ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 74/25

Causal Inference Using Antidotal Variables

Tirthatanmoy Das, Solomon W. Polachek

www.rfberlin.com SEPTEMBER 2025

Causal Inference Using Antidotal Variables

Authors

Tirthatanmoy Das, Solomon W. Polachek

Reference

JEL Codes: C18, C36, I38, J18, J38

Keywords: Bias, California Paid Family Leave, Causality, Nullifying Effect, Spillover Effect

Recommended Citation: Tirthatanmoy Das, Solomon W. Polachek (2025): Causal Inference Using Antidotal Variables. RFBerlin Discussion Paper No. 74/25

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

Causal Inference Using Antidotal Variables

Tirthatanmoy Das
Economics Area
Indian Institute of Management Bangalore and IZA
tirthatanmoy.das@iimb.ac.in

Solomon W. Polachek
Department of Economics
State University of New York at Binghamton and IZA
polachek@binghamton.edu

September 2025

ABSTRACT

This paper shows that incorporating what we call antidotal variables (AV) into a causal treatment effects analysis can with one cross-sectional regression identify the causal effect, the spillover effect, as well as possible biases from selectivity. We apply the AV technique to analyze leave taking arising from the California Paid Family Leave (CPFL) program. Our analysis yields between a 55% and 70% larger treatment effect than the traditional DID methods, which we attribute to confounding effects and spillovers, neither of which are found in traditional studies.

Key words: Bias, California Paid Family Leave, Causality, Nullifying Effect, Spillover Effect

1. INTRODUCTION

This paper presents a new method to estimate treatment effects, spillover effects, and selection bias using a single cross-sectional dataset. Spillover effects—often overlooked—go by many names, including violations of the "stable unit treatment value assumption" (SUTVA) (Rubin, 1986), interference (Sobel, 2006), peer effects (Adamopoulou, 2012), neighborhood effects (Christafore and Leguizamon, 2019), herd effects (Benjamin-Chung et al., 2018), and network effects (DiMaggio and Garip, 2012). Ignoring spillovers can bias treatment effect estimates, as shown in studies on police deterrence (Sherman and Weisburd, 1995), job training returns (Ashenfelter, 1978), and paid family leave (Kang et al., 2022). Spillovers also matter independently; for example, understanding peer effects in job training helps policymakers optimize seat assignments (Baird et al., 2023). Our method focuses on spillovers where untreated individuals are indirectly affected by treatment, leading to biased treatment effect estimates.

Statisticians mainly approach spillovers experimentally (Forastiere et al., 2021), typically through two-stage randomized trials. These divide the population into clusters, with treatment randomly assigned to varying proportions within each cluster. Simply put, differences in outcomes among untreated individuals across clusters enable one to compute spillovers. This approach has a large literature (Hudgens and Halloran 2008; Liu et al. 2016; Liu 2019a and 2019b; Sävje et al. 2021; and Tchetgen and VanderWeele 2012). More recent work is able to get at treatment selectivity, namely noncompliance within each cluster (Wilke, Green, and Cooper 2020; Imai and Jiang 2020; Imai et al, 2021 and DiTraglia et al. 2023).

Not always can one subdivide a population into clusters with varying proportions of treated individuals. An alternative is a placebo design, in which the proportion of subjects treated does not vary, but in which a randomly selected cluster receives a fake (different or no) treatment unrelated to the outcome (Wilke, Green and Cooper 2020; Imai and Jiang 2020; Huber and Steinmayr 2021). Here, the treatment effect is measured by comparing the outcome of the treated

group with the outcome of the placebo group, and the spillover is measured by comparing the outcome of the non-treated groups in both clusters.

