Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 5, *pp.* 729–752

Augmented designs to assess principal strata direct effects

Alessandra Mattei and Fabrizia Mealli

University of Florence, Italy

[Received May 2010. Revised March 2011]

Summary. Many studies involving causal questions are often concerned with understanding the causal pathways by which a treatment affects an outcome. Thus, the concept of 'direct' *versus* 'indirect' effects comes into play. We tackle the problem of disentangling direct and indirect effects by investigating new augmented experimental designs, where the treatment is randomized, and the mediating variable is not forced, but only randomly encouraged. There are two key features of our framework: we adopt a principal stratification approach, and we mainly focus on principal strata effects, avoiding involving *a priori* counterfactual outcomes. Using non-parametric identification strategies, we provide a set of assumptions, which allow us to identify partially the causal estimands of interest: the principal strata direct effects. Some examples are shown to illustrate our design and causal estimands of interest. Large sample bounds for the principal strata average direct effects are provided, and a simple hypothetical example is used to show how our augmented design can be implemented and how the bounds can be calculated. Finally our augmented design is compared and contrasted with a standard randomized design.

Keywords: Augmented designs; Bounds; Causal inference; Direct and indirect effects; Principal stratification

1. Introduction

Many research questions involving causal effects are often concerned with understanding the causal pathways by which an exposure or a treatment affects an outcome. Thus, the concept of 'direct' *versus* 'indirect' effects comes into play. The use of this concept is common, not only in statistics, but also in many areas of social, economic and political sciences as well as in biomedical and pharmacological sciences, where the closely related concepts of 'biomarkers' and 'surrogate outcomes' are often used (e.g. Joffe and Greene (2009) and Gilbert and Hudgens (2008)).

A classical example, which also illustrates the policy-making implications of direct and indirect effects, involves a drug treatment having side effects (Pearl, 2001). Patients who suffer from these side effects might tend to take additional rescue medication, which in turn may affect the response to the treatment. Therefore, the total effect of the drug treatment will be a combination of the direct effect of the treatment on the outcome and the indirect effect mediated by the rescue medication. To understand the mechanistic pathways by which the drug acts to cause or prevent a disease, the total treatment effect must be decomposed into direct and indirect effects. Untying the direct and mediated effects may help understanding, e.g. what the effect of the treatment would be if its side effects were taken away, and so answers policy-related questions of practical significance (for example, the drug manufacturer might consider ways of eliminating the adverse side effects of the drug; doctors might deem it helpful to suggest or prevent the use of rescue

Address for correspondence: Alessandra Mattei, Department of Statistics, University of Florence, Viale Morgagni 59, 50134 Florence, Italy.
E-mail: mattei@ds.unifi.it

 1369–7412/11/73729

medication). Another example is in biomedical sciences, where the focus is on assessing the effect of physical activity on circulation diseases, not channelled through the body mass index (Sjölander *et al.*, 2009). Assessing whether physical activity has a protective direct effect against circulation diseases is relevant for policies that are aimed at reducing the risk of circulation diseases. For instance, if a suitable level of physical activity can prevent circulation diseases, even if it fails to prevent obesity, a policy maker may consider promoting physical activity programmes. In economics, Flores and Flores-Lagunes (2010) focused on understanding to what extent the effects of a training programme on participants' employment and earnings is mediated by the achievement of a secondary educational degree (high school, general educational development or vocational degree). The direct effect of the programme on employment and earnings may help a policy maker to decide whether the completion of a secondary educational degree should be encouraged or discouraged during the training programme, and how to design future interventions, where the focus can be on increasing job-related skills *versus* accumulating human capital.

Disentangling direct and indirect effects may be a difficult task, because the intermediate outcome is generally not under experimental control. For instance, in the above examples, the use of rescue medication, obesity and completion of a secondary educational degree cannot be, in general, controlled. Traditional analyses of scientific problems where treatment comparisons need to be adjusted for post-treatment confounded variables are typically based on a standard method that directly controls for (conditions on) observed values of those post-treatment variables, resulting in estimates that generally lack causal interpretation (e.g. Rosenbaum (1984)).

The use of alternative study designs has proved to be useful for addressing post-treatment complications (e.g. Follmann (2006) and Baker *et al.* (2007)). In this paper, we shall investigate new augmented designs, where the treatment is randomized, and the mediating variable is not forced, but only randomly encouraged. We argue that this source of exogenous variation may help to identify and estimate direct and indirect effects. These designs will be feasible in some clinical and social experiments, when partial control of the intermediate variable can be conceived. In the drug treatment example that was previously described, side effects of the drug, and thus the use of rescue medication, cannot be directly controlled; however, the use of rescue medication can be encouraged (or discouraged), for instance by offering rescue medication to randomly selected patients. Analogously, in the physical activity example, the body mass index represents a biomarker, and it is not obvious how to conceptualize interventions on such a variable. However, a suitable level of the body mass index might be encouraged, for instance by suggesting following a specific diet to randomly selected patients. In the programme evaluation study, completion of a secondary educational degree cannot be, in general, directly manipulated, but researchers may encourage (or discourage) randomly selected participants to achieve a secondary educational degree, for instance by offering them a grant. (Some augmented designs have been recently proposed in the vaccine trial literature (e.g. Follmann (2006) and Qin *et al.* (2008)); however, they focus on simplified settings, where the surrogate response in the absence of treatment is constant, and so they can be viewed as special cases of our design.) We show that, under our augmented design, inferences can be drawn about direct causal effects by focusing on specific subsets of the population.

The definition of direct and indirect effects is straightforward in linear equation systems but can be rather controversial in non-linear systems, although recent work provides some clarification in this regard (e.g. Robins and Greenland (1992), Pearl (2001), Flores and Flores-Lagunes (2009a), VanderWeele (2009) and Imai *et al.* (2010a)). The problem of defining, identifying and estimating direct and indirect effects has been tackled extensively in the causal inference literature, and a variety of identification and estimation strategies have been developed, by

using different approaches for causal inference. In general, alternative approaches focus on different causal estimands, and full agreement on what the relevant estimands should be and how we should estimate them is still lacking. In this paper, we focus on the potential outcomes framework (Rubin, 1974, 1977, 1978), also referred to as the Rubin causal model (Holland, 1986), and use the concept of principal stratification (Frangakis and Rubin, 2002) for addressing the topic of direct and indirect causal effects (Mealli and Rubin, 2003; Rubin, 2004).

Principal stratification with respect to a post-treatment intermediate variable is a cross-classification of subjects into latent classes defined by the joint potential values of that post-treatment variable under each of the treatments being compared, so principal strata comprise units having the same values of the intermediate potential outcomes. Frangakis and Rubin (2002) defined a principal causal effect (PCE) as the comparison of potential outcomes under different treatment levels within a principal stratum (or union of principal strata). The key property of principal strata is that they are not affected by treatment assignment. Therefore, a PCE is always a causal effect and does not suffer from the complications of standard post-treatment-adjusted estimands.

In this view of causal inference, PCEs naturally provide information on the extent to which a causal effect of the treatment on the primary outcome occurs together with a causal effect of the treatment on the intermediate outcome. Specifically, a principal strata direct effect (PSDE) of the treatment, after controlling for the intermediate outcome, exists if there is a causal effect of the treatment on the primary outcome for subjects belonging to principal strata where the mediator is not affected by the treatment.

Principal stratification is one of several possible ways to conceptualize the mediatory role of an intermediate variable in the treatment–outcome relationship (Joffe *et al.*, 2007). An alternative approach, which is usually applied in the causal graph framework to causal inference (Pearl, 2000), focuses on what would happen to the treatment–outcome relationship under interventions on the intermediate variable and defines direct and indirect causal effects by using the concept of *a priori* counterfactual values of outcomes that would have been observed under assignment to a given treatment level and if the post-treatment variable were somehow simultaneously forced to attain a predetermined value (Robins and Greenland, 1992; Pearl, 2001). This framework, with its *a priori* counterfactual estimands, needs to assume that the intermediate variables can be controlled and fixed by an external intervention, or it is at least conceivable to do so. This may be a reasonable assumption when the mediators represent additional treatments, which could, at least in principle, be randomized. Imai *et al.* (2010c) proposed two new designs that can be used to identify direct and indirect effects, involving *a priori* counterfactual outcomes, when the intermediate variable cannot be perfectly controlled: the parallel encouragement design and the crossover encouragement design. These designs allow for imperfect manipulation of the mediator, by assuming that researchers can only encourage (rather than force) experimental subjects to take a particular value of the mediator. The parallel encouragement design and the crossover encouragement design are similar in spirit to our augmented design; the idea of encouraging, rather than randomly forcing, the intermediate variable links the two approaches. However, our augmented design differs from those proposed by Imai *et al.* (2010c) in important ways that will be discussed in Section 8.

Our approach has various advantages as well as some drawbacks. An important insight is that it identifies average direct treatment effects for well-defined subpopulations: the subpopulations of subjects for whom the treatment does not affect the intermediate variable. As a result, our estimands, the PSDEs, are generally different from the average direct treatment effect for the overall population and do not decompose the total average effect into a direct and an indirect effect, unless, for instance, the individual PSDEs are assumed constant over the population.

Although in some cases the PSDEs may not be the average direct effects (ADEs) that researchers would like to estimate, the 'internal' validity (the sensibleness of the causal interpretation for the population that the study refers to) of these estimands may be much higher than that of other estimands, involving *a priori* counterfactuals.

In this paper, we shall focus on designs with two binary treatments and one binary intermediate variable. In principle, however, the principal stratification approach, and our augmented design, can be also extended to address the problem of disentangling direct and indirect effects when the intermediate variable is continuous. In fact, principal stratification analysis is not impossible with continuous mediating variables, although continuous intermediate outcomes induce an infinite number of possible principal strata, in theory, leading to substantial complications in both inference and interpretation (e.g. Jin and Rubin (2008) and Schwartz *et al.* (2010)). Therefore, extensions to general intermediate variables are a valuable topic for future research.

The paper is organized as follows. In Section 2 we give a brief overview of competing frameworks for defining the concept of direct and indirect effects. In Section 3 we describe our framework and in Section 4 we present our design's structural assumptions. We derive large sample bounds for PSDEs in Section 5. Calculation of these bounds is then illustrated in Section 6 with a numerical example. In Section 7 our augmented randomized design is compared and contrasted with a standard randomized design with respect to the accuracy of large sample bounds for an average (overall) direct effect. We conclude in Section 8, providing some discussion and suggesting directions for future research.

2. Alternative concepts of direct and indirect effects

In this section we briefly review some of the several alternative ways to define and formalize the concept of direct and indirect effects.

Consider a random sample of units, indexed by $i = 1, \ldots, n$. Each unit i can be potentially assigned either a standard treatment C or a new treatment T. Let Z denote the treatment variable. The objective is to assess the causal effect of the T- *versus* the C-treatment on an outcome Y. Let S stand for an intermediate variable which is on the causal pathway between the treatment and the main end point Y. Let $Y_i(z)$ and $S_i(z)$ denote the potential outcomes of Y and S respectively, if unit i was assigned treatment $Z = z$, $z = C, T$. Finally, let $Y_i(z, s)$ denote the (*a priori*) counterfactual value for Y if, possibly contrary to fact, Z was set to z and S was set to s: the potential outcomes $Y_i\{z, S_i(z) = s\}$ are *a priori* counterfactuals for units assigned to treatment z which exhibit a value of the intermediate outcome S not equal to s because, in one specific experiment, they can be never observed for such a type of units (Rubin, 2004). Note that we assume that the potential values $S_i(z)$, $Y_i(z)$ and $Y_i(z, s)$ for individual i do not depend on the treatments that are received by other individuals (the stable unit treatment value assumption (SUTVA); Rubin (1978, 1980, 1990)).

