UNIVERSITÀ DEGLI STUDI FIRENZE

DISIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

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

Michela Baccini,
Alessandra Mattei, Fabrizia Mealli

DISIA WORKING PAPER
2015/06

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

Michela Baccini
email: baccini@disia.unifi.it
Dipartimento di Statistica, Informatica, Applicazioni, University of Florence
Biostatistics Unit, ISPO Cancer Prevention and Research Institute

Alessandra Mattei, Fabrizia Mealli
email: mattei@disia.unifi.it, mealli@disia.unifi.it
Dipartimento di Statistica, Informatica, Applicazioni, University of Florence

Abstract

Principal stratification and mediation analysis are two ways to conceptualize the mediating role of an intermediate variable in the causal pathways by which a treatment affects an outcome. They are often viewed as competing frameworks, and their role in dealing with issues concerning causal mechanisms has often fired up glowing discussions. However a thoughtful comparative analysis, highlighting the substantive differences between the two frameworks is still lacking. We aim at filling this gap conducting both principal stratification and mediation analysis using, as a motivating example, a prospective, randomized, double-blind study to investigate to which extent the positive overall effect of treatment on postoperative pain control is mediated by postoperative self administration of intra-venous analgesia by patients. Using the Bayesian approach for inference, we estimate both associative and dissociative principal strata effects arising in principal stratification analysis, as well as natural effects and controlled direct 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 identifying assumptions. We discuss these aspects along the results arising from our analyses.

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

1 Introduction

Principal stratification and casual mediation analysis are two ways to conceptualize the mediating role of an intermediate variable in the causal pathways between treatment and outcome that have received increasing attention in the last years. However, they are often viewed as competing frameworks. One exception is VanderWeele (2008), who compares the concepts of associative and dissociative principal strata effects arising in a principal stratification framework (Frangakis and Rubin, 2002) and the notions of direct and indirect effects from mediation analysis (Robins and Greenland, 1992; Pearl, 2001), showing 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.

In this paper, we aim at filling 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, to which we refer as "the morphine study" throughout the article, involved adult patients who were undergoing an elective open colorectal abdominal surgery. Patients were randomized to received before surgery the experimental treatment or an active placebo and the outcomes of primary interest was postoperative pain intensity, measured using a visual analogue scale. According to the medical guidelines for pain control, after surgery, patients received an IntraVenous Patient Controlled Analgesia (IVPCA) system programmed to give off fixed doses of morphine sulphate upon patient demand. 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 the channeled (indirect) effect mediated through postoperative self-administration of morphine sulphate, and the unchanneled (direct) effect (that is, the effect not mediated through postoperative self-administration of morphine sulphate) from one another. Borracci et al. (2013), who first analyzed data from the morphine study, face the problem by conditioning on the observed number of self-administered doses of morphine sulphate, including that post-treatment intermediate variable as a covariate in regression models. The comparison of pain intensity between treated and control patients adjusted for the observed value of the post-treatment variable may provide some insight into the treatment mechanism, but lacks a causal interpretation, unless the treatment has no effect on the post-treatment variable (Rosenbaum, 1984). This does not seem to be the case in the morphine study, where preoperative administration of oral morphine sulphate is effective in reducing postoperative self-administration of morphine sulphate (Borracci et al., 2013). In this article we use principal stratification and mediation analysis to get some information on the extent to which the effect of the treatment on the outcome 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 glowing discussions in the causal community. In this article, we aim at smoothing these controversies over, using the morphine study to highlight that mediation analysis and principal stratification analysis generally focus on different causal estimands, answer different questions and involve different sets of identifying assumptions, which lead to use the information provided by the data in a substantially different way.

Principal stratification focuses on local causal effects, that is, causal effects for specific sub-populations named principal strata. Despite the local nature of principal strata effects, we view the concept of principal stratification as a useful principle for addressing the topic of direct and indirect causal effects. Mediation analysis focuses on disentangling direct and indirect effects, which are generally defined at the individual level and averaged over the whole population, like natural direct and indirect effects and controlled direct effects we consider in our study.

We adopt a Bayesian approach for inference. From a Bayesian perspective, inference is based on the posterior distribution of the causal estimands of interest. The Bayesian analysis yields valid estimates of quantities of interest and also properly accounts for uncertainty about these quantities (Rubin, 1978).

The remainder of the paper 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 propose a Bayesian approach for principal stratification analyses and mediation analysis, specifying our modeling assumptions. We present and discuss the results of the analyses in Section 5 and conclude the article in Section 6.

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 control. 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 Intra-Venous Patient-Controlled Analgesia (IV-PCA). The device was 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 Visual Analogue Scale (VAS) scores at rest and for movement (that is, upon coughing). We also refer to these outcome variables 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 4 hours after the end of surgery (see Borracci et al., 2013, for further details on the study).

Our objective is to measure the causal effect of preoperative medication on pain relief, accounting for postoperative self-administration of morphine sulphate. Postoperative self-administration of morphine sulphate 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 a part of the treatment effect. Indeed, the number of morphine doses administered upon patient demand may cause variation in pain intensity, but at the same time it could vary depending on the preoperative treatment. Therefore, a key issue is about how to extricate the channeled and unchanneled effects from one another.

2.1 Notation and Descriptive Analyses

In order to answer the research question of interest we first introduce some notation. We will frame our discussion in the context of the potential outcome approach to causal inference, also known as the Rubin Causal Model (Rubin, 1974, 1978).

Each patient who participates in the study can either be assigned to the oral morphine group, $z = 1$, or 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 $S_i(1)$ be the number of self-administered post-operative doses of morphine sulphate if the patient is exposed to preoperative oral morphine, and let $S_i(0)$ be the number of self-administered post-operative doses of morphine sulphate if the patient is exposed to the active placebo. 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.

For each patient i, we observe the treatment actually assigned, Z_i, and only one potential outcome for each post-treatment variable depending on the treatment actually assigned. Let $S_i^{obs} = S_i(Z_i)$ be the observed number of self-administered post-operative doses of morphine sulphate, 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, to simplify the notation, we will use $Y_i(z)$ and Y_i^{obs} to denote the potential outcomes and the actual outcomes for pain intensity at rest or for movement, dropping the second subscript, unless necessary to avoid misunderstandings. For each patient we also observe two covariates, X_{i1}, gender, and X_{i2}, age in years. The vectors $\mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}$ are $n-$dimensional vectors with ith elements equal to $Z_i, S_i^{obs}, Y_i^{obs}$, respectively. The $n \times 2$ matrix $\mathbf{X}$ has ith row equal to $\mathbf{X}_i' \equiv (X_{i1}, X_{i2})$.

If the intermediate variable, S, could be, at least in principle, regarded as an additional treatment and could be at least potentially controlled by external interventions, under an appropriate version of SUTVA (see, e.g., Mattei and Mealli, 2011), we could also define

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

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 the level z and the mediator S were set to a specific prefixed value, s; and $Y_i(z, S_i(z'))$ would be the value of the outcome 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 this type include the values of pain intensity under oral morphine, if the number of self-administered doses of morphine sulphate somehow were simultaneously forced to attain a specific value s, or the value it would have taken under the active placebo.

Even if we are willing to regard the treatment variable, Z, and the intermediate variable, S, as a multivariate treatment variable, (Z, S), and to hypothesize the existence of potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$, the intermediate variable is indeed a post-treatment variable, which can be potentially affected by treatment assignment. Therefore some potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$ are a priori counterfactuals in the experiment, because $Y_i(z, s)$ and $Y_i(z, S_i(z'))$ can be never observed for units for whom $S_i(z) \neq s$ and $S_i(z) \neq S_i(z')$ for $z \neq z'$, respectively. In this specific experiment, the potential outcomes $Y_i(z, s)$ and $Y_i(z, S_i(z'))$ can be never observed for such type of patients. 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.