Not all spillovers can be analyzed experimentally. In the above examples, Kansas City cannot be divided into police presence clusters because crime rates are precinct-specific and all residents are equally treated. Further, one precinct may affect another, as criminals might move to less policed areas. Similarly, comparing mothers in California in a difference-in-differences setting (Kang et al., 2022) to mothers before and after Paid Family Leave's passage (or to mothers in other states) does not allow for clusters with varying treatment intensity – treatment is either present or absent. For this reason we propose an alternative method to estimate spillovers which can be implemented in a quasi-experimental setting.

The approach requires an intercession, defined by an "antidotal variable" (AV), which if received, nullifies both the treatment effect and its spillovers. As such, similar to epidemiological effect modification (VanderWeele and Robins 2007 and VanderWeele 2009), those receiving the antidote are unaffected by the treatment or its spillovers. The approach includes a validity test to ensure the AV is mean independent of unobserved outcome determinants (the error component), a crucial assumption for identification. It also controls for potential confounding from other concurrent policies often assumed away in most DID studies, allowing identification of treatment effects even in the presence of another concomitant treatment. The approach works in a single cross-section either in an experimental or observational setting. When the AV only partially nullifies the treatment and spillover effects, estimates can be bounded. To our knowledge, this is the first unified framework to identify the average treatment effect (*ATE)* or the average treatment effect on the treated (*ATT*), the average spillover bias (the SUTVA violation) on the untreated (*ASEU*), and the selectivity bias (*SELB*).

The logic is as follows: Instead of two groups (treated and untreated), there are now four. First, the treated group splits into those who receive the antidote and those who don't. Since the antidote negates the treatment effect, the outcome difference between these two constitutes the

treatment effect. If the antidote and treatment are mean independent of the parameters, this difference equals the average treatment effect (ATE). If not, it represents the average treatment effect of the treated (ATT) because the treated sample is innately different than the untreated sample. Second, the control group also splits into those with and without the antidote. The outcome difference here yields the average spillover effect on the untreated (*ASEU*), as the antidote cancels spillovers on the untreated. Third, those getting the antidote split into those that get the treatment and those that do not get the treatment. The outcome difference between these is the selectivity bias (*SELB*), since neither group experiences the treatment or spillovers, but they differ due to treatment selection. In short, the antidotal variable enables one to identify the *ATE* or *ATT* as well as the *ASEU* and *SELB* by nullifying both the treatment effect within the treated as well as the spillover effects within the control.

Antidotal variables differ from traditional instrumental variables. While both are unrelated to the error term, an instrument affects the treatment itself, whereas an *AV* leaves the treatment intact but negates its effect. Unlike a negative control (Lipsitch et al. 2010), an AV doesn't eradicate the treatment and is administered to both the treatment and control subgroups. Importantly, it nullifies the treatment effect in the treatment group and the spillover effects in the control group, enabling one to identify the *ASEU*.

An antidotal variable also differs from a placebo. This is crucial because it highlights how the approach identifies the bias arising from *SUTVA* violations. Placebo recipients do not actually get the treatment but can still experience spillovers from the treated. In contrast, those getting an antidote are protected from spillover effects, whether or not they receive the treatment or spillover. This distinction enables the *AV* approach to identify the *ASEU*, which a placebo approach would not.

Antidotal variables can be actively administered, such as using earplugs, as described in the intuitive hypothetical loud music example discussed in the next section. Alternatively, an antidotal variable could be immutable, for example an innate characteristic in a population

subsample that makes them immune to the treatment. In this latter case, the subgroup incurs no treatment effect despite being treated. In the application we discuss later in the paper, this immunity likely applies to middle aged California women covered by the California Paid Family Leave (CPFL) program, who typically do not benefit from paid family leave because they neither have young children nor sufficiently old parents, despite being covered.

The antidotal variable approach also differs from standard DID methods. Both compute differences. However, the DID compares changes between a treatment and control group, usually over a given time, *before and after* a treatment, which can be biased if there are spillovers. In contrast the AV approach computes *cross-sectional* differences, thus not requiring time-series or panel data. We compare estimates from both methods when assessing the impact of the California Paid Family Leave (CPFL) program.

