Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2010/04

Augmented Designs to Assess Principal Strata Direct Effects

Alessandra Mattei,
Fabrizia Mealli

Università degli Studi
di Firenze

Augmented Designs to Assess Principal Strata Direct Effects

By A. Mattei and F. Mealli
Department of Statistics, University of Florence,
Viale Morgagni 59, 50134 Florence, Italy
mattei@ds.unifi.it mealli@ds.unifi.it

Summary

Many research questions involving causal inference are often concerned with understanding the causal pathways by which a treatment affects an outcome. Thus, the concept of 'direct' versus 'indirect' effects comes to play. Disentangling direct and indirect effects may be a difficult task, because the intermediate outcome is generally not under experimental control. We tackle this problem 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 to involve a priori counterfactual outcomes. Using non parametric identification strategies, we provide a set of assumptions, which allow us to partially identify the causal estimands of interest, the Principal Strata Direct Effects. Large sample bounds for various 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 with and contrasted to a standard randomized design.

Some key words: Augmented Designs; Bounds; Causal Inference; Direct and Indirect Effects; Principal Stratification.

1. Introduction

Many research questions involving causal inference are often concerned with understanding the causal pathways by which an exposure or a treatment affects an outcome. Researchers want to know not only if the treatment is effective, but also how the treatment effect on the outcome is mediated by intermediate posttreatment variables. Thus, the concept of 'direct' versus 'indirect' effects comes to play. The use of this concept is common not only in statistics, but also in many area of social, economic and political sciences as well as in biomedical and pharmacological sciences, where they are the closely related concepts of 'biomarkers' and 'surrogate outcomes' (e.g., Joffe & Greene (2009); Gilbert & 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. In order to understand the mechanistic pathways by which the drug acts to cause or prevent a disease, the total treatment effect has to 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 would be taken away, and so answers policy-related questions of practical significance (e.g., the drug manufacturer might consider ways of eliminating the adverse side-effects of the drug; doctors might deem helpful to suggest or prevent the use of rescue medication).

Disentangling direct and indirect effects may be a difficult task, because the intermediate outcome is generally not under experimental control. For instance, in the drug treatment example, side-effects, and so the use of a rescue medication, cannot be, in general, controlled. Traditional analyses of scientific problems where treatment comparisons need to be adjusted for posttreatment confounded variables are typically based on a standard method that directly controls for (conditions on) observed values of those posttreatment variables, resulting in estimates that generally lack causal interpretation (e.g., Cochran (1957); Rosenbaum (1984); Prentice (1989); Freedman et al. (1992); Lin et. al (1997); Buyse & Molenberghs (1998); Buyse et al. (2000)). On one other hand, a common problem in the existing literature attempting to estimate causal mechanisms of a treatment is that the estimands are not clearly defined, or are defined within the context of the estimation procedure used (e.g., OLS, matching), and the assumptions needed for a causal interpretation of the estimates are not always made explicit.

The definition of direct and indirect effects is straightforward in linear equation systems, but is rather contrived in non-linear systems. 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. Currently, predominant frameworks to causal inference include the decision theoretic approach, which is grounded in statistical decision theory (Dawid, 2000, 2002), the causal graph or structural models framework (Pearl, 1995, 2000), and the potential outcomes framework, originally introduced by Neyman (1923) for randomized experiments and randomization-based inference, and generalized and extended by Rubin (1974, 1977, 1978) for nonrandomized studies and alternative forms of inference. In general, these approaches focus on different causal estimands, and full agreement on what the relevant estimands should be and how one should estimate them is still lacking.

In this paper we focus on the potential outcomes framework, also referred to as the Rubin Causal Model (RCM, Holland (1986)), and use the concept of the principal stratification (Frangakis & Rubin, 2002) for addressing the topic of direct and indirect causal effects. Fundamentally, the potential outcome perspective views causal inference as a problem of missing data with explicit mathematical modelling of the assignment mechanism as a process for revealing the observed data. One of the advantages of the framework is that it allows for heterogeneity of treatment effects. In addition, causal estimands can be defined and assumptions stated without specifying parametric models. Specifically, the RCM allows one to make explicit the assumptions necessary for valid causal inference, and to clearly define the structural behavioural assumptions, which are the ones that make the estimands of interest identifiable. Finally, by separating and defining the critical assumptions, the RCM allows for a clear assessment of the consequences of violations of these assumptions through sensitivity analysis.

In this setting, the use of principal stratification is key to understanding the meaning of direct and indirect causal effects (Mealli & Rubin, 2003; Rubin, 2004). Principal stratification with respect to a posttreatment intermediate variable is a cross-classification of subjects into latent classes defined by the joint potential values of that posttreatment variable under each of the treatments being compared, so principal strata comprise units having the same values of the intermediate potential outcomes. Frangakis & Rubin (2002) define 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. As a result, the central property of a PCE is that it is always a causal effect and does not suffer from the complications of standard posttreatment-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 causal 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. On the other hand, if there is no causal effect of treatment on the outcome for these subjects, then 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 causal effect of treatment on the posttreatment intermediate variable.

Principal stratification is one of the several possible ways to conceptualize the mediatory role of an intermediate variable in the treatment-outcome relationship Joffe et al. (2007). An alternative approach, usually applied in the causal graph framework to causal inference, 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 posttreatment variable were somehow simultaneously forced to attain a predetermined value (Robins & 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. In the studies we consider, however, we prefer to relax this assumption, and focus on the principal stratification approach, which classifies the population into groups according to potential mediator behavior, without assuming any manipulation of mediators.

PCE analysis is challenging, due to the latent nature of principal strata. Identification and estimation strategies must generally involve techniques for incomplete data, which usually require strong structural or modelling assumptions. Some of these assumptions can be weakened by using alternative study designs (e.g., Follman (2006); Baker et al. (2007)). In this paper, in order to ease identification and estimation of PCEs, we will 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 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 a rescue medication to randomly selected patients. In biomedical and pharmacological sciences, Sjölander et al. (2009) focus on assessing the effect of physical activity on circulation diseases, not channeled through body mass index. Body mass index represents a biomarker, and it is not obvious how to conceptualize interventions on such a variable. However, suitable level of body mass index might be encouraged for instance, by suggesting to follow a specific diet to randomly selected patients[1].

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 4 we present our design's structural assumptions and we derive large sample bounds for $PSDE$s 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 with and contrasted to 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 ($z = C$) or a new treatment ($z = T$). 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 the set of all intermediate variables which are on the causal pathway between the treatment and the main endpoint, Y. Henceforth, for simplicity of notation, we assume that S is a single binary variable (e.g., assuming values 0 and 1). Let $Y_i(z)$ and $S_i(z)$ denote the potential outcomes of Y and S, respectively if treatment Z was set, possibly contrary to fact, to the value 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 priori counterfactuals for units assigned to treatment z who exhibit a value of the intermediate outcome S not equal to s, because in one specific experiment, they can be never observed for such 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 received by other individuals (Stable Unit Treatment Value Assumption: SUTVA; Rubin (1978, 1980, 1990)).

Robins & Greenland (1992) and Pearl (2001) give definitions for controlled direct effects and natural direct and indirect effects based on interventions on the intermediate variable, using a priori counterfactuals. The (average) Controlled Direct Effect (CDE) of the treatment Z on the outcome Y, setting S to s, is defined by $CDE(s) = E[Y_i(T, s) - Y_i(C, s)]$, and measures the effect of Z on Y not mediated through S, that is, the effect of Z on Y after intervening to fix the mediator, S, to some value s. The (average) Natural Direct Effect (NDE) also measures the effect of the treatment Z on the outcome Y not mediated through the intermediate variable S, but now the mediator is hold fixed at whatever level it would have taken under a predetermined level z of the treatment:

[1] Some augmented designs have been recently proposed in the vaccine trial literature (e.g., Follman (2006); 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.

$NDE(z) = E[Y_i(T, S_i(z)) - Y_i(C, S_i(z))]$, $z = C, T$. Corresponding to the NDE is the concept of Natural Indirect Effect (NIE), which assesses the extent to which an intervention affects the outcome through the mediator. The NIE measures the effect on the outcome Y of intervening to set the mediator to what it would have been if the treatment was fixed at value T in contrast to what it would have been if the treatment was fixed at C: $NIE(z) = E[Y_i(z, S_i(T)) - Y_i(z, S_i(C))]$, $z = C, T$. These effects provide an intuitive decomposition of the Average total Treatment Effect ($ATE = E[Y_i(T) - Y_i(C)]$) into the sum of a natural direct effect and a natural indirect effect:

$$ATE \equiv E[Y_i(T) - Y_i(C)] \equiv E[Y_i(T, S_i(T)) - Y_i(C, S_i(C))]$$
$$= E[Y_i(T, S_i(T)) - Y_i(T, S_i(C))] + E[Y_i(T, S_i(C)) - Y_i(C, S_i(C))] = NIE(T) + NDE(C),$$

or

$$ATE \equiv E[Y_i(T) - Y_i(C)] \equiv E[Y_i(T, S_i(T)) - Y_i(C, S_i(C))]$$
$$= E[Y_i(T, S_i(T)) - Y_i(C, S_i(T))] + E[Y_i(C, S_i(T)) - Y_i(C, S_i(C))] = NDE(T) + NIE(C).$$

Flores & Flores-Lagunes (2009a,b) make a similar distinction of direct and indirect effects by introducing the concepts of Mechanism Average Treatment Effect (MATE), and Net Average Treatment Effect (NATE). Analogously, Didelez et al. (2006) and Geneletti (2007) apply similar concepts using a decision theoretic approach, where relationships between variables are encoded by using conditional independence statements, but without using counterfactuals.