Table 1 presents some summary statistics for the sample, classified by treatment assignment, Z_i. As can be seen in Table 1, there is some evidence that preoperative administration of oral morphine sulphate reduces the number of postoperative self-administrated doses of morphine sulphate and reduces pain intensity, both at rest and for movement.

3 Causal Estimands

In the potential outcome approach to causal inference a causal effect of the treatment Z on an outcome Y is defined as a comparison of the potential outcomes on a common set of units. Here we focus on the average causal effect of the preoperative treatment on pain intensity, defined as the expected difference between potential outcomes for the study population:

$$ACE = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]. \tag{1}$$

However, this causal estimand does not account for postoperative self-administration of analgesia, S. In order to get some insight on the causal pathways by which preoperative administration of oral morphine sulphate affects pain intensity, we use both principal stratification and mediation analysis.

Principal stratification analysis may provide useful information by looking at the joint value of the mediating variable under treatment and under control, $(S_i(0), S_i(1))$. The joint potential value of $S_i(0)$ and $S_i(1)$ is essentially a characteristic of a subject, describing how an individual reacts to the treatment. The framework of principal stratification focuses on local causal effects, that is, causal effects for specific subpopulations (principal strata), therefore it does not always answer the causal question of primary interest, but often provides useful insights, and has the advantage to avoid a priori counterfactuals. In the principal stratification framework causal effects are defined using only potential outcomes of the form $Y_i(z)$ and $S_i(z)$.

Causal mediation analysis focuses on disentangling direct and indirect effects, which are generally defined at the individual level and averaged over the whole population. To formalize the concepts of direct and indirect effects mediation analysis usually involves potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$.

3.1 Principal Stratification and Principal Strata Effects

Principal stratification uses the joint value of the potential intermediate values to define a stratification of the population into principal strata. Formally, the *basic* principal stratification with respect to a post-treatment variable S is the partition of subjects into sets such that all subjects in the same set have the same vector $(S_i(0); S_i(1))$. A *principal stratification* with respect to the post-treatment variable S is a partition of units whose sets are unions of sets in the basic principal stratification (Frangakis and Rubin, 2002).

In the morphine study principal strata are defined by the joint potential values of the number of self-administrated doses of morphine under the oral morphine treatment and under the active placebo treatment. The intermediate variable takes on several values, thus the basic principal stratification leads to classify units into several principal strata. Given the reduced sample size, here we prefer to focus on a simplified setting with a binary intermediate variable. Therefore we apply the principal stratification approach by dichotomizing the intermediate variable, considering a binary variable equal to 1 if patients use 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 a function taking the value one if its argument is true and the value zero otherwise. It should be noticed that all the conceptual issues surrounding the comparison between principal stratification analysis and causal mediation analysis are captured also using a binary version of the intermediate variable within the principal stratification framework.

The basic principal stratification with respect to the binary intermediate variable S^* partitions patients into four latent groups: (1) patients who would self-administer morphine sulphate at a low level irrespective of their treatment assignment $00 = \{i :$

$S_i^*(0) = 0, S_i^*(1) = 0\}$, whom we label as "pain-tolerant patients"; (2) patients who would self-administer morphine sulphate at a high level under the active placebo, but would self-administer morphine sulphate at a low level under oral morphine $10 = \{i : S_i^*(0) = 1, S_i^*(1) = 0\}$, whom we label as "normal patients"; (3) patients who would self-administer morphine sulphate at high level irrespective of their treatment assignment $11 = \{i : S_i^*(0) = 1, S_i^*(1) = 1\}$, whom we label as "pain-intolerant patients"; and (4) patients who would self-administer morphine sulphate at a low level under the active placebo, but would self-administer morphine at a high level under oral morphine $01 = \{i : S_i^*(0) = 0, S_i^*(1) = 1\}$, whom we label as "special patients."

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

$$PCE(s_0, s_1) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid S_i^*(0) = s_0, S_i^*(1) = s_1\right]. \qquad (2)$$

Due to the fact that principal strata are not affected by treatment assignment by definition, principal effects are always well-defined causal effects.

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

The average total effect is the weighted average of principal causal effects 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 $\{i : S_i^*(0) = s_0, S_i^*(1) = s_1\}$. However, it should be stressed that associative and dissociative principal causal effects do not in general allow one to decompose the total effect into overall direct and indirect effects, unless additional assumptions are made.

Principal stratification makes it clear that only in strata where the intermediate variable is unaffected by the treatment can we hope to learn something about the direct effect of the treatment. Dissociative principal causal effects naturally provide information on the existence of a direct 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 (Mealli and Rubin, 2003). If dissociative principal causal effects are all zero, then there is no evidence on the unchanneled (direct) effect of the treatment after controlling for the mediator (Rubin, 2004; Mattei and Mealli, 2011). This does not mean that there is no direct effect of the treatment because associative effects generally combine unchanneled (direct) and channeled effects (e.g., VanderWeele, 2008).

3.2 Natural Direct and Indirect Effects and Controlled Direct Effects

We conduct mediation analysis focusing on the concepts of controlled direct effects and natural direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001). Formally, the average controlled direct effect of the treatment Z on the outcome Y, setting S to s, is defined as follows:

$$CDE(s) = \mathbb{E}\left[Y_i(1, s) - Y_i(0, s)\right]. \tag{3}$$

$CDE(s)$ measures the effect of Z on Y after intervening to fix the mediator, S, to a prefixed value s. In other word, the controlled direct effect measures the effect of preoperative oral morphine sulphate on pain intensity that is not mediated through the number of self-administrated doses of morphine sulphate, which is assumed to be fixed to a specific value. Controlled direct effects are prescriptive in the sense that the intermediate outcome is fixed at a prescriptive value: $CDE(s)$ measures the causal effect of preoperative oral administration of morphine sulphate on pain intensity if a prescribed number of postoperative doses of morphine sulphate, s, were administered to all patients in the population.

The average natural direct and indirect effects are defined as follows:

$$\begin{aligned} NDE(z) &= \mathbb{E}\left[Y_i(1, S_i(z)) - Y_i(0, S_i(z))\right] && (4)\\ NIE(z) &= \mathbb{E}\left[Y_i(z, S_i(1)) - Y_i(z, S_i(0))\right], && (5)\end{aligned}$$

for $z = 0, 1$. $NDE(z)$ measures the effect on the outcome Y of intervening to fix the mediator to the value it would have taken if Z had been set to z, that is, it measures the effect of the administration of preoperative oral morphine sulphate on pain intensity not mediated through self-administration of morphine sulphate. $NIE(z)$ measures the effect on the outcome Y of intervening to set the mediator to what it would have been if Z were $z = 1$ in contrast to what it would have been if Z were $z = 0$, that is, it measures the extent to which the administration of preoperative oral morphine sulphate affects pain intensity, through the number of postoperative self-administrated doses of morphine sulphate. Natural direct effects are the part of the effect of the administration of preoperative oral morphine sulphate on pain intensity that is not due to a change in the number of self-administrated doses of morphine sulphate, while natural indirect effects measure the effect on pain intensity of a change in the number of postoperative self-administrated doses of morphine sulphate, which is due to the administration of preoperative oral morphine sulphate. Pearl (2001) originally defines natural effects as *descriptive* tools, in the sense that they describe the part of the effect attributable to the treatment itself and the part of the effect attributable to an intervention on the intermediate variable which reproduces *natural conditions*; this interpretation has received some criticisms though (see Imai et al., 2013, and discussion).