The California Paid Family Leave program began in 2004. It allows parents up to six weeks paid leave to take care of young children. A common analysis compares utilization rates in California before and after passage of the law relative to utilization rates in other comparable states over the same time period. However, this *DID* approach has at least two potential biases. One is the spillover bias, if those in neighboring states respond to the California policy despite not being covered. Another, is the selectivity bias which can potentially change with time, especially if there are confounders. The usual assumption is that the control states and California have a constant non-changing selectivity component before and after the law. However, many other policies may be implemented concurrently, 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 also enacted the Private Attorneys General Act (PAGA) which helps low-wage workers enforce labor rights by allowing class-action lawsuits for Labor Code violations.

The antidotal variable method deals with both spillover (a *SUTVA* violation) and selectivity issues. Although CPFL applies to all Californians, young childbearing aged women are the main

beneficiaries, not older women without young children (Rossin-Slater, Ruhm, and Waldfogel 2011; Baum and Ruhm 2014). As such, age 45-55 can serve as an antidotal variable to help identify the *ATE* or *ATT*, along with the *ASEU* and *SELB*. Even if this group is partially affected (an imperfect antidote), the AV method is still able to identify the bounds for these effects.

We apply this antidotal variable approach (in Section 5) using two measures of leave utilization. We find at least a 55% increase in the probability of leave taking and an 70% increase in leave taking hours. This holds despite a selectivity bias indicating that Californians are in general about 40% less likely to take leave. We find insignificant *ASEU* (*SUTVA* bias), which is reasonable when comparing California to the rest of the country. Similarly, we find equivalent CPFL effects when comparing California to its three neighboring states, but here we detect a negative *ASEU*, meaning leave taking decreased in those neighboring states after the CPFL was instituted.

2. A HYPOTHETICAL EXAMPLE

To lay an intuitive foundation for the antidotal variable approach, consider (Figure 1) a hypothetical example of noise pollution based on the overpowering boomboxes prevalent in the 1980s.

Imagine one wants to determine the impact of loud music (D, the treatment) on mental well-being (Y, the outcome). Individuals with initially *high-medium* stress (well-being level $Y = 2$), the treatment group (D=1), listen to loud music to reduce their stress and gain well-being so that $Y = 4$. Others with *low-medium* stress (well-being level $Y = 3$), the control group (D=0), may not need to, but may find the resulting loud boombox music (to them noise) *highly stressful resulting in* well-being $Y = 1$. The difference in well-being associated with 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), after the boombox is played (i.e., 4 -1=3), 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, or welfare*

$Y = 2$) differs from the control group's (*low-medium, or* $Y = 3$). The second results from violation of *SUTVA* caused by treatment spillover. 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 lowering the well-being (i.e., 1-3 = -2, the spillover effect). These are illustrated in the middle column of Figure 1.

	Antidote: Earplugs	
Treatment: Loud Music	**W=1 (no earplug) No Antidote**	**W=0 (earplug) Antidote**
D=0 No Treatment (no loud music)	• *Without Treatment*: low-medium stress $\bar{Y}n_2$ = 3 • No direct loud music (but overhears loud music from others), no earplugs • *With Spillover*: high stress $\bar{Y}n_2$ = 1 • (Subsample: n_2)	• *Without Treatment*: low-medium stress $\bar{Y}n_4$ = 3 • No loud music, earplugs • *With Treatment*: low-medium stress $\bar{Y}n_4$ = 3 • (Subsample: n_4)
D=1 Treatment (loud music)	• *Without Treatment:* High-medium stress $\bar{Y}n_1 = 2$ • Direct loud music, no earplug • *With Treatment:* Low stress $\bar{Y}n_1 = 4$ • (Subsample: n_1)	• *Without Treatment:* High-medium stress $\bar{Y}n_3 = 2$ • Direct loud music, earplugs • *With Treatment:* High-medium stress $\bar{Y}n_3 = 2$ • (Subsample: n_3)

