Das, Tirthatanmoy; Polachek, Solomon

Working Paper

The Econometrics of Antidotal Variables

IZA Discussion Papers, No. 15558

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Das, Tirthatanmoy; Polachek, Solomon (2022) : The Econometrics of Antidotal Variables, IZA Discussion Papers, No. 15558, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/265779

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15558

The Econometrics of Antidotal Variables

Tirthatanmoy Das
Solomon W. Polachek

SEPTEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15558

The Econometrics of Antidotal Variables

Tirthatanmoy Das
Indian Institute of Management Bangalore and IZA

Solomon W. Polachek
State University of New York at Binghamton and IZA

SEPTEMBER 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Econometrics of Antidotal Variables*

Some interventions or population attributes negate the effects of a treatment. This paper shows that incorporating these, what we call antidotal variables (AV), into a causal treatment effects analysis can with one cross-sectional regression identify the true causal effect, in addition to possible biases from selectivity and SUTVA violations. Whereas we apply the AV technique to analyze the California Paid Family Leave program, it has applications beyond this example.

JEL Classification: C18, C36, I38, J18, J38
Keywords: antidotal variables, causality, CPFL

Corresponding author:
Solomon W. Polachek
Department of Economics
State University of New York at Binghamton
Binghamton, NY 13902
USA
E-mail: polachek@binghamton.edu

* We thank Joshua Angrist, Robert Basmann, Alfonso Flores-Lagunes, Ivan Korolev, and David Slichter and participants in Camp Econometrics, 2021 for extremely valuable comments on an earlier version Das and Polachek (2019). We again thank Alfonso Flores-Lagunes, Ivan Korolev, David Slichter, as well as Marlon Tracey for insightful comments on this version. Of course, any errors are our own responsibility.

1. Introduction

The treatment effects literature deals with estimating the causal impact of an intervention. Ideally, one randomly divides a population into treatment and control groups. The difference in the average outcomes between the two constitutes the treatment effect. Whereas random assignment has emerged as the gold standard, evaluating public policies via random assignment is often unrealistic and frequently impossible, especially in observational settings. Also, in reality, especially for policy related issues, there are complications. First, treatments are not necessarily administered randomly. Nonrandom assignment typically leads to a selectivity bias. Second, simultaneous confounding shocks (as well as possible multiple concurrent treatments) are often assumed away. Third, random assignment cannot ensure one satisfies the stable unit treatment value assumption (SUTVA), namely that there are no spillovers such that the untreated group remains unaffected by the assignment of treatment to the treated.

Standard instrumental variables (IV) methods are not always a panacea. First, IV approaches can identify a treatment effect when SUTVA is satisfied, but in actuality many applications violate the SUTVA assumption. Decreasing disease through inoculations indirectly protects the uninoculated because a disease spreads less easily in a sufficiently large vaccinated population; recipients of advanced education might transmit knowledge to others; and literally any policy intervention can run the risk of spillover externality effects. Second, the exclusion restriction in the IV approach requires the instrument to have a causal effect only on the treatment, but not on the outcome directly. As such, there can be no confounding effects of the instrument on the outcome, thereby implying the instrument must be uncorrelated with the error term. But for this reason, the validity of an exclusion restriction usually remains untested because the condition involves an unobservable error.[1] Third, potential concomitant treatments are typically excluded since the instrument is assumed to affect one and only one treatment. Fourth, IV imposes a monotonicity requirement so that there are no defiers. Fifth, IV typically assumes no sorting on gains (essential heterogeneity), meaning that those opting into the treatment are the ones who benefit the most from treatment. Again, these latter three assumptions cannot be tested, and thus must be assumed. As such, any method that relies solely on the random assignments of treatment (e.g., DID IV, PSM, RDD) could yield biased estimates if these assumptions are not met.

We examine an alternative method to identify a treatment effect. While also not perfect, it has the advantage of identifying the treatment effect, the selection bias, as well as possible SUTVA spillovers. The approach requires an intercession, defined by what can be called an antidotal variable (AV), if received, that nullifies the effect of the original treatment. In addition, the approach encompasses a validity test, namely whether the antidotal variable is mean independent of the error. The approach removes potential confounding effects caused by other concurrent policies that are usually assumed away in most DID studies. As such, the treatment effect can be identified even in the presence another concomitant treatment. An attractive feature of the method is that neither the treatment nor the antidotal variable need be random, though in this case the antidotal variable needs be mean independent of the error, independent of the treatment, and mean independent of the parameters in the estimating

[1] One exception is Madestam et al. (2013).

equation. Further, the approach can be used in a single cross-section either in an experimental or observational setting. Finally, the estimates can be bounded when the antidotal variable does not completely nullify the treatment effect. To our knowledge, this is the first formulation where the treatment effect, the SUTVA bias and the selection bias are identified within a unified framework.

The logic is as follows: There are now four groups instead of two. First, the original treatment group comprises one group of the treated and a second group of those within the treated group that obtain the antidote. Because the antidote negates the effect of the treatment, the difference in outcome between these two groups constitutes the treatment effect.[2] Second, the original control group now makes up one group that gets the antidote, and the second that does not. The difference in outcomes between those in this control group that get the antidote, and those that do not, yields the SUTVA bias because neither receives the treatment, and the antidote nullifies any treatment spillover effect from the treated to the untreated. But now, one can divide those getting the antidote into two groups: those that originally got the treatment (and also the antidote) and those that did not get the treatment (but got the antidote). The difference in outcomes for these is the selectivity bias, because neither group has a treatment effect, but one group (the one that originally got the treatment) is different from the group that originally did not get the treatment.

As surveyed in Forastiere (2021) a recent literature is emerging on estimating treatment effects in the presence of SUTVA spillover externalities from the treated to the untreated, as earlier described by Cox (1958) and Rubin (1980). Among these, Van der Laan (2014), Aronow and Samii (2017) assume SUTVA interference externalities occur only for immediate neighbors, but diminish with distance based on a hypothetical assignment rule. Related, Tchetgen and VanderWeele (2012) propose inverse probability-weighted estimators based on group-level propensity scores. Liu et al. (2016) extend this by utilizing a joint propensity score based both on individual binary-treatment and neighborhood multivalued treatment propensity scores. These approaches do not use an antidotal variable, but instead assume an arbitrary function defining how spillovers relate to control group characteristics.

Antidotal variables differ from traditional instrumental variables. A traditional instrument is related to the *treatment*, but unrelated to the error term and unrelated to the treatment effect. An antidotal variable is unrelated to the treatment, unrelated to the error term, but related to the treatment *effect*. Also, the approach differs from standard DID methods. Whereas the AV approach computes differences, they emanate from cross-sectional dissimilarities rather than from comparing changes in one group compared to another over time. We present an intuitive hypothetical example of this related to loud music in the next section, but given the available data, later in the paper we apply the approach to the California Paid Family Leave (CPFL) program.

[2] If the antidotal variable and the treatment variable are mean independent of the parameters, this is the same average treatment effect (ATE) estimated in the treatment effects literature. It is the average treatment effect of the treated (ATT) if the treatment is not mean independent of the parameters, because the treated sample is innately different than the untreated sample.

The CPFL program became effective in 2004. California's paid family leave allows parents to take up to six weeks paid leave to take care of young children. A common analysis entails comparing utilization rates in California before and after passage of the law relative to utilization rates in other comparable states, also in years before and after California implemented the law. However, this DID approach has at least two potential biases. One is SUTVA, the fact that those in neighboring states might emigrate to California, and thus affect leave taking in the control states. Another, is selectivity which can potentially change with time, especially if there are confounding effects. The usual assumption is that the control states and California have a constant selectivity component before and after the introduction of the new law. However, in fact, many other policies may be implemented simultaneously with CPFL, either in California, or in other states, or in both. In such cases, the selectivity component will be different before and after the implementation of the law. For example, in 2004 California enacted the Private Attorneys General Act (PAGA) designed to increase enforcement of the Labor Code, a law useful for low wage workers to enforce their labor rights by filing lawsuits on behalf of a group or "class" of employees who have suffered Labor Code violations.

The antidotal variable method deals with both these SUTVA and selectivity issues. CPFL applies to the whole California population, but in reality, young childbearing aged women are the prime beneficiaries as they are the group predominantly taking advantage of the program, rather than older women with no young children (Rossin-slater, Ruhm, and Waldfogel, 2011; Baum and Ruhm, 2014). As such, the variable that identifies those aged 45-55 can serve as an antidotal variable which may then be used to identify the average treatment effect along with the selectivity and SUTVA biases. We apply this antidotal variable approach (in Section 6) using two measures of leave utilization. We find a minimum of 50% increase in the probability of leave taking and an 80% increase in leave taking hours. This is independent of a selectivity bias indicating that Californians are in general about 40% less likely to take leave. We find no SUTVA bias, which is reasonable when we compare California to the rest of the country. Similarly, we find equivalent effects when comparing California to its three neighboring states, but now in this case we detect a negative SUTVA, meaning those in neighboring states reduce leave taking after CPFL was instituted. One reason may be because some of those prone to taking a leave emigrated to California.

The paper progresses as follows. Section 2 presents a hypothetical example. Section 3 explains the mechanics of the approach. Section 4 discusses identification under various assumptions. Section 5 provides a simulation showing the effectiveness of the approach. Section 6 applies the approach to analyze the California Paid Family Leave program. Finally, Section 7 concludes.

2. A Hypothetical Example

To lay a foundation for the antidotal variable approach, consider an intuitive hypothetical example.

Imagine one wants to determine the impact of loud music on mental stress.[3] Individuals with *high-medium* stress, the treatment group (D=1), listen to loud music (the treatment) to reduce their stress. Others with *low-medium* stress, the control group (D=0), may not need to, but may find the resulting loud boombox music (to them noise) *highly stressful*. The difference in stress levels between these two groups, namely those who listen to the boombox music and those who do not intend to (*low* stress minus *high* stress), provides a biased estimate of the treatment effect for two reasons. The first stems from the sample selection process. Participants in the treatment group, those who listen to the music, have themselves selected into the treatment group, meaning their pre-treatment average stress level (*high-medium*) differs from the control group's (*low-medium*). The second results from violation of SUTVA. Control group members forced to listen to the loud music, but do not wish to do so, are adversely affected, thereby increasing their stress level from *low-medium* to *high* stress. These are illustrated in the first column of Figure 1.

Figure 1 The Impact of Loud Boombox Music on Stress*

	Antidote: Earplugs	
Treatment: Loud Music	**W=1 (no earplug) No Antidote**	**W=0 (earplug) Antidote**
D=0 No Treatment (no loud music)	• *Before*: low-medium stress • No direct loud music (but overhears loud music from others), no earplugs • *After*: high stress • (Subsample: n_2)	• *Before*: low-medium stress • No loud music, earplugs • *After*: low-medium stress • (Subsample: n_4)
D=1 Treatment (loud music)	• *Before:* High-medium stress • Direct loud music, no earplug • *After:* Low stress • (Subsample: n_1)	• *Before:* High-medium stress • Direct loud music, earplugs • *After:* High-medium stress • (Subsample: n_3)

* Subsample classifications n_1, n_2, n_3, and n_4 are defined in the text below.

Now suppose earplugs were distributed randomly to a subsample of individuals, and everyone receiving them uses them.[4] If earplugs negate the effect of the loud music, then those wearing earplugs are not affected by the loud music. As such, we consider earplugs to be an antidote to the loud music. We denote this antidotal variable as W=0 when one has earplugs, and W=1 when one does not have earplugs.[5] This characterization results in four groups: First is the *medium-high* stressed group who listen to loud music to destress (subsample n_1 in Figure 1). They now have *low* stress. Second are *low-medium* stress individuals who do not intentionally listen, but now must endure the ambient loud music, what they consider noise (subsample n_2 in Figure 1). Their stress level is now *high*. Third are *high-medium* stressed

[3] Remember the overpowering boomboxes prevalent in the 1980s. This isn't a farfetched example given the recent interest in noise pollution (e.g., Fried and Cohen, 2021).

[4] We deal with noncompliance later in the paper.

[5] This counterintuitive notation of setting "no antidote" to one and "receiving the antidote" to zero will be explained in the next section when we formally define an antidotal variable.

individuals who normally would have listened to the loud music, but are now wearing earplugs (subsample n_3 in Figure 1). Their stress level remains the same (*high-medium*). Fourth are those with *low-medium* stress who do not listen to the loud music, but are wearing earplugs anyway (subsample n_4 in Figure 1) to protect themselves from overhearing ambient loud music. They remain *low-medium* stressed.

Now consider differences between the subsamples based on the antidotal variable approach. The difference in average stress levels between groups n_1 and n_3 (*low* and *high-medium* stress) is the effect of listening to loud music for those who listen to loud music. If distribution of earplugs is random, this difference would represent the average treatment effect (ATE). Group n_3 and n_4 do not hear loud music at all since they use earplugs. The difference in their stress levels would reflect the difference in their pre-treatment averages. This represents the selectivity bias (*high-medium* minus *low-medium*). Finally, the difference between groups n_2 and n_4 is the SUTVA bias (*high* minus *low-medium* stress). This is because group n_2 does not intentionally listen to the loud music, but instead is forced to, while group n_4 is completely unaffected. Thus, this earplug intercession enables one to identify the treatment effect as well as both the selectivity and SUTVA biases.

