioning on covariates is not required by randomization, however, they can be used to improve estimation because they may be predictive of the principal stratum membership and address confounding due to residual imbalance between treatment groups in finite samples.

5.1. *Bayesian inference for PCEs*

Bayesian inference for PCEs requires the specification of two sets of models: a model for the conditional distribution of the principal stratum membership given the pretreatment variables, and a model for the conditional distribution of potential outcomes given pretreatment variables and principal stratum membership. Let $G_i \equiv (S_i^*(0), S_i^*(1))$ denote the principal stratum membership for unit i, and let $S_i^* = S_i^*(Z_i)$ be the observed value of the binary intermediate outcome S^*. Under Assumption 4.2, $G_i \in \{00, 10, 11\}$, for $i = 1, \ldots, n$. Let $\mathbf{S}^*$ and $\mathbf{G}$ be n-dimensional vectors with ith elements equal to S_i^* and G_i, respectively.

Under exchangeability, we can assume that conditional on a general parameter, denoted by $\boldsymbol{\theta}$, with prior distribution $p(\boldsymbol{\theta})$, the model has an independent and identical distribution (i.i.d.) structure. We specify a Normal outcome model for pain intensity: $Y_i(z) \mid G_i = g, \mathbf{X}_i; \boldsymbol{\theta} \sim N(\mu_{i,g,z}, \sigma^2_{g,z})$, where $\mu_{i,g,z} = \beta_1^{(g,z)} + \boldsymbol{\beta}_X^{(g,z)'}\mathbf{X}_i$, for $g \in \{00, 10, 11\}$ and $z = 0, 1$. We assume that conditional on $\mathbf{X}_i$ and $\boldsymbol{\theta}$, the two outcomes $Y_i(0)$ and $Y_i(1)$ are independent[1] . For the distribution of principal stratum membership we use two conditional probit models, defined using indicator variables $\mathbb{I}\{G_i = 00\}$ and $\mathbb{I}\{G_i = 11\}$ for whether patient i is a pain-tolerant patient or a pain-intolerant patient. Formally, define $G^*_{i,00} = \alpha_1^{(00)} + \boldsymbol{\alpha}_X^{(00)'}\mathbf{X}_i + \epsilon_{i,00}$ and $G^*_{i,11} = \alpha_1^{(11)} + \boldsymbol{\alpha}_X^{(11)'}\mathbf{X}_i + \epsilon_{i,11}$, where $\epsilon_{i,00} \sim N(0,1)$ and $\epsilon_{i,11} \sim N(0,1)$ and independent. Then, $\mathbb{I}\{G_i = 00\} = \mathbb{I}\{G^*_{i,00} \leq 0\}$ and $\mathbb{I}\{G_i = 11\} = \mathbb{I}\{G^*_{i,00} > 0\} \cdot \mathbb{I}\{G^*_{i,11} \leq 0\}$.

Let $\boldsymbol{\alpha}^{(g)} = (\alpha_1^{(g)}, \boldsymbol{\alpha}_X^{(g)})$, $g = 00, 11$ and $\boldsymbol{\beta}^{(g,z)} = (\beta_1^{(g,z)}, \boldsymbol{\beta}_X^{(g,z)})$, $g = 00, 10, 11$; $z = 0, 1$. Given the relatively small sample size in the morphine study, we impose prior equality of the slope coefficients and variances in the outcome regressions across principal strata: $\boldsymbol{\beta}_X^{(00,z)} = \boldsymbol{\beta}_X^{(10,z)} = \boldsymbol{\beta}_X^{(11,z)} \equiv \boldsymbol{\beta}_X^{(z)}$, and $\sigma^2_{00,z} = \sigma^2_{10,z} = \sigma^2_{11,z} \equiv \sigma^2_z$, $z = 0, 1$. We assume that parameters are a priori independent and use proper but low-informative prior distributions (see Web Appendix B available at *Biostatistics* online).