Natural effects provide a decomposition of the average total causal effect into the sum of a natural direct effect and a natural indirect effect and thus, should, at least in principle, discover causal mechanisms: $ACE = NDE(0)+NIE(1)$ and $ACE = NDE(1)+NIE(0)$. Conversely, controlled direct effects do not in general allow one to decompose the total effect into overall direct and indirect effects.

4 Structural Assumptions

In the morphine study patients are assigned to either the oral morphine group or the active placebo group according to a completely randomized experiment. Randomization implies that oral morphine is assigned independently of all potential outcomes and pretreatment covariates. Formally,

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

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

Under Assumption 1 we can easily identify the total average causal effect, ACE, but here interest focuses on principal strata effects and direct and indirect effects. The assumptions that allow us to identify principal strata effects and direct and indirect effects are of a different nature and a careful evaluation of their plausibility is crucial.

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 strata: $Pr\left(Z_i \mid S_i(0), S_i(1), Y_i(0), Y_i(1), \mathbf{X}_i\right) = Pr\left(Z_i \mid S_i(0), S_i(1)\right)$, 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 which a subject belongs to, because we cannot directly observe both $S_i(0)$ and $S_i(1)$ for any subject. Observed groups are typically mixtures of principal strata. 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 principal strata effects in principal stratification analysis a challenging task. In principle, we can avoid the introduction of structural assumptions using a fully Bayesian approach for inference. In fact, the fully Bayesian approach does not need full identification (e.g., Imbens and Rubin, 1997). Under randomization, without additional structural assumptions, models, and thus principal causal effects, are weakly identified in the sense that their posterior distributions

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	$01 \cup 11$

are proper (and this is is always true with proper priors) but have substantial regions of flatness.

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

Assumption 2 *Monotonicity of Morphine Consumption. For each* $i = 1, \ldots, n$, $S_i^*(1) \leq S_i^*(0)$.

Assumption 2 rules out the presence of special patients who would self-administer a low number of morphine doses under the active placebo treatment but would self-administer a high number of morphine doses under the oral morphine treatment (01 principal stratum). Although this assumption is not necessary for Bayesian inference, it helps sharpen inference, because under Assumption 2 we can identify the principal stratum proportions. Assumption 2 is not directly verifiable, and it is a strong assumption, which may not be satisfied. We thoroughly discussed it with physicians and experts, who found it substantially plausible due to the pharmacological characteristics of the active placebo. Indeed, underlying Assumption 2 is the clue that although patients may use morphine sulphate at high level after surgery upon receipt of the oral morphine sulphate treatment, because, e.g., they are highly sensitive to pain, they are unlikely to use morphine sulphate at high level after surgery upon receipt of oral morphine sulphate if they would have used morphine sulphate at low level under the active placebo treatment.

4.2 Structural Assumptions in Mediation Analysis

The definition of natural direct and indirect effects as well as the definition of controlled direct effects involve potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$. If we admit the existence of these potential outcomes, we need to incorporate them in the assumption about ignorability of treatment assignment. Formally:

Assumption 3 *Ignorability of treatment assignment in the presence of potential outcomes of the form* $Y_i(z, s)$. *For each* $i = 1, \ldots, n$,

$$Pr\left(Z_i \mid S_i(0), S_i(1), Y_i(0, s), Y_i(1, s), \mathbf{X}_i\right) = Pr\left(Z_i\right) \text{ for each } s \in \mathcal{S}$$

Note that potential outcomes of the form $Y_i(z, S_i(z'))$ and $Y_i(z)$ can be viewed as specific values of $Y_i(z, s)$, where the intermediate variable is forced to attained the value $s = S_i(z')$ and $s = S_i(z)$, respectively.

Moreover, due to the fact that potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$ are never observed in this specific experiment for some patients, in order to identify natural direct and indirect effects and controlled direct effects we need to introduce additional structural assumptions that allow us to extrapolate information on $Y_i(z, s)$ and $Y_i(z, S_i(z'))$ from the observed data. 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. Alternative sets of assumptions have been proposed in the literature (see, e.g., Pearl, 2001; Robins and Greenland, 1992; Robins, 2003; Petersen et al., 2006; VanderWeele and Vansteelandt, 2009). Here, we focus on the assumptions proposed by VanderWeele and Vansteelandt (2009).

Specifically, in order to identify controlled direct effects, we assume that the number of self-administered doses of morphine sulphate is assigned independently of potential outcomes for pain intensity, $Y_i(z, s)$, given the observed treatment and pretreatment variables:

Assumption 4 *Sequential Ignorability 1. For each $i = 1, \dots, n$,*

$$Pr\left(S_i^{obs} \mid Y_i(z, s), Z_i, \mathbf{X}_i\right) = Pr\left(S_i^{obs} \mid Z_i, \mathbf{X}_i\right) \text{ for each } s \in \mathcal{S} \text{ and } z \in \{0, 1\}$$

In order to identify natural direct and indirect effects we also need to impose an additional ignorability assumption, which implies that potential outcomes for the number of self-administrated doses of morphine sulphate are independent of potential outcomes for pain intensity given pretreatment variables:

Assumption 5 *Sequential Ignorability 2. For each $i = 1, \dots, n$,*

$$Pr\left(S_i(z') \mid Y_i(z, s), \mathbf{X}_i\right) = Pr\left(S_i(z') \mid \mathbf{X}_i\right) \text{ for each } s \in \mathcal{S}, \text{ and } z', z \in \{0, 1\}$$

Assumptions 4 and 5 are not verifiable and are strong assumptions, which may be not satisfied in many cases, even if the intermediate variable can be reasonably regarded as an additional treatment. The plausibility of these assumptions rests heavily on the amount of information contained in the covariates, $\mathbf{X}$: the higher the dimension of $\mathbf{X}$, the more plausible we might consider Assumptions 4 and 5 to be. In the morphine study, we only have information on two covariates, gender and age, so Assumptions 4 and 5 might be arguable, and results from mediation analysis might not be defensible.

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 assumptions (Assumptions 4 and 5) in mediation analysis. In principal stratification analysis, conditioning on covariates is not required by randomization, however they can be used to improve efficiency of estimation and address confounding due to residual unbalance between treatment groups in finite samples.

Bayesian Inference for Principal Causal Effects

Bayesian inference for principal causal effects requires to specify two sets of models: a model for the conditional distribution of the principal stratum membership given pretreatment variables, and a model for the conditional distribution of potential outcomes given pretreatment variables and principal stratum membership (Imbens and Rubin, 1997). 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 2 (monotonicity of morphine consumption), $G_i \in \{00, 10, 11\}$, 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. Formally, denote $\pi_{i,g} = Pr(G_i = g \mid \mathbf{X}_i; \boldsymbol{\theta})$ and $f_{i,g,z} = p(Y_i(z) \mid G_i = g, \mathbf{X}_i; \boldsymbol{\theta})$. Then, the posterior distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y}^{obs}, \mathbf{S}^*, \mathbf{X}) \propto p(\boldsymbol{\theta}) \times \prod_{i:Z_i=0,S_i^*=0} \pi_{i,00} \cdot f_{i,00,0} \times \prod_{i:Z_i=1,S_i^*=1} \pi_{i,11} \cdot f_{i,11,1} \times$$
$$\prod_{i:Z_i=0,S_i^*=1} \left[\pi_{i,10} \cdot f_{i,10,0} + \pi_{i,11} \cdot f_{i,11,0}\right] \times \prod_{i:Z_i=1,S_i^*=0} \left[\pi_{i,00} \cdot f_{i,00,1} + \pi_{i,10} \cdot f_{i,10,1}\right],$$

where the sum in the likelihood is because patients with $(Z_i = 0, S_i^* = 1)$ are mixture of normal patients and pain-intolerant patients, and patients with $(Z_i = 1, S_i^* = 0)$ are mixture of normal patients and pain-tolerant patients.