3. The Mechanics

A Potential Outcomes Framework Incorporating Selectivity and *SUTVA* violations

Suppose unit i is seen in one of the two potential treatment states: treated and untreated. Let D indicate these two states so that $D_i = 1$ when the unit receives the treatment and $D_i = 0$ when the unit remains untreated. Define i's potential outcomes in these two states to be $Y_{(D=1)i}$ if treated and $Y_{(D=0)i}$ if not treated. Let these potential outcomes be expressed as follows:

$$Y_{(D=1)i} = \mu_1 + \omega_{(D=1)i} + \widetilde{\theta_i} \tag{1a}$$

$$Y_{(D=0)i} = \mu_0 + \omega_{(D=0)i} + \tilde{\delta}_i \tag{1b}$$

where μ_1 and μ_0 are constants reflecting average outcomes with and without the treatment; $\omega_{(D=1)i}$ and $\omega_{(D=0)i}$ are deviations from respective average outcomes with $E\left[\omega_{(D=1)i}\right] = 0$ and $E\left[\omega_{(D=0)i}\right] = 0$. $\widetilde{\theta_i}$ is person i's unique additional gain from selecting into the treatment. As such, it represents the selectivity bias. $\tilde{\delta}_i$ is the additional outcome the untreated group reaps because of possible treatment spillovers from those choosing treatment. It represents the bias arising from a *SUTVA* violation.[6]

A priori, it is not known whether i would ultimately be treated or not. Hence, i's observed outcome at any given point in time can be expressed by a switching regression formulation (Quandt, 1958) as follows:

$$Y_i = Y_{(D=1)i} D_i + Y_{(D=0)i}(1 - D_i) \tag{2}$$

[6] Here we define a specific treatment spillover from the treated to the untreated. However, the approach can identify this treatment spillover even in the presence of other treatment spillovers, namely from the treated to the treated, and from the untreated to the untreated.

Rearranging (2) yields

$$Y_i = Y_{(D=0)i} + \left(Y_{(D=1)i} - Y_{(D=0)i}\right)D_i \tag{3}$$

where $\left(Y_{(D=1)i} - Y_{(D=0)i}\right)$ represents i's total difference in outcome when moving between the non-treated and treated status. This difference constitutes the treatment effect plus any included biases comprising all the outcome differences that are not the effect of the treatment, namely selectivity and SUTVA violations.

The true treatment effect is the change in outcome absent any confounding effects caused by sample selection and $SUTVA$ biases. As such, i's treatment effect can be expressed as

$$\tilde{\beta}_{Ti} = \left(Y_{(D=1)i} - Y_{(D=0)i} + \tilde{\delta}_i - \tilde{\theta}_i\right) = (\mu_1 - \mu_0) + (\omega_{(D=1)i} - \omega_{(D=0)i}) \tag{4}$$

or,

$$\tilde{\beta}_{Ti} = \beta_T + \eta_i \tag{5}$$

where $\beta_T = (\mu_1 - \mu_0)$ is the constant gain caused by the treatment and $\eta_i = \left(\omega_{(D=1)i} - \omega_{(D=0)i}\right)$ is the idiosyncratic outcome gain caused by the treatment. The latter term is unit specific, hence carries an i subscript.

Substituting (5) and (1b) into (3) yields a simplified version of the observed outcome:

$$Y_i = \mu_0 + \left[\tilde{\beta}_{Ti} - \tilde{\delta}_i + \tilde{\theta_\iota}\right]D_i \; + \tilde{\delta}_i + \omega_{(D=0)i} \tag{6}$$

Rearranging terms yields

$$Y_i = \mu_0 + \tilde{\beta}_{Ti}D_i + \tilde{\delta}_i(1 - D_i) + \tilde{\theta_\iota}D_i + \omega_{(D=0)i} \tag{7}$$

The observed outcome depends on the effect of the treatment ($\tilde{\beta}_{Ti}$) and any biases from selectivity ($\tilde{\theta_\iota}$) and $SUTVA$ violations ($\tilde{\delta}_i$).

Defining an Antidotal Variable

We consider a design to identify the treatment effect, the selection bias, as well as the SUTVA violation bias in an integrated framework.

Suppose one can devise a specific intervention that fully negates the effect of the treatment, without affecting any other concurrent treatments administered to the treatment group. Call this an antidote.[7] Let W_i be the antidotal variable. Assume application of the antidote

[7] Because the antidote is not a treatment, it cannot have an independent effect on the outcome. It is only pertinent in that it nullifies the treatment effect and its spillovers for those who receive the antidote. Giving the antidote is not equivalent to withdrawing the treatment because the untreated without the

abrogates the effect of any treatment, either for those directly treated or for those indirectly affected through treatment spillover SUTVA violations. Also, let W_i be independent of D_i as well as mean independent of β_{Ti}.[8] When the antidote is not applied, the treatment recipients continue to experience their regular treatment effect. As such, the antidotal variable can be thought of as creating another control group. For computational convenience we define $W_i = 0$ when the antidote is applied, and $W_i = 1$ when antidote is not applied.

Based on this definition, we now present an integrated framework for the potential treatment effect and the potential SUTVA bias.

The Potential Treatment Effect

An antidotal variable nullifies the *effect* of a treatment. Thus, unlike the regular instrumental variable framework, which designates a variable to proxy potential treatment *participation*, the antidotal variable abrogates the potential *effect* of a treatment. As such, this approach differs from the standard IV method in that the antidotal variable represents the *impact*, that is the eradication of a treatment's *effect*, but is not standing in for the treatment itself, as in standard IV. When $W_i = 1$, the antidote is *not* administered and the treatment effect remains intact, assuming i fully complies. When $W_i = 0$ the antidote is administered and the treatment effect goes to zero, again assuming compliance. Thus, given full compliance, such that the unit has to take the antidote when administered, and would not take it when not administered, the potential treatment is specified as

$$\beta_{T(W=1)i} = \beta_{Ti} \tag{8a}$$

$$\beta_{T(W=0)i} = 0 \tag{8b}$$

The observed or effective treatment effect is $\beta_{T(W=0)i} = 0$.[9] As such,

$$\tilde{\beta}_{Ti} = \beta_{T(W=1)i} W_i + \beta_{T(W=0)i}(1 - W_i) = \beta_{T(W=1)i} W_i = \beta_{Ti} W_i \tag{9}$$

Assumption 1: The assignment of the antidote is done independent of unit's treatment effect β_{Ti} and treatment D_i, i.e., $W_i \perp \beta_{Ti}$ and $W_i \perp D_i$. This implies $W_i \perp \eta_i$.

---

antidote still get the spillover from the treated with the antidote. (Those without earplugs hear the loud noise spilling over from of those with earplugs still playing loud music.)

[8] In the earlier boombox example (Figure 1), this means those not owning a boombox (the nontreated) do not hear the loud music when given earplugs (subsample n_4). Similarly, given earplugs (group n_3), those boombox owners hear no music. In the application used later in the paper, middle aged women, presumably with no new born children, are the antidotal group since they do not benefit from paid family leave either in California (group n_3) or in other states (group n_4).

[9] Note, in this framework we assume the antidotal variable completely nullifies the treatment effect. This includes all channels through which the treatment and spillovers operate. Shortly, we show how to bound $\beta_{T(W=0)i}$ when the antidote is partially effective.

With this assumption, (9) collapses to

$$\tilde{\beta}_{Ti} = \beta_{Ti} W_i = \beta_T W_i + \eta_i W_i \tag{10}$$

where $\eta_i = \left(\omega_{(D=1)i} - \omega_{(D=0)i}\right)$. This means $E[\eta_i] = 0$ since $E\left[\omega_{(D=1)i}\right] = 0$ and $E\left[\omega_{(D=0)i}\right] = 0$.

The Potential *SUTVA* Bias

The SUTVA bias arises when the treatment effect spills over to the nontreated. But suppose the antidotal variable nullifies spillover effects in the same way it eradicates treatment effects of those directly treated. In the boombox loud music example, the (untreated) non-boombox owners who were assigned earplugs (the antidote) do not hear loud music. If such is the case, the antidote negates the effect of the spillover effect. As such, those untreated persons who receive the antidote will not experience the indirect treatment effects they receive as a result of the spillover. With this framework, spillover effects for those not receiving the antidote ($W_i = 1$) remain δ_i and are zero for those receiving the antidote ($W_i = 0$). Thus,

$$\delta_{(W=1)i} = \delta_i \tag{11a}$$

$$\delta_{(W=0)i} = 0 \tag{11b}$$

The observed $SUTVA$ bias is

$$\widetilde{\delta_i} = \delta_{(W=1)i} W_i + \delta_{(W=0)i}(1 - W_i) = \delta_i W_i \tag{12}$$

Define $\delta_i = \delta + v_i$, where v_i is a random shock component representing possible heterogeneity in the SUTVA bias, such that $E[v_i] = 0$, transforms (12) into

$$\tilde{\delta}_i = \delta W_i + v_i W_i \tag{13}$$

Assumption 2: The assignment of the treatment and antidote is independent of the unit's SUTVA violation effect δ_i.

This assumption means that $v_i \perp W_i, D_i$, i.e., $E[v_i | W_i, D_i] = 0$.

The Corresponding Regression Framework

We derive the final regression equation by substituting both the observed treatment effect (eq. 10) and observed $SUTVA$ bias (eq. 13) in the observed outcome equation (eq. 7).

$$Y_i = \mu_0 + (\beta_T W_i + \eta_i W_i) D_i + (\delta W_i + v_i W_i)(1 - D_i) + \widetilde{\theta_i} D_i + \omega_{(D=0)i} \tag{14}$$

Rearranging terms yields

$$Y_i = \mu_0 + \beta_T W_i D_i + \widetilde{\theta_i} D_i + \delta W_i (1 - D_i) + \omega_{(D=0)i} + \eta_i W_i D_i + v_i W_i (1 - D_i) \quad (15)$$

Defining $\widetilde{\theta_i} = \theta + \phi_i$, where $E[\phi_i] = 0$, one can rewrite the (15) as

$$Y_i = \mu_0 + \beta_T W_i D_i + \theta D_i + \delta W_i (1 - D_i) + \omega_{(D=0)i} + \phi_i D_i + \eta_i W_i D_i + v_i W_i (1 - D_i) \quad (16)$$

Assumption 3: The assignment of the treatment and antidote are independent of ϕ_i, i.e. $\phi_i \perp W_i, D_i$.

To further simplify, we define $u_i = \omega_{(D=0)i} + \phi_i D_i + \eta_i W_i D_i + v_i W_i (1 - D_i)$, which when inserted into (16) yields

$$Y_i = \mu_0 + \beta_T W_i D_i + \theta D_i + \delta W_i (1 - D_i) + u_i \quad (17)$$

This is the regression equation we estimate to identify the three effects, namely the treatment effect, the selectivity bias and the bias due to $SUTVA$ violation.

We now focus on u_i to interpret the effects estimated from (17). As defined,

$$u_i = \omega_{(D=0)i} + \phi_i D_i + \eta_i W_i D_i + v_i W_i (1 - D_i) \quad (18)$$

Here $E[\omega_{(D=0)i}] = E[\phi_i] = E[\eta_i] = E[v_i] = 0$, implying that $E[u_i] = 0$. Whether u_i is independent of W_i and D_i, or not, depends on whether the components of u_i are independent of W_i and D_i. By *Assumptions 1,2,3*, $\omega_{(D=0)i}, \eta_i, v_i, \phi_i \perp W_i$, implying $u_i \perp W_i$. Moreover, by construction, $\omega_{(D=0)i}$ is a noise component, that is independent of the treatment D_i i.e., $\omega_{(D=0)i} \perp D_i$. Essential heterogeneity (selection on the gain) arises when η_i and D_i are not mean independent, implying u_i and D_i are not mean independent. For now we assume no essential heterogeneity, but later we relax this assumption.

The Common Regression Framework

The typical regression model is a special case of the generalized potential outcome model. Consider the case in which $\tilde{\delta}_i = 0$ and $\widetilde{\theta_i} = 0$. With these modifications, (17) reduces to

$$Y_i = \mu_0 + \beta_T W_i D_i + u_i' \quad (19)$$

where $u_i' = \omega_{(D=0)i} + \eta_i W_i D_i$

Conditions $\tilde{\delta}_i = 0$ and $\widetilde{\theta_i} = 0$ indicate no selection at the level and no SUTVA violation. However, OLS regression based on (19) cannot identify β_T if these parameters take non-zero values.

Identification

To distinguish the antidotal variable (AV) from traditional IV identification strategies, we first examine how treatment effects are defined. In the absence of selectivity and SUTVA the average treatment effect is

$$ATE = E[\beta_T + \eta_i] = E[\beta_T] + E[\eta_i] = \beta_T$$

In this case the η_i is the individual specific component which causes heterogeneity of treatment effect across units. When treatment selection is based on η_i, i.e., selection on the gains (essential heterogeneity), $E[\eta_i|D_i = 1] \neq E[\eta_i|D_i = 0]$. When limited to using data only on the treated, $E[\beta_{Ti}|D_i = 1]$,

$$ATT = E[(\beta_T + \eta_i)|D_i = 1] = E[\beta_T|D_i = 1] + E[\eta_i|D_i = 1] = \beta_T + E[\eta_i|D_i = 1]$$

Similarly, the antidotal variable method can identify both ATE and ATT depending on the presence of essential heterogeneity. To see how, consider the four subpopulations defined earlier. We specify mean outcomes in each, then derive the treatment effects based on appropriate differencing of outcomes between the groups.