Various identification strategies for controlled direct effects and natural direct and indirect effects have been developed (e.g., Robins & Greenland (1992); Pearl (2001); Robins (2003); Petersen et al. (2006); Geneletti (2007); Flores & Flores-Lagunes (2009a,b)), and most of them involves causal graphical models: an exception can be found in Flores & Flores-Lagunes (2009a,b), who present identification and estimation strategies within the RCM. 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 strong conditions such as constant-effect, parametric and/or conditional independence assumptions. Conversely, a principal stratification approach does not require to use estimation strategies based on extrapolation methods, because it does not use a priori counterfactuals, but addresses the topic of direct and indirect effects by focussing on subsets of the target population - the principal strata - which can naturally provide information on the causal pathways by which the treatment affects the outcome.

Formally, a principal stratum with respect to the posttreatment variable S is a group of individuals who have the same vector $(S_i(C), S_i(T))$, and a principal causal effect is a comparison between the potential outcomes $Y_i(C)$ and $Y_i(T)$ within a particular stratum (or union of principal strata): $PCE(s_C, s_T) = E[Y_i(T) - Y_i(C) | S_i(C) = s_C, S_i(T) = s_T]$. Since S is supposed to be a binary variable, units can be classified into four latent groups: subjects who would exhibit a zero value of the intermediate outcome under both treatment arms ($1 = \{i : S_i(C) = S_i(T) = 0\}$); subjects who would exhibit a positive value of the intermediate outcome under control but would exhibit a zero value of the intermediate outcome under treatment ($2 = \{i : S_i(C) = 1, S_i(T) = 0\}$); subjects who would exhibit a zero value of the intermediate outcome under control but would exhibit a positive value of the intermediate outcome under treatment ($3 = \{i : S_i(C) = 0, S_i(T) = 1\}$); and subjects who would exhibit a positive value of the intermediate outcome under both treatment arms ($4 = \{i : S_i(C) = S_i(T) = 1\}$).

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, $S_i(C) = S_i(T)$ (i.e., principal strata 1 and 4). Specifically, the $PSDE$ of Z on Y at level s, $s \in \{0, 1\}$, can be formally defined as

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

If $PSDE(s) = 0$, for each $s = 0, 1$, all effect is indirect, that is, mediated by the posttreatment variable S, because if the treatment cannot change S, it cannot affect the primary outcome Y.

PCEs (and $PSDE$s) 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:

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

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 to the a priori counterfactual approaches, where direct and indirect causal effects are defined for each individual and average over the entire population. For this reason, whereas a priori counterfactual direct and indirect effects provide a natural decomposition of the total effect, principal stratification does not in general allow one to decompose the total effect into direct and indirect effects, unless additional assumptions are made: the PCEs for units belonging to principal strata where the posttreatment variable is affected by treatment combine direct and indirect effects.

VanderWeele (2008) studies the conceptual relations between $PSDE$s and controlled and natural direct effects, assuming knowledge on all potential outcomes. He shows that if there are no controlled direct effects or no natural direct effects then there can be no principal strata direct effects. However the absence of principal strata direct effects does not imply the absence of natural direct effects and does not necessarily even imply that the average controlled or natural direct effect is zero. These relationships, however, immediately follow from the definition of $PSDE$s and controlled direct effects and natural direct effects: $PSDE$s are ‘local’ effects (i.e., they are causal effects within principal strata), whereas controlled direct effects and natural direct effects are defined for each unit. Although these are relevant theoretical results, they do not help identification and estimation of the causal estimands, using the data usually available. The observed data contain information on the a priori counterfactuals $Y_i(z, S_i(z'))$, $z \neq z'$ only for those units who receive the treatment z and for which the treatment does not affect the intermediate variable ($S_i(C) = S_i(T)$). Analogously, information on the a priori counterfactuals $Y_i(z, s)$ can be only drawn for those units who receive the treatment z and for which $S_i(z) = s$. This result implies that, in the presence of heterogeneous effects, estimation of average controlled direct effects and average natural direct effects for other subpopulations (including the entire population) can only be based on extrapolations of the a priori counterfactuals $Y_i(z, S_i(z'))$, $z \neq z'$ to those units for which the treatment affects the mediator, and extrapolations of the a priori counterfactuals $Y_i(z, s)$ to those units for which $S_i(z) \neq s$, since their potential outcomes are never observed. In this paper, we are not willing to use extrapolation methods, therefore we prefer to focus on the identification of PCEs, by concentrating on partial identification of $PSDE$s.

Various (full and partial) identification strategies, estimation methods and applications for the concepts of principal stratification and principal effects have been considered (e.g., Barnard et al. (2003); Cheng et al. (2009); Frangakis et al. (2007); Gallop et al. (2009); Imai (2008); Lee (2009); Lyncn et al. (2008); Mattei & Mealli (2007); Sjölander et al. (2009); Zhang & Rubin (2003); Zhang et al. (2009)). In this paper we contribute to this literature developing a novel approach to the identification and estimation of principal stratum direct and indirect effects effects based on new augmented designs. Focus will be on the assumptions characterizing the encouragement variable, which allows one to partially identify causal estimands for specific subpopulations, and usually derive tighter bounds than those derived in standard randomized experiments.

3. The Principal Stratification Framework with a Treatment and an Encouragement Variable

Inference about principal causal effects, which involves prediction of the subjects' missing memberships to the principal strata, as well as prediction of the subjects' missing potential outcomes, requires that some identifying assumptions are made. In order to clearly define the critical assumptions, it is crucial to think very carefully about the hypothetical randomized experiment that led to the observed data. With 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. Actually, 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, henceforth referred to as encouragement variable, which affects the primary outcome, only through its effect on the intermediate outcome. Using the econometric language (Reiersol, 1941; Haavelmo, 1943), this additional variable would play the role of an instrument. 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[2].

Formally, each unit i in the sample can be potentially either encouraged or not encouraged to exhibit a specific value of the intermediate outcome, S: 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 $\boldsymbol{Z}$ and $\boldsymbol{W}$ denote the n-dimensional vectors with ith element Z_i and W_i, respectively. Next, let $S_i(\boldsymbol{Z}, \boldsymbol{W})$ and $Y_i(\boldsymbol{Z}, \boldsymbol{W})$ 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, $\boldsymbol{Z}$ and $\boldsymbol{W}$.

This notation emphasizes that the potential outcomes of any subject can be affected by the treatment and encouragement assignments of every other subject. In addition, there might be different forms of each treatment and encouragement level for each unit.

[2] This parallelism between a standard encouragement design and a randomized experiment involving an encouragement design for the mediating variable should not lead to equalize the two frameworks. Indeed, alternative identifying strategies and assumptions have to be made in order to draw inference about direct and indirect casual effects by using this new augmented design.

Table 1. *Principal strata with two binary treatments and a binary intermediate variable*

G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$	G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$
1	0	0	0	0	9	0	1	1	0
2	1	0	0	0	10	0	1	0	1
3	0	1	0	0	11	0	0	1	1
4	0	0	1	0	12	1	1	1	0
5	0	0	0	1	13	1	1	0	1
6	1	1	0	0	14	1	0	1	1
7	1	0	1	0	15	0	1	1	1
8	1	0	0	1	16	1	1	1	1

Consistently with the approach we adopted in the previous section, we generalize the Stable Unit Treatment Value Assumption (SUTVA; Rubin (1978, 1980, 1990)) by assuming that (i) the potential outcomes for any unit do not vary with the treatments and the encouragements assigned to any other units; and (ii) for each unit there are no different versions of each treatment and encouragement level. Formally,

Assumption 1 (Stable Unit Treatment Value Assumption, SUTVA). *If $Z_i = Z_i'$ and $W_i = W_i'$, then $S_i(\boldsymbol{Z},\boldsymbol{W}) = S_i(\boldsymbol{Z}',\boldsymbol{W}')$ and $Y_i(\boldsymbol{Z},\boldsymbol{W}) = Y_i(\boldsymbol{Z}',\boldsymbol{W}')$*

SUTVA allows one to write $S_i(\boldsymbol{Z},\boldsymbol{W})$ and $Y_i(\boldsymbol{Z},\boldsymbol{W})$ 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: $S_i(C,e), S_i(C,E), S_i(T,e), S_i(T,E)$, and four potential values for the response variable: $Y_i(C,e), Y_i(C,E), Y_i(T,e), Y_i(T,E)$. SUTVA is an important restriction, and situations where this assumption is not plausible require to carefully investigate the potentially complex interactions between units and the entire set of treatment and encouragement levels which a unit might receive.

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 posttreatment variable S is the partition of units $i = 1,\dots,n$ 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 posttreatment variable S is a partition of the units whose sets are unions of sets in the basic principal stratification P_0 (Frangakis & Rubin, 2002).

Because the posttreatment variable is binary, units can be classified into sixteen basic principal strata as shown in Table 1. For instance, the principal stratum $1 = \{i : S_i(C,e) = 0, S_i(C,E) = 0,\ S_i(T,e) = 0,\ S_i(T,E) = 0\}$ comprises units who would exhibit a zero value of the mediating variable under each arm defined by the joint values of the treatment and encouragement variables, and the principal stratum $10 = \{i : S_i(C,e) = 0, S_i(C,E) = 1\ \ S_i(T,e) = 0, S_i(T,E) = 1\}$ comprises units who 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, $g \in \{1,2,\dots,16\}$ comprises a proportion, π_g of all units. Let G_i represent the principal stratum indicator for subject i: $G_i \in \{1,2,\dots,16\}$.