We assume 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 $\epsilon_{i,11} \sim N(0,1)$ independently. Then, $\mathbb{I}\{G_i = 00\} = \mathbb{I}\{G^*_{i,00} \leq 0\}$ and $\mathbb{I}\{G_i = 11\} = \mathbb{I}\{G^*_{i,00} > 0\} \cdot \mathbb{I}\{G^*_{i,11} \leq 0\}$. It is worth noting that although the order of listing the principal strata, and so the choice of the baseline principal stratum, is irrelevant from a theoretical perspective, the performances of the algorithm we use to derive the posterior distribution of the parameters of the principal stratum sub-model improve in terms of convergence rate using the group of normal patients (the 01 group) as reference group. This result is intuitive because normal patients are always observed in a mixture with another group (either pain-tolerant or pain-intolerant patients), whereas under the monotonicity assumption, the observed

[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 principal causal effects in this hypothetical population. Super-population average principal causal effects 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).

groups $(Z_i = 0, S_i^* = 0)$ and $(Z_i = 1, S_i^* = 1)$ only include pain-tolerant patients and pain-intolerant patients, respectively.

Let $\boldsymbol{\alpha}^{(g)} = (\alpha_1^{(g)}, \boldsymbol{\alpha}_X^{(g)})$, $g = 00, 11$ and $\boldsymbol{\beta}^{(g,z)} = (\beta_1^{(g,z)}, \boldsymbol{\beta}_X^{(g,z)})$, $g = 00, 10, 11$, $z = 0, 1$. The full parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{(00)}, \boldsymbol{\alpha}^{(11)}, \boldsymbol{\beta}^{(00,0)}, \sigma^2_{00,0}, \boldsymbol{\beta}^{(00,1)}, \sigma^2_{00,1}, \boldsymbol{\beta}^{(10,0)}, \sigma^2_{10,0}, \boldsymbol{\beta}^{(10,1)}, \sigma^2_{10,1}, \boldsymbol{\beta}^{(11,0)}, \sigma^2_{11,0}, \boldsymbol{\beta}^{(11,1)}, \sigma^2_{11,1})$, which consists of 30 elements. 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: $\boldsymbol{\beta}_X^{(00,0)} = \boldsymbol{\beta}_X^{(10,0)} = \boldsymbol{\beta}_X^{(11,0)} \equiv \boldsymbol{\beta}_X^{(0)}$, $\sigma^2_{00,0} = \sigma^2_{10,0} = \sigma^2_{11,0} \equiv \sigma^2_0$, and $\boldsymbol{\beta}_X^{(00,1)} = \boldsymbol{\beta}_X^{(10,1)} = \boldsymbol{\beta}_X^{(11,1)} \equiv \boldsymbol{\beta}_X^{(1)}$, and $\sigma^2_{00,1} = \sigma^2_{10,1} = \sigma^2_{11,1} \equiv \sigma^2_1$, reducing the number of parameters to 18.

We assume that parameters are a priori independent and use proper but uninformative prior distributions. The prior distributions for the principal stratum model are $\boldsymbol{\alpha}^{(00)} \sim N\left(\underline{\mu}_{\boldsymbol{\alpha}^{(00)}}, \underline{\Sigma}_{\boldsymbol{\alpha}^{(00)}}\right)$ and $\boldsymbol{\alpha}^{(11)} \sim N\left(\underline{\mu}_{\boldsymbol{\alpha}^{(11)}}, \underline{\Sigma}_{\boldsymbol{\alpha}^{(11)}}\right)$ where $\underline{\mu}_{\boldsymbol{\alpha}^{(00)}}$ and $\underline{\mu}_{\boldsymbol{\alpha}^{(11)}}$ are null vectors and $\underline{\Sigma}_{\boldsymbol{\alpha}^{(00)}}$ and $\underline{\Sigma}_{\boldsymbol{\alpha}^{(11)}}$ are set at $10^6\mathbb{I}_3$, where $\mathbb{I}_k$ is the identity matrix of order k. The prior distributions for the outcome models are $\beta_1^{(g,z)} \sim N\left(\underline{\mu}_{\beta_1^{(g,z)}} = 0, \underline{\sigma}^2_{\beta_1^{(g,z)}} = 10^6\right)$, $g \in \{00, 10, 11\}$, $z = 0, 1$; $\boldsymbol{\beta}_X^{(z)} \sim N\left(\underline{\mu}_{\boldsymbol{\beta}_X^{(z)}} = \mathbf{0}, \underline{\Sigma}_{\boldsymbol{\beta}_X^{(z)}} = 10^6\mathbb{I}_2\right)$, $z = 0, 1$; and $\sigma^2_z \sim \mathrm{Inv} - \chi^2_{\underline{\nu}_z}(\underline{s}^2_z)$, with $\underline{\nu}_z = 0.002$ and $\underline{s}^2_z = 1$, $z = 0, 1$.

Bayesian Inference for Direct and Indirect Effects

Under Assumptions 3, 4 and 5 we can use regression models for estimate natural direct and indirect effects and average controlled direct effects. Following VanderWeele and Vansteelandt (2009), in the morphine study we use linear regression models including a product term between the mediator and the treatment indicator in the model for the outcomes, allowing the exposure to interact in its effect on the outcomes with the mediator:

$$
\begin{aligned}
S_i^{obs} &= \alpha_1 + \alpha_2 Z_i + \boldsymbol{\alpha}'_X \mathbf{X}_i + \epsilon_{S,i} && (6)\\
Y_i^{obs} &= \beta_1 + \beta_2 Z_i + \beta_3 S_i^{obs} + \beta_4 Z_i S_i^{obs} + \boldsymbol{\beta}'_X \mathbf{X}_i + \epsilon_{Y,i}, && (7)
\end{aligned}
$$

where $\epsilon_{S,i} \sim N(0, \sigma^2_S)$ and $\epsilon_{Y,i} \sim N(0, \sigma^2_Y)$ independently.

We conduct Bayesian inference under exchangeability, Assumptions 3, 4 and 5 and the linear models specified in Equations (6) and (7). The full parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \sigma^2_S, \boldsymbol{\beta}, \sigma^2_Y)$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \boldsymbol{\alpha}_X)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \boldsymbol{\beta}_X)$, for a total of 12 parameters. We assume that parameters are a priori independent with prior distributions $\boldsymbol{\alpha} \sim N\left(\underline{\mu}_{\boldsymbol{\alpha}}, \underline{\Sigma}_{\boldsymbol{\alpha}}\right)$, $\sigma^2_S \sim \mathrm{Inv} - \chi^2_{\underline{\nu}_S}(\underline{s}^2_S)$, $\boldsymbol{\beta} \sim N\left(\underline{\mu}_{\boldsymbol{\beta}}, \underline{\Sigma}_{\boldsymbol{\beta}}\right)$, $\sigma^2_Y \sim \mathrm{Inv} - \chi^2_{\underline{\nu}_Y}(\underline{s}^2_Y)$. In order to specify flat priors, we set $\underline{\mu}_{\boldsymbol{\alpha}}$ and $\underline{\mu}_{\boldsymbol{\beta}}$ to be null vectors, $\underline{\Sigma}_{\boldsymbol{\alpha}}$ and $\underline{\Sigma}_{\boldsymbol{\beta}}$ to be $10^6\mathbb{I}_4$ and $10^6\mathbb{I}_6$, respectively, $\underline{\nu}_S = 0.002$, $\underline{s}^2_S = 1$, $\underline{\nu}_Y = 0.002$ and $\underline{s}^2_Y = 1$. The posterior distribution for the parameter vector $\boldsymbol{\theta}$ is

$$
p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{Y}^{obs}, \mathbf{X}) \propto p(\boldsymbol{\alpha}) p(\sigma^2_S) p(\boldsymbol{\beta}) p(\sigma^2_Y) \times \textstyle\prod_{i=1}^n \left[\frac{1}{\sqrt{\sigma^2_S}} \phi\left(\frac{S_i^{obs} - (\alpha_1 + \alpha_2 Z_i + \boldsymbol{\alpha}'_X \mathbf{X}_i)}{\sqrt{\sigma^2_S}} \right) \right.
$$

$$
\left. \times \frac{1}{\sqrt{\sigma^2_Y}} \phi\left(\frac{Y_i^{obs} - (\beta_1 + \beta_2 Z_i + \beta_3 S_i^{obs} + \beta_4 Z_i S_i^{obs} + \boldsymbol{\beta}'_X \mathbf{X}_i)}{\sqrt{\sigma^2_Y}} \right) \right],
$$

where $\phi(\cdot)$ is the probability density function of a standard normal distribution.