The first group consists of those units that received treatment without receiving an antidote ($D_i = 1; W_i = 1$). Let $n_1$be the number of units in this group. The second group includes units which are neither treated nor treated with an antidote ($D_i = 0; W_i = 1$). The number of units in the second group is n_2. The third group consists of units that receive both the antidote and treatment ($D_i = 1; W_i = 0$). The size of this group is n_3. The fourth group is made up of units that do not receive the treatment, but receive the antidote ($D_i = 0; W_i = 0$). The size of this group is n_4.[10]

The main challenge of identification emerges from the structure of the error u_i. As (18) shows, u_i comprises the random shock $\omega_{(D=0)i}$ and various heterogeneity components (i.e., η_i, ϕ_i, v_i). Identification of the ATE, the selectivity and $SUTVA$ biases depend on the relationships between these heterogeneity components and D_i and W_i. To see this, consider the mean of u_i for each of these four subgroups. The conditional mean of u_i can be expressed as $E[u_i|D_i, W_i]$

Expanding this using (18) we obtain

$$E[u_i|D_i, W_i] = E[\{\omega_{(D=0)i} + \phi_i D_i + \eta_i W_i D_i + v_i W_i(1 - D_i)\}|D_i, W_i]$$

Rearranging terms yields

[10] The AV approach should not be confused with a DID with multiple control groups. In our case the whole population is subject to spillover effects, but only one segment of the population is subject to spillovers in the multiple control DID case. More specifically, in the AV approach the $D_i = 1; W_i = 0$ group (which we denote as n_3 in our earlier example) yields spillover effects because it gets the treatment, whereas this group would not induce spillovers in a multiple control setting since it would not get the treatment. Thus, the multiple control DID may bias the SUTVA spillover estimate. Further, whereas in the AV approach the selectivity bias can be identified, one cannot identify the selectivity bias if this group does not get the treatment.

$$E[u_i|D_i, W_i] = E[\omega_{(D=0)i}|D_i, W_i] + E[\phi_i D_i|D_i, W_i] + E[\eta_i W_i D_i|D_i, W_i] + E[v_i W_i(1 - D_i)|D_i, W_i]$$

The mean of the composite error u_i in subgroup n_1 is

$$E[u_i|D_i = 1, W_i = 1] = E[\omega_{(D=0)i}|D_i = 1, W_i = 1] + E[\phi_i D_i|D_i = 1, W_i = 1] + E[\eta_i D_i W_i|D_i = 1, W_i = 1]$$

The assumption $\omega_{(D=0)i} \perp D_i, W_i$,implies $E[\omega_{(D=0)i}|D_i = 1, W_i = 1] = E[\omega_{(D=0)i}] = 0$. Thus,

$$E[u_i|D_i = 1, W_i = 1] = E[\phi_i D_i|D_i = 1, W_i = 1] + E[\eta_i W_i D_i|D_i = 1, W_i = 1] \quad (20a)$$

Similarly, the mean of the composite error u_i for subgroup n_2 is

$$E[u_i|D_i = 0, W_i = 1] = E\left[\omega_{(D=0)i}\,|D_i = 0, W_i = 1\right] + E[\phi_i D_i|D_i = 0, W_i = 1] + E[\eta_i W_i D_i|D_i = 0, W_i = 1] + E[v_i W_i(1 - D_i)|D_i = 0, W_i = 1]$$

By construction $E[\omega_{(D=0)i}|D_i] = 0$ and by assumption $E[\omega_{(D=0)i}|W_i] = 0$ jointly imply

$$E[u_i|D_i = 0, W_i = 1] = E[v_i W_i(1 - D_i)|D_i = 0, W_i = 1] \quad (20b)$$

The mean of the composite error u_i for subgroups n_3

$$E[u_i|D_i = 1, W_i = 0] = E\left[\omega_{(D=0)i}\,|D_i = 1, W_i = 0\right] + E[\phi_i D_i|D_i = 1, W_i = 0] + E[\eta_i W_i D_i|D_i = 1, W_i = 0] + E[v_i W_i(1 - D_i)|D_i = 1, W_i = 0]$$

$E[\omega_{(D=0)i}|D_i] = 0$ and $E[\omega_{(D=0)i}|W_i] = 0$ and *assumption 3* jointly imply

$$E[u_i|D_i = 1, W_i = 0] = E[\phi_i D_i|D_i = 1, W_i = 0] = 0 \quad (20c)$$

The mean of the composite error u_i for subgroups n_4

$$E[u_i|D_i = 0, W_i = 0] = E\left[\omega_{(D=0)i}\,|D_i = 0, W_i = 0\right] + E[\phi_i D_i|D_i = 0, W_i = 0] + E[\eta_i W_i D_i|D_i = 0, W_i = 0] + E[v_i W_i(1 - D_i)|D_i = 0, W_i = 0]$$

Again since $E[\omega_{(D=0)i}|D_i] = 0$ and $E[\omega_{(D=0)i}|W_i] = 0$,

$$E[u_i|D_i = 0, W_i = 0] = 0 \quad (20d)$$

Group Averages

Given the above conditional u_i average and regression equation (17), we now formulate the average outcome Y for each group.

Under the independence assumptions, the average of observed outcomes for the first group $(D_i = 1; W_i = 1)$ is defined as $\bar{Y}_1$ where

$$\bar{Y}^1 = E[Y_i|D_i = 1, W_i = 1] = \mu_0 + \beta_T + \theta + E[u_i|D_i = 1, W_i = 1]$$

Substituting conditional expectations $E[u_i|D_i = 1, W_i = 1]$ from (20a) yields

$$\bar{Y}^1 = \mu_0 + \beta_T + \theta + E[\phi_i D_i|D_i = 1, W_i = 1] + E[\eta_i D_i W_i|D_i = 1, W_i = 1] \qquad (21a)$$

where $i \in n_1$.

For the second group, where $(D_i = 0; W_i = 1)$, define the average $\bar{Y}_2$ as

$$\bar{Y}^2 = E[Y_i|D_i = 0, W_i = 1] = \mu_0 + \delta + E[u_i|D_i = 0, W_i = 1]$$

Substituting conditional expectations $E[u_i|D_i = 0, W_i = 1]$ from (20b)

$$\bar{Y}^2 = E[Y_i|D_i = 0, W_i = 1] = \mu_0 + \delta + E[v_i W_i(1 - D_i)|D_i = 0, W_i = 1] \qquad (21b)$$

where $i \in n_2$.

For the third group, where $(D_i = 1; W_i = 0)$, specify the average $\bar{Y}_3$ as

$$\bar{Y}^3 = E[Y_i|D_i = 1, W_i = 0] = \mu_0 + \theta + E[u_i|D_i = 1, W_i = 0]$$

Substituting conditional expectations $E[u_i|D_i = 1, W_i = 0]$ from (20c)

$$\bar{Y}^3 = E[Y_i|D_i = 1, W_i = 0] = \mu_0 + \theta + E[\phi_i D_i|D_i = 1, W_i = 0] \qquad (21c)$$

where $i \in n_3$.

For the fourth group, where $(D_i = 0; W_i = 0)$, let the average be $\bar{Y}_4$ where

$$\bar{Y}^4 = E[Y_i|D_i = 0, W_i = 0] = \mu_0 + E[u_i|D_i = 0, W_i = 0]$$

Substituting conditional expectations $E[u_i|D_i = 0, W_i = 0]$ from (20d)

$$\bar{Y}^4 = E[Y_i|D_i = 0, W_i = 0] = \mu_0 + 0 \qquad (21d)$$

where $i \in n_4$.

Identifying Parameters:

Proposition 1: In absence of essential heterogeneity, the difference between $\bar{Y}^1 - \bar{Y}^3$ identifies average treatment effect (ATE).

By definition

$$ATE = E[\beta_{Ti}] = E[\beta_T + \eta_i] = \beta_T + E[\eta_i] = \beta_T$$

Collecting all components from (21a) and (21c) yields

$$\bar{Y}^1 - \bar{Y}^3 = (\mu_0 + \beta_T + \theta + E[\phi_i D_i | D_i = 1, W_i = 1] + E[\eta_i D_i W_i | D_i = 1, W_i = 1]) \\ -(\mu_0 + \theta + E[\phi_i D_i | D_i = 1, W_i = 0]) \quad (22a)$$

In the absence of essential heterogeneity $E[\eta_i | D_i = 1] = E[\eta_i | D_i = 0] = 0$. Since η_i and W_i are independent by assumption, $E[\eta_i D_i W_i | D_i = 1, W_i = 1] = 0$. Similarly, since ϕ_i, D_i and W_i are pairwise independent by assumption $E[\phi_i D_i | D_i = 1, W_i = 1] = E[\phi_i D_i | D_i = 1, W_i = 0] = E[\phi_i] = 0$. Substituting these terms into the above equations yields

$$\bar{Y}^1 - \bar{Y}^3 = (\mu_0 + \beta_T + \theta + 0 + 0) - (\mu_0 + \theta + 0) = \beta_T = ATE$$

Proposition 2: In the presence of essential heterogeneity, the difference $\bar{Y}^1 - \bar{Y}^3$ identifies the average treatment effect on the treated (ATT).

In the presence of heterogeneity $E[\eta_i | D_i = 1]$ no longer equals $E[\eta_i | D_i = 0]$, nor is it equal to 0. Given that $E[\phi_i | D_i = 1, W_i = 1] = E[\phi_i | D_i = 1, W_i = 0] = 0$, (22a) reduces to

$$\bar{Y}^1 - \bar{Y}^3 = (\mu_0 + \beta_T + \theta + E[\eta_i D_i W_i | D_i = 1, W_i = 1]) \\ -(\mu_0 + \theta)$$

or,

$$\bar{Y}^1 - \bar{Y}^3 = \beta_T + E[\eta_i D_i W_i | D_i = 1, W_i = 1] = \beta_T + E[\eta_i | D_i = 1] = ATT$$

Proposition 3: The difference between $\bar{Y}^3 - \bar{Y}^4$ identifies the average selection bias.

By definition

$$Selection = E[\theta_i] = E[\theta + \phi_i] = \theta + E[\phi_i] = \theta$$

The difference

$$\bar{Y}^3 - \bar{Y}^4 = (\mu_0 + \theta + E[\phi_i D_i | D_i = 1, W_i = 0]) - (\mu_0) = \theta + E[\phi_i D_i | D_i = 1, W_i = 0]$$

The assumption that ϕ_i, D_i and W_i are mutually independent and $E[\phi_i] = 0$ imply $E[\phi_i | D_i = 1, W_i = 0] = 0$, which means

$$\bar{Y}^3 - \bar{Y}^4 = \theta = selectivity\ bias$$

Proposition 4: The difference between $\bar{Y}^2 - \bar{Y}^4$ identifies the average $SUTVA$ bias.

The difference is

$$\bar{Y}^2 - \bar{Y}^4 = (\mu_0 + \delta + E[v_i W_i (1 - D_i) | D_i = 0, W_i = 1]) - (\mu_0) \\ = \delta + E[v_i W_i (1 - D_i) | D_i = 0, W_i = 1]$$

The assumption that v_i, D_i and W_i are mutually independent and $E[v_i] = 0$, imply $E[v_i W_i (1 - D_i) | D_i = 0, W_i = 1] = E[v_i] = 0$. This implies

$$\bar{Y}^2 - \bar{Y}^4 = \delta + 0 = \delta = SUTVA\ bias$$

Note that the presence of essential heterogeneity does not impact the identification of the selection or $SUTVA$ biases.

The Imperfect Antidote Case

An antidote may sometimes be imperfect. As such, the antidote need not completely nullify the effects of treatment. In our prior example, a defective earplug fails to provide complete noise protection.

Here, the point estimates of the treatment effect, the selection bias, and the SUTVA violations cannot be identified. However, by using additional assumptions, one can bound the effects (Manski (1997) and Manski and Pepper (2000). We employ two widely used sets of assumptions and present the resultant bounds.

Assumption Set 1: Positive Monotone Treatment Response (MTR) and $Y_i \geq 0$

The assumption $Y \geq 0$ asserts the outcome cannot be negative. This assumption holds true for a wide variety of applications that measure outcomes as a positive number. As a result,

$$E[Y_i|D, W] \geq 0$$

The positive MTR assumption asserts the treatment cannot lower Y whether the subject is in the treatment group or the control group. As such,

$$E[Y_i(T = 1)|D, W] \geq E[Y_i(T = 0)|D, W]$$

This relationship applies to milder versions of the treatment, for instance when the treatment is caused by a spillover or partially weakened by antidotes.

Assumption Set 2: Monotone Treatment Selection (MTS) and Optimal Treatment Selection (OTS)

The MTS assumption asserts that $E[Y(T = 1)|D = 1] \geq E[Y(T = 1)|D = 0]$ and $E[Y(T = 0)|D = 1] \geq E[Y(T = 0)|D = 0]$. This means the treatment group achieves better outcomes than the control group either when both receive treatment or when both remain untreated.