This partition of the units can be viewed as a generalization of the idea of principal stratification (Frangakis & Rubin, 2002) to multiple treatments. Generally, a principal

Table 2. *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	$PSDE_1(0;e,e)$	$PSDE_1(0;E,E)$	$PSDE_1(0;e,E)$	$PSDE_1(0;E,e)$
2	1	0	0	0		$PSDE_2(0;E,E)$		$PSDE_2(0;E,e)$
3	0	1	0	0	$PSDE_3(0;e,e)$		$PSDE_3(0;e,E)$	
4	0	0	1	0		$PSDE_4(0;E,E)$	$PSDE_4(0;e,E)$	
5	0	0	0	1	$PSDE_5(0;e,e)$			$PSDE_5(0;E,e)$
6	1	1	0	0				
7	1	0	1	0	$PSDE_7(1;e,e)$	$PSDE_7(0;E,E)$		
8	1	0	0	1			$PSDE_8(1;e,E)$	$PSDE_8(0;E,e)$
9	0	1	1	0			$PSDE_9(0;e,E)$	$PSDE_9(1;E,e)$
10	0	1	0	1	$PSDE_{10}(0;e,e)$	$PSDE_{10}(1;E,E)$		
11	0	0	1	1				
12	1	1	1	0	$PSDE_{12}(1;e,e)$			$PSDE_{12}(1;E,e)$
13	1	1	0	1		$PSDE_{13}(1;E,E)$	$PSDE_{13}(1;e,E)$	
14	1	0	1	1	$PSDE_{14}(1;e,e)$		$PSDE_{14}(1;e,E)$	
15	0	1	1	1		$PSDE_{15}(1;E,E)$		$PSDE_{15}(1;E,e)$
16	1	1	1	1	$PSDE_{16}(1;e,e)$	$PSDE_{16}(1;E,E)$	$PSDE_{16}(1;e,E)$	$PSDE_{16}(1;E,e)$

stratification with a binary treatment and a binary encouragement generates the following $PSDE$s:

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

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

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

Note that,

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

where the outer expectations are 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 $S_i(T,w') = S_i(C,w) = s$. Therefore, each $PSDE$, which involves units belonging to the union of different sets in the basic principal stratification, can be decomposed into 'basic' $PSDE$s, namely direct effects within sets of the basic principal stratification. In our setting with a binary intermediate outcome, eight $PSDE$s 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 2.

As stated previously, 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

Table 3. *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

data are

$$Z_i^{\text{obs}}, W_i^{\text{obs}}, S_i^{\text{obs}}\left(Z_i^{\text{obs}}, W_i^{\text{obs}}\right), Y_i^{\text{obs}}\left(Z_i^{\text{obs}}, W_i^{\text{obs}}\right),$$

which we will denote by $Z_i^{\text{obs}}, W_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}}$, $i = 1, \ldots, n$. Therefore, what we can observe are the eight groups reported in Table 3, where the latent principal strata associated with each observed group are shown. Each subject is observed to fall into one of these groups. If all sixteen principal strata existed, that is, if $\pi_g > 0$, for each $g \in 1, 2, \ldots, 16$, each observed group would be a mixture of eight principal strata.

Throughout the paper, let $OBS(z, w, s)$ denote the observed group 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 $P_{s|z,w} = Pr\left(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w\right)$ be the conditional distribution of the observed intermediate outcome given the treatment status and the encouragement status.

4. Structural Assumptions

A key component in a causal analysis is the assignment mechanism: the process that determines which units receives which treatments, hence which potential outcomes are observed, and which are missing. 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,*

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

Assumption 2 implies that

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

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

In order 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 for subpopulations of treated and control units. Specifically, we assume that within each treatment arm z, $z = C, T$, the distributions of two potential outcomes $Y_i(z, e)$ and $Y_i(z, E)$ are the same for units who 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,

Assumption 3. (Conditional Stochastic Exclusion Restrictions w.r.t. the Encouragement).

$$\begin{aligned} & Pr\left(Y_i\left(z,E\right) \mid S_i(z,E)=S_i(z,e), S_i(z',E), S_i(z',e)\right) \\ & = Pr\left(Y_i\left(z,e\right) \mid S_i(z,E)=S_i(z,e), S_i(z',E), S_i(z',e)\right) \qquad \textit{for } z \neq z' \in \{C,T\} \end{aligned}$$

Assumption 3 implies that some of the basic $PSDE$s, $PSDE_{G_{s,w,w'}}(s,w,w')$, take the same value, which only depends on the value of the intermediate potential outcomes. Specifically, $PSDE_{G_{s,e,w'}}(s;e,w') = PSDE_{G_{s,E,w'}}(s;E,w')$, 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 $S_i(T,w') = s$, $w' \in \{e,E\}$. Analogously, $PSDE_{G_{s,w,e}}(s;w,e) = PSDE_{G_{s,w,E}}(s;w,E)$, 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 $S_i(C,w) = s$, $w \in \{e,E\}$.

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

Assumption 4 (Nonzero Average Causal Effect of W on S). *The average causal effect of W on S*

$$E[(S_i\left(z,E\right) - S_i\left(z,e\right)] \qquad \textit{for } z = C,T$$

is not equal to zero.

This Assumption warrants that there is at least one stratum where the behavior with respect to the intermediate variable S is different with and without encouragement.

In order to identify (even partially, Manski (1990, 2003)) the proportion of each principal stratum and the corresponding $PSDE$s additional assumptions are required. Alternative sets of assumptions, which allow us to either reduce the number of strata or state the equivalence of the distribution of Y across some strata, can be proposed. Here we focus on a specific set of assumptions, which leads to partially identify the causal estimands of interest. We will show how the presence of an encouragement variable can be exploited to derive large sample bounds for these causal estimands, which are narrower than those we would derive in the absence of the encouragement variable.

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

Assumption 5 (Monotonicity of S with respect to W). *For all i*

$$(i) \quad S_i(C,e) \leq S_i(C,E) \qquad \textit{and} \qquad S_i(T,e) \leq S_i(T,E)$$

or

$$(ii) \quad S_i(C,e) \geq S_i(C,E) \qquad \textit{and} \qquad S_i(T,e) \geq S_i(T,E)$$

Assumption 5 relates to the mediating variable, S, with respect to the encouragement variable W. Without loss of generality, let $S_i(C,e) \leq S_i(C,E)$ and $S_i(T,e) \leq S_i(T,E)$ for all i. Therefore, Assumption 5 implies that for a fixed value of the treatment variable, units who exhibit a positive value of S when $W = e$, would exhibit a positive value of S also when $W = E$. Unfortunately, the data can never provide any direct evidence

Table 4. *Principal strata (Table on the left) and observed groups with associated possible latent principal strata (Table on the right) under Assumption 5(i)*

G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$
1	0	0	0	0
3	0	1	0	0
5	0	0	0	1
6	1	1	0	0
10	0	1	0	1
11	0	0	1	1
13	1	1	0	1
15	0	1	1	1
16	1	1	1	1

Z_i^{obs}	W_i^{obs}	S_i^{obs}	Latent Strata (G_i)					
C	e	0	1	3	5	10	11	15
C	e	1	6	13	16			
C	E	0	1	5	11			
C	E	1	3	6	10	13	15	16
T	e	0	1	3	5	6	10	13
T	e	1	11	15	16			
T	E	0	1	3	6			
T	E	1	5	10	11	13	15	16

against this Assumption, so that it is not testable without auxiliary information. However, Assumption $5(i)$ may be made plausible by design, for instance, encouraging units to exhibit a positive value of the mediating variable. Assumption $5(i)$ rules out the existence of seven out of the sixteen principal strata $(2, 4, 7-9, 12$, and $14)$, as shown in Table 4.

We also make one additional assumption, which implies that for a fixed encouragement level, units who exhibit a positive value of S when exposed to the active treatment, would exhibit a positive value of S also when randomly assigned to the standard treatment. Formally,

Assumption 6 (Monotonicity of S with respect to Z).

$$S_i(C,e) \leq S_i(T,e) \qquad \text{and} \qquad S_i(C,E) \leq S_i(T,E).$$

Together, the monotonicity Assumptions 5 and 6 rule out the existence of many principal strata $(2-4,\ 6-9,\ 12-14)$, leading to a classification of units across principal strata, which allows us to more easily investigate the benefits of our augmented randomized design with respect to a standard treatment randomized design. Under Assumptions 1 through 6, we can point identify the proportion of units who belong to the first and the last principal stratum (see Table 5):

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

and derive large sample bounds for the other principal stratum proportions and the $PSDE$ estimands. The partial identification strategy we pursue is similar in spirit to those in Cai et al. (2008); Flores & Flores-Lagunes (2009b); Imai (2008); Lee (2009), and Zhang & Rubin (2003). However, our general set up has peculiar features, stemming from the presence of the encouragement variable for the intermediate outcome. In addition, the causal estimands of interest are different: Imai (2008); Lee (2009) and Zhang & Rubin (2003) aim at identifying the average treatment effects in the presence of truncation by death, which can be viewed as a special type of posttreatment variable. Cai et al. (2008) and Flores & Flores-Lagunes (2009b) focus on an a priori counterfactual estimands: the average controlled direct effect and the net average treatment effect, respectively. These effects are defined as the average on the entire population of unit-level direct treatment effects. Indeed, our focus is on $PSDE$s as defined in equation (2), which are local effects, being defined for particular subpopulations or principal strata.