Results

In this section, we show results for the causal estimands of interest in principal stratification analysis and mediation analysis. The posterior distributions of the parameters are obtained from Markov chain Monte Carlo (MCMC) methods. Note that the causal estimands are not parameters of the models, but rather are functions of parameters and data. Details on the MCMC algorithms we adopt and the functions defining the causal estimands of interest are given in Appendix.

5.1 Results from Principal Stratification 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 $s^* = 14$. About 60% of patients self-administered a number of morphine doses greater than 8, and 35% of patients self-administered a number of morphine doses 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, and for the principal causal effects and the proportions of patients belonging to each stratum.

The qualitative conclusions are similar, regardless the cut-off. Approximately, the average total effects indicate a 19 points reduction in static VAS and 22 points reduction in dynamic VAS due to the administration of oral morphine before surgery.

More than 70% of patients are pain-tolerant or pain-intolerant, that is, patients whose postoperative consumption of morphine sulphate is unaffected by preoperative administration of oral morphine sulphate. The remaining 30% are normal patients, who would lower postoperative morphine consumption as a consequence of receiving oral morphine sulphate before surgery.

Dissociative effects, which provide information on the presence of an unchanneled (direct) effect, appear to be heterogeneous: The effect of preoperative oral morphine sulphate for pain-tolerant patients, $PCE(0,0)$, is stronger than for pain-intolerant patients, $PCE(1,1)$. For instance, if we consider the principal causal effects 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% posterior 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% posterior intervals always cover zero. Heterogeneity of the causal effect between pain-tolerant and pain-intolerant patients arises also for static VAS, 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 0.

Table 3: Principal stratification analysis: Posterior means, standard deviations and 95% posterior credible intervals of principal strata proportions, principal causal effects 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
$PCE(1,0)$	-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

The associative effect $PCE(1,0)$ estimates the causal effect of preoperative oral morphine in normal patients. If the cut-off is set to 12 or 14 self-administered doses of morphine sulphate, a larger reduction in pain intensity is estimated for normal patient than for pain-tolerant and pain-intolerant patients. If the cut-off is set to 8, the value of $PCE(1,0)$ is intermediate between the two dissociative effects.

5.2 Results from Mediation Analysis

Table 4 presents summary statistics of the posterior distributions for the average total causal effect, and for natural direct and indirect effects, and Figure 1 shows the posterior means and the 95% posterior credible intervals for controlled direct effects calculated fixing the number of self-administered doses of morphine sulphate at different values s,

Table 4: Mediation analysis: Posterior means, standard deviations and 95% posterior credible intervals of 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

$s = 2, 4, 6, \ldots, 32, 34, 36$.

The estimated natural direct and indirect effects show that preoperative administration of 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 and dynamic VAS, respectively). Conversely the natural indirect effects are small and their 95% posterior credible intervals cover zero, indicating that the part of the treatment effect channeled by the number of self-administered doses of morphine sulphate is negligible.

Results on controlled direct effects suggest that the direct effect of preoperative administration of oral morphine sulphate does not vary very much with the number of self-administered doses of morphine sulphate for static VAS. On the contrary, for dynamic VAS, pain reduction attributable to the administration of oral morphine is lower the higher the self-administration of morphine sulphate after surgery is. For both outcomes, controlled direct effects are clearly different from zero if the number of self-administered doses of morphine sulphate is lower than 24 (credible intervals do not include zero).

6 Conclusions

Even if principal stratification analysis and mediation analysis focus on different causal estimands and answer different causal questions, in this specific application they both suggest that there exist a strong unchanneled effect of preoperative administration of oral morphine on pain intensity after surgery, which is through other pathways other than the postoperative number of self-administered doses of morphine sulphate.

While in the case of mediation analysis, this conclusion directly derives from the fact that natural indirect effects are negligible, in the case of principal stratification analysis, evidence on the existence of a unchanneled effect is only for the subsets of pain-intolerant and pain-tolerant patients, according to the size of the dissociative effects. Regarding normal patients, we are not able to draw the same conclusions unless additional assumptions are made, because associative effects are a mixture of unchanneled and channeled effects. In this sense, principal stratification could not answer to the causal question of primary

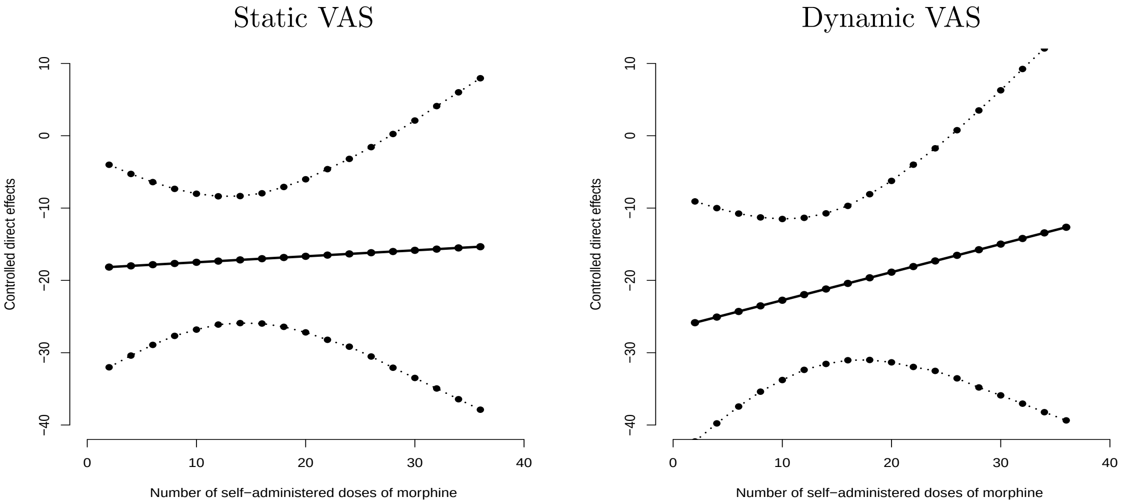

Figure 1: Mediation analysis: Posterior means (solid lines) and 95% posterior credible intervals (dotted lines) of controlled direct effects

interest, even if usually provides useful insights.

Principal stratification and mediation analysis rely on assumptions of a different nature and a careful evaluation of their plausibility is crucial. We conduct Bayesian principal stratification analysis under the randomization assumption (Assumption 1), which holds by design in the morphine study, and a monotonicity assumption (Assumption 2). The monotonicity assumption, which is arguable in general, appears very plausible in the morphine study, due to the characteristics of the active placebo.