Robins and Greenland (1992) and Pearl (2001) gave definitions for controlled direct effects (CDEs) and natural direct and indirect effects based on interventions on the intermediate variable, thus using *a priori* counterfactuals. The (average) CDE of the treatment Z on the outcome Y, setting S to s, is defined by $\mathrm{CDE}(s) = E[Y_i(T, s) - Y_i(C, s)]$. The (average) natural direct effect (NDE) measures the effect of the treatment Z on the outcome Y when the mediator is set to the potential value that it would have taken under treatment status z: $\mathrm{NDE}(z) = E[Y_i\{T, S_i(z)\} - Y_i\{C, S_i(z)\}]$, $z = C, T$. Corresponding to NDE is the concept of natural indirect effect (NIE): $\mathrm{NIE}(z) = E[Y_i\{z, S_i(T)\} - Y_i\{z, S_i(C)\}]$, $z = C, T$. These effects provide an intuitive decomposition of the average total treatment effect $\mathrm{ATE} = E[Y_i(T) - Y_i(C)]$

into the sum of an NDE and an NIE: $\text{ATE} = \text{NIE}(T) + \text{NDE}(C)$ or $\text{ATE} = \text{NIE}(C) + \text{NDE}(T)$. Geneletti (2007) applied similar concepts by using a decision theoretic approach, where relationships between variables are encoded by using conditional independence statements, but without using counterfactuals (Dawid, 2000).

Various identification and estimation strategies for CDEs and NDEs and NIEs have been developed (e.g. Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen *et al.* (2006), Geneletti (2007), Goetgeluk *et al.* (2008), Flores and Flores-Lagunes (2009a, b, 2010) and Imai *et al.* (2010b)). A drawback of these methods is that estimation can only be based on extrapolations, because data can never provide direct evidence on *a priori* counterfactual values, and extrapolation typically involves relatively strong conditions such as constant effect, parametric and/or conditional independence assumptions.

Alternatively, a principal stratification approach can be used. Formally, a principal stratum with respect to the post-treatment variable S (with support $\mathcal{S}$) is a group of individuals who have the same vector $(S_i(C), S_i(T))$. For instance, if S is a binary variable, units can be classified into four latent groups: 1, $\{i : S_i(C) = S_i(T) = 0\}$; 2, $\{i : S_i(C) = 1, S_i(T) = 0\}$; 3, $\{i : S_i(C) = 0, S_i(T) = 1\}$; 4, $\{i : S_i(C) = S_i(T) = 1\}$. A PCE is a comparison between the potential outcomes $Y_i(C)$ and $Y_i(T)$ within a particular stratum (or union of principal strata): $\text{PCE}(s_C, s_T) = E[Y_i(T) - Y_i(C) | S_i(C) = s_C, S_i(T) = s_T]$.

Evidence on the direct effect of the treatment on the primary outcome is provided by principal strata where the intermediate variable is unaffected by treatment, i.e. $S_i(C) = S_i(T)$. Specifically, the PSDE of Z on Y at level s, $s \in \mathcal{S}$, can be formally defined as

$$\text{PSDE}(s) = E[Y_i(T) - Y_i(C) | S_i(T) = S_i(C) = s]. \tag{1}$$

If $\text{PSDE}(s) = 0$, for each $s \in \mathcal{S}$, there is no direct effect of treatment after controlling for the mediator, because the causal effect of treatment on the outcome exists only in the presence of a causal effect of treatment on the intermediate variable.

PCEs (and PSDEs) are causal effects for specific subpopulations (principal strata), and so the total effect of the treatment Z on the outcome Y is the weighted average of PCEs across units belonging to different principal strata:

$$\text{ATE} = \sum_{(s_C, s_T)} \text{PCE}(s_C, s_T)\pi_{s_C, s_T} = \sum_{s_C = s_T = s} \text{PSDE}(s)\pi_s + \sum_{s_C \neq s_T} \text{PCE}(s_C, s_T)\pi_{s_C, s_T},$$

where π_{s_C, s_T} is the proportion of subjects belonging to principal stratum $\{i : S_i(C) = s_C, S_i(T) = s_T\}$, and $\pi_s = \pi_{s,s}$. This result is in contrast with other approaches, where direct and indirect causal effects are defined for each individual and averaged over the entire population. For this reason, whereas NDEs and NIEs provide a decomposition of the total effect, principal stratification does not in general allow us to decompose the total effect into direct and indirect effects, unless additional assumptions are made. In fact, the PCEs for units belonging to principal strata where the post-treatment variable is affected by treatment combine direct and indirect effects. However, estimation of PSDEs does not, in general, require extrapolation of potential outcomes for subgroups where those are never observed.

VanderWeele (2008) studied the conceptual relationships between PSDEs and CDEs and NDEs, showing that if there are no CDEs or no NDEs then there can be no PSDEs. He also showed that the absence of PSDEs does not imply the absence of NDEs and does not necessarily even imply that the CDE or NDE is zero. These relationships, however, immediately follow from the definition of these effects: PSDEs are 'local' effects (i.e. they are causal effects within principal strata), whereas CDEs and NDEs are defined for each unit. Moreover, although these

are relevant theoretical results, they do not help identification and estimation of the causal estimands, using the data that are usually available.

Various (full and partial) identification strategies and estimation methods for principal effects have been developed (e.g. Barnard *et al.* (2003), Cheng and Small (2006), Cheng *et al.* (2009), Frangakis *et al.* (2007), Gallop *et al.* (2009), Imai (2008), Lee (2009), Lynch *et al.* (2008), Mattei and Mealli (2007), Sjölander *et al.* (2009), Zhang and Rubin (2003) and Zhang *et al.* (2009)). In this paper we contribute to this literature, developing a novel approach to the identification and estimation of PSDEs based on a new augmented design. Focus will be on the assumptions, which allow us to identify partially (Manski, 2003) causal estimands for specific subpopulations, and usually to derive tighter bounds than those derived in standard randomized experiments.

3. Principal stratification in augmented designs

To draw inference about PCEs, it is crucial to think very carefully about the hypothetical randomized experiment that led to the observed data. In this respect, the encouragement design—a special quasi-experimental design, where the only experimental manipulation is exposure to the encouragement conditions—can be used as a template to address issues of direct and indirect causal relationships. Although it is in general unreasonable to assume that the experimenter can directly control the administration of the mediating variable, it could be plausible to think about the existence of an additional variable, which is henceforth referred to as the encouragement variable, which affects the primary outcome, only through its effect on the intermediate outcome. Throughout, we therefore assume that, in addition to the treatment, whose causal effect on the outcome is still our primary interest, units are exposed to an additional treatment, which is related to the mediating variable, but unrelated to the outcome.

Formally, each sample unit i can be potentially either encouraged or not encouraged to exhibit a specific value of the intermediate outcome, S. Henceforth, we assume that the intermediate variable is binary (e.g. taking on values 0 and 1). Let W_i denote the indicator variable assuming value E if unit i is encouraged, for example, to exhibit a positive value of the intermediate outcome, and e otherwise. Let $\mathbf{Z}$ and $\mathbf{W}$ denote the n-dimensional vectors with ith element Z_i and W_i respectively. Then, let $S_i(\mathbf{Z}, \mathbf{W})$ and $Y_i(\mathbf{Z}, \mathbf{W})$ be the potential indicator for whether unit i would exhibit a positive value of S, and the potential response for unit i, given the vectors of treatment and encouragement assignments, $\mathbf{Z}$ and $\mathbf{W}$.

We generalize the SUTVA (Rubin, 1978, 1980, 1990) by assuming that

(a) the potential outcomes for any unit do not vary with the treatments and the encouragements that are assigned to any other unit and
(b) for each unit there are no different versions of each treatment and encouragement level.

Formally we have the following assumption.

Assumption 1 (SUTVA). If $Z_i = Z_i'$ and $W_i = W_i'$, then $S_i(\mathbf{Z}, \mathbf{W}) = S_i(\mathbf{Z}', \mathbf{W}')$ and $Y_i(\mathbf{Z}, \mathbf{W}) = Y_i(\mathbf{Z}', \mathbf{W}')$.

The SUTVA allows us to write $S_i(\mathbf{Z}, \mathbf{W})$ and $Y_i(\mathbf{Z}, \mathbf{W})$ as $S_i(Z_i, W_i)$ and $Y_i(Z_i, W_i)$ respectively. Therefore, for each unit i, there are four potential values for the mediating variable, $S_i(C, e)$, $S_i(C, E)$, $S_i(T, e)$ and $S_i(T, E)$, and four potential values for the response variable, $Y_i(C, e)$, $Y_i(C, E)$, $Y_i(T, e)$ and $Y_i(T, E)$. Principal strata are now defined according to the joint values of the potential variables $(S_i(C, e), S_i(C, E), S_i(T, e), S_i(T, E))$.

Definition 1. The basic principal stratification P_0 with respect to post-treatment variable S is

Table 1. Principal strata with two binary treatments and a binary intermediate variable: PSDEs by encouragement status under standard and active treatment and value of the intermediate outcome

G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$	$PSDE_{G_{s,w,w'}}(s;w,w')$			
1	0	0	0	0	$\text{PSDE}_1(0;e,e)$	$\text{PSDE}_1(0;E,E)$	$\text{PSDE}_1(0;e,E)$	$\text{PSDE}_1(0;E,e)$
2	1	0	0	0		$\text{PSDE}_2(0;E,E)$		$\text{PSDE}_2(0;E,e)$
3	0	1	0	0	$\text{PSDE}_3(0;e,e)$		$\text{PSDE}_3(0;e,E)$	
4	0	0	1	0		$\text{PSDE}_4(0;E,E)$	$\text{PSDE}_4(0;e,E)$	
5	0	0	0	1	$\text{PSDE}_5(0;e,e)$			$\text{PSDE}_5(0;E,e)$
6	1	1	0	0				
7	1	0	1	0	$\text{PSDE}_7(1;e,e)$	$\text{PSDE}_7(0;E,E)$		
8	1	0	0	1			$\text{PSDE}_8(1;e,E)$	$\text{PSDE}_8(0;E,e)$
9	0	1	1	0			$\text{PSDE}_9(0;e,E)$	$\text{PSDE}_9(1;E,e)$
10	0	1	0	1	$\text{PSDE}_{10}(0;e,e)$	$\text{PSDE}_{10}(1;E,E)$		
11	0	0	1	1				
12	1	1	1	0	$\text{PSDE}_{12}(1;e,e)$			$\text{PSDE}_{12}(1;E,e)$
13	1	1	0	1		$\text{PSDE}_{13}(1;E,E)$	$\text{PSDE}_{13}(1;e,E)$	
14	1	0	1	1	$\text{PSDE}_{14}(1;e,e)$		$\text{PSDE}_{14}(1;e,E)$	
15	0	1	1	1		$\text{PSDE}_{15}(1;E,E)$		$\text{PSDE}_{15}(1;E,e)$
16	1	1	1	1	$\text{PSDE}_{16}(1;e,e)$	$\text{PSDE}_{16}(1;E,E)$	$\text{PSDE}_{16}(1;e,E)$	$\text{PSDE}_{16}(1;E,e)$

the partition of units $i=1,\ldots,n$ such that all units within any set of P_0 have the same vector $(S_i(C,e), S_i(C,E), S_i(T,e), S_i(T,E))$. A principal stratification P with respect to post-treatment variable S is a partition of the units whose sets are unions of sets in the basic principal stratification P_0 (Frangakis and Rubin, 2002).