Table 5. *Principal strata (Table on the left) and observed groups with associated possible latent principal strata (Table on the right) 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

Z_i^{obs}	W_i^{obs}	S_i^{obs}	Latent Strata (G_i)				
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 *PSDE*s

The two equations in (3) imply that

$$\pi_5 + \pi_{10} = P_{1|TE} - P_{1|Te} \tag{4}$$
$$\pi_5 + \pi_{11} = P_{1|TE} - P_{1|CE} \tag{5}$$
$$\pi_{10} + \pi_{15} = P_{1|CE} - P_{1|Ce} \tag{6}$$
$$\pi_{11} + \pi_{15} = P_{1|Te} - P_{1|Ce} \tag{7}$$

In order for Equations (4) - (7) to hold, the differences on their right must be non negative. Note that $\left(P_{1|TE} - P_{1|CE}\right)$ is the average causal effect of the treatment on the intermediate outcome among units randomly encouraged; $\left(P_{1|CE} - P_{1|Ce}\right)$ and $\left(P_{1|TE} - P_{1|Te}\right)$ are the average causal effects of the encouragement on the intermediate outcome among units randomly assigned to the standard and active treatment, respectively; and $\left(P_{1|Te} - P_{1|Ce}\right)$ is the average causal effect of the treatment on the intermediate outcome among units who are not encouraged. Therefore, Assumptions 5 and 6 are not falsified by the data if in large sample these causal effects are non negative.

Using Equations (3), (4) and (5), and taking into account that the principal strata proportions need to add up to one ($1 = \pi_1 + \pi_5 + \pi_{10} + \pi_{11} + \pi_{15} + \pi_{16}$), we have

$$\pi_{10} = P_{1|TE} - P_{1|Te} - \pi_5 \tag{8}$$
$$\pi_{11} = P_{1|TE} - P_{1|CE} - \pi_5 \tag{9}$$
$$\pi_{15} = \pi_5 + \left(P_{1|CE} - P_{1|Ce}\right) - \left(P_{1|TE} - P_{1|Te}\right) \tag{10}$$

Equations (8), (9) and (10) hold for any π_5 such that

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

We now establish large sample bounds on the *PSDE*s. In order to formally write these bounds we introduce some extra notation.

Let $\pi_{g|zws}$ denote the conditional probability that a unit belongs to the principal strata g, $g = 1, 5, 10, 11, 15, 16$, given that the unit is observed to belong to the $OBS\,(z, w, s)$ group, $z = C, T; w = e, E; s = 0, 1$:

$$\pi_{g|zws} = Pr\left(G_i = g | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right).$$

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 using Equation (11). For instance,

$$\pi_{10|CE1} = \frac{\pi_{10}}{\pi_{10} + \pi_{15} + \pi_{16}} = \frac{\left(P_{1|TE} - P_{1|Te}\right) - \pi_5}{P_{1|CE}}.$$

Therefore, from Equation (11), we have

$$\min_{\pi_5} \frac{\left(P_{1|TE} - P_{1|Te}\right) - \pi_5}{P_{1|CE}} \le \pi_{10|CE1} \le \max_{\pi_5} \frac{\left(P_{1|TE} - P_{1|Te}\right) - \pi_5}{P_{1|CE}}.$$

Each $OBS(z,w,s)$ group is the $\pi_{g|zws}$ mixture of some principal strata g, $g = 1,5,10,11,15,$ 16. For instance, the $OBS(C,E,1)$ group is the mixture of the principal strata 10, 15, and 16 with weights $\pi_{10|CE1} = \pi_{10}/\left(\pi_{10}+\pi_{15}+\pi_{16}\right)$, $\pi_{15|CE1} = \pi_{15}/\left(\pi_{10}+\pi_{15}+\pi_{16}\right)$, and $\pi_{16|CE1} = \pi_{16}/\left(\pi_{10}+\pi_{15}+\pi_{16}\right)$.

We now establish the bounds on the $PSDE$s.

PROPOSITION 1. *Let $\mathcal{Y}$ be the sample space of Y. Define $y^{\alpha}_{zws} = \inf\{y : Pr\left(Y_i^{\text{obs}} \le y | Z_i^{\text{obs}} = z, \quad W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right) \ge \alpha\}$ if $0 < \alpha < 1$, $y^{\alpha}_{zws} = \inf\{y : y \in \mathcal{Y}\}$ if $\alpha \le 0$, and $y^{\alpha}_{zws} = \sup\{y : y \in \mathcal{Y}\}$ if $\alpha \ge 1$. Let*

$$\begin{aligned}
E_{zws}\left[Y_i^{\text{obs}}\right] &= E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right]\\
E_{zws}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y^{\alpha}_{zws}\right] &= E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s, Y_i^{\text{obs}} \le y^{\alpha}_{zws}\right]\\
E_{zws}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y^{1-\alpha}_{zws}\right] &= E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s, Y_i^{\text{obs}} \ge y^{1-\alpha}_{zws}\right].
\end{aligned}$$

Then, under Assumption 1 through 6, the following bounds can be derived:

$$\begin{aligned}
&E_{Te0}\left[Y_i^{\text{obs}}\right] - E_{Ce0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{Ce0}^{1-\pi_{1,5,10|Ce0}}\right] \le PSDE(0,e,e) \le\\
&\qquad E_{Te0}\left[Y_i^{\text{obs}}\right] - E_{Ce0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{Ce0}^{\pi_{1,5,10|Ce0}}\right]
\end{aligned} \tag{12}$$

$$\begin{aligned}
&\max\left\{E_{Te1}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{Te1}^{\pi_{16|Te1}}\right]; E_{TE1}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{TE1}^{\pi_{16|TE1}}\right]\right\} - E_{Ce1}\left[Y_i^{\text{obs}}\right]\\
&\qquad \le PSDE(1,e,e) = PSDE(1,e,E) \le\\
&\min\left\{E_{Te1}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{Te1}^{1-\pi_{16|Te1}}\right]; E_{TE1}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{TE1}^{1-\pi_{16|TE1}}\right]\right\} - E_{Ce1}\left[Y_i^{\text{obs}}\right]
\end{aligned} \tag{13}$$

$$\begin{aligned}
&E_{TE0}\left[Y_i^{\text{obs}}\right] - \min\left\{E_{Ce0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{Ce0}^{1-\pi_{1|Ce0}}\right]; E_{CE0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{CE0}^{1-\pi_{1|CE0}}\right]\right\}\\
&\qquad \le PSDE(0,e,E) = PSDE(0,E,E) \le\\
&E_{TE0}\left[Y_i^{\text{obs}}\right] - \max\left\{E_{Ce0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{Ce0}^{\pi_{1|Ce0}}\right]; E_{CE0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{CE0}^{\pi_{1|CE0}}\right]\right\}
\end{aligned} \tag{14}$$

$$\begin{aligned}
&E_{TE1}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{TE1}^{\pi_{10,15,16|TE1}}\right] - E_{CE1}\left[Y_i^{\text{obs}}\right] \le PSDE(1,E,E) \le\\
&\qquad E_{TE1}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{TE1}^{1-\pi_{10,15,16|TE1}}\right] - E_{CE1}\left[Y_i^{\text{obs}}\right]
\end{aligned} \tag{15}$$

$$\begin{aligned}
&\min\nolimits_{\pi_5}\left\{E_{Te0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{Te0}^{\pi_{1,5|Te0}}\right] - E_{CE0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{CE0}^{1-\pi_{1,5|CE0}}\right]\right\}\\
&\qquad \le PSDE(0,E,e) \le\\
&\max\nolimits_{\pi_5}\left\{E_{Te0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \ge y_{Te0}^{1-\pi_{1,5|Te0}}\right] - E_{CE0}\left[Y_i^{\text{obs}} | Y_i^{\text{obs}} \le y_{CE0}^{\pi_{1,5|CE0}}\right]\right\}
\end{aligned} \tag{16}$$

and

$$\begin{aligned}\min\nolimits_{\pi_5}\left\{E_{Te1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{Te1}^{\pi_{15,16|Te1}}\right] - E_{CE1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{CE1}^{1-\pi_{15,16|CE1}}\right]\right\}& \\ \le PSDE(1,E,e) \le & \qquad (17)\\ \max\nolimits_{\pi_5}\left\{E_{Te1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{Te1}^{1-\pi_{15,16|Te1}}\right] - E_{CE1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{CE1}^{\pi_{15,16|CE1}}\right]\right\},&\end{aligned}$$

where $\pi_{1,5,10|Ce0} \equiv \sum_{g\in\{1,5,10\}}\pi_{g|Ce0} = \frac{1-P_{1|Te}}{1-P_{1|Ce}}$, $\pi_{16|Te1} = \frac{P_{1|Ce}}{P_{1|Te}}$, $\pi_{1|Ce0} = \frac{1-P_{1|TE}}{1-P_{1|Ce}}$, $\pi_{16|TE1} = \frac{P_{1|Ce}}{P_{1|TE}}$, $\pi_{1|CE0} = \frac{1-P_{1|TE}}{1-P_{1|CE}}$, $\pi_{10,15,16|TE1} \equiv \sum_{g\in\{10,15,16\}}\pi_{g|TE1} = \frac{P_{1|CE}}{P_{1|TE}}$, $\pi_{1,5|Te0} \equiv \pi_{1|Te0} + \pi_{5|Te0} = \frac{(1-P_{1|TE})+\pi_5}{1-P_{1|Te}}$, $\pi_{1,5|CE0} \equiv \pi_{1|CE0} + \pi_{5|CE0} = \frac{(1-P_{1|TE})+\pi_5}{1-P_{1|CE}}$, $\pi_{15,16|Te1} \equiv \pi_{15|Te1} + \pi_{16|Te1} = \frac{P_{1|CE}-(P_{1|TE}-P_{1|Te})+\pi_5}{P_{1|Te}}$, *and* $\pi_{15,16|CE1} \equiv \pi_{15|CE1} + \pi_{16|CE1} = \frac{P_{1|CE}-(P_{1|TE}-P_{1|Te})+\pi_5}{P_{1|CE}}$.