5.2. *Bayesian inference for natural direct and indirect effects*

We conduct mediation analysis using estimators based on linear structural equation models, which are not the only possible estimators, but are still widely used in applied mediation analysis. Formally, in the morphine study we specify the following linear structural models including a product term between the mediator and the treatment status in the model for the outcome:

$$S_i(z) = \alpha_1 + \alpha_2 z + \boldsymbol{\alpha}'_X\mathbf{X}_i + \epsilon_{S,i} \tag{5.1}$$

$$Y_i(z,s) = \beta_1 + \beta_2 z + \beta_3 s + \beta_4 zs + \boldsymbol{\beta}'_X\mathbf{X}_i + \epsilon_{Y,i}, \tag{5.2}$$

with $\epsilon_{S,i} \sim N(0, \sigma_S^2)$ and $\epsilon_{Y,i} \sim N(0, \sigma_Y^2)$ and independent. Note that the structural model for the intermediate outcome, S, in Equation (5.1) implies monotonicity of the mediator with respect to z: $S_i(0) \leq S_i(1)$ for each i if $\alpha_2 \geq 0$, and $S_i(0) \geq S_i(1)$ for each i if $\alpha_2 \leq 0$. Under Assumptions 4.3 and 4.4, the parameters of the structural models in Equations (5.1) and (5.2) are the same as the parameters of the following associational models: $\mathbb{E}[S_i^{obs}|Z_i, \mathbf{X}_i] = \alpha_1 + \alpha_2 Z_i + \boldsymbol{\alpha}'_X\mathbf{X}_i$ and $\mathbb{E}[Y_i^{obs}|Z_i, S_i^{obs}, \mathbf{X}_i] = \beta_1 + \beta_2 Z_i + \beta_3 S_i^{obs} + \beta_4 Z_i S_i^{obs} + \boldsymbol{\beta}'_X\mathbf{X}_i$, therefore, we can use regression methods to estimate natural direct and indirect effects. (e.g. VanderWeele and Vansteelandt, 2009). We conduct Bayesian inference under exchangeability, Assumptions 4.3 and 4.4 and the linear structural models specified in Equations (5.1) and (5.2) (see Web Appendix B available at *Biostatistics* online for details).

6. Results

The posterior distributions of the causal estimands of interest are obtained from Markov chain Monte Carlo (MCMC) methods (see Web Appendix B available at *Biostatistics* online for details).

6.1. *Results from principal stratification and mediation analysis*

We conduct principal stratification analysis using three cut-off points to dichotomize the intermediate variable S: $s^* = 8$, $s^* = 12$ (the overall study sample median), and $s^* = 14$. It is worth noting that the three cut-off points define different binary mediators, and the joint potential values of each binary mediator define a different sub-classification of units into principal strata. Therefore, PCEs corresponding to alternative cut-off points/binary mediators are different causal estimands, that is, causal effects for different sup-populations.

In the sample, about 60% of patients self-administered a number of doses of morphine greater than 8, and 35% of patients self-administered a number of doses of morphine greater than 14. For each cut-off, Table 3 presents posterior mean, standard deviation and 95% posterior credible interval for the average total causal effect, the PCEs, and the proportions of patients belonging to each stratum.

The qualitative conclusions are similar, regardless the cut-off, suggesting that averaging heterogeneous effects over different sup-populations implied by different thresholds leads to similar average treatment effects. Approximately, the average total effects indicate a 19 points reduction in static VAS and 22 points reduction in dynamic VAS due to preoperative oral morphine administration. Around 70% of patients are pain-tolerant or pain-intolerant. The remaining 30% of patients are normal patients, who would lower postoperative morphine consumption as a consequence of receiving oral morphine before surgery.

Dissociative effects appear to be highly heterogeneous: The effect of oral morphine for pain-tolerant patients, $PCE(0, 0)$, is stronger than for pain-intolerant patients, $PCE(1, 1)$. For instance, if we consider the PCEs for dynamic VAS, a reduction greater than 24.8 points in pain intensity on movement is estimated for pain-tolerant patients under all cut-off points, with the associated 95% credible intervals being large, but located far from zero. Conversely, for pain-intolerant patients, the estimated reduction in pain intensity on movement varies from 5.7 (when the cut-off is set to 12) to 14.4 points (when the cut-off is set to 8). In this case, the 95% credible intervals always cover zero. Also dissociative effects on static VAS are heterogeneous between pain-tolerant and pain-intolerant patients, although the differences between the posterior means of $PCE(0, 0)$ and $PCE(1, 1)$ are smaller, and the 95% posterior intervals for $PCE(0, 0)$ are close to zero or cover zero. The associative effects $PCE(1, 0)$ estimate the causal effects of oral morphine in normal patients. If the cut-off is set to eight doses of morphine, a smaller reduction in pain intensity is estimated for normal patients than for pain-tolerant patients. Because normal patients use a larger number of doses of morphine under the active placebo, this result could indicate a non-negligible channeled effect of positive sign for these patients. If the cut-off is set to 12 or 14 doses of morphine, the large reduction in pain intensity estimated for normal patients could be indicative of a strong un-channeled effect due to the active treatment only.