Because the intermediate variable is binary, units can be classified into 16 basic principal strata as shown in the first five columns of Table 1, where G_i is the principal stratum indicator for subject i: $G_i \in \{1,2,\ldots,16\}$. For instance, principal stratum 10 is $\{i: S_i(C,e)=0, S_i(C,E)=1, S_i(T,e)=0, S_i(T,E)=1\}$ and comprises units which would exhibit positive values of the mediating variable under encouragement and would exhibit a zero value of the mediating variable without encouragement, irrespective of the treatment level. Each principal stratum g, $g \in \{1,2,\ldots,16\}$, comprises a proportion π_g of all units.

This partition of the units can be viewed as a generalization of the idea of principal stratification (Frangakis and Rubin, 2002) to multiple treatments. Generally, a principal stratification with a binary treatment and a binary encouragement generates the following PSDEs.

Definition 2. The average PSDE of Z on Y at level s, $s \in \{0,1\}$, which is denoted $\text{PSDE}(s;w,w')$, is defined as

$$\text{PSDE}(s;w,w') = E[Y_i(T,w') - Y_i(C,w) | S_i(T,w') = S_i(C,w) = s] \tag{2}$$

for $w, w' \in \{e, E\}$.

Note that,

$$\begin{aligned}\text{PSDE}(s;w,w') &= E[Y_i(T,w') - Y_i(C,w) | S_i(T,w') = S_i(C,w) = s] \\ &= E[E[Y_i(T,w') - Y_i(C,w) | S_i(T,w') = S_i(C,w) = s, S_i(T,w), S_i(C,w')]] \\ &= E[\text{PSDE}(s;w,w') | S_i(T,w), S_i(C,w')] \equiv E[\text{PSDE}_{G_{s,w,w'}}(s;w,w')],\end{aligned}$$

where the outer expectation is over the joint distribution of the potential outcomes $S_i(T,w)$ and $S_i(C,w')$, and $G_{s,w,w'}$ is the principal stratum comprising units with $S_i(T,w') = S_i(C,w) = s$.

Table 2. Group classification based on observed data ($Z_i^{\text{obs}}, W_i^{\text{obs}}, S_i^{\text{obs}}$) and associated latent principal strata

Z_i^{obs}	W_i^{obs}	S_i^{obs}	*Latent strata* G_i							
C	e	0	1	3	4	5	9	10	11	15
C	e	1	2	6	7	8	12	13	14	16
C	E	0	1	2	4	5	7	8	11	14
C	E	1	3	6	9	10	12	13	15	16
T	e	0	1	2	3	5	6	8	10	13
T	e	1	4	7	9	11	12	14	15	16
T	E	0	1	2	3	4	6	7	9	12
T	E	1	5	8	10	11	13	14	15	16

Therefore, each PSDE, which involves units belonging to the union of different sets in the basic principal stratification, can be decomposed into 'basic' PSDEs, namely direct effects within sets of the basic principal stratification. In our setting with a binary intermediate outcome, eight PSDEs can be defined and evidence about each of them is provided by the union of different sets in the basic principal stratification as shown in Table 1.

Unfortunately, we cannot in general observe the principal stratum to which a subject belongs, because we cannot directly observe each potential intermediate value $S_i(z, w)$, $z = C, T$, $w = e, E$, for any subject. If we indicate with Z_i^{obs} the observed treatment assignment, and with W_i^{obs} the observed encouragement indicator, the observed data are, Z_i^{obs}, W_i^{obs}, $S_i^{\text{obs}} = S_i^{\text{obs}}(Z_i^{\text{obs}}, W_i^{\text{obs}})$ and $Y_i^{\text{obs}} = Y_i^{\text{obs}}(Z_i^{\text{obs}}, W_i^{\text{obs}})$, $i = 1, \ldots, n$. Therefore, what we can observe are the eight groups that are reported in Table 2, where the latent principal strata that are associated with each observed group are also shown. Each subject is observed to fall into one of these groups. If all 16 principal strata existed, i.e. if $\pi_g > 0$, for each $g \in \{1, 2, \ldots, 16\}$, each observed group would be a mixture of eight principal strata.

Let $\text{OBS}(z, w, s)$ denote the observed group that is defined by $Z_i^{\text{obs}} = z$, $W_i^{\text{obs}} = w$ and $S_i^{\text{obs}} = s$, $z = C, T$, $w = e, E$ and $s = 0, 1$, and let $P_{s|z,w} = \Pr(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = z,\ W_i^{\text{obs}} = w)$ be the conditional distribution of the observed intermediate outcome given the treatment status and the encouragement status.

4. Structural assumptions

Throughout this paper, we assume that the treatment and the encouragement are randomly assigned.

Assumption 2 (randomization of the treatment and the encouragement). For all i,

$$(S_i(C, e), S_i(C, E), S_i(T, e), S_i(T, E), Y_i(C, e), Y_i(C, E), Y_i(T, e), Y_i(T, E)) \perp\!\!\!\perp (Z_i, W_i).$$

Assumption 2 implies that

$$(Y_i(C, e), Y_i(C, E), Y_i(T, e), Y_i(T, E)) \perp\!\!\!\perp (Z_i, W_i) | (S_i(C, e), S_i(C, E), S_i(T, e), S_i(T, E))$$

so potential outcomes are independent of both the treatment and the encouragement given the principal strata.

To characterize W as an encouragement variable, we introduce an exclusion–restriction type of assumption, which rules out direct effects of the encouragement W on the primary outcome

Y. Specifically, we assume that the distributions of two potential outcomes $Y_i(z, e)$ and $Y_i(z, E)$, for each $z = C, T$, are the same for units which would exhibit the same value of the intermediate outcome regardless of the encouragement. Although this assumption is not directly testable, it can be made plausible by design. Formally, we have the following assumption.

Assumption 3 (stochastic exclusion restriction with respect to W).

$$\Pr\{Y_i(z, E)|S_i(z, E)=S_i(z, e), S_i(z', E), S_i(z', e)\}=\Pr\{Y_i(z, e)|S_i(z, E)=S_i(z, e), S_i(z', E), S_i(z', e)\} \quad \text{for } z \neq z' \in \{C, T\}.$$

Assumption 3 implies that some of the basic PSDEs, $\text{PSDE}_{G_{s,w,w'}}(s; w, w')$, take the same value, which depends only on the value of the intermediate potential outcomes. Specifically, $\text{PSDE}_{G_{s,e,w'}}(s; e, w') = \text{PSDE}_{G_{s,E,w'}}(s; E, w')$, for each principal stratum $G_{s,e,w'} = G_{s,E,w'} = \{i : S_i(C, e) = S_i(C, E) = s, S_i(T, e) = s_{Te}, S_i(T, E) = s_{TE}\}$ where $s_{Tw'} = s$ for $w' \in \{e, E\}$. Analogously, $\text{PSDE}_{G_{s,w,e}}(s; w, e) = \text{PSDE}_{G_{s,w,E}}(s; w, E)$, for each principal stratum $G_{s,w,e} = G_{s,w,E} = \{i : S_i(C, e) = s_{Ce}, S_i(C, E) = s_{CE}, S_i(T, e) = S_i(T, E) = s\}$, where $s_{Cw} = s$, for $w \in \{e, E\}$.

We also require the encouragement variable W to have some effect on the intermediate outcome S.

Assumption 4 (non-zero average causal effect of W on S). The average causal effect of W on S, $E[S_i(z, E) - S_i(z, e)]$, is not equal to 0 for $z = C, T$.

This assumption warrants that there is at least one stratum where the behaviour regarding the intermediate variable S is different with and without encouragement.

To identify (even partially; Manski (2003)) the proportion of each principal stratum and the corresponding PSDEs additional assumptions are required. Alternative sets of assumptions, which allow us either to reduce the number of strata or to state the equivalence of the distribution of Y across some strata, can be proposed. Here we focus on a specific set of assumptions, which lead to identify partially the causal estimands of interest.

An assumption—which can be made plausible by designing an appropriate encouragement—requires monotonicity of S with respect to the encouragement variable W. Formally, we have the following assumption.

Assumption 5 (monotonicity of S with respect to W). For all i,

(a) $S_i(C, e) \leqslant S_i(C, E)$ and $S_i(T, e) \leqslant S_i(T, E)$

or

(b) $S_i(C, e) \geqslant S_i(C, E)$ and $S_i(T, e) \geqslant S_i(T, E)$.

Assumption 5 relates to the mediating variable S with respect to the encouragement variable W. Without loss of generality, let $S_i(C, e) \leqslant S_i(C, E)$ and $S_i(T, e) \leqslant S_i(T, E)$ for all i (assumption 5, part (a)). Therefore, assumption 5 implies that, for a fixed value of the treatment variable, units that exhibit a positive value of S when $W = e$ would exhibit a positive value of S also when $W = E$. In the drug treatment example, this assumption states that patients who would take additional rescue medication when non-encouraged would take additional rescue medication also when encouraged, irrespective of the treatment. Although the data can never provide any direct evidence against this assumption, it can be made plausible by design, for instance, encouraging units to exhibit a positive value of the mediating variable. Importantly, assumption 5 has some testable implications that can be used to falsify it. For instance, assumption 5, part (a), implies that the encouragement has a non-negative effect on the intermediate variable S

under both the standard treatment and the active treatment. Therefore, assumption 5, part (a), is not falsified by the data if, in large samples, the estimates of these causal effects are both non-negative. Assumption 5, part (a), rules out the existence of seven of the 16 principal strata (2, 4, 7–9, 12 and 14).

To sharpen inference on the principal strata proportions and the causal estimands of interest, we also make one additional assumption, which implies that, for a fixed encouragement level, units which exhibit a positive value of S when exposed to the standard treatment would exhibit a positive value of S also when randomly assigned to the active treatment. Formally, we have the following assumption.

Assumption 6 (monotonicity of S with respect to Z). For all i,

$$S_i(C, e) \leqslant S_i(T, e) \quad \text{and} \quad S_i(C, E) \leqslant S_i(T, E).$$

Assumption 6 implies that the treatment has a non-negative effect on the intermediate variable S irrespective of the encouragement. As with assumption 5, it is not directly testable, but the data can falsify it when, in large samples, the estimated treatment effect on the intermediate variable is negative among either encouraged or non-encouraged units. Assumption 6 may be controversial and untenable in a particular application. Actually, its plausibility may depend not only on the study design but also on the scientific theory behind any particular application. However, it appears to be plausible in many contexts. For instance, in the drug treatment example, assumption 6 states that the new drug has a non-negative effect on taking additional rescue medication both for patients who are encouraged to use some additional rescue medication and for patients who are not. Since the drug treatment has side effects, we believe that assumption 6 is likely to be satisfied. In the physical activity example, dichotomizing the body mass index (the intermediate variable) as 'not obese' ($S = 1$) *versus* 'obese' ($S = 0$), assumption 6 implies that subjects who would be not obese without doing exercise would be not obese also doing exercise, irrespective of whether they are or are not encouraged to follow a specific diet. This assumption appears to be reasonable; in fact, although it is conceivable that physical activity has little or no effect on the reduction of obesity, it is more difficult to understand how it may cause an increase in obesity. In the programme evaluation example, assumption 6 may be made plausible by appropriately designing the training programme. For instance, the training programme might ease the achievement of a secondary educational degree, by providing participants with academic and vocational skills. In such a case, it is reasonable to believe that the training programme has a non-negative effect on the achievement of a secondary educational degree regardless of whether a subject is encouraged or not.