A sketch of the proof is given in the Appendix.

The sampling process allows us to identify the conditional distributions, $\widehat{P}_{s|z,w}$, the conditional expected values $E_{zws}\left[Y_i^{\text{obs}}\right]$, the quantile y_{zws}^{α}, and the conditional lower and upper trimmed means $E_{zws}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{zws}^{\alpha}\right]$ and $E_{zws}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{zws}^{1-\alpha}\right]$. Therefore finding estimators for the bounds defined in Proposition 1 is relatively straightforward. For instance, the following estimators can be used, where $\mathbb{1}(.)$ is the indicator function:

$$\widehat{P}_{1|zw} = \frac{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = 1)}{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)}$$

$$\widehat{P}_{0|zw} = \frac{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = 0)}{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)} = 1 - \widehat{P}_{1|zw}$$

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

$$\begin{aligned}&\widehat{E}_{zws}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{zws}^{\alpha}\right] = \\ &\frac{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)\mathbb{1}(Y_i^{\text{obs}} \le \hat{y}_{zws}^{\alpha})Y_i^{\text{obs}}}{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)} \equiv \overline{Y}_{zws}^{\le\alpha}\end{aligned}$$

$$\begin{aligned}&\widehat{E}_{zws}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{zws}^{1-\alpha}\right] = \\ &\frac{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)\mathbb{1}(Y_i^{\text{obs}} \ge \hat{y}_{zws}^{1-\alpha})Y_i^{\text{obs}}}{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)} \equiv \overline{Y}_{zws}^{\ge 1-\alpha}\end{aligned}$$

and

$$\hat{y}_{zws}^{\alpha} = \min\left\{y : \frac{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)\mathbb{1}(Y_i^{\text{obs}} \le \hat{y}_{zws}^{\alpha})}{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)} \ge \alpha\right\} \text{ if } 0 < \alpha < 1$$

$$\hat{y}_{zws}^{\alpha} = \min_i Y_i^{\text{obs}} \quad \text{if} \quad \alpha \le 0 \qquad \text{and} \qquad \hat{y}_{zws}^{\alpha} = \max_i Y_i^{\text{obs}} \quad \text{if} \quad \alpha \ge 1$$

$$\hat{y}_{zws}^{1-\alpha} = \min\left\{y : \frac{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)\mathbb{1}(Y_i^{\text{obs}} \le \hat{y}_{zws}^{1-\alpha})}{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(W_i^{\text{obs}} = w)\mathbb{1}(S_i^{\text{obs}} = s)} \ge 1-\alpha\right\} \text{ if } 0 < \alpha < 1$$

Table 6. *Full Hypothetical Data under Assumptions 3 through 6 (Upper Panel) and Corresponding PSDEs (Bottom Panel)*

						Expected Values			
G_i	$S_i(C,e)$	$S_i(C,E)$	$S_i(T,e)$	$S_i(T,E)$	π_g	$Y_i(C,e)$	$Y_i(C,E)$	$Y_i(T,e)$	$Y_i(T,E)$
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

			$G_i : \{S_i(C,e), S_i(C,E), S_i(T,e), S_i(T,E)\}$					
$PSDE_{G_{s,w,w'}}(s;w,w')$			$G_i = 1$	$G_i = 5$	$G_i = 10$	$G_i = 15$	$G_i = 16$	$PSDE(s;w,w')$
s	w	w'	$\{0,0,0,0\}$	$\{0,0,0,1\}$	$\{0,1,0,1\}$	$\{0,1,1,1\}$	$\{1,1,1,1\}$	Mean
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

$$\hat{y}^{1-\alpha}_{zws} = \min_i Y_i^{\text{obs}} \quad \text{if} \quad \alpha \geq 1 \qquad \text{and} \qquad \hat{y}^{1-\alpha}_{zws} = \max_i Y_i^{\text{obs}} \quad \text{if} \quad \alpha \leq 0.$$

6. An Illustrative Example

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

Table 6 shows the full (hypothetical) data and the corresponding $PSDE$ under Assumptions 3 through 6, given in the previous section. The sixth column shows the principal strata proportions: each principal stratum comprises a proportion of 16% of all patients, except principal stratum 11 = $\{S_i(C,e) = 0, S_i(C,E) = 0, S_i(T,e) = 1, S_i(T,E) = 1\}$, which comprises a proportion of 20% of all patients.

From Table 6, we can see that if everyone were assigned treatment and encouraged, 84% (= 16% + 16% + 20% + 16% + 16%) would take a rescue medication, whereas if everyone were assigned control and encouraged, 48% (= 16% + 16% + 16%) would take a rescue medication. Analogously, if everyone were assigned treatment and not encouraged, 52% (= 20% + 16% + 16%) would take a rescue medication, whereas if everyone were assigned control and not encouraged, only 16% would take a rescue medication. Thus, both the

Table 7. *Summary Statistics of Hypothetical Observed Data*

			Observed	Mean	
Z_i^{obs}	W_i^{obs}	S_i^{obs}	Proportions	Rescue Medication Usage	Disease Status
				(S_i^{obs})	(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

new drug treatment and the encouragement have a quite strong causal effect on rescue medication usage.

The total effect of the treatment Z on the primary outcome Y is 0.420 for the encouraged units, and 0.388 for the not-encouraged units:

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

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

$PSDE$s for patients who would use a rescue medication under both treatment arms range from 0.5 to 0.6, and are higher than $PSDE$s for patients who would not use a 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 getting the drug treatment and being encouraged to take a rescue medication; the second getting the drug treatment and being not encouraged to take a rescue medication; the third getting the placebo treatment and being encouraged to take a rescue medication; and the forth getting the placebo treatment and being not encouraged to take a rescue medication.

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}. This Table provides a simple mediation analysis based on standard methods which directly control for observed values of the posttreatment variable. Specifically, we can easily estimate net treatment effects of assignment (Z, W) adjusting for the observed value of the post-treatment variable S^{obs}:

$$E_{Te0}\left[Y_i^{\text{obs}}\right] - E_{Ce0}\left[Y_i^{\text{obs}}\right] = 0.171 \qquad E_{TE0}\left[Y_i^{\text{obs}}\right] - E_{Ce0}\left[Y_i^{\text{obs}}\right] = 0.038$$

$$E_{Te0}\left[Y_i^{\text{obs}}\right] - E_{CE0}\left[Y_i^{\text{obs}}\right] = 0.195 \qquad E_{TE0}\left[Y_i^{\text{obs}}\right] - E_{CE0}\left[Y_i^{\text{obs}}\right] = 0.062,$$

and

$$E_{Te1}\left[Y_i^{\text{obs}}\right] - E_{Ce1}\left[Y_i^{\text{obs}}\right] = 0.492 \qquad E_{TE1}\left[Y_i^{\text{obs}}\right] - E_{Ce1}\left[Y_i^{\text{obs}}\right] = 0.419$$

$$E_{Te1}\left[Y_i^{\text{obs}}\right] - E_{CE1}\left[Y_i^{\text{obs}}\right] = 0.492 \qquad E_{TE1}\left[Y_i^{\text{obs}}\right] - E_{CE1}\left[Y_i^{\text{obs}}\right] = 0.419.$$

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	

			Proportions of units with	
Estimand	Lower	Upper	$S_i(T,w') = S_i(C,w) = s$	
$PSDE(s;w,w')$	Bound	Bound	Lower Bound	Upper Bound
$PSDE(0;e,e)$	0.052	0.333	0.16	0.80
$PSDE(0;E,E) = PSDE(0;e,E)$	-0.239	0.200	0.16	
$PSDE(0;E,e)$	-0.439	0.951	0.16	0.48
$PSDE(1;e,e) = PSDE(1;e,E)$	0.025	0.700	0.16	
$PSDE(1;E,e)$	-0.529	1.000	0.16	0.48
$PSDE(1;E,E)$	0.208	0.700	0.16	0.80

Integrating out the observed encouragement variable W^{obs}, we have

$$E_{T0}\left[Y_i^{\text{obs}}\right] - E_{C0}\left[Y_i^{\text{obs}}\right] = 0.147, \quad \text{and} \quad E_{T1}\left[Y_i^{\text{obs}}\right] - E_{C1}\left[Y_i^{\text{obs}}\right] = 0.447.$$

The standard interpretation of these results would be that the new drug treatment has a positive effect on disease status, and this effect appears to be higher among patients who take a rescue medication against headache ($i : S_i^{\text{obs}} = 1$). Although these results do not clash with the real $PSDE$s, the differences between the average outcome among subjects who take a rescue medication when assigned new versus standard treatment are lower than the $PSDE$s for patients who would use a rescue medication under both treatment arms. In addition, we have to keep in mind that the net treatment effects lack of a causal interpretation, because they involve comparisons between sets of potential outcomes on different sets of units, the observed groups $OBS(z,w,s)$, $z = C,T$, $w = e,E$ and $s = 0,1$, which are mixtures of more 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 the bounds for π_5 are $\max\left\{0; \left(P_{1|TE} - P_{1|Te}\right) - \left(P_{1|CE} - P_{1|Ce}\right)\right\} \leq \pi_5 \leq \min\left\{\left(P_{1|TE} - P_{1|CE}\right); \left(P_{1|TE} - P_{1|Te}\right)\right\}$, that is, $0 \leq \pi_5 \leq 0.32$. Table 8 shows the bounds for the $PSDE$s, calculated using the results in Proposition 1. All our bounds contain the actual $PSDE$s and provide useful information about the direct effect of the treatment on the outcome. The estimated bounds for $PSDE(1,e,e) = PSDE(1;e,E)$, and $PSDE(1;E,E)$ cover only positive regions and are relatively narrow, suggesting that there exists a positive direct effect of the drug treatment on the disease for patients who would take a rescue medication under both treatment arms. Some uncertainty is on the sign of the $PSDE(1;E,e)$, although the positive region covered by the bounds is larger than the negative one. The drug treatment seems to have a positive although lower direct effect also for patients belonging to principal strata where $S_i(T,e) = S_i(C,e) = 0$. On the contrary, the data does not provide decisive evidence on the direct effects $PSDE(0,E,E) = PSDE(0,e,E)$ and $PSDE(0;E,e)$.