Figure 1 The Impact of Loud Boombox Music on Stress: Subsample classifications n_1, n_2, n_3, and n_4 are defined in the text. Treatment (ATT): $\beta_T = \bar{Y}n_1 - \bar{Y}n_3 = 4 - 2 = 2$; Selectivity: $\theta = \bar{Y}n_3 - \bar{Y}n_4 = 2 - 3 = -1$; Spillover: $\delta = \bar{Y}n_2 - \bar{Y}n_4 = 1 - 3 = -2$

Now suppose earplugs were distributed randomly to a subsample of individuals, and for this example, everyone receiving them uses them. 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. 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 in which $Y = 2$). They now have *low* stress ($Y = 4$). Second are *low-medium* ($Y = 3$) 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 (* $Y = 1$*)*. Third are *high-medium* stressed

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, or* $Y = 2$). 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 *or* $Y = 3$.

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, i.e., $Y_{n_1} - Y_{n_3} = 4 - 2$ (*low* and *high-medium* stress) is the effect of listening to loud music for those who listen to loud music. If the distribution of earplugs is random, this difference would represent the ATT (or the ATE if there is no essential heterogeneity in the treatment effect). 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 non-treatment averages. This represents the *SELB* (i.e., *high-medium* minus *low-medium, or* $Y_{n_3} - Y_{n_4} = 2 - 3 = -1$). Finally, the difference between groups n_2 and n_4 is the *ASEU* (i.e., *high* minus *low-medium* stress, *or* $Y_{n_2} - Y_{n_4} = 1 - 3 = -2$). 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 *ATE* or *ATT* as well as both the *ASEU* and *SELB*. It is noteworthy that, unlike the *AV* approach, the naïve simple difference in outcomes between the treated and no-treatment groups (4-1=3), and the DID type treatment effect estimate, namely the difference in the potential treatment – no-treatment differences for the treatment and no-treatment groups *((4-2)-(1-3)=4)* are clearly biased.

3. A SPILLOVER AUGMENTED POTENTIAL OUTCOME FRAMEWORK

Consider a cross-sectional setting where some units receive treatment and others do not. The untreated units may experience spillovers—either fully or in attenuated form—from treated peers. Thus, lack of direct treatment does not imply lack of exposure. Similarly, the treated units may also receive additional exposure

through spillovers from other treated peers, causing their outcomes to reflect both direct and indirect effects. Spillovers can therefore occur from treated to untreated units and between treated units themselves. Each unit i thus occupies one of four states: (1) treated without spillover, (2) treated with spillover, (3) untreated with spillover, and (4) untreated without spillover.

Let $D_i \in \{0,1\}$ denote unit i's treatment status, where $D_i = 1$ indicates receipt of treatment and $D_i = 0$ indicates no treatment. Let $S_i \in \{0,1\}$ denote spillover exposure, where $S_i = 1$ implies exposure to spillover and $S_i = 0$ implies no such exposure. Each unit therefore belongs to one of four treatment-spillover states, represented by the ordered pair $((D_i, S_i) \in \{(1,0),(1,1),(0,1),(0,0)\}$. The corresponding potential outcomes are $Y_i(1,1)$, $Y_i(1,0)$, $Y_i(0,1)$, and $Y_i(0,0)$, where, $Y_i(1,1)$ denotes the outcome when unit i receives both treatment and spillover; $Y_i(1,0)$ when it only receives the treatment; $Y_i(0,1)$ when it only receives the spillover; and $Y_i(0,0)$ when it receives neither. For notational simplicity lets define $Y_i(1^+) = Y_i(1,1)$, $Y_i(1) = Y_i(1,0)$, $Y_i(1^S) = Y_i(0,1)$ and $Y_i(0) = Y_i(0,0)$.