Together, the monotonicity assumptions 5 and 6 rule out the existence of principal strata 2–4, 6–9 and 12–14, leading to a classification of units across principal strata, which allows us to investigate the benefits of our augmented randomized design with respect to a standard randomized design. Under assumptions 1–6, we can point identify the proportion of units which belong to principal strata 1 and 16 (Tables 3 and 4),

$$\pi_1 = 1 - P_{1|TE} \quad \text{and} \quad \pi_{16} = P_{1|Ce}, \tag{3}$$

and derive large sample bounds for the other principal strata proportions and the PSDE estimands. The partial identification strategy that we pursue is similar in spirit to those in Flores and Flores-Lagunes (2009b), Imai (2008), Lee (2009) and Zhang and Rubin (2003). However, our general set-up has peculiar features, stemming from the encouragement variable for the intermediate outcome. In addition, the causal estimands of interest are different.

Table 3. Principal strata under assumptions 5 and 6

G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$
1	0	0	0	0
5	0	0	0	1
10	0	1	0	1
11	0	0	1	1
15	0	1	1	1
16	1	1	1	1

Table 4. Observed groups with associated possible latent principal strata under assumptions 5 and 6

Z_i^{obs}	W_i^{obs}	S_i^{obs}	*Latent strata*				
C	e	0	1	5	10	11	15
C	e	1			16		
C	E	0		1	5	11	
C	E	1		10	15	16	
T	e	0		1	5	10	
T	e	1		11	15	16	
T	E	0			1		
T	E	1	5	10	11	15	16

5. Large sample bounds for principal strata direct effects

Assumptions 1–6, along with the two equations in expression (3), imply that

$$\pi_5 + \pi_{10} = P_{1|TE} - P_{1|Te}, \tag{4}$$

$$\pi_5 + \pi_{11} = P_{1|TE} - P_{1|CE}, \tag{5}$$

$$\pi_{10} + \pi_{15} = P_{1|CE} - P_{1|Ce}, \tag{6}$$

$$\pi_{11} + \pi_{15} = P_{1|Te} - P_{1|Ce}. \tag{7}$$

For equations (4)–(7) to hold, the differences on their right-hand sides must be non-negative. $P_{1|TE} - P_{1|CE}$ is the average causal effect of the treatment on the intermediate outcome among units randomly encouraged; $P_{1|CE} - P_{1|Ce}$ and $P_{1|TE} - P_{1|Te}$ are the average causal effects of the encouragement on the intermediate outcome among units randomly assigned to the standard and the active treatment respectively and $P_{1|Te} - P_{1|Ce}$ is the average causal effect of the treatment on the intermediate outcome among units which are not encouraged. Therefore, assumptions 5 and 6 are not falsified by the data if, in large samples, these causal effects are non-negative.

Using equations (3)–(5), and taking into account that the principal strata proportions need to add up to 1 ($1 = \pi_1 + \pi_5 + \pi_{10} + \pi_{11} + \pi_{15} + \pi_{16}$), we have

$$\pi_{10} = P_{1|TE} - P_{1|Te} - \pi_5, \tag{8}$$

$$\pi_{11} = P_{1|TE} - P_{1|CE} - \pi_5, \tag{9}$$

$$\pi_{15} = \pi_5 + P_{1|CE} - P_{1|Ce} - (P_{1|TE} - P_{1|Te}). \tag{10}$$

Equations (8)–(10) hold for any π_5 such that

$$\max\{0; P_{1|TE} - P_{1|Te} - (P_{1|CE} - P_{1|Ce})\} \leqslant \pi_5 \leqslant \min(P_{1|TE} - P_{1|CE}; P_{1|TE} - P_{1|Te}). \tag{11}$$

We now establish large sample bounds on the PSDEs. To write these bounds formally we introduce some extra notation. Let $\pi_{g|zws}$ denote the conditional probability that a unit belongs to principal stratum g, $g = 1, 5, 10, 11, 15, 16$, given that the unit is observed to belong to the OBS(z, w, s) group, $z = C, T$, $w = e, E$, $s = 0, 1$: $\pi_{g|zws} = \Pr(G_i = g | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s)$. Each OBS$(z, w, s)$ group is the $\pi_{g|zws}$ mixture of some principal strata g, $g = 1, 5, 10, 11, 15, 16$. The conditional probabilities $\pi_{g|zws}$ cannot be point identified (except for $\pi_{1|zws}$ and $\pi_{16|zws}$). However, large sample bounds can be easily derived minimizing (or maximizing) their expressions over the possible range of π_5 (see equation (11)). Throughout, let $\min_{\pi_5}\{\psi(\pi_5)\}$ and $\max_{\pi_5}\{\psi(\pi_5)\}$ denote the minimum and the maximum of the function ψ with respect to π_5.

Proposition 1. Define $E_{zws}[Y_i^{\text{obs}}] = E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s]$ and let $E_{zws}^{\leqslant \alpha}[Y_i^{\text{obs}}]$ and $E_{zws}^{\geqslant \alpha}[Y_i^{\text{obs}}]$ be the conditional expectations of Y^{obs} in the α- $(0 < \alpha < 1)$ fraction of the observed group OBS(z, w, s) with the smallest and largest values of the outcome variable Y respectively. Then, under assumptions 1–6, the following bounds can be derived:

$$E_{Te0}[Y_i^{\text{obs}}] - E_{Ce0}^{\geqslant \pi_{1,5,10|Ce0}}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(0, e, e) \leqslant E_{Te0}[Y_i^{\text{obs}}] - E_{Ce0}^{\leqslant \pi_{1,5,10|Ce0}}[Y_i^{\text{obs}}], \tag{12}$$

$$\begin{gathered} E_{TE0}[Y_i^{\text{obs}}] - \min(E_{Ce0}^{\geqslant \pi_{1|Ce0}}[Y_i^{\text{obs}}]; E_{CE0}^{\geqslant \pi_{1|CE0}}[Y_i^{\text{obs}}]) \\ \leqslant \text{PSDE}(0, e, E) = \text{PSDE}(0, E, E) \leqslant \\ E_{TE0}[Y_i^{\text{obs}}] - \max(E_{Ce0}^{\leqslant \pi_{1|Ce0}}[Y_i^{\text{obs}}]; E_{CE0}^{\leqslant \pi_{1|CE0}}[Y_i^{\text{obs}}]), \end{gathered} \tag{13}$$

$$\begin{gathered} \min_{\pi_5}(E_{Te0}^{\leqslant \pi_{1,5|Te0}}[Y_i^{\text{obs}}] - E_{CE0}^{\geqslant \pi_{1,5|CE0}}[Y_i^{\text{obs}}]) \\ \leqslant \text{PSDE}(0, E, e) \leqslant \\ \max_{\pi_5}(E_{Te0}^{\geqslant \pi_{1,5|Te0}}[Y_i^{\text{obs}}] - E_{CE0}^{\leqslant \pi_{1,5|CE0}}[Y_i^{\text{obs}}]), \end{gathered} \tag{14}$$

$$\begin{gathered} \max(E_{Te1}^{\leqslant \pi_{16|Te1}}[Y_i^{\text{obs}}]; E_{TE1}^{\leqslant \pi_{16|TE1}}[Y_i^{\text{obs}}]) - E_{Ce1}[Y_i^{\text{obs}}] \\ \leqslant \text{PSDE}(1, e, e) = \text{PSDE}(1, e, E) \leqslant \\ \min(E_{Te1}^{\geqslant \pi_{16|Te1}}[Y_i^{\text{obs}}]; E_{TE1}^{\geqslant \pi_{16|TE1}}[Y_i^{\text{obs}}]) - E_{Ce1}[Y_i^{\text{obs}}], \end{gathered} \tag{15}$$

$$\begin{gathered} \min_{\pi_5}(E_{Te1}^{\leqslant \pi_{15,16|Te1}}[Y_i^{\text{obs}}] - E_{CE1}^{\geqslant \pi_{15,16|CE1}}[Y_i^{\text{obs}}]) \\ \leqslant \text{PSDE}(1, E, e) \leqslant \\ \max_{\pi_5}(E_{Te1}^{\geqslant \pi_{15,16|Te1}}[Y_i^{\text{obs}}] - E_{CE1}^{\leqslant \pi_{15,16|CE1}}[Y_i^{\text{obs}}]), \end{gathered} \tag{16}$$

$$\begin{gathered} E_{TE1}^{\leqslant \pi_{10,15,16|TE1}}[Y_i^{\text{obs}}] - E_{CE1}[Y_i^{\text{obs}}] \leqslant \\ \text{PSDE}(1, E, E) \leqslant E_{TE1}^{\geqslant \pi_{10,15,16|TE1}}[Y_i^{\text{obs}}] - E_{CE1}[Y_i^{\text{obs}}] \end{gathered} \tag{17}$$

where

$$\pi_{1,5,10|Ce0} \equiv \sum_{g \in \{1,5,10\}} \pi_{g|Ce0} = \frac{1 - P_{1|Te}}{1 - P_{1|Ce}},$$

$$\pi_{1|Ce0} = \frac{1 - P_{1|TE}}{1 - P_{1|Ce}},$$

$$\pi_{1|CE0} = \frac{1 - P_{1|TE}}{1 - P_{1|CE}},$$

$$\pi_{1,5|CE0} \equiv \pi_{1|CE0} + \pi_{5|CE0} = \frac{1 - P_{1|TE} + \pi_5}{1 - P_{1|CE}},$$

$$\pi_{1,5|Te0} \equiv \pi_{1|Te0} + \pi_{5|Te0} = \frac{1 - P_{1|TE} + \pi_5}{1 - P_{1|Te}},$$

$$\pi_{16|Te1} = \frac{P_{1|Ce}}{P_{1|Te}},$$

$$\pi_{16|TE1} = \frac{P_{1|Ce}}{P_{1|TE}},$$

$$\pi_{15,16|Te1} \equiv \pi_{15|Te1} + \pi_{16|Te1} = \frac{P_{1|CE} - (P_{1|TE} - P_{1|Te}) + \pi_5}{P_{1|Te}},$$

$$\pi_{15,16|CE1} \equiv \pi_{15|CE1} + \pi_{16|CE1} = \frac{P_{1|CE} - (P_{1|TE} - P_{1|Te}) + \pi_5}{P_{1|CE}},$$

and

$$\pi_{10,15,16|TE1} \equiv \sum_{g \in \{10,15,16\}} \pi_{g|TE1} = \frac{P_{1|CE}}{P_{1|TE}}.$$

The proof is given in Appendix A, where we also provide simple estimators for the bounds.