Table 4 presents summary statistics of the posterior distributions for the average total causal effect, and for natural direct and indirect effects. The estimated natural direct and indirect effects show that oral morphine has a strong direct effect in reducing pain intensity both at rest and on movement. The size of the estimated natural direct effects is similar to the size of the total effects (−17.7 and −21.4 for static

Table 3. *Principal stratification analysis: Posterior means, standard deviations and 95% posterior credible intervals for principal strata proportions, PCEs and the average total causal effect*

	Static VAS				Dynamic VAS			
Estimand	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
$S_i^* = \mathbb{I}\{S_i > 8\}$								
π_{10}	0.29	0.09	0.13	0.46	0.26	0.09	0.10	0.44
π_{00}	0.25	0.07	0.12	0.39	0.26	0.07	0.14	0.41
π_{11}	0.46	0.07	0.33	0.60	0.47	0.07	0.34	0.62
$PCE(1,0)$	−20.59	14.09	−47.02	7.36	−15.69	19.33	−49.76	24.38
$PCE(0,0)$	−23.78	11.88	−46.78	−1.12	−36.82	12.45	−61.72	−12.57
$PCE(1,1)$	−13.41	8.09	−29.66	2.35	−14.41	9.64	−32.70	4.71
ACE	−18.13	5.06	−27.89	−8.08	−21.17	5.73	−32.41	−10.03
$S_i^* = \mathbb{I}\{S_i > 12\}$								
π_{10}	0.29	0.08	0.15	0.46	0.28	0.08	0.14	0.44
π_{00}	0.38	0.07	0.24	0.52	0.40	0.07	0.26	0.54
π_{11}	0.32	0.06	0.21	0.45	0.32	0.06	0.21	0.44
$PCE(1,0)$	−27.55	12.70	−53.34	−3.62	−34.09	13.18	−60.04	−8.66
$PCE(0,0)$	−17.49	10.10	−37.08	1.44	−26.01	10.37	−46.44	−5.43
$PCE(1,1)$	−11.79	8.89	−28.79	6.18	−5.69	9.97	−25.04	14.25
ACE	−18.60	4.83	−27.97	−9.24	−21.72	5.77	−33.12	−10.69
$S_i^* = \mathbb{I}\{S_i > 14\}$								
π_{10}	0.28	0.08	0.14	0.44	0.25	0.08	0.11	0.42
π_{00}	0.47	0.07	0.33	0.61	0.50	0.08	0.35	0.65
π_{11}	0.25	0.06	0.14	0.37	0.25	0.06	0.14	0.38
$PCE(1,0)$	−26.59	14.08	−53.51	−0.01	−31.05	16.10	−61.37	0.32
$PCE(0,0)$	−17.74	9.84	−37.14	−0.09	−24.81	9.72	−44.33	−6.18
$PCE(1,1)$	−15.99	11.51	−39.70	6.90	−8.41	12.88	−33.27	17.31
ACE	−19.78	5.10	−29.90	−9.66	−22.34	5.97	−34.39	−10.52

Table 4. *Mediation analysis: Posterior means, standard deviations and 95% posterior credible intervals for natural direct and indirect effects and the average total causal effect*

	Static VAS				Dynamic VAS			
Estimand	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
$NDE(0)$	−17.02	4.66	−26.14	−7.91	−20.55	5.56	−31.33	−9.63
$NIE(1)$	−0.69	1.72	−4.59	2.49	−0.84	2.04	−5.46	2.92
$NDE(1)$	−17.36	4.66	−26.51	−8.12	−22.15	5.55	−33.00	−11.32
$NIE(0)$	−0.35	1.59	−3.93	2.82	0.76	1.93	−2.81	5.12
ACE	−17.71	4.45	−26.39	−8.95	−21.39	5.27	−31.69	−10.95

and dynamic VAS, respectively). Conversely the natural indirect effects are small and their 95% credible intervals cover zero, indicating that the part of the treatment effect channeled by the number of doses of morphine is negligible.