We focus on three parameters: the average treatment effect (*ATE*), the average spillover effect on the untreated (*ASEU*), and the selectivity bias (*SELB*)). While selection on gains, arising from essential heterogeneity, as illustrated in the Loud Boombox example, may be present, it is not a parameter of interest. We account for it but do not estimate it. Accordingly, the parameters of interest are:

$$ATE = E[Y_i(1) - Y_i(0)]$$

$$ASEU = E[(Y_i(1^S) - Y_i(0))|D_i = 0] = E[Y_i(1^S)|D_i = 0] - E[Y_i(0)|D_i = 0]$$

$$SELB = E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

3.1 Corresponding Regression Formulation

To specify the regression model, we express the potential outcomes as a linear function of covariates, with regression errors capturing unobserved determinants and idiosyncratic shocks. Let

$$Y_i(1) = \mu_0 + \tilde{\beta}_{Ti} W_i + \omega_i^0 \quad (1a)$$

$$Y_i(1^S) = \mu_0 + \widetilde{\delta_\iota} W_i + \omega_i^0 \quad (1b)$$

$$Y_i(0) = \mu_0 + \omega_i^0 \quad (1c)$$

$$Y_i(1^+) = \mu_0 + \tilde{\beta}_{Ti} W_i + \widetilde{\delta_\iota} W_i + \omega_i^0 \quad (1d)$$

The coefficient μ_0 denotes the intercept, $\tilde{\beta}_{Ti}$ captures unit i's treatment effect, and $\widetilde{\delta_i}$ represents the spillover effect. The error term ω_i^0 absorbs unobserved determinants, with $E[\omega_i^0] = 0$. We include W_i to denote unit i's antidote exposure, where $W_i = 1$ indicates no antidote and $W_i = 0$ nullifies both treatment and spillover effects. For purpose of this derivation, we here assume $W_i = 1$, the no antidote status deferring a detailed treatment of the $W_i = 0$ case to later sections.

The observed outcome takes the following form: if individual i receives the treatment ($D_i = 1$), then two cases arise—either the individual is also exposed to spillovers ($S_i = 1$) or not ($S_i = 0$). The observed outcome reflects the corresponding potential outcome from this treatment–spillover combination.

$$Y_i(1,S) = Y_i(1,1)S_i + Y_i(1,0)(1 - S_i)$$

Substituting the values from (1a, 1d)

$$Y_i(1,S) = (\mu_0 + \tilde{\beta}_{Ti}W_i + \widetilde{\delta_i}W_i + \omega_i^0)S_i + (\mu_0 + \tilde{\beta}_{Ti}W_i + \omega_i^0)(1 - S_i)$$

Simplifying

$$Y_i(1,S) = \mu_0 + \tilde{\beta}_{Ti}W_i + \widetilde{\delta_i}W_iS_i + \omega_i^0 \tag{2}$$

Similarly, if individual i does not receive the treatment ($D_i = 0$), two cases arise: either i experiences a spillover ($S_i = 1$) or not ($S_i = 0$). The observed outcome corresponds to the potential outcome associated with the spillover status.

$$Y_i(0,S) = Y_i(0,1)S_i + Y_i(0,0)(1 - S_i)$$

Substituting the values from (1b-1c)

$$Y_i(0,S) = (\mu_0 + \widetilde{\delta_i}W_i + \omega_i^0)S_i + (\mu_0 + \omega_i^0)(1 - S_i)$$