The OTS assumption posits the treatment group selects treatment because these group members gain from the treatment, whereas the control group does not select treatment because control group members lose from the treatment. This means,

$$E[Y_i(T = 1)|D = 1] \geq E[Y_i(T = 0)|D = 1]$$

$$E[Y_i(T=0)|D=0] \geq E[Y_i(T=1)|D=0]$$

The bounds based on these assumptions are given in Figure 2 and formally derived in Appendix B.[11]

Figure 2: Assumptions and identified bounds

Assumptions	Bounds
Positive MTR and $Y \geq 0$ Positive MTR: $E[Y_i(1)\|D=1] \geq E[Y_i(0)\|D=1]$ $E[Y_i(1)\|D=0] \geq E[Y_i(0)\|D=0]$	β_T: UB= $E[Y_i(T=1)\|D=1, W=1]$ LB=$[Y_i(T=1)\|D=1, W=1] -$ $E\left[Y_i\left(\hat{T}=1\right)\middle\|D=1, W=0\right]$ δ: UB= $E\left[Y_i\left(\tilde{T}=1\right)\middle\|D_i=0, W_i=1\right]$ LB=$E\left[Y_i\left(\tilde{T}=1\right)\middle\|D_i=0, W_i=1\right] -$ $E[Y_i(\bar{T}=1)\|D_i=0, W_i=0]$ θ: UB= $E\left[Y_i\left(\hat{T}=1\right)\middle\|D=1, W=0\right]$ LB= Not identified
MTS and OTS: MTS: $E[Y_i(1)\|D=1] \geq E[Y_i(1)\|D=0]$ $E[Y_i(0)\|D=1] \geq E[Y_i(0)\|D=0]$ OTS: $E[Y_i(1)\|D=1] \geq E[Y_i(0)\|D=1]$ $E[Y_i(0)\|D=0] \leq E[Y_i(0)\|D=0]$	β_T: UB=$E[Y_i(T=1)\|D_i=1, W_i=1] -$ $E[Y_i(\bar{T}=1)\|D_i=0, W_i=0]$ $LB = [Y_i(T=1)\|D_i=1, W_i=1]$ $- E\left[Y_i\left(\hat{T}=1\right)\middle\|D_i=1, W_i=0\right]$ δ:UB=$E[Y_i\left(\tilde{T}=1\right)\|D_i=0, W_i=1] -$ $E[Y_i(\bar{T}=1)\|D_i=0, W_i=0]$ LB= $E[Y_i\left(\tilde{T}=1\right)\|D_i=0, W_i=1] -$ $E\left[Y\left(\hat{T}=1\right)\middle\|D=1, W=0\right]$ θ: UB= $E\left[Y_i\left(\hat{T}=1\right)\middle\|D=1, W=0\right] -$ $E[Y_i(\bar{T}=1)\|D=0, W=0]$ LB= Not identified

Notes: MTR: Monotone Treatment Response; MTS: Monotone Treatment Selection; OTS: Optimum Treatment Selection.

Testing Whether the Antidote Assignment (W) is Random

[11] Other assumptions could also be used to construct bounds, but these are the most common. Thus, we limit ourselves to the two sets of assumptions presented.

If W_i is not mean independent of u_i, one generally cannot identify the treatment effect (β_T), the selectivity bias (θ), and the bias arising from a SUTVA violation (δ) from a single cross-section. However, the advantage of the antidotal variable method is that it is flexible enough to allow one to test whether W_i is correlated with u_i. If one can obtain data on the same variables for just another pre-treatment cross-section, one can test this correlation by checking whether the average Y_i for $W_i = 1$ differs from average Y_i for $W_i = 0$. If these averages do not differ, one can apply the antidotal variable method and successfully retrieve the parameter estimates. If they do, one should be careful to see which parameters are identified and which ones are not.

Consider two subsamples, n_1 and n_3 before the treatment is assigned. Since treatment is not assigned,

$$\begin{aligned}\bar{Y}^1 - \bar{Y}^3 &= E[Y_i|D_i = 1, W_i = 1] - E[Y_i|D_i = 1, W_i = 0] \\ &= (\mu_0 + \theta + E[u_i|D = 1, W = 1]) - (\mu_0 + \theta + E[u|D = 1, W = 0])\end{aligned}$$

Since $D_i = 1$ in both subsamples, the above expression reduces to

$$\bar{Y}^1 - \bar{Y}^3 = E[u_i|D_i = 1, W_i = 1] - E[u_i|D_i = 1, W_i = 0]$$

In this context mean independence of u_i and W_i implies that $E[u|D_i = 1, W_i = 1] = E[u_i|D_i = 1, W_i = 0]$, so that these terms cancel each other. Thus, with mean independence of W_i, and with no essential heterogeneity, the identification in the post treatment period arises from the following equation

$$\beta_T = E[Y_i|D_i = 1, W_i = 1] - E[Y_i|D_i = 1, W_i = 0]$$

When W_i is not mean independent of u_i, $E[u_i|D_i = 1, W_i = 1] \neq E[u_i|D_i = 1, W_i = 0]$, meaning $E[Y_i|D_i = 1, W_i = 1] - E[Y_i|D_i = 1, W_i = 0] = \beta_T + (E[u_i|D_i = 1, W_i = 1] - E[u_i|D_i = 1, W_i = 0]) \neq \beta_T$. Since $E[u_i|D_i = 1, W_i = 1] - E[u_i|D_i = 1, W_i = 0]$ is unknown, β_T is not identified. With just one cross-section this problem cannot be solved.

However, if one has a cross-section from the pre-treatment period, one can test the mean independence assumption of W_i. Consider two subsamples who will be treated and not treated (i.e., $D_i = 1$ and $D_i = 0$) once the treatment rolls out. Because it is the pre-treatment period, $\beta_T = 0$. Thus, the difference in mean of Y_i for $W_i = 1$ and $W_i = 0$ is only due to the difference between $E[u_i|D_i = 1, W_i = 1]$ and $E[u_i|D_i = 1, W_i = 0]$ or $E[u_i|D_i = 0, W_i = 1]$ and $E[u_i|D_i = 0, W_i = 0]$. Thus, whether these means are different, or not, can be tested with the simple regressions within the treated/untreated groups below

$$Y_{i(D=1)} = \gamma_{0(D=1)} + \gamma_{1(D=1)} W_{i(D=1)} + \zeta_{i(D=1)} \tag{26a}$$

$$Y_{i(D=0)} = \gamma'_{0(D=0)} + \gamma'_{1(D=0)} W_{i(D=0)} + \zeta'_{i(D=0)} \tag{26b}$$

If the regression results suggest that $\gamma_{1D=1} = 0$ and $\gamma'_{1D=0} = 0$, one can infer that non-randomness of W_i does not affect the parameter identification. Hence, one can proceed with the identification as outlined above. If $\gamma_{1D=1} \neq 0$ and $\gamma'_{1D=0} = 0$, then one identifies the bias

from a SUTVA violation (δ) only. The treatment effect and selectivity bias (β_T and θ) cannot be identified. Conversely, if $\gamma_{1D=1} = 0$ and $\gamma'_{1D=0} \neq 0$, then one identifies β_T, but cannot identify the SUTVA bias (δ) and the selectivity bias (θ). If ($\gamma_{1D=1} \neq 0$ and $\gamma_{1D=0} \neq 0$, then none of the parameters of interest are identified.[12]

4. Antidote Compliance

There is an important difference between a typical instrument and an antidotal variable. Standard instruments, those correlated with the treatment but not the outcome, require monotonicity, a framework where no defiers are assumed in the data.[13] [14] Yet defiers can, and often, readily exist with good reason, especially when the treatment recipient's motivations differ based on anticipated effects of the treatment. For example, using two-children of the same sex as an instrument for having more children, might fail for those families with either boy or girl preferences (Dahl and Moretti, 2008) as these families would not be motivated to continue childbirth. Yet as de Chaisemartin (2017) and Swanson et al. (2015) show, there are numerous other examples.

Antidotal variables are different. They are not linked to the treatment, but instead to eradicating the treatment's impact. Thus they negate the *effect* of a treatment, rather than the treatment itself. As such, there is a smaller susceptibility to defy because there is less incentive to do so.

Take the loud music example considered earlier. Defiers would comprise those who would seek earplugs to nullify the effect of the noise if not given them *and* would not use earplugs if they were received. Although those choosing treatment (D = 1) who are given earplugs (W = 0) likely will seek not to use them, since the earplugs keeps them at high-medium stress instead of low stress; those choosing loud music (the treated where D = 1) have no incentive to seek earplugs, as earplugs will make them worse off (high-medium stress with earplugs versus low stress without earplugs). Thus, there likely should be no defiers among the treated.

Also, there are likely no defiers among the non-treated. Those not choosing the boombox treatment (D = 0) might seek out earplugs in order to attain low-medium rather than high stress if they were not given them (W = 1). However, those among the non-treated who were given earplugs (D = 0 and W = 1) presumably will always use them, again to lower stress. Thus, here too, there are likely no defiers among the non-treated.

In this sense, as long as treatment selection is endogenous, the antidotal approach has weaker underlying assumptions than the traditional instrumental variables approach, since no

[12] Typical IV validity cannot be tested in the same way using pre-treatment data because there cannot be any change in the IV. Any change in the pre-treatment IV would imply a weak instrument.

[13] Angrist, Imbens, and Rubin (1996). JASA

[14] There are exceptions. For example, de Chaisemartin (2017) shows one can obtain a local average treatment effect (LATE) by replacing the no-defiers condition by a "compliers-defiers" condition in which "a subgroup of compliers accounts for the same percentage of the population as defiers and has the same LATE" (p. 368). Dahl et al (2019) uses a strictly weaker local monotonicity condition to identify LATE for compliers and defiers.

monotonicity type assumption is needed to interpret the estimates. With full compliance, when neither always affected groups and never affected groups exist, the antidotal procedure yields an average treatment effect (ATE). With always affected and/or never affected groups, the procedure identifies the effects for those participants bound by the antidote, what we call the binding average treatment effect (BATE). BATE is similar to the traditional IV local average treatment effect (LATE).

5. Simulations

To validate the approach and test for consistency, we conduct several simulation exercises. First, we simulate data based on a process where $W = 0{,}1$ is randomly assigned. We generate the treatment variable D through a uniform distribution. A value of $D = 1$ indicates the unit receives the treatment, and $D = 0$ indicates no treatment. Similarly, we independently create the antidotal variable W through a uniform distribution. The value $W = 0$ indicates the antidotal intervention nullifies the effect of the treatment, and $W = 1$ indicates no antidotal intervention so that the treatment remains effective. This process essentially divides the sample into four subsamples: $\{D = 1, W = 1\}, \{D = 0, W = 1\}, \{D = 1, W = 0\}$, and $\{D = 0, W = 0\}$. Based on these we generate the outcome variable y in the following manner:

$$y_i = \beta_{0i} + \beta_{Ti} W_i D_i + \theta_i D_i + \delta_i W_i (D_i - 1) + u_i$$

where the parameters β_{0i}, β_{Ti}, θ_i, δ_i and u_i are as previously defined. To keep the simulation simple, we assume these parameters follow normal distributions with different means and standard deviations. Thus, each individual receives an assigned value based on random draws. This process ensures parameter heterogeneity as well as no essential heterogeneity. We experiment with various parameter values to check the generality of the results. First, we generate data based on the following schemes: $\beta_{0i} = \mathrm{N}(.3, .1)$; $\beta_{Ti} = \mathrm{N}(.7, .2)$; $\theta_i = \mathrm{N}(.4, .4)$; $\delta_i = \mathrm{N}(.8, .3)$. Then, we replicate the same exercise with $\beta_{0i} = \mathrm{N}(.3, .1)$; $\beta_{Ti} = \mathrm{N}(.2, .2)$; $\theta_i = \mathrm{N}(.15, .4)$; $\delta_i = \mathrm{N}(.35, .3)$. We then estimate the parameters from the simulated data for different number of observations (100, 1000, 10000, 100000, 1000000).

Table 1 reports the results. As can be seen, the estimates are close to the parameter values used to create the data, more so when the number of observations is large. All three estimates, the treatment effect, the selectivity bias, and the bias due to $SUTVA$ violation approach to the true parameter values when the sample size increases. Thus the simulation exercise confirms the consistency property of the estimators. A similar set of results emerges when we alter the parameter values while creating the data.

6. An Application

For illustrative purposes, to demonstrate the antidotal variable method, we examine take-up rates for the California's paid family leave (CPFL) program. Initiated in 2004, CPFL allows employees to take up to 6 weeks of paid leave, usually for child care responsibilities, though it could be used for taking care of ailing parents. Past analyses used difference-in-differences

(DID) techniques to estimate CPFL's effect on leave take-up (Rossin-Slater et al., 2013; Baum and Ruhm, 2014; and Das and Polachek, 2015). Typically, this type analysis estimated the difference in take-up from before to after the law in California relative to a control.[15] We replicate this type analysis, and then present results based on an antidotal variable approach. In the process, we point out how the antidotal variable approach alleviates a number of the biases inherent in DID.