Mediation analysis requires additional assumptions on the intermediate variable (such as Assumptions 4 and 5), which can be very critical in situations where the intermediate variable can not be seen as a treatment or the information on the pretreatment variables is too poor to make those assumptions plausible in the study. In the morphine study, only two pretreatment variables are observed, so Assumptions 4 and 5 might be questionable, and a principal stratification analysis, which only requires the randomization assumption (Assumption 1), might be preferable, although it only provides information on local effects. Principal stratification may also provide useful insight on the plausibility of Assumptions 4 and 5. Specifically, mediation analysis extrapolates information on potential outcomes of the form $Y_i(z, s)$ and $Y_i(z, S_i(z'))$ from the observed data, by mixing information across principal strata, which may be inappropriate if effects are heterogeneous across principal strata (Mealli and Mattei, 2012). This is the case of the morphine study, especially when dynamic VAS is considered.

In the principal stratification analysis, relevant information could also be obtained looking at the distribution of baseline characteristics within each principal stratum. While beyond the scope of the current paper, further analyses aimed at investigating the role of covariates to explain the heterogeneity of the effects across principal strata, are at the top of our research agenda.

Acknowledgments

The authors wish to thank Fabio Picciafuochi (Azienza Ospedaliero-Universitaria Careggi, Firenze, Italy) for providing the data and for helpful discussions on the interpretation of the results.

References

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

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

Imai, K., D. Tingley, and T. Yamamoto (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*, 5–51.

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

Mattei, A. and F. Mealli (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society, B 73*(5), 729–752.

Mealli, F. and A. Mattei (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics 8*(1), 1–19.

Mealli, F. and D. B. Rubin (2003). Assumptions allowing the estimation of direct causal effects. *Journal of Econometrics 112*(1), 79–87.

Pearl, J. (2001). Direct and indirect effects. In J. S. Breese and D. Koller (Eds.), *17th Conference on Uncertainty in Artificial Intelligence*, pp. 411–420.

Petersen, M., S. E. Sinisi, and M. van der Laan (2006). Estimation of direct causal effects. *Epidemiology 17*, 276–284.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. In. *Highly Structured Stochastic Systems*, (eds. P. Green, N. Hjort and S. Richardson), pp. 70–81, Oxford: Oxford University Press.

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

Rosenbaum, P. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, A 147*, 656–666.

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*(2), 161–170.

VanderWeele, T. L. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters 78*(17), 2957–2962.

VanderWeele, T. L. and S. Vansteelandt (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Inference 2*, 457–468.

Appendix

Details of Calculation

The posterior distributions of the parameters are obtained from Markov chain Monte Carlo (MCMC) methods (see details below). For each model, we ran three independent chains from different starting values for 12 500 iterations with the first 2 500 used as burn-in, saving every $5th$ iteration. The chains for the various models appear to converge very well with Gelman-Rubin diagnostic statistics approximately equal to 1.

MCMC for Principal Stratification Analysis

The MCMC algorithm that we adopt uses Gibbs sampler with data augmentation to impute at each step the missing principal strata indicators $S_i^*(1-Z_i)$. Specifically, we first obtain the joint posterior distribution of $(\boldsymbol{\theta}, S_i^*(1-Z_i))$ from a Gibbs sampler by iteratively sampling from $p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y}^{obs}, \mathbf{G}, \mathbf{X})$ and $p(\mathbf{S}^*(1-Z_i) \mid \mathbf{Z}, \mathbf{Y}^{obs}, \mathbf{X}, \boldsymbol{\theta})$, which in turn provides the marginal posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y}^{obs}, \mathbf{S}^*, \mathbf{X})$. The key to the posterior computation is the evaluation of the complete intermediate-data posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y}^{obs}, \mathbf{G}, \mathbf{X})$, which has the following simple form:

$$\begin{aligned}
&p(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y}^{obs}, \mathbf{G}, \mathbf{X}) \propto p(\boldsymbol{\theta}) \times \\
&\quad \prod_{i:Z_i=0,G_i=00} \pi_{i,00} \cdot f_{i,00,0} \times \prod_{i:Z_i=0,G_i=10} \pi_{i,10} \cdot f_{i,10,0} \times \prod_{i:Z_i=0,G_i=11} \pi_{i,11} \cdot f_{i,11,0} \times \\
&\quad \prod_{i:Z_i=1,G_i=00} \pi_{i,00} \cdot f_{i,00,1} \times \prod_{i:Z_i=1,G_i=10} \pi_{i,10} \cdot f_{i,10,1} \times \prod_{i:Z_i=1,G_i=11} \pi_{i,11} \cdot f_{i,11,1}.
\end{aligned}$$

The MCMC algorithm can be described as follows. Let $\tilde{\mathbf{X}} = [\mathbf{1}, \mathbf{X}]$ be the $n \times 3$ matrix with ith row equal to $\tilde{\mathbf{X}}_i' = (1, X_{i1}, X_{i2})$. Let $(\mathbf{G}^t, \theta^{(t)})$ denote the state of the chain at time t. The state of the chain at time $t+1$ follows from applying the following steps.

1. Sample $\mathbf{G}^{(t+1)}$ according to $Pr(\mathbf{G} \mid \mathbf{X}, \mathbf{Z}, \mathbf{S}^*, \mathbf{Y}^{obs}; \boldsymbol{\theta})$. Conditional on $\boldsymbol{\theta}$ and $\mathbf{X}_i$, Z_i, S_i^*, and Y_i^{obs}, G_i is independent of G_j, Z_j, Y_j^{obs}, S_j^*, $\mathbf{X}_j$, for all $j \neq i$. Then, by the monotonicity assumption

$$\begin{aligned}
Pr(G_i = 00 \mid \mathbf{X}_i, Z_i = 0, S_i^* = 0, Y_i^{obs}) &= 1 \\
Pr(G_i = 11 \mid \mathbf{X}_i, Z_i = 1, S_i^* = 1, Y_i^{obs}) &= 1,
\end{aligned}$$

   and for subjects with $Z_i = 0, S_i^* = 1$ and $Z_i = 1, S_i^* = 0$,

$$\begin{aligned}
Pr(G_i = 11 \mid \mathbf{X}_i, Z_i = 0, S_i^* = 1, Y_i^{obs}) &\propto \frac{\pi_{i,11} \cdot f_{i,11,0}}{\pi_{i,10} \cdot f_{i,10,0} + \pi_{i,11} \cdot f_{i,11,0}} \\
Pr(G_i = 00 \mid \mathbf{X}_i, Z_i = 1, S_i^* = 0, Y_i^{obs}) &\propto \frac{\pi_{i,00} \cdot f_{i,00,1}}{\pi_{i,10} \cdot f_{i,10,1} + \pi_{i,00} \cdot f_{i,00,1}}
\end{aligned}$$

2. Sample the latent variables $G_{i,00}^*$ and $G_{i,11}^*$:

(a) Sample the latent variable $G^*_{i,00}$ from $N(\alpha_1^{(00)}+\boldsymbol{\alpha}_X^{(00)'}\mathbf{X}_i, 1)$ truncated to $(-\infty, 0)$ if $G_i = 00$ and to $(0, \infty)$ if $G_i \neq 00$.

(b) Sample the latent variable $G^*_{i,11}$ from $N(\alpha_1^{(11)}+\boldsymbol{\alpha}_X^{(11)'}\mathbf{X}_i, 1)$ truncated to $(-\infty, 0)$ if $G_i = 11$ and to $(0, \infty)$ if $G_i \neq 11$.