Table 9. *Standard Randomized Design: Full Data (Upper Panel), Observed Data (Bottom Panel on the Left) and Estimated Bounds (Bottom Panel on the Right)*

				Expected Values		
$\widetilde{G}_i$	$\widetilde{S}_i(C)$	$\widetilde{S}_i(T)$	$\widetilde{\pi}_i$	$\widetilde{Y}_i(C)$	$\widetilde{Y}_i(T)$	$PSDE(s)$
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

		Observed	Mean	
Z_i^{obs}	S_i^{obs}	Proportions	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

Estimand	Lower Bound	Upper Bound
$\widetilde{\pi}_1$	0.48	
$\widetilde{\pi}_3$	0.36	
$\widetilde{\pi}_4$	0.16	
$PSDE(0)$	0.051	0.333
$PSDE(1)$	0.019	0.700

In order to investigate the benefit of the presence versus the absence of an encouragement variable for the intermediate outcome, the estimated bounds in Table 8 are now compared with and contrasted to 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 ($S_i(T) \geq S_i(C)$), which implies that the principal stratum $\{i : S_i(C) = 1, S_i(T) = 0\}$ is empty. We also assume that in 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 $S_i(C) = S_i(C, e)$, $S_i(T) = S_i(T, e)$, and $Y_i(C) = Y_i(C, e)$, and $Y_i(C) = Y_i(T, e)$.

Table 9 shows 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)[3]. 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.

The major gain of our augmented randomized design with respect to the standard one can be observed by comparing the estimated bounds for $PSDE(1; E, E)$ with those for $PSDE(1)$: the first one is narrower and more informative. Specifically, the estimated bounds for $PSDE(1; E, E)$ ($[0.208; 0.700]$) suggest that there exists a positive and quite strong direct effect of the drug treatment on the disease for subjects who would take a rescue medication irrespective of the treatment under encouragement. The bounds for $PSDE(1)$ ($[0.019; 0.700]$) also show some evidence that the treatment has a positive

[3] Following Manski (1990), large sample bounds for $PSDE(s)$, $s = 0, 1$, can be easily derived. Specifically, let $\mathcal{Y}$ the sample space of Y. Define $y_{zs}^{\alpha} = \inf\{y : Pr\left(Y_i^{\text{obs}} \leq y | Z_i^{\text{obs}} = z, S_i^{\text{obs}} = s\right) \geq \alpha\}$ if $0 < \alpha < 1$, $y_{zs}^{\alpha} = \inf\{y : y \in \mathcal{Y}\}$ if $\alpha \leq 0$, and $y_{zs}^{\alpha} = \sup\{y : y \in \mathcal{Y}\}$ if $\alpha \geq 1$. Then, under SUTVA, randomization and the monotonicity assumption $S_i(T) \geq S_i(C)$, we have $E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = T, S_i^{\text{obs}} = 0] - E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = C, S_i^{\text{obs}} = 0, Y_i^{\text{obs}} \geq y_{C0}^{1-\widetilde{\pi}_{1|C0}}] \leq PSDE(0) \leq E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = T, S_i^{\text{obs}} = 0] - E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = C, S_i^{\text{obs}} = 0, Y_i^{\text{obs}} \leq y_{C0}^{\widetilde{\pi}_{1|T0}}]$, and $E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = T, S_i^{\text{obs}} = 1, Y_i^{\text{obs}} \leq y_{T1}^{\widetilde{\pi}_{4|T1}}] - E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = C, S_i^{\text{obs}} = 1] \leq PSDE(1) \leq E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = T, S_i^{\text{obs}} = 1, Y_i^{\text{obs}} \geq y_{T1}^{1-\widetilde{\pi}_{4|T1}}] - E[Y_i^{\text{obs}} | Z_i^{\text{obs}} = C, S_i^{\text{obs}} = 1]$, where $\widetilde{\pi}_{1|C0} = Pr\left(S_i(C) = S_i(T) = 0 | Z_i^{\text{obs}} = C, S_i^{\text{obs}} = 0\right)$ and $\widetilde{\pi}_{4|T1} = Pr\left(S_i(C) = S_i(T) = 1 | Z_i^{\text{obs}} = T, S_i^{\text{obs}} = 1\right)$.

direct effect for subjects who would take a 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 are not generally able to disentangle direct and indirect effects if a standard randomized experiment was conducted. Specifically, in a standard randomized experiment, information on $PSDE$s 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 $\widetilde{G} = 3$ in our example), give no direct information on the existence of $PSDE$s. However, such 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 behavior which might be useful in order to draw inference about $PSDE$s. For instance, in our setting, the encouragement would split the principal stratum $\{i : S_i(c) = 0, S_i(T) = 1\}$ into two principal strata: $11 = \{i : S_i(C, e) = 0, S_i(C, E) = 0, S_i(T, e) = 1, S_i(T, E) = 1\}$ and $15 = \{i : S_i(C, e) = 0, S_i(C, E) = 1, S_i(T, e) = 1, S_i(T, E) = 1\}$, and the last one provides information on $PSDE(1; E, E)$ and $PSDE(1; E, e)$.

7. Augmented Designs versus Standard Designs

In the previous section we empirically showed the potential benefits of our augmented randomized design with respect to a standard randomized design through an illustrative example. We now try to formally compare the two designs. A relatively ease way to face this issue is to analyze direct effects defined for the overall population, rather than for some specific subpopulations. This approach has the advantage that the causal estimand of interest − the overall direct effect − is the same for the two designs, but has the drawback of involving the concept of a priori potential outcomes, as explained in Section 2.

Therefore, in order to formally define overall direct effects we need to extend the theoretical framework underlying our augmented randomized design and the standard randomized design to allow for a priori counterfactual outcomes. First, we focus on a standard randomized design, where the intermediate variable is not randomly encouraged. Under SUTVA, which can be now formalized as

Assumption 7. (Stable Unit Treatment Value Assumption with a priori Counterfactuals).

1. *If* $Z_i = Z_i'$*, then* $S_i(\boldsymbol{Z}) = S_i(\boldsymbol{Z}')$
2. *If* $Z_i = Z_i'$ *and* $S_i = S_i'$*, then* $Y_i(\boldsymbol{Z}, \boldsymbol{S}) = Y_i(\boldsymbol{Z}', \boldsymbol{S}')$,

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 Average Direct Effect, ADE, can be defined as 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:

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

Random assignment of the treatment, Z, implies

Assumption 8. (Randomization of the Treatment with a priori Counterfactuals). *For all i,*

$$\Big(S_i(C), S_i(T), Y_i(C, S_i(C) = 0), Y_i(T, S_i(T) = 0), Y_i(C, S_i(C) = 1), Y_i(T, S_i(T) = 1)\Big) \perp\!\!\!\perp Z_i$$

Assumptions 7 and 8 alone do not lead to point identify ADE; additional strong assumptions, which allow one to extrapolate the behavior of the (a priori counterfactual) potential outcomes, would be required. However, large sample bounds for ADE can be derived.

Proposition 2. *Suppose that, $Y_i\left(z, S_i(z) = s\right)$ is bounded within some known interval $[L_{zs}, U_{zs}]$, where $-\infty < L_{zs} \leq U_{zs} < +\infty$, $z = T, C$ and $s = 0,1$. Then, under Assumptions 7 and 8, the following bounds can be derived:*

$$\begin{aligned}
&\Big(E\left[Y_i^{\text{obs}}|Z_i^{\text{obs}} = T, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s|Z_i^{\text{obs}} = T\right) + L_{Ts} \cdot Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}} = T\right)\Big)\\
&\quad -\Big(E\left[Y_i^{\text{obs}}|Z_i^{\text{obs}} = C, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s|Z_i^{\text{obs}} = C\right) + U_{Cs} \cdot Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}} = C\right)\Big)\\
&\qquad\qquad\qquad\qquad \leq ADE(s) \leq \qquad\qquad\qquad\qquad (19)\\
&\Big(E\left[Y_i^{\text{obs}}|Z_i^{\text{obs}} = T, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s|Z_i^{\text{obs}} = T\right) + U_{Ts} \cdot Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}} = T\right)\Big)\\
&\quad -\Big(E\left[Y_i^{\text{obs}}|Z_i^{\text{obs}} = C, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s|Z_i^{\text{obs}} = C\right) + L_{Cs} \cdot Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}} = C\right)\Big).
\end{aligned}$$

The width of the bounds in Equation (19) is

$$width(s) = \left(U_{Cs} - L_{Cs}\right) Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}} = C\right) + \left(U_{Ts} - L_{Ts}\right) Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}} = T\right),$$

which depends on both the selection probabilities $Pr\left(S_i^{\text{obs}} = 1 - s|Z_i^{\text{obs}}\right)$, as well as the width of the intervals $[L_{zs}, U_{zs}]$, $z = C, T$[4].