We utilize data from CPS-AESC rounds from 2001-2006 collected in March of each year. The CPS-AESC is nationally representative and includes information on individuals' demographics, work and other characteristics. We focus on two measures of leave taking: (1) leave hours and (2) leave incidence.[16]

We divide the data into two time periods: 2004-06, the period when CPFL was in effect; and 2001-2003, the period before CPFL's implementation.[17] Table 2 presents the summary statistics for leave taking hours before and after CPFL became effective. In 2004-2006, women in California took less leave than their counterparts in other states (1.21 hours in 2001-03 vs 1.59 hours). However, in 2004-2006 after CPFL became effective, women in California took more leave (2.06 hours in 2001-03 vs 1.91 hours). This increase predominated in the younger age group, as leave increased for young employees in California, whereas it decreased for all other groups in California as well as in the other states (2.11 versus 1.93 for the young in other states and 1.5 versus 1.06 for the old in California and 1.98 versus 1.88 for the young in other states).

Table 3 presents the summary statistics for leave taking incidence before and after CPFL became effective. This table shows a very similar pattern as observed in Table 2. California's incidence of leave taking rises to 3.3 percent in 2004-06 from 2.7 percent in 2001-03. Other states experience a decline (3.8 percent in 2004-06 from 4.1 percent in 2001-03). As before, only young women in California experienced an increase in leave taking.

In Tables 2 and 3 we denote California as receiving treatment D = 1 and the other states as untreated D = 0. Since CPFL primarily addresses the leave taking need of women of childbearing age, we assume older women 45-55 are unaffected by the paid family leave

[15] Rossin-Slater et al. (2013) define their treatment group to be California mothers with children less than 1 year old compared to a control of all others in California. Baum and Ruhm (2014) use a triple difference to compare California mothers before and after paid family leave relative to those in other states, and Das and Polachek (2015) use a quadruple difference approach to compare young women in California to older women and men in California and the rest of the country before and after the program's implementation.

[16] The actual measures are (1) the difference in a worker's usual weekly work hours and her actual hours of work, and (2) whether a worker has a job but is on leave at the time of interview. It is a binary/dummy variable which takes value 0 if at work, and 1 if on leave. We also consider employment status, gender, and age, and hold constant number of children below 15 years of age and years of completed education. We drop the respondents engaged in military services from this dataset.

[17] Technically the law was implemented in July 2004. However, the law actually passed the state legislature in 2002, both employers and employees most likely anticipated the change and changed their leave taking behaviour earlier in 2004 slightly before the enactment date. Hence, the effect on leave taking may start appearing even before the July 2004, which is why we include 2004 in the post-policy period.

because they typically do not have young children and hence are unlikely to take family leave.[18] We assign W = 0 for women between 45-55 years of age since it is unlikely that women in that age group have young children. We denote W = 1 for those women 25-40.[19] The 45-55 year old W = 0 group is important because the antidotal approach requires a group which is unaffected by the paid family leave. It is likely that this group fits the bill.

DID estimates CPFL's effect by computing the before and after differences between California and the rest of the US. This ATE amount to 0.46 for hours and a 0.009 increase in the incidence of taking a leave. [20]

The antidotal variable approach entails four groups. Group 1 (D = 1 and W = 1) are those who receive treatment and do not have the antidote (age 25-40). Group 2 (D = 0 and W = 1) constitute those not receiving the antidote (age 25-40) and those in the control states. Group 3 (D = 1 and W = 0) comprise those receiving treatment but getting the antidote (age 45-55). Finally, group 4 (D = 0 and W = 0) are those in the control group who get the antidote (are 45 – 55). Define the mean values of leave-taking hours and leave incidence for each group as $\bar{Y}_1$, $\bar{Y}_2$, $\bar{Y}_3$, and $\bar{Y}_4$. Based on Table 2 (hours of leave), $\bar{Y}_1 = 1.97$, $\bar{Y}_2 = 1.93$, $\bar{Y}_3 = 1.06$, and $\bar{Y}_4 = 1.88$. Based on Table 3 (the incidence of leave), these values are $\bar{Y}_1 = 0.039$, $\bar{Y}_2 = 0.039$, $\bar{Y}_3 = 0.025$, and $\bar{Y}_4 = 0.037$. The antidotal variable approach defines the average treatment effect as $\bar{Y}_1 - \bar{Y}_3$, the selectivity bias as $\bar{Y}_3 - \bar{Y}_4$, and the SUTVA bias as $\bar{Y}_2 - \bar{Y}_4$. Thus, the average treatment effect is 0.91 hours, the selectivity bias is -0.82 hours, and the SUTVA bias is -0.05 hours. For incidence, the values are 0.014, -0.012, and 0.02 respectively.

Several differences between the two approaches are noteworthy. First, the DID approach requires two cross-sections spanning two time periods (2001-2003 and 2004-2006) and identifies only one parameter. The antidotal approach requires only one cross-section in one time period (2004-2006) and identifies three parameters. Second, the DID approach assumes that in the absence of treatment, the unobserved differences between treatment and control groups are the same overtime. In our example, this means nothing else should change between California and the control states except paid family leave; otherwise these other interventions can affect the result as new confounders. Changing confounders between the two periods manifest themselves as changes in selectivity, which can be identified in the antidotal variable approach by comparing group mean values $\bar{Y}_3$ and $\bar{Y}_4$ in the earlier 2001-2003 time period. Third, DID assumes no SUTVA violations. In our example, this means California's paid family leave cannot affect the leave taking behavior in the control states.

Interestingly, the average treatment effect we just found differs between the two approaches. The DID estimate (0.46) is about half the antidotal variable estimate (0.91) for hours leave, and

[18] Biases in our estimates could result to the extent older women actually take leave to look after older parents, but according to Wettstein and Zulkarnain (2017), this is more confined to those over 55, which as noted later we drop from the sample. Further, the incidence and amount of parent-motivated leave taking for 45-55 and 25-40 year old adult children is similar. This implies little if any estimation bias, given the antidotal variable technique exploits differences in outcomes between these two groups.

[19] We drop respondents above 55 years of age as they might take leave for other reasons such as decaying general health condition.

[20] The computations are: (1.59-1.21) - (1.91-2.06) = 0.46 and (0.033-0.027) - (0.038-0.041) = 0.09

about 2/3 the size based on incidence (0.009 versus 0.014). The DID approach assumes the selectivity bias remains constant across both periods so that no other policy or comparable changes occur in California relative to the control states once the policy is implemented. In short, there cannot be changes in the confounding effects. Typically, most DID studies spend much time trying to justify this, but do so by relying on institutional considerations, typically without hard evidence. However possible confoundedness can bias the estimates.[21] Evaluating $\bar{Y}_3$ - $\bar{Y}_4$ in 2001-2003 yield -.63 in hours and -.009 in incidence probability. Noteworthy, these are higher than the -.82 and -.012 values in 2004-2006, thus implying changes in the confounding effects. Indeed, these changes in confounding effects explain upwards of 60% of the discrepancy when considering incidence.[22] As we will show shortly, we observe no statistically significant treatment or SUTVA effects when examining 2001-2003, an expected placebo test.

While these statistics indicate that young women in California take more leave, other unincluded covariates can affect the results. For this reason, we now use a regression framework to apply the antidotal variable approach in a more rigorous way based on (17). However, to do so, we first test whether the antidotal variable W is mean independent of u. Mean independence implies we can obtain unbiased estimates of selectivity and SUTVA violations. We utilize hours and incidence of leave as dependent variables in an OLS regression on W using 2001-03 data, the period prior to the policy implementation. For each dependent variable we run an OLS regression on W, once for California and once for the other states. Table 4 presents the results. An insignificant coefficient implies mean independence between the antidotal variable W and the error term u. As illustrated, the coefficients for the antidotal variable W are insignificant in each of these regressions.[23] Thus, the results suggest that W is mean independent of u. Based on this finding we proceed to estimate the treatment effect, the selection bias, and bias emanating from violation of $SUTVA$ using the antidotal variable method.

Table 5 presents the causal effect of CPFL on weekly leave taking hours obtained from regressions based on (17). Young women in California take 0.91 hours more leave.[24] Including control variables household size (to get at the presence of children) and schooling level (education) did not change the coefficients appreciably. The table also shows that the selection bias is in the range of 0.82. As before, we find little evidence of any bias arising from violation of $SUTVA$. This might be expected since California's policy change is unlikely to have significant effect on the rest of the country's labor market.

As indicated earlier, one can bound the estimates if one believes older age (45-55) serves as an imperfect antidote. MTR and $Y_i \geq 0$ are satisfied because leave incidence and leave

[21] Suppose other factors change in California and the control states. These could be specific elected officials, new laws in either state, or any number of other variables. Such observed changes between California and the rest of the country are subsumed in the selectivity bias. The AV approach can detect changes in how California differed from the control group from before to after the policy change.

[22] Computed as (-.009-(-.012))/(.009-.014)=0.60.

[23] Adding covariates does not alter this result.

[24] Note this is the same as previously computed because with no covariates this regression simply reports differences in mean values between the four groups.

taking exceed zero and paid family leave does not lower leave taking when implemented. Accordingly, we construct the following bounds on the ATE, the SUTVA bias, and the selectivity bias using these assumptions along with the means of various subsamples: $0.91 \leq \beta_T \leq 1.97$, $0.05 \leq \delta \leq 1.93$, and $-1.88 \leq \theta \leq 1.06$. The results mean that the ATE and the SUTVA bias are identified, but the selectivity bounds span zero.

One advantage of the antidotal approach is the ability to do a placebo test using the 2001-2003 data. Given CPFL did not occur until 2004, we rerun (17) for 2001-3. We should find no effect of CPFL and no SUTVA bias as neither is present in 2001-3 prior to the policy's implementation. As can be seen in columns (1) and (3), the coefficients of DW (treatment effect) and W(1-D) (SUTVA bias) are both statistically insignificant. However, noteworthy, as seen above, the selectivity coefficient remains significant, but smaller, likely because of changing confounding variables between the two time periods.

Table 6 presents the results on the incidence of leave taking. As before simply looking at the means, young women in California take 1.4 percentage points more leave than young women in the other states. This amounts to 50 percent increase in the probability of leave taking. The selectivity bias in this case is about -1 to – 1.2 percentage points, i.e., young women in California are 40 – 48 percent less likely to take leave than young women in the other states. Again, the SUTVA bias is insignificantly different from zero. Also, as above, both the treatment effect and selectivity bias estimates are zero in the pre-treatment period.

One reason for a virtually zero SUTVA bias is comparing California to the rest of the nation. In the case of a policy like CPFL, one would expect the SUTVA bias, if it exists, to arise because women in the control states move to California to take advantage of the new policy. Typically, those potentially benefiting the most, immigrate. Those remaining would have lower leave taking behavior, thus negating the SUTVA bias. Given moving costs, the direct and the information acquisition costs, one would expect migrants to comprise those living close by, most likely, from neighboring states. Utilizing all states but California as the controls possibly led to no SUTVA bias. For this reason, we repeat the analysis, this time limiting our control states to the three states bordering California: Arizona Nevada and Oregon.

Table 7 and 8 present these results. As before, we observed a positive treatment effect, essentially the same magnitude as before (0.91 for hours and 0.13 for incidence). Selectivity is larger (-1.41 for hours and -.38 for incidence) meaning California differs more from its neighboring states than from the whole US. And as suspected, there is a negative SUTVA effect (-.84 for hours and -.04 for incidence)[25] meaning neighboring states reduce leave taking after the CPFL was instituted. Notably, all coefficients are insignificant in 2001-03 when there was no treatment. Somewhat surprisingly, this includes the coefficient for selectivity. This essentially zero coefficient implies the possibility of other unobserved factors, thus exacerbating the difference between California and its neighbors.

The findings with respect to the incidence of leave taking are similar to that of hours work. The treatment effect and SUTVA bias are zero before CPFL came into effect. The zero

[25] Note a positive coefficient implies a negative SUTVA because (12) has a $-\delta W$ coefficient.

selectivity during 2001-03 also suggests that California, Arizona. Oregon and Nevada are similar in terms of their workers' leave taking. However, in 2004-06, all three are statistically significant. The bias due to SUTVA violation is significant supporting the results obtained for hours of leave taking.

7. Conclusion

The fundamental problem of treatment effect identification is each observation can only be seen in one of two states: treated or untreated. Counterfactual outcomes are not observed. The industry standard strategy to overcome this shortcoming is primarily through random assignment of treatment, but this is not always possible, especially in observational settings. This led to a number of fixes such as IV, DID, RDD, RCT and other methods, but essentially all these solutions are designed to make the treatment and control groups as similar as possible, thereby mimicking randomization as best as can be done. Nevertheless, the threat of a potential SUTVA violation remains. Moreover, these approaches cannot identify the treatment effect in the presence of concomitant treatments.

In this paper we examine another approach. We introduce an antidotal variable (AV) to both treatment and control groups that negates the impact of the treatment for this set of individual observations. Abrogating the treatment effect, as such, separates the sample into four groups, instead of two. From these four groups, we identify the treatment effect, as well as selectivity and SUTVA violation biases. The only requirement is that the antidotal variable be independent of the treatment and mean independent of the outcome variable, which can be tested using pre-treatment data. This is a weaker assumption than standard IV approaches, which require a variable related to the treatment but unrelated to the dependent variable, a condition that for the most part one cannot test.