3. Sample the coefficients $\boldsymbol{\alpha}^{(00)}$ and $\boldsymbol{\alpha}^{(11)}$ given the following prior distributions:

$$\boldsymbol{\alpha}^{(00)} \sim N\left(\underline{\mu}_{\boldsymbol{\alpha}^{(00)}}, \underline{\Sigma}_{\boldsymbol{\alpha}^{(00)}}\right) \qquad \text{and} \qquad \boldsymbol{\alpha}^{(11)} \sim N\left(\underline{\mu}_{\boldsymbol{\alpha}^{(11)}}, \underline{\Sigma}_{\boldsymbol{\alpha}^{(11)}}\right)$$

(a) Sample $\boldsymbol{\alpha}^{(00)}$ from $N(\mu_{\boldsymbol{\alpha}_{(00)}}, \Sigma_{\boldsymbol{\alpha}_{(00)}})$ where

$$\mu_{\boldsymbol{\alpha}^{00}} = \Sigma_{\boldsymbol{\alpha}^{(00)}}\left(\underline{\Sigma}^{-1}_{\boldsymbol{\alpha}^{(00)}}\underline{\mu}_{\boldsymbol{\alpha}^{00}} + \tilde{\mathbf{X}}'\mathbf{G}^*_{00}\right) \quad \text{and} \quad \Sigma_{\boldsymbol{\alpha}^{(00)}} = \left(\underline{\Sigma}^{-1}_{\boldsymbol{\alpha}^{(00)}} + \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}$$

(b) Let $\tilde{\mathbf{X}}^{-00}$ denote the sub-matrix of $\tilde{\mathbf{X}}$ for units with $G_i^{(t+1)} = 01$ or $G_i^{(t+1)} = 11$ and let $\mathbf{G}_{11}^{*,-00}$ be the sub-vector of $\mathbf{G}^*_{11}$ for units with $G_i^{(t+1)} = 01$ or $G_i^{(t+1)} = 11$. Sample $\boldsymbol{\alpha}^{(11)}$ from $N(\mu_{\boldsymbol{\alpha}^{(11)}}, \Sigma_{\boldsymbol{\alpha}^{(11)}})$ where

$$\mu_{\boldsymbol{\alpha}^{(11)}} = \Sigma_{\boldsymbol{\alpha}^{(11)}}\left(\underline{\Sigma}^{-1}_{\boldsymbol{\alpha}^{(11)}}\underline{\mu}_{\boldsymbol{\alpha}^{(11)}} + \tilde{\mathbf{X}}^{-00'}\mathbf{G}_{11}^{*,-00}\right) \quad \text{and} \quad \Sigma_{\boldsymbol{\alpha}^{(11)}} = \left(\underline{\Sigma}^{-1}_{\boldsymbol{\alpha}^{(11)}} + \tilde{\mathbf{X}}^{-00'}\tilde{\mathbf{X}}^{-00}\right)^{-1}$$

4. Define $\mathbf{1} = (1, \ldots, 1)'$ and let $\mathbf{Y}^{obs}_{g,z}$ denote the sub-vector of $\mathbf{Y}^{obs}$ for units with $G_i = g$ and $Z_i = z$. Also let $\tilde{\mathbf{X}}_{g,z}$ denote the sub-matrix of $\tilde{\mathbf{X}}$ for units with $G_i = g$ and $Z_i = z$.

5. For $g = 00, 01, 11$ and $z = 0, 1$, sample the coefficients $\beta_1^{(g,z)}$ given their Normal prior distributions,

$$\beta_1^{(g,z)} \sim N\left(\underline{\mu}_{\beta_1^{(g,z)}} = 0, \underline{\sigma}^2_{\beta_1^{(g,z)}} = 10^6\right),$$

from the normal distributions $N(\mu_{\beta_1^{(g,z)}}, \sigma^2_{\beta_1^{(g,z)}})$, where

$$\mu_{\beta_1^{(z)}} = \sigma^2_{\beta_1^{(g,z)}}\left(\frac{1}{\underline{\sigma}^2_{\beta_1^{(g,z)}}}\underline{\mu}_{\beta_1^{(g,z)}} + \frac{1}{\sigma_z^2}\sum_{i:G_i=g,Z_i=z}\left(Y^{obs}_{i,g,z} - \boldsymbol{\beta}_X^{(z)'}\mathbf{X}_{i,g,z}\right)\right)$$

and

$$\sigma^2_{\beta_1^{(g,z)}} = \left(\frac{1}{\underline{\sigma}^2_{\beta_1^{(g,z)}}} + \frac{N_{g,z}}{\sigma_z^2}\right)^{-1},$$

where $N_{g,z}$ is the number of subjects of type g assigned to treatment z at time $t+1$.

6. For $z = 0, 1$, sample the coefficients $\boldsymbol{\beta}_X^{(z)}$ given their joint Normal prior distributions,

$$\boldsymbol{\beta}_X^{(z)} \sim N\left(\underline{\mu}_{\boldsymbol{\beta}_X^{(z)}} = \mathbf{0}, \underline{\Sigma}_{\boldsymbol{\beta}_X^{(z)}} = 10^6\mathbb{I}_2\right),$$

from the multivariate normal distributions $N(\mu_{\boldsymbol{\beta}_X^{(z)}}, \Sigma_{\boldsymbol{\beta}_X^{(z)}})$, where

$$\mu_{\boldsymbol{\beta}_X^{(z)}} = \Sigma_{\boldsymbol{\beta}_X^{(z)}}\left(\underline{\Sigma}^{-1}_{\boldsymbol{\beta}_X^{(z)}}\underline{\mu}_{\boldsymbol{\beta}_X^{(z)}} + \frac{1}{\sigma_z^2}\sum_{g=00,11,01}\mathbf{X}'_{g,z}\left(\mathbf{Y}^{obs}_{g,z} - \mathbf{1}\beta_1^{(g,z)}\right)\right)$$

and

$$\Sigma_{\boldsymbol{\beta}_X^{(z)}} = \left(\underline{\Sigma}^{-1}_{\boldsymbol{\beta}_X^{(z)}} + \frac{1}{\sigma_z^2} \sum_{g=00,11,01} \mathbf{X}'_{g,z}\mathbf{X}_{g,z} \right)^{-1}.$$

7. Sample the outcome variances σ_z^2 for $z = 0, 1$ given their inverse-χ^2 prior distributions,

$$\sigma_z^2 \sim \text{Inv} - \chi^2_{\underline{\nu}_z}\left(\underline{s}_z^2\right), \quad \text{with} \quad \underline{\nu}_z = 0.002 \quad \text{and} \quad \underline{s}_z^2 = 1,$$

from the inverse-χ^2 distributions $\text{Inv} - \chi^2_{\nu_z}\left(s_z^2\right)$, where

$$\nu_z = \underline{\nu}_z + N_z \qquad \text{and} \qquad s_z^2 = \frac{\underline{\nu}_z \underline{s}_z^2 + \sum_{i:G_i=00,11,01}\left(Y_{i,g,z}^{obs} - \beta_1^{(g,z)} - \boldsymbol{\beta}_X^{(z)'}\mathbf{X}_{i,g,z}\right)^2}{\nu_z},$$

where N_z is the number of subjects assigned to treatment z.