Bounds in Proposition 2 can be estimated using the sample analogs of the parameters. Formally, we have

$$\widehat{P}_{1|z} = \frac{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(S_i^{\text{obs}} = 1)}{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)} \quad \text{and} \quad \widehat{P}_{0|z} = \frac{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(S_i^{\text{obs}} = 0)}{\sum_i \mathbb{1}(Z_i^{\text{obs}} = z)} = 1 - \widehat{P}_{1|z}$$

and

$$\widehat{E}_{zs}\left[Y_i^{\text{obs}}\right] = \frac{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(S_i^{\text{obs}} = s)Y_i^{\text{obs}}}{\sum_{i=1}^n \mathbb{1}(Z_i^{\text{obs}} = z)\mathbb{1}(S_i^{\text{obs}} = s)} \equiv \overline{Y}_{zws}.$$

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

[4] When the primary outcome is a logical yes/no indicator, taking the value one or zero, $L_{zs} = 0$ and $U_{zs} = 1$ for each $z = C, T$ and $s = 0,1$. In such a case, the expected value of a one/zero indicator is the probability that the indicator equals one, so that the bounds in Equation (19) equal to the bounds on the Controlled Direct Effect, derived by Cai et al. (2008), using the symbolic Balke & Pearl (1997) linear programming method.

Assumption 9. (Stable Unit Treatment Value Assumption with a priori Counterfactuals).

1. *If* $Z_i = Z'_i$ *and* $W_i = W'_i$, *then* $S_i(\boldsymbol{Z}, \boldsymbol{W}) = S_i(\boldsymbol{Z}', \boldsymbol{W}')$
2. *If* $Z_i = Z'_i$, $W_i = W'_i$ *and* $S_i = S'_i$, *then* $Y_i(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{W})) = Y_i(\boldsymbol{Z}', \boldsymbol{W}', \boldsymbol{S}'(\boldsymbol{Z}', \boldsymbol{W}'))$

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

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

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,*

$$\Big(S_i(C,e), S_i(C,E), S_i(T,e), S_i(T,E), Y_i(C,e,S_i(C,e)=0), Y_i(C,E,S_i(C,E)=0),$$
$$Y_i(T,e,S_i(T,e)=0), Y_i(T,E,S_i(T,E)=0), Y_i(C,e,S_i(C,e)=1), Y_i(C,E,S_i(C,E)=1),$$
$$Y_i(T,e,S_i(T,e)=1), Y_i(T,E,S_i(T,E)=1)\Big) \perp\!\!\!\perp (Z_i, W_i)$$

Bounds for $ADE(s; w)$ are now derived taking into account the augmented nature of the design. Note that, in order to establish bounds for $ADE(s; w)$ no assumption is required in addition to Assumptions 9 and 10; in particular we do not need to invoke any assumption on the role of the encouragement. However, the Exclusion Restriction Assumption 3 and the Monotonicity Assumption 5(i) or 5(ii) (along with Assumption 4) characterize our augmented design, therefore we maintain these assumptions, by extending them to involve a priori counterfactuals, in order to properly interpret our results and appreciate the benefits of the presence versus the absence of the encouragement. Specifically, Assumption 3 justifies our focus on the causal estimand $ADE(s; w)$, defined as the mean difference between potential outcomes while holding both the intermediate variable and the encouragement fixed at some predetermined level: $S = s$ and $W = w$. Actually if Assumption 3 holds, $Y_i(z, e, S_i(z, e) = s)$ and $Y_i(z, E, S_i(z, E) = s)$, for each $z = C, T$, and $s = 0, 1$, have the same distribution, 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 other words, define $ADE(s; w, w') = E[Y_i(T, w', s)] - E[Y_i(C, w, s)]$, $w, w' \in \{e, E\}$. Assumption 3 implies that $ADE(s; w, w') = ADE(s; w', w) = ADE(s; w, w) \equiv ADE(s; w)$, so that each mean difference $ADE(s; w', w)$ is interpretable as direct causal effect of treatment Z on outcome Y[5]. 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(i) or 5(ii) allows us to justify the focus on either $ADE(0; w)$ or $ADE(1; w)$ according to role of the encouragement.

[5] Without the Exclusion Restriction Assumption a comparison between the potential outcomes $Y_i(T, w', s)$ and $Y_i(C, w, s)$, $w, w' \in \{e, E\}$, while holding the intermediate variable S fixed at s could not be interpreted as direct effect of the treatment Z, because it would depend on the encouragement.

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

$$
\begin{aligned}
&\Big(E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w\right) \\
&\quad + L_{Tws} \cdot Pr\left(S_i^{\text{obs}} = 1 - s | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w\right)\Big) \\
&\quad - \Big(E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w\right) \\
&\quad + U_{Cws} \cdot Pr\left(S_i^{\text{obs}} = 1 - s | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w\right)\Big) \\
&\qquad\qquad\qquad \leq ADE(s; w) \leq \qquad\qquad\qquad (21) \\
&\Big(E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w\right) \\
&\quad + U_{Tws} \cdot Pr\left(S_i^{\text{obs}} = 1 - s | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w\right)\Big) \\
&\quad - \Big(E\left[Y_i^{\text{obs}} | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w, S_i^{\text{obs}} = s\right] \cdot Pr\left(S_i^{\text{obs}} = s | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w\right) \\
&\quad + L_{Cws} \cdot Pr\left(S_i^{\text{obs}} = 1 - s | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w\right)\Big)
\end{aligned}
$$

for $s = 0, 1$.

The width of the bound in Equation (21) is

$$
\begin{aligned}
width(s; w) = {} & (U_{Tws} - L_{Tws})\, Pr\left(S_i^{\text{obs}} = 1 - s | Z_i^{\text{obs}} = T, W_i^{\text{obs}} = w\right) \\
& + (U_{Cws} - L_{Cws})\, Pr\left(S_i^{\text{obs}} = 1 - s | Z_i^{\text{obs}} = C, W_i^{\text{obs}} = w\right).
\end{aligned}
$$

As before, bounds in Proposition 3 can be estimated using the sample analogs of the parameters (see section 5).

As we could 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 $Pr\left(S_i^{\text{obs}} = 1 - s^* | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w^*\right) \leq Pr\left(S_i^{\text{obs}} = 1 - s^* | Z_i^{\text{obs}} = z\right)$, $z = C, T$, then $width(s^*; w^*) \leq width(s^*)$. This result depends on the study design in the sense that the relationship $Pr\left(S_i^{\text{obs}} = 1 - s^* | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = w^*\right) \leq Pr\left(S_i^{\text{obs}} = 1 - s^* | Z_i^{\text{obs}} = z\right)$, $z = C, T$, holds if the encouragement status w^* boosts units to exhibit a value of the intermediate variable S equal to s^*.

To fix the ideas, suppose that a standard randomized experiment and an augmented randomized study are performed in order to estimate the direct effect of the treatment on the outcome if, possibly contrary to fact, the intermediate variable S was set to $s^* = 1$. In line with the objective of the study, suppose that units assigned to the 'active' encouragement ($W_i = E$) are boosted to exhibit a positive value of the intermediate outcome S, so that Assumption 5(i) holds. Focus will be on $ADE(1)$ and $ADE(1; E)$ [6]. Throughout, we assume that $U_{zws^*} = U_{zs^*}$ and $L_{zw^*s^*} = L_{zs^*}$ for $z = C, T$, $w^* = E$ and $s^* = 1$.

[6] Recall that our augmented design is based on Assumptions 3, which implies that $ADE(1; E) = ADE(1; e)$.

In a standard randomized experiment, the bounds for $ADE(1)$ are more informative the higher the proportion of units who would always exhibit a positive value of the intermediate variable regardless of treatment assignment (proportion of units belonging to $\{i : S_i(C) = 1, S_i(T) = 1\}$). With this respect, we can reasonably expect that in the augmented design where the encouragement boosts units to exhibit a positive value of the mediating variable, the proportion of units who 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 who would exhibit a positive value of the mediator S when not encouraged under either the C or the T treatment. Therefore, if the encouragement is able to 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 who would exhibit a positive mediator value regardless treatment assignment when encouraged, then its presence improves the partial estimates of ADE, by tightening the bounds. In other words, in order that the relationship $Pr\left(S_i^{\text{obs}} = 0 | Z_i^{\text{obs}} = z, W_i^{\text{obs}} = E\right) \leq Pr\left(S_i^{\text{obs}} = 0 | Z_i^{\text{obs}} = z\right)$, $z = C, T$, holds, the encouragement must affect the intermediate variable, which is postulated in Assumption 4.

8. Concluding Remarks

In this paper we study identification and estimation of causal mediation effects. We introduce new augmented designs, 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 direct and indirect effects. There are two key features of our framework: we adopt a principal stratification approach, extending it to include an encouragement variable on the mediator, and we mainly focus on principal strata effects, avoiding to involve a priori counterfactual outcomes.

In order to achieve identification of $PSDE$s, assumptions characterizing the encouragement variable are investigated. Specifically, we provide a set of assumptions leading to partially identify the causal estimands of interest for the case in which the treatment and the encouragement assignment are random and the intermediate variable is binary. Our partial identification results for the $PSDE$s rely on a (stochastic) exclusion restriction – which rules out direct effects of the encouragement on the primary outcome, and two monotonicity assumptions for the effect on the mediator of the encouragement and the treatment, respectively.