Despite the power of the antidotal variable approach, there are limitations. First, one needs to find an antidotal variable that abrogates the treatment effect for a subsample of the data. This could be a direct intervention nullifying the treatment or a characteristic of a subsample of observations for which the effect of the treatment is nullified. In some applications, antidotal variables may be difficult to find. Second, the antidotal variable should be independent of treatment effects before the application of the antidote. This likely holds if the antidote is randomly administered. It also holds if the non-antidoted group would have behaved similarly to the antidoted group if they did not receive the antidote, a weaker condition. Third, the antidote is assigned to both treated and untreated groups. Violation of this latter assumption simply makes it impossible to identify the SUTVA bias. Fourth, the antidote needs to abrogate the treatment spillover effects. Finally, we rule out any spillovers from the treated to others in the treated group.

To validate the approach and test for consistency, we simulated data based on randomly assigning treatment and antidotes. In all cases estimated coefficients converged relatively quickly to the true parameter values.

Also, we applied the approach to estimate the impact of paid family leave. A simple DID approach found CPFL increased leave taking by about ½ hour per week and leave incidence by about 1 percentage point. The antidotal variable approach yielded about 0.9 hours and 1.4%, most of the differences arising because of selectivity. DID assumes selectivity (the difference between California and the other states) remain constant from before and to after the law's implementation. But the antidotal variable approach showed this not to be the case, as it was able to pick up other factors, such as California's simultaneously implemented 2004 Private Attorneys General Act (PAGA), that could affect leave taking. In addition, the approach showed SUTVA spillover effects between California and its neighboring states Arizona, Nevada, and Oregon, arising after the law's implementation, possibly implying some migration to California by those inclined towards leave taking.

Whereas we apply the AV technique to analyze the California Paid Family Leave program, it potentially has applications beyond this example.

Table 1: Estimates from the simulated data (random W)

	Actual Value β_T = 0.7	Actual Value θ=0.4	Actual Value δ = 0.8	Actual Value β_T=0.2	Actual Value θ = 0.15	Actual Value δ = 0.35
	Estimated values			Estimated values		
No. of obs	β_T	θ	δ	β_T	θ	δ
100	0.48	0.62	0.82	-0.02	0.37	0.37
1000	0.67	0.52	0.80	0.17	0.27	0.35
10000	0.68	0.40	0.81	0.18	0.15	0.36
100000	0.69	0.40	0.79	0.19	0.15	0.34
1000000	0.70	0.40	0.80	0.20	0.15	0.35

Source: Simulated data and authors' own computations.

Table 2: Hours of Leave Taking (The number of hours worked less than usual hours)

		2001-2003			2004-2006		
		W=0	W=1		W=0	W=1	
D		45-55	25-40	ALL	45-55	25-40	ALL
0	Other states	1.98	2.11	2.06	1.88	1.93	1.91
1	Calif	1.35	1.12	1.21	1.06	1.97	1.59
	Total	1.91	1.99	1.96	1.79	1.93	1.87

Source: IPUMS-CPS (AESC rounds); authors computations.

Table 3: Incidence of leave

		2001-2003			2004-2006		
		W=0	W=1		W=0	W=1	
D		45-55	25-40	ALL	45-55	25-40	ALL
0	Other states	0.040	0.042	0.041	0.037	0.039	0.038
1	Calif	0.031	0.025	0.027	0.025	0.039	0.033
	Total	0.039	0.040	0.040	0.036	0.039	0.038

Source: IPUMS-CPS (AESC rounds); authors computations.

Table 4: 2001-03 Testing mean independence of the antidotal variable W

VARIABLES	(1) Hrsabsnt D=1	(2) Hrsabsnt D=0	(3) Dleave D=1	(4) Dleave D=0
W	-0.229	0.131	-0.00590	0.00275
	(0.284)	(0.0994)	(0.00555)	(0.00201)
Constant	1.348***	1.978***	0.0306***	0.0396***
	(0.231)	(0.0750)	(0.00454)	(0.00151)
Observations	3,809	49,173	4,082	52,973
R-squared	0.000	0.000	0.000	0.000

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 5: Regression results for the United States

VARIABLES	(1) hrsabsnt 2001-03	(2) hrsabsnt 2004-06	(3) hrsabsnt 2001-03	(4) hrsabsnt 2004-06
DW	-0.229	0.908***	-0.274	0.888***
	(0.284)	(0.287)	(0.285)	(0.288)
D	-0.630***	-0.821***	-0.644***	-0.819***
	(0.243)	(0.223)	(0.244)	(0.222)
W(1-D)	-0.131	-0.0475	-0.0929	-0.0268
	(0.0994)	(0.0943)	(0.101)	(0.0960)
Household Size			0.0432**	0.00947
			(0.0205)	(0.0204)
Education			0.00727***	0.00766***
			(0.00217)	(0.00201)
Constant	1.978***	1.879***	1.231***	1.179***
	(0.0750)	(0.0707)	(0.207)	(0.194)
Observations	52,982	53,416	52,982	53,416
R-squared	0.001	0.000	0.001	0.001

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 6: Regression Results for the United States

VARIABLES	(1) dleave 2001-03	(2) dleave 2004-06	(3) dleave 2001-03	(4) dleave 2004-06
DW	-0.00590	0.0139**	-0.00732	0.0131**
	(0.00555)	(0.00577)	(0.00556)	(0.00579)
D	-0.00899*	-0.0120***	-0.00949**	-0.0125***
	(0.00479)	(0.00426)	(0.00480)	(0.00427)
W(1-D)	-0.00275	-0.00184	-0.00152	-0.000831
	(0.00201)	(0.00191)	(0.00202)	(0.00194)
Household Size			0.00144***	0.000844**
			(0.000417)	(0.000425)
Education			0.000224***	0.000186***
			(4.30e-05)	(3.99e-05)
Constant	0.0396***	0.0372***	0.0164***	0.0187***
	(0.00151)	(0.00141)	(0.00423)	(0.00392)
Observations	57,055	57,748	57,055	57,748
R-squared	0.001	0.000	0.001	0.001

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7: Regression Results for California and Neighboring States

VARIABLES	(1) Hrsabsnt 2001-03	(2) Hrsabsnt 2004-06	(3) Hrsabsnt 2001-03	(4) Hrsabsnt 2004-06
DW	-0.229	0.908***	-0.243	0.903***
	(0.284)	(0.287)	(0.290)	(0.294)
D	-0.0563	-1.403***	-0.0576	-1.408***
	(0.394)	(0.437)	(0.397)	(0.437)
W(1-D)	-0.574	0.833*	-0.577	0.839*
	(0.436)	(0.467)	(0.441)	(0.470)
Household Size			0.0102	0.00957
			(0.0489)	(0.0543)
Education			0.00369	-0.000633
			(0.00498)	(0.00474)
Constant	1.404***	2.461***	1.056*	2.490***
	(0.319)	(0.383)	(0.552)	(0.577)
Observations	6,172	5,927	6,172	5,927
R-squared	0.001	0.003	0.001	0.003

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 8: Regression Results for California and Neighboring States

VARIABLES	(1) dleave 2001-03	(2) dleave 2004-06	(3) Dleave 2001-03	(4) dleave 2004-06
DW	-0.00590	0.0139**	-0.00644	0.0133**
	(0.00555)	(0.00578)	(0.00557)	(0.00585)
D	-0.000968	-0.0379***	-0.00139	-0.0383***
	(0.00784)	(0.00967)	(0.00793)	(0.00970)
W(1-D)	-0.000185	0.0300***	0.000358	0.0306***
	(0.00811)	(0.0103)	(0.00820)	(0.0104)
Household Size			0.000656	0.000851
			(0.000951)	(0.00111)
Education			3.28e-05	2.62e-05
			(9.18e-05)	(9.78e-05)
Constant	0.0316***	0.0631***	0.0271**	0.0584***
	(0.00639)	(0.00880)	(0.0106)	(0.0123)
Observations	6,663	6,447	6,663	6,447
R-squared	0.000	0.003	0.000	0.004

Robust standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

References

Angrist, J., G. W. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91:444-472.

Aronow, P. M., and C. Samii (2017) "Estimating Average Causal Effects Under General Interference," arXiv no. 1305.6156.

Baum, C. L., and C. J. Ruhm (2014) "The Effects of Paid Family Leave in California on Labor Market Outcomes." IZA Discussion Paper No. 8390.

Cox, D. R. (1958) Planning of Experiments, New York: Wiley.

Dahl, Christian M., Martin Huber, and Giovanni Mellace (2019): "It's Never Too LATE: A New Look at Local Average Treatment Effects with or without Defiers," https://www.bc.edu/content/dam/bc1/schools/mcas/economics/pdf/seminars/LATEwm2019.pdf

Dahl, Gordon and Enrico Moretti (2008), "The Demand for Sons," *The Review of Economic Studies*, 75(4): 1085–1120, https://doi.org/10.1111/j.1467-937X.2008.00514.x.

Das, Tirthatanmoy and Solomon Polachek (2015) "Unanticipated Effects of California's Paid Family Leave Program," *Contemporary Economic Policy*, 33(4): 619-635.

Das, Tirthatanmoy and Solomon Polachek (2019). "A New Strategy to Identify Causal Relationships: Estimating a Binding Average Treatment Effect," IZA Discussion Paper No. 12766.

de Chaisemartin, Clément (2017) "Tolerating Defiance? Local Average Treatment Effects Without Monotonicity" *Quantitative Economics* 8: 367–396.

Forastiere, Laura, Edoardo M. Airoldi and Fabrizia Mealli (2021) "Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks," *Journal of the American Statistical Association*, 116:534, 901-918.

Friedt, Felix and Jeffrey Cohen (2021) Valuation of Noise Pollution and Abatement Policy: Evidence from the Minneapolis-St. Paul International Airport," *Land Economics* 97(1): 107-136.

Heckman, James, Sergio Urzua, and Edward Vytlacil (2006) "Understanding Instrumental Variables in Models with Essential Heterogeneity," *The Review of Economics and Statistics* 88 (3): 389–432.

Liu, L., M. G. Hudgens, and S. Becker-Dreps, (2016) "On Inverse Probability-Weighted Estimators in the Presence of Interference," *Biometrika*, 103, 829–842.

Madestam, Andreas, Daniel Shoag, Stan Veuger, David Yanagizawa-Drott (2013) "Do Political Protests Matter? Evidence from the Tea Party Movement" *The Quarterly Journal of Economics*, 128(4): 1633–1685.

Manski, C. F. (1997) "Monotone Treatment Response. *Econometrica*. 65: 1311–1334.

Manski, C. F, and J. V. Pepper (2000) "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*. 68: 997–1010.

Quandt, R. E., (1958) "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Association*, 53(284): 873-880.

Rossin-Slater, M., C. J. Ruhm, and J. Waldfogel (2013) "The Effects of California's Paid Family Leave Program on Mothers' Leave-Taking and Subsequent Labor Market Outcomes." *Journal of Policy Analysis and Management*, 32(2): 224–45.

Rubin, Donald B. (1980) "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75(371): 591-593.

Swanson, S. A., M. Miller, J. M. Robins and M. A. Hernán (2015). Definition and evaluation of the monotonicity condition for preference-based instruments. *Epidemiology*, 26(3): 414-20.

Tchetgen, E. J. Tchetgen and T. J. VanderWeele (2012) "On Causal Inference in the Presence of Interference," *Statistical Methods in Medical Research*, 21, 55–75.

Van der Laan, M. J. (2014) "Causal Inference for a Population of Causally Connected Units," *Journal of Causal Inference*, 2, 13–74.

Wettstein, Gal and Alice Zulkarnain (2017) "How Much Long-Term Care Do Adult Children Provide," Center for Retirement Research at Boston College Research Paper 17-11.