6. Illustrative example

In this section we apply our results to a hypothetical study example, which has been adapted from Pearl (2001). Suppose that we are interested in assessing the causal effect of a new drug treatment having headache as a side effect. Patients who suffer from headache tend to take rescue medication, which, in turn, may have its own effect on the disease or may strengthen (or weaken) the effect of the drug on the disease. To assess the causal effect of the new drug treatment on the primary outcome, and also to decide whether the use of rescue medication should be encouraged or discouraged during the treatment, a study is planned, where each patient can be potentially assigned either the new drug treatment ($Z_i = T$) or the standard (placebo) treatment ($Z_i = C$). Simultaneously, each patient can be either encouraged ($W_i = E$) or not encouraged ($W_i = e$) to take rescue medication against headache.

Tables 5 and 6 respectively show the full (hypothetical) data and the corresponding PSDEs under the SUTVA and assumptions 3–6, which are given in Section 4. The sixth column of Table 5 shows the principal strata proportions: each principal stratum comprises a proportion of 16% of all patients, except principal stratum 11, which comprises a proportion of 20% of all patients.

From Table 5, we can see that, if everyone were assigned treatment and encouraged, 84% ($= 16\% + 16\% + 20\% + 16\% + 16\%$) would take rescue medication, whereas, if everyone were assigned control and encouraged, 48% ($= 16\% + 16\% + 16\%$) would take rescue medication. Analogously, if everyone were assigned treatment and not encouraged, 52% ($= 20\% + 16\% + 16\%$) would take rescue medication, whereas, if everyone were assigned control and not encouraged, only 16% would take rescue medication. Thus, both the new drug treatment and the encouragement have quite a strong positive causal effect on usage of rescue medication.

The total effect of the treatment Z on the primary outcome Y is

$$E[Y_i(T, E) - Y_i(C, E)] = \sum_{g=1,5,10,11,15,16} \pi_g \, E[Y_i(T, E) - Y_i(C, E) | G_i = g] = 0.420$$

Table 5. Full hypothetical data under assumptions 3–6

G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$	π_g	*Expected values*			
						$Y_i(C,e)$	$Y_i(C,E)$	$Y_i(T,e)$	$Y_i(T,E)$
1	0	0	0	0	0.16	0.1	0.1	0.2	0.2
5	0	0	0	1	0.16	0.1	0.1	0.3	0.5
10	0	1	0	1	0.16	0.2	0.3	0.5	0.7
11	0	0	1	1	0.20	0.2	0.2	0.7	0.7
15	0	1	1	1	0.16	0.2	0.3	0.8	0.8
16	1	1	1	1	0.16	0.3	0.3	0.9	0.9

Table 6. PSDEs correponding to Table 5

$PSDE_{G_{s,w,w'}}(s;w,w')$			G_i: $\{S_i(C,e), S_i(C,E), S_i(T,e), S_i(T,E)\}$					$PSDE(s;w,w')$, *mean*
s	w	w'	$G_i=1$, $\{0,0,0,0\}$	$G_i=5$, $\{0,0,0,1\}$	$G_i=10$, $\{0,1,0,1\}$	$G_i=15$, $\{0,1,1,1\}$	$G_i=16$, $\{1,1,1,1\}$	
0	e	e	0.1	0.2	0.3			0.20
0	E	E	0.1					0.10
0	e	E	0.1					0.10
0	E	e	0.1	0.2				0.15
1	e	e					0.6	0.60
1	e	E					0.6	0.60
1	E	e				0.5	0.6	0.55
1	E	E			0.4	0.5	0.6	0.50

for encouraged units, and

$$E[Y_i(T,e)-Y_i(C,e)]=\sum_{g=1,5,10,11,15,16}\pi_g\,E[Y_i(T,e)-Y_i(C,e)|G_i=g]=0.388$$

for not-encouraged units. PSDEs for patients who would use rescue medication under both treatment arms range from 0.5 to 0.6 and are higher than PSDEs for patients who would not use rescue medication under both treatment arms, which range from 0.1 to 0.2.

Now suppose that an experiment is conducted where the sample is randomly divided into four groups, with the first receiving $Z=T$ and $W=E$, the second receiving $Z=T$ and $W=e$, the third receiving $Z=C$ and $W=E$, and the fourth receiving $Z=C$ and $W=e$.

Table 7 presents some summary statistics for the sample, classified by treatment assignment Z_i^{obs}, encouragement assignment W_i^{obs} and rescue medication usage S_i^{obs}. Table 7 provides a simple mediation analysis based on standard methods which directly control for observed values of the post-treatment variable. Specifically, we can easily estimate net treatment effects of assignment (Z,W) adjusting for the observed value of the post-treatment variable S_i^{obs}: $E_{Tes}[Y_i^{\text{obs}}]-E_{Ces}[Y_i^{\text{obs}}]=0.171$ $(s=0), 0.492$ $(s=1)$; $E_{TEs}[Y_i^{\text{obs}}]-E_{Ces}[Y_i^{\text{obs}}]=0.038$ $(s=0), 0.419$ $(s=1)$; $E_{Tes}[Y_i^{\text{obs}}]-E_{CEs}[Y_i^{\text{obs}}]=0.195$ $(s=0), 0.492$ $(s=1)$; $E_{TEs}[Y_i^{\text{obs}}]-E_{CEs}[Y_i^{\text{obs}}]=0.062$ $(s=0), 0.419$ $(s=1)$. Integrating out the observed encouragement variable W^{obs}, we have $E_{T0}[Y_i^{\text{obs}}]-E_{C0}[Y_i^{\text{obs}}]=0.147$ and $E_{T1}[Y_i^{\text{obs}}]-E_{C1}[Y_i^{\text{obs}}]=0.447$. The standard interpretation of these results would be that the new drug treatment has a positive effect on the disease status,

Table 7. Summary statistics of hypothetical observed data

Z_i^{obs}	W_i^{obs}	S_i^{obs}	*Observed proportions*	*Means*	
				Rescue medication usage (S_i^{obs})	*Disease status* (Y_i^{obs})
C	e		0.25	0.16	0.184
C	E		0.25	0.48	0.216
T	e		0.25	0.52	0.572
T	E		0.25	0.84	0.636
C	e	0	0.21	0	0.162
C	e	1	0.04	1	0.300
C	E	0	0.13	0	0.138
C	E	1	0.12	1	0.300
T	e	0	0.12	0	0.333
T	e	1	0.13	1	0.792
T	E	0	0.04	0	0.200
T	E	1	0.21	1	0.719

and this effect appears to be higher among patients who take rescue medication against headache ($i : S_i^{obs} = 1$). Although these results do not clash with the real PSDEs, the differences between the average outcome among subjects who take rescue medication when assigned new *versus* standard treatment are lower than the PSDEs for patients who would use rescue medication under both treatment arms. In addition, we must keep in mind that the net treatment effects lack a causal interpretation, because they involve comparisons between sets of potential outcomes on different sets of units, the observed groups $\mathrm{OBS}(z, w, s)$, $z = C, T$, $w = e, E$ and $s = 0, 1$, which are mixtures of different principal strata.

From the observed data in Table 7, we immediately have $P_{1|Ce} = 0.16$, $P_{1|CE} = 0.48$, $P_{1|Te} = 0.52$ and $P_{1|TE} = 0.84$, so, from equation (11), the bounds for π_5 are $0 \leqslant \pi_5 \leqslant 0.32$.

Tables 8 and 9 show the bounds for the PSDEs, calculated by using the results in proposition 1. All our bounds contain the actual PSDEs and provide useful information about the direct effect of the treatment on the outcome. The estimated bounds for $\mathrm{PSDE}(1, e, e) = \mathrm{PSDE}(1; e, E)$ and $\mathrm{PSDE}(1; E, E)$ cover only positive regions and are relatively narrow, suggesting that there is a positive direct effect of the drug treatment on the disease for patients who would take rescue medication under both treatment arms. The drug treatment seems to have a positive, although lower, direct effect also for patients belonging to principal strata where $S_i(T, e) = S_i(C, e) = 0$: $0.050 \leqslant \mathrm{PSDE}(0; e, e) \leqslant 0.333$. Some uncertainty is in the sign of the causal effects $\mathrm{PSDE}(0; E, e)$ and $\mathrm{PSDE}(1; E, e)$, although the positive region that is covered by the bounds is larger than the negative region. In contrast, the data do not provide decisive evidence on the direct effects $\mathrm{PSDE}(0, E, E) = \mathrm{PSDE}(0, e, E)$.

To investigate the benefit of the presence *versus* the absence of an encouragement variable for the intermediate outcome, the estimated bounds in Tables 8 and 9 are now compared and contrasted with the bounds which would be derived in a standard randomized experiment where the intermediate variable is not randomly encouraged. To make the two designs comparable, we assume monotonicity of the intermediate outcome with respect to the treatment ($S_i(T) \geqslant S_i(C)$), which implies that the principal stratum $\{i : S_i(C) = 1, S_i(T) = 0\}$ is empty. We also assume that in

Table 8. Estimated bounds

Principal strata proportions	*Lower bound*		*Upper bound*
π_1		0.16	
π_5	0.00		0.32
π_{10}	0.00		0.32
π_{11}	0.04		0.36
π_{15}	0.00		0.32
π_{16}		0.16	

Table 9. Estimated bounds

Estimand, PSDE(s; w, w′)	*Lower bound*	*Upper bound*	*Proportions of units with $S_i(T, w') = S_i(C, w) = s$*	
			Lower bound	*Upper bound*
PSDE(0; e, e)	0.050	0.333	0.16	0.80
PSDE(0; E, E) = PSDE(0; e, E)	−0.250	0.200	0.16	
PSDE(0; E, e)	−0.450	0.975	0.16	0.48
PSDE(1; e, e) = PSDE(1; e, E)	0.025	0.700	0.16	
PSDE(1; E, e)	−0.550	1.000	0.16	0.48
PSDE(1; E, E)	0.208	0.700	0.16	0.80

the standard randomized study subjects behave as they would behave in the augmented randomized study when assigned to not be encouraged. This assumption implies that $S_i(z) = S_i(z, e)$, and $Y_i(z) = Y_i(z, e)$, for $z = C, T$.

Tables 10–12 show the hypothetical full and observed data of the standard randomized experiment, and the corresponding estimated bounds for PSDE(0) and PSDE(1) defined in equation (1). Consistently with the above assumptions, these bounds are similar to those for PSDE(0; e, e) and PSDE(1; e, e) = PSDE(1; e, E) respectively. (Following Manski (2003), large sample bounds for PSDE(s), $s = 0, 1$, can be easily derived. Specifically, define $E_{zs}[Y_i^{\text{obs}}] = E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = z, S_i^{\text{obs}} = s]$, and let $E_{zs}^{\leqslant\alpha}[Y_i^{\text{obs}}]$ and $E_{zs}^{\geqslant\alpha}[Y_i^{\text{obs}}]$ be the conditional expectations of Y^{obs} for the α- $(0 < \alpha < 1)$ fraction of smallest and largest values of Y respectively for the observed group with $Z_i^{\text{obs}} = z$, and $S_i^{\text{obs}} = s$. Then, under the SUTVA, randomization and the monotonicity assumption $S_i(T) \geqslant S_i(C)$, we have $E_{T0}[Y_i^{\text{obs}}] - E_{C0}^{\geqslant\pi_{1|C0}}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(0) \leqslant E_{T0}[Y_i^{\text{obs}}] - E_{C0}^{\leqslant\pi_{1|C0}}[Y_i^{\text{obs}}]$ and $E_{T1}^{\leqslant\pi_{4|T1}}[Y_i^{\text{obs}}] - E_{C1}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(1) \leqslant E_{T1}^{\geqslant\pi_{4|T1}}[Y_i^{\text{obs}}] - E_{C1}[Y_i^{\text{obs}}]$, where $\pi_{1|C0} = \Pr\{S_i(C) = S_i(T) = 0 | Z_i^{\text{obs}} = C, S_i^{\text{obs}} = 0\}$ and $\pi_{4|T1} = \Pr\{S_i(C) = S_i(T) = 1 | Z_i^{\text{obs}} = T, S_i^{\text{obs}} = 1\}$. In large samples, $\pi_{1|C0} = (1 - P_{1|T})/(1 - P_{1|C})$ and $\pi_{4|T1} = P_{1|C}/P_{1|T}$, where $P_{s|z} = \Pr(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = z)$, $z = C, T$, $s = 0, 1$.)