We empirically show that our bounds on the $PSDE$s are narrower and more informative than those we would derive in a standard randomized experiment. The benefits of the presence with respect to the absence of an encouragement for the intermediate outcome are also formally shown, focussing on an average direct effect for the entire population. As we expect, these results strongly depends 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 will be required. Future research will focus on addressing this issue. If pre-treatment variables are available, auxiliary information from them can be used to enhance efficiency of estimation and to sharp the bounds. Indeed, although covariates do not enter the treatment/encouragment 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 through prediction of the missing intermediate potential outcomes. Further sharpening of the bounds will be pursued exploiting (semi-)parametric models within a Bayesian framework. Our preference is for Bayesian methods because we believe that a Bayesian model-based approach is the most direct and flexible mode of inference for causal effects: Bayesian analysis is formally clear about the role played by the treatment assignment mechanism and the complications arising when drawing inference about direct and indirect effects.

Another direction for extensions is to use our augmented design as a template for the analysis of direct and indirect causal effects in observational studies. Carefully designed augmented observational studies, where unconfoundedness of both the treatment and the encouragement can be reasonably assumed, might be a powerful tool for mediation analysis in many setting, where randomized experiments cannot be conducted.

Finally, focus will be on the planning phase of our augmented designs, in order to develop 'optimal' augmented designs, which allow one to achieve a required precision for estimating $PSDE$s, and minimize the study's cost (e.g., Frangakis & Baker (2001)).

References

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

Baker, S. G, Frangakis, C. E, & Lindeman, K. S. (2007). Estimating efficacy in a proposed randomized trial with initial and later noncompliance. *J. R. Statist. Soc. C* **56**, 211-221.

Barnard, J., Frangakis, C. E., Hill, J. L. & 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. Amer. Statist. Assoc.* **98**, 299–323.

Buyse, M. & Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. & Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–68.

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

Cheng, J., Small, D. S, Tan, Z. & Ten Have, T. R (2008). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika* **96**, 19–36.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics* **13**, 261–281.

Dawid, A. P. (2000). Causal inference wihout countefactuals (with discussion). *J. Amer. Statist. Assoc.* **95**, 407–448.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *Int. Statist. Rev.* **2**, 161–189.

Didelez, V., Dawid, A. P. & Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proc. 22nd A. Conf. Uncertainty in Artifical Intelligence*, 138–146. Arlington: Association for Uncertainty in Artificial Intelligence Press.

Flores, C. A. & Flores-Lagunes, A. (2009a). Identification and estimation of causal mechanisms and net effects of a treatment. *IZA Discussion Paper* No. 4237.

Flores, C. A. & Flores-Lagunes, A. (2009b). Nonparametric partial and point identification of net or direct causal effects. *American Economic Association, Annual Meeting Paper 2009.*

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

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

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

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

Freedman, L. S., Graubard, B. I. & Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statist. Med.* **11**, 167–178.

Gallop, R., Small, D., Lin, J. Y., Elliot M. R., Joffe, M. M. & Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statist. Med.* **28**(7), 1108–1130.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Statist. Soc. B* **69**, Part 2, 199–215.

Gilbert, P. B., & Hudgens, M. G. (2008). Evaluating Candidate Principal Surrogate Endpoints. *Biometrics* **64**, 1146–1154.

Joffe, M. M. & Greene T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65**(2), 530–539.

Joffe, M. M., Small, D. & Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statist. Sci.* **22**, 74–97.

Haavelmo, T. (1943). Statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.

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

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

Lee, D. S. (2009) Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev. Econ. Stud.* **76**(3), 1071–1102.

Lin, D. Y., Fleming, T. R. & De Gruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statist. Med.* **16**, 1515–1527.

Lynch, K. G., Cary, M., Gallop, R. & Ten Have, T. R. (2008). Causal mediation analysis for randomized trial. *Health. Serv. Outcome Res. Meth.* **8**, 57–76.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *Amer. Econ. Rev.* **80**(2), 319–323.

Manski, C. F. (2003). *Partial identification of probabilities distributions.* New York and Heidelberg: Springer.

Mattei, A. & Mealli, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63**, 437–446.

Mealli, F. & 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 Adams et al., *J. Econometrics* **112**, 79 – 87.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on principles, Section 9. Translated in *Statist. Sci.* **5**, 465–480, 1990.

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 & D. Koller), 411–420. Morgan Kaufman, San Francisco, CA.

Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika* **82**, 669–688.

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

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statist. Med.* **8**, 431–440.

Qin, L., Gilbert, P. B., Follmann, D. & Li, D. (2008). Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the cox model. *Ann. Appl. Statist.* **2**, 386–407.

Reiersol, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* **9**, 1–24.

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), 70–81. Oxford: Oxford University Press.

Robins, J. M. & Greenland, S. (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. *J. R. Statist. Soc. A* **147**, 656–66.

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. Amer. Statist. Assoc.* **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 and reply). *Scand. J. Statist.* **31**, 161–170; 196–198.

Sjölander, A. Humphreys, K., Vansteelandt, S., Bellocco, R. & 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.

Zhang, J. L. & 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. & Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification *J. Amer. Statist. Assoc.* **104**(485), 166–176.

VanderWeele, T. L. (2008) Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.* **78**(17), 2957–2962.

9. Appendix

We briefly sketch the proof of Proposition 1. Consider the principal stratum direct effect for units who exhibit a zero value of the intermediate outcome under both treatment arms if not encouraged: $PSDE(0,e,e)$. These units may belong to either stratum $1 = \{i : S_i(C,e) = 0, S_i(C,E) = 0, S_i(T,e) = 0, S_i(T,E) = 0\}$, or stratum $5 = \{i : S_i(C,e) = 0, S_i(C,E) = 0, S_i(T,e) = 0, S_i(T,E) = 1\}$, or stratum $10 = \{i : S_i(C,e) = 0, S_i(C,E) = 1, S_i(T,e) = 0, S_i(T,E) = 1\}$. The $OBS(T,e,0)$ group only includes units belonging to one of these three principal strata, so in large sample $E\left[Y_i(T,e)|G_i \in \{1,5,10\}\right] = E_{Te0}\left[Y_i^{\text{obs}}\right]$. The $OBS(C,e,0)$ 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/her membership to the $OBS(C,e,0)$ group is $\sum_{g\in\{1,5,10\}} \pi_{g|Ce0} = \frac{1-P_{Te0}}{1-P_{Ce1}} \equiv \pi_{1,5,10|Ce0}$. Therefore, $E_{Ce0}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{Ce0}^{\pi_{1,5,10|Ce0}}\right] \le E\left[Y_i(C,e)|G_i \in \{1,5,10\}\right] \le E_{Ce0}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{Ce0}^{1-\pi_{1,5,10|Ce0}}\right]$. As a result, the large sample bounds for $PSDE(0,e,e)$ in (12) follow immediately. A similar proof leads to derive the large sample bounds for $PSDE(1,E,E)$ in equation (15), and the following large sample bounds for $PSDE(1,e,e)$, $PSDE(0,E,e)$ $PSDE(1,E,e)$, and $PSDE(0,E,E)$:

$$E_{Te1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{Te1}^{\pi_{16|Te1}}\right] - E_{Ce1}\left[Y_i^{\text{obs}}\right] \le PSDE(1,e,e) \le \\ E_{Te1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{Te1}^{1-\pi_{16|Te1}}\right] - E_{Ce1}\left[Y_i^{\text{obs}}\right] \tag{A1}$$

$$E_{TE1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{TE1}^{\pi_{16|TE1}}\right] - E_{Ce1}\left[Y_i^{\text{obs}}\right] \le PSDE(1,e,E) \le \\ E_{TE1}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{TE1}^{1-\pi_{16|TE1}}\right] - E_{Ce1}\left[Y_i^{\text{obs}}\right] \tag{A2}$$

$$E_{TE0}\left[Y_i^{\text{obs}}\right] - E_{Ce0}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{Ce0}^{1-\pi_{1|Ce0}}\right] \le PSDE(0,e,E) \le \\ E_{TE0}\left[Y_i^{\text{obs}}\right] - E_{Ce0}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{Ce0}^{\pi_{1|Ce0}}\right] \tag{A3}$$

$$E_{TE0}\left[Y_i^{\text{obs}}\right] - E_{CE0}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \ge y_{CE0}^{1-\pi_{1|CE0}}\right] \le PSDE(0,E,E) \le \\ E_{TE0}\left[Y_i^{\text{obs}}\right] - E_{CE0}\left[Y_i^{\text{obs}}|Y_i^{\text{obs}} \le y_{CE0}^{\pi_{1|CE0}}\right] \tag{A4}$$

Assumptions 5 and 6 imply that $PSDE(0,e,E) = PSDE_1(0,e,E)$, $PSDE(0,E,E) = PSDE_1(0,E,E)$, $PSDE(1,e,e) = PSDE_{16}(1,e,e)$, and $PSDE(1,e,E) = PSDE_{16}(1,e,E)$. At the same time, Assumption 3 implies that $PSDE_{16}(1,e,e) = PSDE_{16}(1,e,E)$ and $PSDE_1(0,e,E) = PSDE_1(0,e,E)$, therefore the bounds in equations (13) and (14) can be immediately derived by combining the bounds in equations (A1) and (A2), and in equations (A3) and (A4), respectively. Finally, the proof of equations (16) and (17) is analogous, but we have an additional source of variation because the proportions of units for which $PSDE(0,E,e)$ and $PSDE(1,E,e)$ are defined depend on π_5, which is unknown and can be only partially identified (see equation (11)).