Simplifying

$$Y_i(0,S) = \mu_0 + \widetilde{\delta_i}W_iS_i + \omega_i^0 \tag{3}$$

Given the two potential outcomes of i with treatment ($D_i = 1$) and without treatment ($D_i = 0$), the observed outcome is

$$Y_i = Y_i(1,S)D_i + Y_i(0,S)(1 - D_i) \tag{4}$$

$$Y_i = (\mu_0 + \tilde{\beta}_{Ti}W_i + \widetilde{\delta_i}W_iS_i + \omega_i^0)D_i + (\mu_0 + \widetilde{\delta_i}W_iS_i + \omega_i^0)(1 - D_i) \tag{5}$$

$$Y_i = \mu_0 + \tilde{\beta}_{Ti}W_iD_i + \widetilde{\delta_i}W_iS_i + \omega_i^0 \tag{6}$$

where the observed outcome depends on the treatment, spillover, and antidote exposures. The treatment and spillover enter the outcome equation independently so that the treatment status provides no information about the spillover exposure. To identify the relevant parameters we impose the following assumptions.

Assumption 1: *Spillovers only occur from the treated units to the untreated units (treated-to-untreated), implying* $S_i = (1 - D_i)$.

Assumption 2 (consistency and overlap): *For each unit* i*, the observed outcome satisfies* $Y_i = Y_i(D_i, S_i)$*, meaning the observed outcome corresponds to the potential outcome under the treatment and spillover actually received; and for every* i*, the probabilities of the treatment and spillovers are strictly between 0 and 1, i.e.,* $0 < P(D_i), P(S_i) < 1$.

Assumption 1 imposes a specific structure on the spillover that implies the untreated units always receive a spillover. As a result, the potential outcome $Y_i(0,0)$, an untreated unit without a spillover is excluded. The assumption also rules out spillovers among treated units, eliminating $Y_i(1,1)$. Thus, only two potential outcomes remain: $Y_i(1) = Y_i(1,0)$, the outcome under treatment without spillover; $Y_i(1^S) = Y_i(0,1)$, the outcome under spillover without treatment

Under *Assumptions 1 and 2*, equation (6) can also be expressed as

$$Y_i = \mu_0 + \tilde{\beta}_{Ti} W_i D_i + \tilde{\delta}_i W_i (1 - D_i) + \omega_i^0 \qquad (6')$$

3.1.1 Identification

With *Assumptions 1 and 2*, the observed outcome can be expressed in a switching regression form:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(1^S)$$

Rearranging the equation above yields:

$$Y_i = Y_i(1^S) + [Y_i(1) - Y_i(1^S)] D_i$$

Simplifying

$$Y_i = Y_i(1^S) + \Delta_i D_i$$

where $\Delta_i = [Y_i(1) - Y_i(1^S)]$ represents the naïve difference in i's outcome with and without the *direct* treatment. Note that Δ_i does not represent the treatment effect for unit i.

As illustrated in *Appendix A* (***Theorem A.1***), the naïve difference in observed averages outcomes between the treated and untreated groups (Δ) can be expressed as

$$\Delta = ATE - ASEU + (1 - \pi)\ (ATT - ATU) + SELB$$

where π is the proportion of units that receive treatment, *ATT* and *ATU* are the average treatment effects on the treated and untreated, defined as $ATT = E[Y_i(1)|D = 1] - E[Y_i(0)|D_i = 1]$ and $ATU = E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0]$. Here Δ differs from the common causal inference framework used to identify *ATE* because it contains the *ASEU* term since *SUTVA* is not satisfied given that there are spillovers. Hence, standard causal inference methods yield a biased and inconsistent estimator of *ATE*. Moreover, standard inference methods cannot identify the *ASEU* and *SELB*.

3.2 Heterogeneous Effects and Selectivity

The coefficients $\tilde{\beta}_{Ti}$ and $\tilde{\delta}_i$ in (6) and (6') vary by unit i. Let η_i and v_i represent the heterogeneity in the treatment effect and spillover effect, respectively, such that:

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

$$\tilde{\delta}_i = \delta + v_i$$

where $E[\eta_i] = 0$ and $E[v_i] = 0$. With these, (6) can be rewritten as

$$Y_i = \mu_0 + \beta_T W_i D_i + \eta_i W_i D_i + \delta W_i S_i + v_i W_i S_i + \omega_i^0 \quad (7)$$