The major gain of our augmented randomized design with respect to the standard design can be observed by comparing the estimated bounds for PSDE(1; E, E) with those for PSDE(1): the first bound is narrower and more informative. Specifically, the estimated bounds for PSDE(1; E, E) ([0.208; 0.700]) suggest that there is a positive and quite strong direct effect of the drug treatment on the disease for subjects who would take rescue medication irrespective

Table 10. Standard randomized design: full data

G_i	$S_i(C)$	$S_i(T)$	π_i	*Expected values*		$PSDE(s)$
				$Y_i(C)$	$Y_i(T)$	
1	0	0	0.48	0.13	0.33	0.2
3	0	1	0.36	0.20	0.74	
4	1	1	0.16	0.30	0.90	0.6

Table 11. Standard randomized design: observed data

Z_i^{obs}	S_i^{obs}	*Observed proportions*	*Means*	
			Rescue medication usage (S_i^{obs})	*Disease status* (Y_i^{obs})
C		0.50	0.16	0.184
T		0.50	0.52	0.604
C	0	0.42	0	0.162
C	1	0.08	1	0.300
T	0	0.24	0	0.333
T	1	0.26	1	0.792

Table12. Standard randomized design: estimated bounds

Estimand	*Lower bound*		*Upper bound*
π_1		0.48	
π_3		0.36	
π_4		0.16	
PSDE(0)	0.050		0.333
PSDE(1)	0.019		0.700

of the treatment under encouragement. The bounds for PSDE(1) ([0.019; 0.700]) also show some evidence that the treatment has a positive direct effect for subjects who would take rescue medication irrespective of the treatment, but they are not informative on the size of this effect, allowing for a somewhat small effect.

These results might be at least partially justified, thinking carefully about what kind of information is given by the two designs. A key feature of our augmented randomized design is that it may provide information about the direct effect of the treatment also for subjects who would belong to principal strata where we would not generally be able to disentangle direct and indirect effects if a standard randomized experiment was conducted. Specifically, in a standard

randomized experiment, information on PSDEs is only provided by units belonging to either the principal stratum $\{i : S_i(C) = S_i(T) = 0\}$ or the principal stratum $\{i : S_i(C) = S_i(T) = 1\}$. Units belonging to the other principal strata, $\{i : S_i(C) = s, S_i(T) = 1 - s\}$, $s = 0, 1$ (principal stratum $G = 3$ in our example; see Table 10), give no direct information on the existence of PSDEs. However, such a type of units could potentially provide some information on the direct causal effects of the treatment if an encouragement design for the intermediate outcome was applied. In other words, each principal stratum defined by $(S_i(C), S_i(T))$ might be split into more principal strata under an encouragement design for the intermediate outcome, digging out some individual behaviour which might be useful to draw inference about PSDEs.

7. Augmented designs *versus* standard designs

We now try formally to compare our augmented randomized design and a standard randomized design. For this we need a common causal estimand, which is defined as an overall direct effect for the entire population, which has the drawback of involving the concept of *a priori* counterfactual outcomes, as explained in Section 2. Therefore, to define overall direct effects formally we need to extend the theoretical framework underlying our augmented randomized design and the standard randomized design to allow for *a priori* counterfactual outcomes.

Assumption 7 (SUTVA with *a priori* counterfactuals).

(a) If $Z_i = Z_i'$, then $S_i(\mathbf{Z}) = S_i(\mathbf{Z}')$.
(b) If $Z_i = Z_i'$ and $S_i = S_i'$, then $Y_i(\mathbf{Z}, \mathbf{S}) = Y_i(\mathbf{Z}', \mathbf{S}')$.

First, we focus on a standard randomized design, where the intermediate variable is not randomly encouraged. Under the SUTVA, which has now been formalized as assumption 7, we have two potential intermediate outcomes $S_i(C)$ and $S_i(T)$, and four potential primary outcomes, $Y_i\{C, S_i(C) = 0\}$, $Y_i\{T, S_i(T) = 0\}$, $Y_i\{C, S_i(C) = 1\}$ and $Y_i\{T, S_i(T) = 1\}$. The ADE can be defined as the mean difference between $Y_i\{T, S_i(T) = s\}$ and $Y_i\{C, S_i(C) = s\}$ while holding the mediator fixed at some level s:

$$\mathrm{ADE}(s) = E[Y_i\{T, S_i(T) = s\}] - E[Y_i\{C, S_i(C) = s\}] \qquad s = 0, 1. \tag{18}$$

Random assignment of the treatment Z implies the following assumption.

Assumption 8 (randomization of the treatment with *a priori* counterfactuals). For all i,

$$\{S_i(z), Y_i\{z, S_i(z) = 0\}, Y_i\{z, S_i(z) = 1\}\}_{z=C,T} \perp\!\!\!\perp Z_i.$$

Assumptions 7 and 8 alone do not lead to point-identify the ADE. However, large sample bounds for the ADE can be derived following Manski (2003). The following proposition provides a simple generalization to bounded outcomes of the bounds on the ADE for binary outcomes derived by Cai *et al.* (2008), using the symbolic Balke–Pearl method (Balke and Pearl, 1997).

Proposition 2. Suppose that $Y\{z, S(z) = s\}$ is bounded within some known interval $[L_{zs}, U_{zs}]$, where $-\infty < L_{zs} \leqslant U_{zs} < \infty$, $z = C, T$ and $s = 0, 1$. Define $E_{zs}[Y_i^{\mathrm{obs}}] = E[Y_i^{\mathrm{obs}} | Z_i^{\mathrm{obs}} = T, S_i^{\mathrm{obs}} = s]$ and $P_{s|z} = \Pr(S_i^{\mathrm{obs}} = s | Z_i^{\mathrm{obs}} = z)$, $z = C, T$ and $s = 0, 1$. Then, under assumptions 7 and 8, the following bounds can be derived:

$$\begin{aligned} &E_{Ts}[Y_i^{\mathrm{obs}}]P_{s|T} + L_{Ts}(1 - P_{s|T}) - \{E_{Cs}[Y_i^{\mathrm{obs}}]P_{s|C} + U_{Cs}(1 - P_{s|C})\} \\ &\quad \leqslant \mathrm{ADE}(s) \leqslant E_{Ts}[Y_i^{\mathrm{obs}}]P_{s|T} + U_{Ts}(1 - P_{s|T}) - \{E_{Cs}[Y_i^{\mathrm{obs}}]P_{s|C} + L_{Cs}(1 - P_{s|C})\} \end{aligned} \tag{19}$$

for $s = 0, 1$.

The width of the bounds in equation (19) is

$$\text{width}(s) = (U_{Cs} - L_{Cs})(1 - P_{s|C}) + (U_{Ts} - L_{Ts})(1 - P_{s|T}),$$

which depends on both the selection probabilities $1 - P_{s|z}$ as well as the width of the intervals $[L_{zs}, U_{zs}]$, $z = C, T$.

When an encouragement design for the intermediate variable is combined with a treatment randomized experiment, assumption 7 (the SUTVA) and the definition of the ADE in equation (18) change slightly. Assumption 7 turns into assumption 9.

Assumption 9 (the SUTVA with *a priori* counterfactuals).

(a) If $Z_i = Z_i'$ and $W_i = W_i'$, then $S_i(\mathbf{Z}, \mathbf{W}) = S_i(\mathbf{Z}', \mathbf{W}')$.
(b) If $Z_i = Z_i'$, $W_i = W_i'$ and $S_i = S_i'$, then $Y_i\{\mathbf{Z}, \mathbf{W}, \mathbf{S}(\mathbf{Z}, \mathbf{W})\} = Y_i\{\mathbf{Z}', \mathbf{W}', \mathbf{S}'(\mathbf{Z}', \mathbf{W}')\}$.

Alternative definitions of the ADE can be considered. For the reasons that are discussed below, we focus on the following estimand:

$$\text{ADE}(s; w) = E[Y_i(T, w, s)] - E[Y_i(C, w, s)] \qquad w = e, E, \quad s = 0, 1. \tag{20}$$

Since the augmented design involves two randomly assigned treatments (the primary treatment variable Z and the encouragement variable W), the following ignorability assumption holds.

Assumption 10 (randomization of the treatment and the encouragement with *a priori* counterfactuals). For all i,

$$\{S_i(z, w), Y_i\{z, w, S_i(z, w) = 0\}, Y_i\{z, w, S_i(z, w) = 1\}\}_{z=C,T;w=e,E} \perp\!\!\!\perp (Z_i, W_i).$$

Bounds for ADE$(s; w)$ are now derived taking into account the augmented nature of the design. To establish bounds for ADE$(s; w)$ no assumption is required in addition to assumptions 9 and 10. However, the exclusion restriction assumption 3 and the monotonicity assumption 5, part (a), or 5, part (b), (along with assumption 4) characterize our augmented design; therefore we maintain these assumptions, by extending them to involve *a priori* counterfactuals. Assumption 3 justifies our focus on the causal estimand ADE$(s; w)$. If assumption 3 holds, $Y_i\{z, e, S_i(z, e) = s\}$ and $Y_i\{z, E, S_i(z, E) = s\}$ have the same distribution, for each $z = C, T$, and $s = 0, 1$; therefore average (overall) direct effects could be defined irrespective of the encouragement as the causal effect on the outcome Y of the (T, w') *versus* the (C, w) treatment $(w, w' \in \{e, E\})$, while holding the intermediate variable S fixed at some predetermined level, s. In addition, if the encouragement has no direct effect on the outcome, ADE$(s; w)$ and ADE(s) can be reasonably viewed as the same estimand, measuring the effect of the treatment Z on the outcome Y not mediated through the intermediated variable S, and can thus be compared. Finally, as we show below, the monotonicity assumption 5, part (a), or 5, part (b), allows us to justify the focus on either ADE$(0; w)$ or ADE$(1; w)$ according to the role of the encouragement.

Proposition 3. Suppose that $Y\{z, w, S(z, w) = s\}$ is bounded within some known interval $[L_{zws}, U_{zws}]$, where $-\infty < L_{zws} \leqslant U_{zws} < \infty$, $z = C, T$, $W = e, E$ and $s = 0, 1$. Then, under assumptions 9 and 10, the following bounds can be derived:

$$\begin{aligned} &E_{Tws}[Y_i^{\text{obs}}]P_{s|Tw} + L_{Tws}(1 - P_{s|Tw}) - \{E_{Cws}[Y_i^{\text{obs}}]P_{s|Cw} + U_{Cws}(1 - P_{s|Cw})\} \\ &\quad \leqslant ADE(s; w) \leqslant E_{Tws}[Y_i^{\text{obs}}]P_{s|Tw} + U_{Tws}(1 - P_{s|Tw}) - \{E_{Cws}[Y_i^{\text{obs}}]P_{s|Cw} + L_{Cws}(1 - P_{s|Cw})\} \end{aligned} \tag{21}$$

for $s = 0, 1$ and $w = e, E$.