APPENDIX

A. The Antidotal Variable Model

The econometric model

$$Y_i = \mu_0 + \beta_T W_i D_i + \theta D_i + \delta W_i(1 - D_i) + u_i \qquad (A.1)$$

where $u_i = \omega_{(D=0)i} + \phi_i D_i + \eta_i W_i D_i + v_i W_i(1 - D_i)$

Define $z_i = \begin{bmatrix} 1 \\ W_i D_i \\ D_i \\ W_i(D_i - 1) \end{bmatrix}$ and $\Gamma = \begin{bmatrix} \mu_0 \\ \beta_T \\ \theta \\ \delta \end{bmatrix}$

With this, (A.1) can be expressed as

$$Y_i = z_i'\Gamma + u_i \qquad (A.2)$$

Define $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_N \end{bmatrix}$, $\mathrm{Z} = \begin{bmatrix} z_1' \\ z_2' \\ . \\ z_N' \end{bmatrix}$ and $u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_N \end{bmatrix}$

Based on these, matrix representation is

$$Y = Z\Gamma + u \qquad (A.3)$$

The OLS estimator of the parameters is

$$\hat{\Gamma} = (\mathrm{Z}'Z)^{-1}Z'Y \qquad (A.4)$$

Unbiasedness

Proof:

The expected value of the estimator $\mathrm{E}[\hat{\Gamma}]$

$$\begin{aligned} E\left[\hat{\Gamma}\right] &= E[Z'Z]^{-1}Z'(Z\Gamma + u) \\ &= E[(\mathrm{Z}'Z)^{-1}Z'Z\Gamma + (\mathrm{Z}'Z)^{-1}Z'u \\ &= \Gamma + (\mathrm{Z}'Z)^{-1}E[Z'u] \end{aligned} \qquad (A.5)$$

As such, $Z'u$ can be expressed as

$$Z'u = \begin{bmatrix} 1 & 1 & & 1 & 1 \\ W_1D_1 & W_2D_2 & \cdots & W_{n-1}D_{n-1} & W_nD_n \\ D_1 & D_2 & & D_{n-1} & D_n \\ W_1D_1 - W_1 & W_2D_2 - W_2 & & W_{n-1}D_{n-1} - W_{n-1} & W_nD_n - W_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ .. \\ u_n \end{bmatrix}$$

Simplifying

$$Z'u = \begin{bmatrix} \sum_i u_i \\ \sum_i W_iD_iu_i \\ \sum_i D_iu_i \\ \sum_i W_iD_iu_i - \sum_i W_iu_i \end{bmatrix} \tag{A.6}$$

Taking expectation on both sides

$$E[Z'u] = E\begin{bmatrix} \sum_i u_i \\ \sum_i W_iD_iu_i \\ \sum_i D_iu_i .. \\ \sum_i W_iD_iu_i - \sum_i W_iu_i \end{bmatrix} = \begin{bmatrix} E[\sum_i u_i] \\ E[\sum_i W_iD_iu_i] \\ E[\sum_i D_iu_i] \\ E[\sum_i W_iD_iu_i] - E[\sum_i W_iu_i] \end{bmatrix}$$

With further simplification

$$E[Z'u] = E\begin{bmatrix} \sum_i u_i \\ \sum_i W_iD_iu_i \\ \sum_i D_iu_i .. \\ \sum_i W_iD_iu_i - \sum_i W_iu_i \end{bmatrix} = \begin{bmatrix} \sum_i E[u_i] \\ \sum_i E[W_iD_iu_i] \\ \sum_i E[D_iu_i] \\ \sum_i E[W_iD_iu_i] - \sum_i E[W_iu_i] \end{bmatrix}$$

When W_i and D_i and u_i are pairwise and jointly mean independent, by the law of iterated expectations the above matrix simplifies to

$$E[Z'u] = \begin{bmatrix} \sum_i E[u_i] \\ \sum_i E_{WD}[W_i D_i E[u_i|W_i D_i]] \\ \sum_i E_D[D_i E[u_i|D_i]] \\ \sum_i E_{WD}[W_i D_i E[u_i|W_i D_i]] - \sum_i E_W[W_i E[u_i|W_i]] \end{bmatrix}$$

With $E[u_i] = 0$, $E[u_i|W_i] = 0$, $E[u_i|D_i] = 0$, and $E[u_i|W_i D_i] = 0$ since ϕ_i, η_i and v_i are independent of W_i and D_i, and $E[\phi_i] = 0$, $E[\eta_i] = 0$, $E[v_i] = 0$, each of the element in the above matrix is reduces to zero, meaning,

$$E[Z'u] = 0 \qquad (A.7)$$

Substituting $E[Z'u] = 0$ in (A.6) into (A.4)

$$E\left[\hat{\Gamma}\right] = \Gamma + (\mathrm{Z}'Z)^{-1}E \times 0 = \Gamma \qquad (A.8)$$

Hence, under the mean independence assumption, $E\left[\hat{\Gamma}\right] = \Gamma$, i.e., unbiased.… Q.E.D.

Consistency

Proof:

$$plim\ \hat{\Gamma} = plim\ \Gamma + plim[\left(\frac{\mathrm{Z}'Z}{N}\right)^{-1}(Z'u/N)]$$

By the product rule of probability limit

$$= plim\ \Gamma + plim\left(\frac{\mathrm{Z}'Z}{N}\right)^{-1} plim\left(\frac{Z'u}{N}\right) \qquad (A.9)$$

Here $\mathrm{Z}'Z$ is a positive definite matrix. Thus, consistency requires $plim\left(\frac{Z'u}{N}\right) = 0$

As shown above

$$Z'u = \begin{bmatrix} \sum_i u_i \\ \sum_i W_i D_i u_i \\ \sum_i D_i u_i .. \\ \sum_i W_i D_i u_i - \sum_i W_i u_i \end{bmatrix}$$

Thus,

$$\frac{Z'u}{N} = \frac{1}{N}\begin{bmatrix} \sum_i u_i \\ \sum_i W_i D_i u_i \\ \sum_i D_i u_i .. \\ \sum_i W_i D_i u_i - \sum_i W_i u_i \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_i u_i \\ \frac{1}{N}\sum_i W_i D_i u_i \\ \frac{1}{N}\sum_i D_i u_i .. \\ \frac{1}{N}\sum_i W_i D_i u_i - \frac{1}{N}\sum_i W_i u_i \end{bmatrix}$$

When $N \to \infty$, and Z' and u are i.i.d draws, by the weak law of large numbers one can write

$$plim \begin{bmatrix} \frac{1}{N}\sum_i u_i \\ \frac{1}{N}\sum_i W_i D_i u_i \\ \frac{1}{N}\sum_i D_i u_i .. \\ \frac{1}{N}\sum_i W_i D_i u_i - \frac{1}{N}\sum_i W_i u_i \end{bmatrix} = E[Z'u] \qquad (A.10)$$

Since $E[Z'u] = 0$ in (A.7), and $E\ (Z'Z)^{-1}$is finite, they imply

$$plim\ \hat{\Gamma} = \Gamma$$

Hence, $\hat{\Gamma}$ is a consistent estimator of Γ

QED

Asymptotic normality

As already shown,

$$\hat{\Gamma} = \Gamma + (Z'Z)^{-1}(Z'u)$$

or

$$\hat{\Gamma} - \Gamma = (Z'Z)^{-1}(Z'u)$$

$$\sqrt{N}(\hat{\Gamma} - \Gamma) = \left(\frac{Z'Z}{N}\right)^{-1}(Z'u/\sqrt{N})$$

The matrix $\mathrm{Z}'Z$ is positive definite, thus invertible. Hence the asymptotic distribution of $(\hat{\Gamma} - \Gamma)$ depends on the limiting properties of $(Z'u/\sqrt{N})$

As such $(Z'u/\sqrt{N})$ can be expressed as

$$\frac{Z'u}{\sqrt{N}} = \frac{1}{\sqrt{N}} \begin{bmatrix} \sum_i u_i \\ \sum_i W_i D_i u_i \\ \sum_i D_i u_i .. \\ \sum_i W_i D_i u_i - \sum_i W_i u_i \end{bmatrix}$$

Since u_i is mean independent of W_i, D_i, by the weak law of large numbers, when $N \to \infty$, $plim\ \frac{1}{N}\sum_i u_i = E[u_i]$; $plim\ \frac{1}{N}\sum_i W_i D_i u_i = E[W_i D_i u_i]$; $plim\ \frac{1}{N}\sum_i D_i u_i = E[D_i u_i]$; $plim\ \frac{1}{N}\sum_i W_i u_i = E[W_i u_i]$ and $plim\ \left(\frac{Z'Z}{N}\right)^{-1} = E[Z'Z]^{-1}$.

Since, as shown above, $E[u_i] = 0$, $E[W_i D_i u_i] = 0$, $E[D_i u_i] = 0$, $E[W_i u_i] = 0$, by Lindeberg-Feller central limit theorem one can write

$$\sqrt{N}\left(\hat{\Gamma} - \Gamma\right) \to^{\mathrm{d}} N[0, \mathrm{E}[\mathrm{Z}'Z]^{-1}\mathrm{Var}[Z'u]\mathrm{E}[\mathrm{Z}'Z]^{-1}] \qquad (A.11)$$

B: The imperfect antidote case

An imperfect antidote does not completely negate the effects of the treatment and its spillover effects. Consequently, β_T, θ and δ are not generally identified. The parameters can, however, be set-identified with additional assumptions. To illustrate, we first redefine the treatment and spillover statuses as follows:

$T = 0$: Treatment not received (not observed in any subsample)
$\bar{T} = 1$: Treatment not received, but there is a spillover with an *imperfect* antidote (subsample n4)
$\tilde{T} = 1$: Treatment not received, but there is a spillover without antidote (subsample n2)
$T = 1$: Received treatment without antidote (subsample n1)
$\hat{T} = 1$: Received treatment with an *imperfect* antidote (subsample n3)

Consider the regression equation (17)

$$Y_i = \mu_0 + \beta_T W_i D_i + \theta D_i + \delta W_i (1 - D_i) + u_i \qquad (B.1)$$

where $u_i = \omega_{(D=0)i} + \eta_i W_i D_i + \phi_i D_i + v_i W_i (1 - D_i)$

Recall the original model assumptions

A.B1) $W_i \perp u_i$ meaning $W_i \perp \omega_{(D=0)i}, W_i \perp \eta_i,\ W_i \perp \phi_i,\ W_i \perp v_i$.
A.B2) $D_i \perp \omega_{(D=0)i}, D_i \perp \phi_i,\ D_i \perp v_i$.
A.B3) $W_i \perp D_i$.
A.B4) Also, $\omega_{(D=0)i}, \eta_i, \phi_i, v_i$ are mutually pairwise independent.

As outlined in the text, we make two sets of additional assumptions in order to deal with the imperfect antidote. Below we show the bounds for each of these set of assumptions. Before going any further, it is useful to express ATE, $SUTVA$ and selection bias in terms of the regression parameters.

Average treatment effect (β_T)

Based on (22a) in the text, without any essential heterogeneity, one can show that the ATE (β_T) is

$$\beta_T = E[Y_i(T=1)|D_i=1, W_i=1] - E[Y_i(T=0)|D_i=1, W_i=0] \qquad (B.2)$$

With a fully effective antidote, $E[Y_i(T=0)|D_i=1, W_i=0] = E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1, W_i=0\right]$. By substituting $E[Y_i(T=0)|D_i=1, W_i=0]$ with $E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1, W_i=0\right]$ (which is observed) into (B.2), one obtains the ATE.

Average bias due to SUTVA violation(δ)

Similarly, *Proposition 4* in the text implies that the average bias due to $SUTVA$ violation or δ is

$$\delta = E\left[Y_i\left(\tilde{T}=1\right)\middle|D_i=0, W_i=1\right] - E[Y_i(T=0)|D_i=0, W_i=0] \qquad (B.3)$$

With a fully effective antidote, $E[Y_i(T=0)|D_i=0, W_i=0] = E[Y_i(\bar{T}=1)|D_i=0, W_i=0]$. By substituting $E[Y_i(T=0)|D_i=0, W_i=0]$ with $E[Y_i(\bar{T}=1)|D_i=0, W_i=0]$ (which is observed) into (B.3), one obtains the δ.

Selectivity bias (θ)

Similarly, *Proposition 3* in the text implies that the average selectivity bias (θ) is

$$\theta = E[Y_i(T=0)|D_i=1, W_i=0] - E[Y_i(T=0)|D_i=0, W_i=0] \qquad (B.4)$$

With a fully effective antidote, $E[Y_i(T=0)|D_i=1, W_i=0] = E\left[Y_i\left(\hat{T}=0\right)\middle|D_i=1, W_i=0\right]$ and $E[Y_i(T=0)|D_i=0, W_i=0] = E[Y_i(\bar{T}=1)|D_i=0, W_i=0]$. Note that both $E\left[Y_i\left(\hat{T}=0\right)\middle|D_i=1, W_i=0\right]$ and $E[Y_i(\bar{T}=1)|D_i=0, W_i=0]$ are observed. As such substituting $E\left[Y_i\left(\hat{T}=0\right)\middle|D_i=1, W_i=0\right]$ and $E[Y_i(\bar{T}=1)|D_i=0, W_i=0]$ into (B.4), one obtains the θ.

Case 1: Imperfect antidote with no essential heterogeneity, MTR and $Y \geq 0$

Bounds on β_T

When the antidote is not fully effective, $E[Y_i(T=0)|D_i=1,W_i=0] \neq E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]$. Thus, (B.2) no longer identifies β_T. However, with MTR *and* $Y \geq 0$ one can put bounds on β_T.

Lower Bound of β_T

The MTR and $Y \geq 0$ assumptions imply that

$$E[Y_i(T=1)|D_i=1,W_i=0] \geq E[Y_i(T=0)|D_i=1,W_i=0] \qquad (B.5)$$

However, with a less effective antidote, the treatment ($\hat{T}=1$) is a milder version of the original treatment. Due to this property, one can write

$$E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \geq E[Y_i(T=0)|D_i=1,W_i=0] \qquad (B.6)$$

where both $E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]$ and $E[Y_i(T=0)|D_i=1,W_i=0]$ are non-negative.

As such, this inequality shows that $E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]$ is the upper bound estimate of $E[Y_i(T=0)|D_i=1,W_i=0]$. This upper bound provides the lower bound estimate of β_T, that is

$$\beta_T \geq E[Y_i(T=1)|D_i=1,W_i=1] - E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \qquad (B.7)$$

Upper bound of β_T

The assumption that $Y_i \geq 0$ also implies that $[Y_i(T=0)|D_i=1,W_i=0] \geq 0$. This provides a lower bound estimate for $E[Y_i(T=0)|D_i=1,W_i=0]$. This condition therefore implies that the upper bound estimate of β_T is

$$\beta_T \leq E[Y_i(T=1)|D_i=1,W_i=1] \qquad (B.8)$$

Thus, the bounds for β_T are

$$E[Y_i(T=1)|D_i=1,W_i=1] - E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \leq \beta_T \leq E[Y_i(T=1)|D_i=1,W_i=1] \qquad (B.9)$$

***Bounds on* $\boldsymbol{\delta}$**

When the antidote is not fully effective, $E[Y_i(T=0)|D_i=0,W_i=0] \neq E[Y_i(\overline{T}=1)|D_i=0,W_i=0]$. Thus, (B.3) no longer identifies δ. However, with MTR and $Y_i \geq 0$, one can bound δ.