It is possible that those who select into the treatment differ in average outcome from those who do not in absence of the treatment or spillover. This can lead to a selectivity bias. To incorporate this bias, we decompose unit i's ω_i^0into two parts: θ_i which reflects the inherent difference from those not treated and ψ_i, the remaining component. As such

$$\omega_i^0 = \theta_i D_i + \psi_i \quad (8)$$

where the first term $\theta_i D_i$is the selectivity component and the latter term ψ_i is the remainder such that $E[\psi_i] = 0$. Further assume, $\theta_i = \theta + \phi_i$.

As such, equation (7) can be written as

$$Y_i = \mu_0 + \beta_T W_i D_i + \eta_i W_i D_i + \delta W_i S_i + v_i W_i S_i + \theta D_i + \phi_i D_i + \psi_i \quad (9)$$

Further simplification and redefining $u_i = \eta_i W_i D_i + v_i W_i S_i + \phi_i D_i + \psi_i$ *in* (9) yield the corresponding estimable regression equation.

$$\begin{aligned} Y_i &= \mu_0 + \beta_T W_i D_i + \delta W_i S_i + \theta D_i + u_i \\ &= \mu_0 + \beta_T W_i D_i + \delta W_i (1 - D_i) + \theta D_i + u_i \end{aligned} \quad (10)$$

3.3 The Effect Parameters and Regression Coefficients

Given the structure in (10), the effects of interest (*ATE or ATT*, *ASEU*, *SELB*) can be expressed as regression parameters. To maintain consistency, we include S_i in the conditioning. This does not alter the interpretation since D_i fully determines S_i. However, this specification helps identify whether the subsample of interest belongs to the spillover exposure group. As such, in a framework without any *AV* (equivalent to $W_i = 1$ for all units), the parameters in equation (10) can be used to express various effects in the following way (see *Appendix A, Lemma A1-A3*).

$$ATE = E[Y_i(1)] - E[Y_i(0)] = \beta_T W_i \tag{11a}$$

$$\begin{aligned} ASEU &= E[Y_i(1^S)|D = 0, S_i = 1, W_i = 1] - E[Y_i(0)D = 0, S_i = 1, W_i = 1] \\ &= \delta W_i + E[v_i W_i|D = 0, S_i = 1, W_i = 1] \end{aligned} \tag{11b}$$

$$ATT = \beta_T W_i + E[\eta_i W_i D_i|D = 1, S_i = 0\ W_i = 1] \tag{11c}$$

$$SELB = E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1]$$

$$E[\theta_i D_i|D = 1, S_i = 0\ W = 1] - E[\theta_i D_i|D = 0, S_i = 0, W_i = 1] = E[\theta_i|D_i = 1] = 0 \tag{11d}$$

Also (10) allows us to express the naïve observed difference Δ as

$$\begin{aligned} \Delta &= \mathrm{E}[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1] \\ &= \beta_T - (\delta - E[v_i W_i(1 - D_i)|D = 0, S_i = 1, W_i = 1]) + E[\eta_i W_i D_i|D = 1, S_i = 0, W_i = 1] + \theta \end{aligned} \tag{12}$$

Appendix B also shows, $E[\eta_i W_i D_i|D = 1, S_i = 0, W_i = 1] = (1 - \pi)\ \{E[\eta_i|D = 1, S_i = 0, W_i = 1] - E[\eta_i|D = 0, S_i = 1, W_i = 1]\}$. Substituting in (*11a-11d)* into yields

$$\begin{aligned} &E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1] = \Delta \\ &= ATE - ASEU + (1 - \pi)[ATT - ATU] + \theta \end{aligned} \tag{13}$$

Equation (10) implies that the difference in observed outcomes does not identify any of the effects.