We omit the proof of this proposition, which is relatively simple, but it is available on request from the authors. The width of the bound in equation (21) is

$$\text{width}(s; w) = (U_{Cws} - L_{Cws})(1 - P_{s|Cw}) + (U_{Tws} - L_{Tws})(1 - P_{s|Tw}).$$

As we might expect, the expressions for the bound widths width(s) and width($s; w$) suggest that the benefits of our augmented design *versus* a standard randomized design depend on the role of the encouragement. Specifically, suppose that, for fixed values of $W = w^*$ and $S = s^*$, $U_{zw^*s^*} = U_{zs^*}$ and $L_{zw^*s^*} = L_{zs^*}$ for $z = C, T$. If $1 - P_{s^*|zw^*} \leqslant 1 - P_{s^*|z}$, $z = C, T$, then width($s^*; w^*$) $\leqslant$ width(s^*). This result depends on the study design in the sense that the relationship $1 - P_{s^*|zw^*} \leqslant 1 - P_{s^*|z}$, $z = C, T$, holds if the encouragement status w^* boosts units to exhibit a value of the intermediate variable S equal to s^*, which is postulated.

To fix the ideas, suppose that $s^* = 1$, and focus on ADE(1) and ADE(1; E). Throughout, we assume that $U_{zws^*} = U_{zs^*}$ and $L_{zw^*s^*} = L_{zs^*}$ for $z = C, T$, $w^* = E$ and $s^* = 1$. In a standard randomized experiment, the bounds for ADE(1) are more informative the higher the proportion of units that would always exhibit a positive value of the intermediate variable regardless of treatment assignment (the proportion of units belonging to $\{i : S_i(C) = 1, S_i(T) = 1\}$). In this respect, we can reasonably expect that, in an augmented design where the encouragement boosts units that exhibit a positive value of the mediating variable, the proportion of units that would exhibit a positive value of the mediator S when encouraged under either the C- or the T-treatment is higher than the proportion of units that would exhibit a positive value of the mediator S when not encouraged under either the C- or the T-treatment. Intuitively, therefore, if the encouragement can move units from the group of those who would show a zero value of the mediator under either the standard treatment or the active treatment to the group of units that would exhibit a positive mediator value regardless of treatment assignment when encouraged, then its presence improves the partial estimates of the ADE, by tightening the bounds.

8. Concluding remarks

In this paper we propose a new augmented design, where the treatment is randomized, and the mediating variable is not forced, but only randomly encouraged, and show how this source of exogenous variation may help to identify and estimate causal PSDEs.

Our augmented design is similar in spirit to the encouragement designs that were proposed by Imai *et al.* (2010c), although there are key differences between the two approaches. Imai *et al.* (2010c) focused on different estimands, namely direct and indirect effects defined at the individual level, and their encouragement designs, the parallel and crossover encouragement designs, allow them to identify averages of these effects for specific subpopulations. In these designs, the encouragement aims at relaxing sequential ignorability assumptions, which are usually made to identify direct and indirect effects in standard randomized experiments (e.g. Robins (2003) and Imai *et al.* (2010b)). Conversely, we focus on PSDEs, which are causal direct effects for specific subpopulations. Therefore, our approach does not require involving parallel and crossover designs, but we must 'simply' augment the standard randomized experiment with an encouragement variable for the intermediate outcome. In our augmented design, the encouragement leads to a finer partition of the population into principal strata, allowing us to obtain information on direct effects for types of units, which would, in general, provide no information in a standard randomized experiment.

In this paper, we provide a set of assumptions leading to identify PSDEs partially for the case in which the treatment and the encouragement are randomly assigned and the intermediate

variable is binary. We empirically show that our bounds on the PSDEs are narrower and more informative than those which we would derive in a standard randomized experiment. The benefits of the presence of an encouragement for the intermediate outcome are also formally shown, focusing on an ADE for the entire population. As we expect, these results strongly depend on the role of the encouragement and the design of the study.

As with any partial identification results, estimated bounds from a given sample may turn out to be uninformative, in which case making additional assumptions may be required. Future research will focus on addressing this issue. If pretreatment variables are available, auxiliary information from them can be used to enhance the efficiency of estimation and to sharpen the bounds. Indeed, although covariates do not enter the treatment or encouragement assignment mechanism in our augmented experimental design, they can improve both prediction of the missing potential outcomes as well as prediction of principal strata membership. Further sharpening of the bounds will be pursued exploiting (semi)parametric models within a Bayesian framework. A Bayesian approach will be also used for assessing the sensitivity of the results to deviations from the key assumptions. The planning phase of our augmented designs will be further investigated, to develop 'optimal' augmented designs, which allow us to achieve a required precision for estimating PSDEs, and to minimize the study's cost (e.g. Frangakis and Baker (2001)). Finally, we propose to study how our augmented design can be used as a template for the analysis of direct and indirect causal effects in observational studies.

Appendix A

A.1. Proof of proposition 1

Assumptions 4–6 imply that

$$\text{PSDE}(0;e,e) = E[Y_i(T,e) - Y_i(C,e)|G_i \in \{1,5,10\}], \tag{22}$$

$$\text{PSDE}(0;e,E) = E[Y_i(T,E) - Y_i(C,e)|G_i = 1] = \text{PSDE}_1(0;e,E), \tag{23}$$

$$\text{PSDE}(0;E,E) = E[Y_i(T,E) - Y_i(C,E)|G_i = 1] = \text{PSDE}_1(0;E,E), \tag{24}$$

$$\text{PSDE}(0;E,e) = E[Y_i(T,e) - Y_i(C,E)|G_i \in \{1,5\}] \tag{25}$$

and

$$\text{PSDE}(1;e,e) = E[Y_i(T,e) - Y_i(C,e)|G_i = 16] = \text{PSDE}_{16}(1;e,e), \tag{26}$$

$$\text{PSDE}(1;e,E) = E[Y_i(T,E) - Y_i(C,e)|G_i = 16] = \text{PSDE}_{16}(1;e,E), \tag{27}$$

$$\text{PSDE}(1;E,e) = E[Y_i(T,e) - Y_i(C,E)|G_i \in \{15,16\}], \tag{28}$$

$$\text{PSDE}(1;E,E) = E[Y_i(T,E) - Y_i(C,E)|G_i \in \{10,15,16\}]. \tag{29}$$

Bounds on $\text{PSDE}(0;e,e)$ can be derived by using information provided by the $\text{OBS}(T,e,0)$ and the $\text{OBS}(C,e,0)$ groups. The $\text{OBS}(T,e,0)$ group includes only units belonging to one of the principal strata 1, 5 and 10; thus, in large samples, $E[Y_i(T,e)|G_i \in \{1,5,10\}] = E_{Te0}[Y_i^{\text{obs}}]$. The $\text{OBS}(C,e,0)$ group is the $\pi_{1|Ce0}$, $\pi_{5|Ce0}$, $\pi_{10|Ce0}$, $\pi_{11|Ce0}$ and $\pi_{15|Ce0}$ mixture of the principal strata 1, 5, 10, 11 and 15. The conditional probability that a unit belongs to either stratum 1, or stratum 5 or stratum 10 given his or her membership to the $\text{OBS}(C,e,0)$ group is

$$\pi_{1,5,10|Ce0} \equiv \sum_{g \in \{1,5,10\}} \pi_{g|Ce0} = \frac{1 - P_{1|Te}}{1 - P_{1|Ce}}.$$

Therefore, $E_{Ce0}^{\leqslant \pi_{1,5,10|Ce0}}[Y_i^{\text{obs}}] \leqslant E[Y_i(C,e)|G_i \in \{1,5,10\}] \leqslant E_{Ce0}^{\geqslant \pi_{1,5,10|Ce0}}[Y_i^{\text{obs}}]$. Then, the large sample bounds for $\text{PSDE}(0;e,e)$ in equation (12) follow immediately by using the equality given in equation (22). Bounds

on PSDE(1; E, E) can be derived in a similar way, using information that is provided by the OBS(T, E, 1) and the OBS(C, E, 1) groups and the equality given in equation (29).

Next, consider PSDE(0; e, E) and PSDE(0; E, E). In large samples, $E[Y_i(T,E)|G_i=1] = E_{TE0}[Y_i^{\text{obs}}]$, $E_{Ce0}^{\leqslant \pi_{1|Ce0}}[Y_i^{\text{obs}}] \leqslant E[Y_i(C,e)|G_i=1] \leqslant E_{Ce0}^{\geqslant \pi_{1|Ce0}}[Y_i^{\text{obs}}]$ and $E_{CE0}^{\leqslant \pi_{1|CE0}}[Y_i^{\text{obs}}] \leqslant E[Y_i(C,E)|G_i=1] \leqslant E_{CE0}^{\geqslant \pi_{1|CE0}}[Y_i^{\text{obs}}]$, where $\pi_{1|Ce0} = (1-P_{1|TE})/(1-P_{1|Ce})$, and $\pi_{1|CE0} = (1-P_{1|TE})/(1-P_{1|CE})$. Therefore, from equations (23) and (24) we have

$$E_{TE0}[Y_i^{\text{obs}}] - E_{Ce0}^{\geqslant \pi_{1|Ce0}}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(0;e,E) \leqslant E_{TE0}[Y_i^{\text{obs}}] - E_{Ce0}^{\leqslant \pi_{1|Ce0}}[Y_i^{\text{obs}}] \tag{30}$$

and

$$E_{TE0}[Y_i^{\text{obs}}] - E_{CE0}^{\geqslant \pi_{1|CE0}}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(0;E,E) \leqslant E_{TE0}[Y_i^{\text{obs}}] - E_{CE0}^{\leqslant \pi_{1|CE0}}[Y_i^{\text{obs}}]. \tag{31}$$

Assumption 3 implies that $\text{PSDE}_1(0;e,E) = \text{PSDE}_1(0;E,E)$; thus the bounds in equation (13) can be immediately derived by combining the bounds in equations (30) and (31). A similar reasoning leads to derive bounds on PSDE(1; e, e) and PSDE(1; e, E), using information provided by the OBS(C, e, 1), the OBS(T, e, 1) and the OBS(T, E, 1) groups, equations (26) and (27), and assumption 3, which implies that $\text{PSDE}_{16}(1;e,e) = \text{PSDE}_{16}(1;e,E)$.