7. Conclusions

We conducted Bayesian principal stratification analysis under the randomization assumption (4.1), which holds by design in the morphine study, and a monotonicity assumption (4.2), which is very plausible in the morphine study and holds also in the mediation analysis under the structural model we consider. In mediation analysis we also invoked a sequential ignorability assumption (4.4). In the morphine study, Assumptions 4.4 might be questionable for various reasons. First, only two pretreatment variables are observed. Second, Assumption 4.4 implies that we can extrapolate information on potential outcomes of the form $Y_i(z, S_i(z'))$ from the observed data, by mixing information across principal strata.

In the morphine study, both analyses, despite focusing on different causal estimands, suggest that there exists a strong un-channeled effect of oral morphine on pain intensity after surgery, which is through other pathways than the number of postoperative doses of morphine. Moreover principal stratification analysis shows evidence that causal effects are highly heterogeneous across principal strata suggesting further investigations for the different types of patients.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank Azienda Ospedaliero-Universitaria Careggi (Firenze, Italy) for providing the data and Dr Fabio Picciafuochi for helpful discussion on the interpretation of the results.
Conflict of Interest: None declared.

Funding

Futuro in Ricerca 2012 grant RBFR12SHVV_003 and PRIN 2012 grant 2012TY8ACC, funded by the Italian Ministry of Research and Higher Education.

Endnotes

[1] In this article, we regard the n subjects in the study as a random sample from a hypothetical super-population, and we focus on super-population causal estimands, that is, average PCEs in this hypothetical population. Super-population average PCEs do not depend on the association between $Y_i(0)$ and $Y_i(1)$, therefore the independence assumption has little inferential effect (Imbens and Rubin, 1997, pages 311–312).

References

Borracci, T., Cappellini, I., Campiglia, L., Picciafuochi, F., Berti, J., Consales, G. and De Gaudio, A.R. (2013). Preoperative medication with oral morphine sulphate and postoperative pain. *Minerva Anestesiologica* **79**, 525–533.

Forastiere, L., Mealli, F. and VanderWeele, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: disentangling bed nets using Bayesian principal stratification. *Journal of the American Statistical Association* **111**, 510–525.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 191–199.

GUSTAFSON, P. (2010). Bayesian inference for partially identified models. *International Journal of Biostatistics* **6**(2), article 17 (17 pages).

HIRANO, K., IMBENS, G. W., RUBIN, D. B. AND ZHOU, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.

IMAI, K., KEELE, L. AND YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* **25**, 51–71.

IMAI, K., TINGLEY, D. AND YAMAMOTO, T. (2013). Experimental designs for identifying causal mechanisms (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **176**, 5–51.

IMBENS, G. W. AND RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* **25**, 305–327.

MATTEI, A. AND MEALLI, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society, B* **73**, 729–752.

MEALLI, F. AND MATTEI, A. (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics* **8**(1), 1–19.

PEARL, J. (2001). Direct and indirect effects. In: Breese, J. S. and Koller, D. (editors), *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 411–420.

ROBINS, J. M. AND GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155.

ROBINS, J. M. AND RICHARDSON, T. S. (2011). Alternative graphical causal models and the identification of direct effects. In: Shrout, P. *and others* (editors), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Oxford University Press, Oxford New York, Chapter 6, pp. 103–158.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* **6**, 34–58.

RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association* **75**, 591–593.

RUBIN, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**, 161–170.

SCHWARTZ, S. L., LI, F. AND MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association* **31**, 949–962.

VANDERWEELE, T. L. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters* **78**, 2957–2962.

VANDERWEELE, T. J. AND VANSTEELANDT, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Inference* **2**, 457–468.

VANDERWEELE, T. J., VANSTEELANDT, S. AND ROBINS, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* **25**, 300–306.

ZIGLER, C. M., DOMINICI, F. AND WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially-correlated multivariate intermediate outcomes. *Biostatistics* **12**, 289–302.

[*Received* May 19, 2015; *revised* January 29, 2017; *accepted for publication* January 31, 2017]