Causal Estimands in Principal Stratification Analysis

$$\begin{aligned}
\pi_{00} &= \frac{1}{N}\sum_{i=1}^{N}\pi_{i,00} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \Phi\left(\alpha_1^{00} + \boldsymbol{\alpha}_X^{(00)'}\mathbf{X}_i\right)\right] \\
\pi_{11} &= \frac{1}{N}\sum_{i=1}^{N}\pi_{i,11} = \frac{1}{N}\sum_{i=1}^{N}\Phi\left(\alpha_1^{00} + \boldsymbol{\alpha}_X^{(00)'}\mathbf{X}_i\right)\cdot\left[1 - \Phi\left(\alpha_1^{11} + \boldsymbol{\alpha}_X^{(11)'}\mathbf{X}_i\right)\right] \\
\pi_{10} &= \frac{1}{N}\sum_{i=1}^{N}\pi_{i,10} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \pi_{i,00} - \pi_{i,11}\right]
\end{aligned}$$

For $z = 0, 1$ and $g = 00, 10, 11$

$$\mu_{z,g} \equiv \mathbb{E}[Y_i(z)|G = g] = \frac{1}{\sum_{i=1}^{N}\pi_{i,g}}\sum_{i=1}^{N}\pi_{i,g}\mu_{i,z,g} = \frac{1}{\sum_{i=1}^{N}\pi_{i,g}}\sum_{i=1}^{N}\pi_{i,g}\left(\beta_1^{(g,z)} + \boldsymbol{\beta}_X^{(z)'}\mathbf{X}_i\right)$$

Therefore, for $(s_0, s_1) \in \{(0,0), (1,0), (1,1)\}$, $g \equiv s_0 s_1$, we have

$$\begin{aligned}
PCE(s_0, s_1) &= \frac{1}{\sum_{i=1}^{N}\pi_{i,g}}\sum_{i=1}^{N}\pi_{i,g}\left[\mu_{i,1,g} - \mu_{i,0,g}\right] \\
&= \frac{1}{\sum_{i=1}^{N}\pi_{i,g}}\sum_{i=1}^{N}\pi_{i,g}\left[\left(\beta_1^{(g,1)} + \boldsymbol{\beta}_X^{(1)'}\mathbf{X}_i\right) - \left(\beta_1^{(g,0)} + \boldsymbol{\beta}_X^{(0)'}\mathbf{X}_i\right)\right]
\end{aligned}$$

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\left(\pi_{i,00}\left[\mu_{i,1,00} - \mu_{i,0,00}\right] + \pi_{i,10}\left[\mu_{i,1,10} - \mu_{i,0,10}\right] + \pi_{i,11}\left[\mu_{i,1,11} - \mu_{i,0,11}\right]\right)$$

MCMC for Mediation Analysis

We conduct Bayesian inference under exchangeability, Assumptions 3, 4 and 5 and the linear models specified in Equations (6) and (7) in the main text. Let $\mathbf{V}$ and $\mathbf{W}$ be matrices with *ith* row equal to $\mathbf{V}_i' = (1, Z_i, \mathbf{X}_i')$ and $\mathbf{W}_i' = (1, Z_i, S_i^{obs}, Z_i \cdot S_i^{obs}, \mathbf{X}_i')$, respectively. The posterior distribution of the parameters of the linear models in Equations (6) and (7) is obtained from MCMC methods using the following algorithm:

1. Sample the coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \boldsymbol{\alpha}_X)$ given their joint Normal prior distribution, $\boldsymbol{\alpha} \sim N\left(\underline{\mu}_{\boldsymbol{\alpha}}, \underline{\Sigma}_{\boldsymbol{\alpha}}\right)$, from the multivariate Normal distribution $N(\mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$ where

$$\mu_{\boldsymbol{\alpha}} = \Sigma_{\boldsymbol{\alpha}}\left(\underline{\Sigma}_{\boldsymbol{\alpha}}^{-1}\underline{\mu}_{\boldsymbol{\alpha}} + \frac{1}{\sigma_S^2}\mathbf{V}'\mathbf{S}\right) \qquad \text{and} \qquad \Sigma_{\boldsymbol{\alpha}} = \left(\underline{\Sigma}_{\boldsymbol{\alpha}}^{-1} + \frac{1}{\sigma_S^2}\mathbf{V}'\mathbf{V}\right)^{-1}$$

2. Sample the variance parameter σ_S^2 given its inverse-χ^2 prior distributions,

$$\sigma_S^2 \sim \mathrm{Inv} - \chi^2_{\underline{\nu}_S}\left(\underline{s}_S^2\right), \quad \text{with } \ \underline{\nu}_S = 0.002 \ \text{ and } \ \underline{s}_S^2 = 1,$$

from the inverse-χ^2 distribution $\mathrm{Inv} - \chi^2_{\nu_S}\left(s_S^2\right)$, where

$$\nu_S = \underline{\nu}_S + n \qquad \text{and} \qquad s_S^2 = \frac{\underline{\nu}_S\underline{s}_S^2 + \sum_{i=1}^n\left[S_i^{obs} - (\alpha_1 + \alpha_2 Z_i + \boldsymbol{\alpha}'_X\mathbf{X}_i)\right]^2}{\nu_S}.$$

3. Sample the coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \boldsymbol{\beta}_X)$ given their joint Normal prior distribution, $\boldsymbol{\beta} \sim N\left(\underline{\mu}_{\boldsymbol{\beta}}, \underline{\Sigma}_{\boldsymbol{\beta}}\right)$, from the multivariate Normal distribution $N(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$ where

$$\mu_{\boldsymbol{\beta}} = \Sigma_{\boldsymbol{\beta}}\left(\underline{\Sigma}_{\boldsymbol{\beta}}^{-1}\underline{\mu}_{\boldsymbol{\beta}} + \frac{1}{\sigma_Y^2}\mathbf{W}'\mathbf{Y}^{obs}\right) \qquad \text{and} \qquad \Sigma_{\boldsymbol{\beta}} = \left(\underline{\Sigma}_{\boldsymbol{\beta}}^{-1} + \frac{1}{\sigma_Y^2}\mathbf{W}'\mathbf{W}\right)^{-1}$$

4. Sample the variance parameter σ_Y^2 given its inverse-χ^2 prior distributions,

$$\sigma_Y^2 \sim \mathrm{Inv} - \chi^2_{\underline{\nu}_Y}\left(\underline{s}_Y^2\right), \quad \text{with } \ \underline{\nu}_Y = 0.002 \ \text{ and } \ \underline{s}_Y^2 = 1,$$

from the inverse-χ^2 distribution $\mathrm{Inv} - \chi^2_{\nu_Y}\left(s_Y^2\right)$, where

$$\nu_Y = \underline{\nu}_Y + n$$

and

$$s_Y^2 = \frac{\underline{\nu}_Y\underline{s}_Y^2 + \sum_{i=1}^n\left[Y_i^{obs} - \left(\beta_1 + \beta_2 Z_i + \beta_3 S_i^{obs} + \beta_4 Z_i S_i^{obs} + \boldsymbol{\beta}'_X\mathbf{X}_i\right)\right]^2}{\nu_Y}.$$

Causal Estimands in Mediation Analysis

If Assumptions 3, 4 and 5 hold and models in Equation (6) and (7) in the main text are correctly specified, total and direct and indirect effects can be estimated from the regression parameters of these models as follows:

$$ACE = \frac{1}{n}\sum_{i=1}^{n}\left[\beta_2 + \beta_3\alpha_2 + \beta_4\left(\alpha_1 + \alpha_2 + \boldsymbol{\alpha}'_X\mathbf{X}_i\right)\right]$$

and

$$CDE = \beta_2 + \beta_4 s$$

$$\begin{array}{rclrcl} NDE(0) & = & \frac{1}{n}\sum\limits_{i=1}^{n}\left[\beta_2 + \beta_4\left(\alpha_1 + \boldsymbol{\alpha}'_X\mathbf{X}_i\right)\right] & NIE(1) & = & \beta_3\alpha_2 + \beta_4\alpha_2 \\ NDE(1) & = & \frac{1}{n}\sum\limits_{i=1}^{n}\left[\beta_2 + \beta_4\left(\alpha_1 + \alpha_2 + \boldsymbol{\alpha}'_X\mathbf{X}_i\right)\right] & NIE(0) & = & \beta_3\alpha_2. \end{array}$$