Finally, consider PSDE(0; E, e) and PSDE(1; E, e). To derive bounds for PSDE(0; E, e) and PSDE(1; E, e), we first assume that the value of π_5 is known. Focus on PSDE(0; E, e). Given a fixed value of π_5, let $\pi_{1,5|CE0} = (1-P_{1|TE}+\pi_5)/(1-P_{1|CE})$ and $\pi_{1,5|Te0} = (1-P_{1|TE}+\pi_5)/(1-P_{1|Te})$. Then, given equation (25) we have that the lower and upper bounds on PSDE(0, E, e) are

$$E_{Te0}^{\leqslant \pi_{1,5|Te0}}[Y_i^{\text{obs}}] - E_{CE0}^{\geqslant \pi_{1,5|CE0}}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(0;E,e) \leqslant E_{Te0}^{\geqslant \pi_{1,5|Te0}}[Y_i^{\text{obs}}] - E_{CE0}^{\leqslant \pi_{1,\,5|CE0}}[Y_i^{\text{obs}}].$$

Analogously, using information provided by the OBS(C, E, 1) and the OBS(T, e, 1) groups and equation (28) we have that the lower and upper bounds on PSDE(1, E, e) are

$$E_{Te1}^{\leqslant \pi_{15,16|Te1}}[Y_i^{\text{obs}}] - E_{CE1}^{\geqslant \pi_{15,16|CE1}}[Y_i^{\text{obs}}] \leqslant \text{PSDE}(1;E,e) \leqslant E_{Te1}^{\geqslant \pi_{15,\,16|Te1}}[Y_i^{\text{obs}}] - E_{CE1}^{\leqslant \pi_{15,\,16|CE1}}[Y_i^{\text{obs}}],$$

where

$$\pi_{15,16|CE1} = \frac{P_{1|CE} - (P_{1|TE} - P_{1|Te}) + \pi_5}{P_{1|CE}}$$

and

$$\pi_{15,16|Te1} = \frac{P_{1|CE} - (P_{1|TE} - P_{1|Te}) + \pi_5}{P_{1|Te}}.$$

Since the value of π_5 is unknown, minimizing the lower bounds and maximizing the upper bounds over the possible range of π_5 gives the desired bounds on the PSDE(0, E, e) and PSDE(1, E, e) in equations (14) and (16) respectively.

A.2. Simple estimators for the bounds derived in proposition 1

The sampling process allows us to identify the conditional distributions $P_{s|z,w}$, the conditional expected values $E_{zws}[Y_i^{\text{obs}}]$ and the conditional lower and upper trimmed means $E_{zws}^{\leqslant \alpha}[Y_i^{\text{obs}}]$ and $E_{zws}^{\geqslant \alpha}[Y_i^{\text{obs}}]$, $0<\alpha<1$. Therefore finding estimators for the bounds that are defined in proposition 1 is relatively straightforward. For instance, the following estimators can be used:

$$\hat{P}_{s|zw} = \frac{\sum_{i=1}^{n} \mathbb{1}(Z_i^{\text{obs}}=z)\,\mathbb{1}(W_i^{\text{obs}}=w)\,\mathbb{1}(S_i^{\text{obs}}=s)}{\sum_{i}^{n} \mathbb{1}(Z_i^{\text{obs}}=z)\,\mathbb{1}(W_i^{\text{obs}}=w)},$$

$$\hat{E}_{zws}[Y_i^{\text{obs}}] = \frac{\sum_{i=1}^{n} \mathbb{1}(Z_i^{\text{obs}}=z)\,\mathbb{1}(W_i^{\text{obs}}=w)\mathbb{1}(S_i^{\text{obs}}=s)Y_i^{\text{obs}}}{\sum_{i=1}^{n} \mathbb{1}(Z_i^{\text{obs}}=z)\,\mathbb{1}(W_i^{\text{obs}}=w)\,\mathbb{1}(S_i^{\text{obs}}=s)} \equiv \bar{Y}_{zws},$$

$$\hat{E}^{\leqslant\alpha}_{zws}[Y_i^{\text{obs}}] = \frac{1}{[n_{zws}\alpha]} \sum_{(i)=1}^{[n_{zws}\alpha]} Y_{(i),\,zws} \equiv \bar{Y}^{\leqslant\alpha}_{zws},$$

$$\hat{E}^{\geqslant\alpha}_{zws}[Y_i^{\text{obs}}] = \frac{1}{[n_{zws}\alpha]} \sum_{(i)=n_{zws}-[n_{zws}\alpha]+1}^{n_{zws}} Y^{\text{obs}}_{(i),\,zws} \equiv \bar{Y}^{\geqslant\alpha}_{zws},$$

$z = C, T$, $w = e, E$ and $s = 0, 1$, where $\mathbb{1}(\cdot)$ is the indicator function,

$$n_{zws} = \sum_{i=1}^{n} \mathbb{1}(Z_i^{\text{obs}} = z)\,\mathbb{1}(W_i^{\text{obs}} = w)\,\mathbb{1}(S_i^{\text{obs}} = s),$$

$[n_{zws}\alpha]$ is the largest integer not greater than $n_{zws}\alpha$ and $Y^{\text{obs}}_{(i),zws}$, $i = 1, \ldots, n_{zws}$, are the ordered statistics within the observed group OBS(z, w, s). In small samples, bounds can be wrapped in confidence bands to account for sampling variability in various ways (Imbens and Manski, 2004).

References

Baker, S. G., Frangakis, C. and Lindeman, K. S. (2007) Estimating efficacy in a proposed randomized trial with initial and later non-compliance. *Appl. Statist.*, **56**, 211–221.

Balke, A. and Pearl, J. (1997) Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Ass.*, **92**, 1171–1176.

Barnard, J., Frangakis, C. E., Hill, J. L. and Rubin, D. B. (2003) A principal stratification approach to broken randomized experiments: a case study of School Choice vouchers in New York City (with discussion). *J. Am. Statist. Ass.*, **98**, 299–323.

Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008) Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, **64**, 695–701.

Cheng, J. and Small, D. S. (2006) Bounds on causal effects in three-arm trials with non-compliance. *J. R. Statist. Soc.* B, **68**, 815–836.

Cheng, J., Small, D. S., Tan, Z. and Ten Have, T. R. (2009) Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika*, **96**, 19–36.

Dawid, A. P. (2000) Causal inference without counterfactuals (with discussion). *J. Am. Statist. Ass.*, **95**, 407–448.

Flores, C. A. and Flores-Lagunes, A. (2009a) Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. *Discussion Paper 4237.* Institute for the Study of Labor, Bonn.

Flores, C. A. and Flores-Lagunes, A. (2009b) Nonparametric partial and point identification of net or direct causal effects. *American Economics Association A. Meet.*, paper 2009.

Flores, C. A. and Flores-Lagunes, A. (2010) Nonparametric partial identification of causal net and mechanism average treatment effects. *Working Paper*. Department of Food and Resource Economics, University of Florida, Gainesville.

Follmann, D. (2006) Augmented designs to assess immune response in vaccine trials. *Biometrics*, **62**, 1161–1169.

Frangakis, C. E. and Baker, S. G. (2001) Compliance subsampling designs for comparative research: estimation and optimal planning. *Biometrics*, **27**, 899–908.

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

Frangakis, C. E., Rubin, D. B., An, M. W. and MacKenzie, E. (2007) Principal stratification designs to estimate input data missing due to death (with discussion). *Biometrics*, **63**, 641–662.

Gallop, R., Small, D., Lin, J. Y., Elliot, M. R., Joffe, M. M. and Ten Have, T. R. (2009) Mediation analysis with principal stratification. *Statist. Med.*, **28**, 1108–1130.

Geneletti, S. (2007) Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Statist. Soc.* B, **69**, 199–215.

Gilbert, P. B. and Hudgens, M. G. (2008) Evaluating candidate principal surrogate endpoints. *Biometrics*, **64**, 1146–1154.

Goetgeluk, S., Vansteelandt, S. and Goetghebeur, E. (2008) Estimation of controlled direct effects. *J. R. Statist. Soc.* B, **70**, 1049–1066.

Holland, P. (1986) Statistics and causal inference (with discussion). *J. Am. Statist. Ass.*, **81**, 945–970.

Imai, K. (2008) Sharp bounds on causal effects in randomized experiments with "truncation-by-death". *Statist. Probab. Lett.*, **78**, 144–149.

Imai, K., Keele, L. and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychol. Meth.*, **15**, 309–334.

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

Imai, K., Tingley, D. and Yamamoto, T. (2010c) Experimental designs for identifying causal mechanisms. *Technical Report*. Department of Politics, Princeton University, Princeton.

Imbens, G. W. and Manski, C. F. (2004) Confidence intervals for partially identified parameters. *Econometrica*, **72**, 1845–1857.
Jin, H. and Rubin, D. B. (2008) Principal stratification for causal inference with extended partial compliance. *J. Am. Statist. Ass.*, **103**, 101–111.
Joffe, M. M. and Greene T. (2009) Related causal frameworks for surrogate outcomes. *Biometrics*, **65**, 530–539.
Joffe, M. M., Small, D. and Hsu, C.-Y. (2007) Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statist. Sci.*, **22**, 74–97.
Lee, D. S. (2009) Training, wages, and sample selection: estimating sharp bounds on treatment effects. *Rev. Econ. Stud.*, **76**, 1071–1102.
Lynch, K. G., Cary, M., Gallop, R. and Ten Have, T. R. (2008) Causal mediation analysis for randomized trial. *Hlth Serv. Outcm. Res. Meth.*, **8**, 57–76.
Manski, C. F. (2003) *Partial Identification of Probabilities Distributions*. New York: Springer.
Mattei, A. and Mealli, F. (2007) Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics*, **63**, 437–446.
Mealli, F. and Rubin, D. B. (2003) Assumptions allowing the estimation of direct causal effects. Commentary on 'Healthy, wealthy, and wise?: tests for direct causal paths between health and socioeconomic status' (by P. Adams, M. D. Hurd, D. McFadden, A. Merrill and T. Ribeiro). *J. Econmetr.*, **112**, 79–87.
Pearl, J. (2000) *Causality*. Cambridge: Cambridge University Press.
Pearl, J. (2001) Direct and indirect effects. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence* (eds J. S. Breese and D. Koller), pp. 411–420. San Francisco: Morgan Kaufmann.
Petersen, M., Sinisi, S. E. and van der Laan, M. (2006) Estimation of direct causal effects. *Epidemiology*, **17**, 276–284.
Qin, L., Gilbert, P. B., Follmann, D. and Li, D. (2008) Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the cox model. *Ann. Appl. Statist.*, **2**, 386–407.
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 Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.
Rosenbaum, P. R. (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Statist. Soc.* A, **147**, 656–666.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1977) Assignment to a treatment group on the basis of a covariate. *J. Educ. Statist.*, **2**, 1–26.
Rubin, D. B. (1978) Bayesian inference for causal effects. *Ann. Statist.*, **6**, 34–58.
Rubin, D. B. (1980) Comment on 'Randomization analysis of experimental data: the Fisher randomization test' (by D. Basu). *J. Am. Statist. Ass.*, **75**, 591–593.
Rubin, D. B. (1990) Comment: 'Neyman (1923) and Causal inference in experiments and observational Studies'. *Statist. Sci.*, **5**, 472–480.
Rubin, D. B. (2004) Direct and indirect causal effects via potential outcomes (with discussion). *Scand. J. Statist.*, **31**, 161–170, 196–198.
Schwartz, S. L., Li, F. and Mealli, F. (2010) A Bayesian semiparametric approach to intermediate variables in causal inference. *Mimeo*. Department of Statistical Science, Duke University, Durham.
Sjölander, A., Humphreys, K., Vansteelandt, S., Bellocco, R. and Palmgren, J. (2009) Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics*, **65**, 514–520.
VanderWeele, T. L. (2008) Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.*, **78**, 2957–2962.
VanderWeele, T. L. (2009) Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, **20**, 18–26.
Zhang, J. L. and Rubin, D. B. (2003) Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educ. Behav. Statist.*, **28**, 353–368.
Zhang, J. L., Rubin, D. B. and Mealli, F. (2009) Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J. Am. Statist. Ass.*, **104**, 166–176.