Lower Bound of δ

The MTR and $Y \geq 0$ assumptions imply that

$$E[Y_i(T=1)|D_i=0,W_i=0] \geq E[Y_i(T=0)|D_i=0,W_i=0] \qquad (B.10)$$

However, with a less effective antidote the treatment spillover ($\bar{T}=1$) is a milder version of the original treatment. Because of this, one can write

$$E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \geq E[Y_i(T=0)|D_i=0,W_i=0] \qquad (B.11)$$

where both $E[Y_i(\bar{T}=1)|D_i=0,W_i=0]$ and $E[Y_i(T=0)|D_i=0,W_i=0]$ are non-negative, i.e., $Y_i \geq 0$.

As such, this inequality shows that $E[Y_i(\bar{T}=1)|D_i=0,W_i=0]$ is the upper bound lower bound estimate of $E[Y_i(T=0)|D_i=0,W_i=0]$. This upper bound estimates provides the lower bound estimate of δ, that is

$$\delta \geq E\left[Y_i(\tilde{T}=1)\middle|D_i=0,W_i=1\right] - E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \qquad (B.12)$$

Upper bound of $\underline{\delta}$

The assumption that $Y_i \geq 0$ also implies that $[Y_i(T=0)|D_i=0,W_i=0] \geq 0$. This provides a lower bound estimate for $E[Y_i(T=0)|D_i=1,W_i=0]$. This condition therefore implies that the upper bound estimate of δ is

$$\delta \leq E\left[Y_i(\tilde{T}=1)\middle|D_i=0,W_i=1\right] \qquad (B.13)$$

Thus, the bounds for δ are

$$E\left[Y_i(\tilde{T}=1)\middle|D_i=0,W_i=1\right] - E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \leq \delta \leq E\left[Y_i(\tilde{T}=1)\middle|D_i=0,W_i=1\right] \qquad (B.14)$$

Bounds on $\underline{\boldsymbol{\theta}}$

As per (B.4)

$$\theta = E[Y_i(T=0)|D_i=1,W_i=0] - E[Y_i(T=0)|D_i=0,W_i=0]$$

In this case, neither $E[Y_i(T=0)|D_i=1,W_i=0]$ nor $E[Y_i(T=0)|D_i=0,W_i=0]$ are observed. However, derivations above shows the upper bounds and lower bounds for each of these terms,

$$0 \leq E[Y_i(T=0)|D_i=1,W_i=0] \leq E\left[Y_i(\hat{T}=1)\middle|D_i=1,W_i=0\right] \qquad (B.15a)$$
$$0 \leq E[Y_i(T=0)|D_i=0,W_i=0] \leq E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \qquad (B.15b)$$

Lower bound of $\underline{\theta}$

The lower bound estimate of $E[Y_i(T=0)|D_i=1,W_i=0]$ and upper bound estimate of $E[Y_i(T=0)|D_i=0,W_i=0]$ provides the lower bound estimate of θ, i.e.,

$$-E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \leq \theta$$

Upper bound of θ

The upper bound estimate of $E[Y_i(T=0)|D_i=1,W_i=0]$ and lower bound estimate of $E[Y_i(T=0)|D_i=0,W_i=0]$ provides the upper bound estimate of θ, i.e.,

$$\theta \leq E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]$$

Thus, the bounds for θ are

$$-E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \leq \theta \leq E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \qquad (B.16)$$

Case 2: Imperfect antidote with no essential heterogeneity, MTS, OTS, $Y_i \geq 0$ *and bigger* Y_i *is preffered*

Bounds on β_T

As per *(B.2)*,

$$\beta_T = E[Y_i(T=1)|D_i=1,W_i=1] - E[Y_i(T=0)|D_i=1,W_i=0] \qquad (B.17)$$

The problem of identification is that $E[Y_i(T=0)|D_i=1,W_i=0]$ is not observed (see in *Case 1* above).

Lower Bound of β_T

Under the OTS assumption, the treatment group prefers the treatment over no-treatment. Because $\left(\hat{T}=1\right)$ represents a milder version of the treatment, we assume that the treatment group would also prefer this over no-treatment. In other words,

$$E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \geq E[Y_i(T=0)|D_i=1,W_i=0] \qquad (B.18)$$

Hence (B.2) can be re written as

$$\begin{aligned}\beta_T = &\left\{E[Y_i(T=1)|D_i=1,W_i=1] - E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]\right\} \\ &+ \left\{E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] - E[Y_i(T=0)|D_i=1,W_i=0]\right\}\end{aligned}$$

so that

$$\begin{aligned}&\beta_T - \left\{E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] - E[Y_i(T=0)|D_i=1,W_i=0]\right\} \\ &= E[Y_i(T=1)|D_i=1,W_i=1] - E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \qquad (B.19)\end{aligned}$$

Equation (B.8) implies

$$\left\{E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]-E[Y_i(T=0)|D_i=1,W_i=0]\right\}\geq 0$$

Thus, (B.19) can be re-written as

$$\beta_T \geq E[Y_i(T=1)|D_i=1,W_i=1]-E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \qquad (B20)$$

Both $E[Y_i(T=1)|D_i=1,W_i=1]$ and $E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right]$ are observed. Hence the lower bound of β_T is identified.

Upper bound of β_T

By the MTS assumption

$$E[Y_i(T=0)|D_i=1,W_i=0]\geq E[Y_i(T=0)|D_i=0,W_i=0] \qquad (B.21)$$

Because treatment spillover ($\bar{T}=1$) is a milder version of treatment ($T=1$), the OTS assumption implies

$$E[Y_i(T=0)|D_i=0,W_i=0]\geq E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \qquad (B.22)$$

Combining (B.21) and (B.22)

$$E[Y_i(T=0)|D_i=1,W_i=0]\geq E[Y_i(\bar{T}=1)|D_i=0,W_i=0]$$

or

$$0\geq E[Y_i(\bar{T}=1)|D_i=0,W_i=0]-E[Y_i(T=0)|D_i=1,W_i=0] \qquad (B.23)$$

Now consider (B.2) again, i.e.,

$$\beta_T=E[Y_i(T=1)|D_i=1,W_i=1]-E[Y_i(T=0)|D_i=1,W_i=0] \qquad (B.24)$$

Rewriting (B.24)

$$\begin{aligned}\beta_T=&\{E[Y_i(T=1)|D_i=1,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0]\}\\&+\{E[Y_i(\bar{T}=1)|D_i=0,W_i=0]-E[Y_i(T=0)|D_i=1,W_i=0]\}\end{aligned}$$

or,

$$\begin{aligned}&\beta_T-\{E[Y_i(\bar{T}=1)|D_i=0,W_i=0]-E[Y_i(T=0)|D_i=1,W_i=0]\}\\&=E[Y_i(T=1)|D_i=1,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0]\end{aligned} \qquad (B.25)$$

By (B.23) and (B.24) one then can write

$$\beta_T\leq E[Y_i(T=1)|D_i=1,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \qquad (B.26)$$

The bounds on β_T are

$$E[Y_i(T=1)|D_i=1,W_i=1]-E\left[Y_i\left(\hat{T}=1\right)\middle|D_i=1,W_i=0\right] \leq \beta_T \leq E[Y_i(T=1)|D_i=1,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \quad (B.27)$$

Bounds on δ

Consider (B.3) again

$$\delta = E[Y_i(\tilde{T}=1)|D_i=0,W_i=1]-E[Y_i(T=0)|D_i=0,W_i=0]$$

Here as well, the problem of identification arises because $E[Y_i(T=0)|D_i=0,W_i=0]$ is not observed.

Upper bound of δ

The OTS assumption implies that those who chose no-treatment, are worse off with treatment than without treatment. Since $\bar{T}=1$ represents a milder version of the actual treatment, the observed average outcome of n_4 with $\bar{T}=1$ is smaller than the average outcome of n_4 without treatment, i.e.,

$$E[Y_i(T=0)|D_i=0,W_i=0] \geq E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \quad (B.28)$$

Now rewrite (B.3) as

$$\delta = \{E[Y_i(\tilde{T}=1)|D_i=0,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0]\} + \{E[Y_i(\bar{T}=1)|D_i=0,W_i=0]-E[Y_i(T=0)|D_i=0,W_i=0]\}$$

Simplifying

$$\delta - \{E[Y_i(\bar{T}=1)|D_i=0,W_i=0]-E[Y_i(T=0)|D_i=0,W_i=0]\} = E[Y_i(\tilde{T}=1)|D_i=0,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \quad (B.29)$$

Note, by (B.28), $E[Y_i(\bar{T}=1)|D_i=0,W_i=0]-E[Y_i(T=0)|D_i=0,W_i=0] \leq 0$, which implies

$$\delta \leq E[Y_i(\tilde{T}=1)|D_i=0,W_i=1]-E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \quad (B.30)$$

Lower bound of δ

By the MTS assumption again,

$$E[Y(T=0)|D=1,W=0] \geq E[Y(T=0)|D=0,W=0] \quad (B.31)$$

By OTS assumption with partial treatment $\hat{T}=1$

$$E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] \geq E[Y(T=0)|D=1,W=0] \quad (B.32)$$

Combing (B.31) and (B.32)

$$E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] \geq E[Y(T=0)|D=0,W=0] \quad (B.33)$$

Rearranging (B.3)

$$\begin{gathered}\delta = E[Y_i\left(\tilde{T}=1\right)|D_i=0,W_i=1] - E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] \\ +E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] - E[Y_i(T=0)|D_i=0,W_i=0] \quad (B.34)\end{gathered}$$

or,

$$\begin{gathered}\delta - \left\{E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] - E[Y_i(T=0)|D_i=0,W_i=0]\right\} \\ = E[Y_i\left(\tilde{T}=1\right)|D_i=0,W_i=1] - E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right]\end{gathered}$$

Note, by (B.33),

$$\left\{E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] - E[Y_i(T=0)|D_i=0,W_i=0]\right\} \geq 0,$$

yielding

$$\delta \geq E[Y_i\left(\tilde{T}=1\right)|D_i=0,W_i=1] - E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] \quad (B.35)$$

Thus, the bounds for δ

$$\begin{gathered}E[Y_i\left(\tilde{T}=1\right)|D_i=0,W_i=1] - E\left[Y\left(\hat{T}=1\right)|D=1,W=0\right] \\ \leq \delta \leq \\ E[Y_i\left(\tilde{T}=1\right)|D_i=0,W_i=1] - E[Y_i(\bar{T}=1)|D_i=0,W_i=0] \quad (B.36)\end{gathered}$$

Bounds of θ

To examine the selectivity with imperfect antidote, consider (B.4), which states

$$\theta = E[Y_i(T=0)|D_i=1,W_i=0] - E[Y_i(T=0)|D_i=0,W_i=0]$$

As stated above, neither $E[Y_i(T=0)|D_i=1,W_i=0]$ nor $E[Y_i(T=0)|D_i=0,W_i=0]$ are observed when antidote is imperfect and there is a treatment spillover.

Upper bound of $\underline{\theta}$

The OTS assumption implies

$$E[Y(T=0)|D=0,W=0] \geq E[Y(\bar{T}=1)|D=0,W=0] \quad (B.37)$$

because the control group ideally prefers no treatment over the full treatment T. If this preference is true, then the control group would also prefer no-treatment over a milder version of the treatment $\bar{T} = 1$.

By the same OTS logic, the treatment group would like treatment over no-treatment. Thus, the treatment group would prefer even a partially effective treatment ($\hat{T} = 1$) than no treatment. Thus

$$E[Y(T = 0)|D = 1, W = 0] \leq E\left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right] \qquad (B.38)$$

Now rewrite (B.4)

$$\begin{aligned}\theta = E[Y(T = 0)|D = 1, W = 0] - E\left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right] + E\left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right] \\ - E[Y(\bar{T} = 1)|D = 0, W = 0] + E[Y(\bar{T} = 1)|D = 0, W = 0] \\ - E[Y(T = 0)|D = 0, W = 0]\end{aligned}$$

implying

$$\begin{aligned}\theta - \left\{E[Y(T = 0)|D = 1, W = 0] - E\left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right]\right\} \\ - \{E[Y(\bar{T} = 1)|D = 0, W = 0] - E[Y(T = 0)|D = 0, W = 0]\} \\ = E\left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right] - E[Y(\bar{T} = 1)|D = 0, W = 0] \qquad (B.39)\end{aligned}$$

Combining (B.39) with (B.37) and (B.38) yields

$$\theta - (\leq 0) - (\leq 0) = \left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right] - E[Y(\bar{T} = 1)|D = 0, W = 0]$$

In other words, the upper bound of θ is

$$\theta \leq E\left[Y\left(\hat{T} = 1\right)|D = 1, W = 0\right] - E[Y(\bar{T} = 1)|D = 0, W = 0] \qquad (B.40)$$

The lower bound for θ is not identified since upper bound of $E[Y(T = 0)|D = 0, W = 0]$ is not identified with the OTS assumption and treatment spillover.