3.4 Introducing the Antidotal Variable (*AV*)

Although W_i is included above, so far we considered the case where no unit receives the antidote, i.e., $W_i = 1$ for all units. We now embed what we call an *AV* into the framework. This means some units now are exposed to the antidote. For those units $W_i = 0$. Those exposed could belong to the treatment group or

control group. As such, an *AV* nullifies the *effect* of the treatment, whether it be the actual treatment effect itself or the spillover. Thus, unlike the regular instrumental variable framework, which designates a variable that determines potential treatment *participation*, the *AV* abrogates the potential *effect* of a treatment. As such, this approach differs from the standard IV method in that the *AV* relates to the *impact*, in that it eradicates the treatment *effect*, but does not determine the treatment status itself, as in standard IV. We incorporate the *AV* as $\beta_{Ti} = \tilde{\beta}_{Ti} W_i$ and $\delta_i = \tilde{\delta}_i W_i$ so that $\beta_{Ti} = 0$ and $\delta_i = 0$ when the antidote is applied, i.e., when $W_i = 0$.

3.5 Identifying Assumptions Regarding the *AV*

To achieve point identification we make four additional assumptions relating to the antidotal variable W_i:
Assumption 3: *Exposure to the AV fully eliminates both the treatment and spillover effects,*

$$\beta_{Ti} = \begin{cases} \tilde{\beta}_{Ti}, & W_i = 1 \\ 0, & W_i = 0 \end{cases} \text{ and } \delta_i = \begin{cases} \tilde{\delta}_i, & W_i = 1 \\ 0, & W_i = 0 \end{cases}$$

Assumption 4 (overlap): *For every i, the probability of receiving antidote assignment W_i is strictly between 0 and 1, i.e., ., $0 < P(W_i) < 1$.*

Assumption 5 (unconditional unconfoundedness or ignorability): *The antidote assignment W_i is independent of potential outcomes $Y_i(1), Y_i(1^S), Y_i(0), Y_i(1^+)$ i.e., $[Y_i(1), Y_i(1^S), Y_i(0), Y_i(1^+)] \perp W_i$.*

Assumption 6 (no essential heterogeneity): *The heterogeneities in the treatment effect, the spillover effect and the selectivity bias are mean independent of treatment assignment (D_i) and antidote assignment (W_i), i.e., $E[\eta_i | D_i, W_i] = E[\eta_i] = 0$, $E[v_i | D_i, W_i] = E[v_i] = 0$ and $E[\phi_i | D_i, W_i] = E[\phi_i] = 0$.*

3.6 Identification of the *ATE or ATT* (β_T), *ASEU* (δ), *SELB* (θ) with the *AV*

Proposition 1a: *Assumptions 1–5 imply that the naïve difference in the average observed outcome between the treatment group without the antidote ($D_i = 1, S_i = 0, W_i = 1$) and the treatment group with the antidote ($D_i = 1, S_i = 0, W_i = 0$) identifies the ATT, i.e.,*

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

The formal proof of this proposition is provided in *Appendix B*. The intuition behind this result is that the difference within the treatment group does not suffer from selectivity bias. Moreover, because *Assumption 1* and *Assumption 5* imply that there is no spillover among treated units, and W_i is unrelated to the respective

potential outcomes, the distribution of heterogeneous treatment effects $\tilde{\beta}_{Ti}$ is the same for the subgroups $(D_i = 1, S_i = 0, W_i = 1)$ and $(D_i = 1, S_i = 0, W_i = 0)$. As a result, the difference in average observed outcomes between these two groups identifies the *ATT*.

Proposition 1b: *Assumptions 1–6 imply that the observed difference in the average observed outcome between the treatment group without the antidote* $(D_i = 1, S_i = 0, W_i = 1)$ and *treatment group with the antidote* $(D_i = 1, S_i = 0, W_i = 0)$ *identifies the ATE, i.e.,*

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

The formal proof of this proposition is provided in *Appendix B*. The intuition follows the same logic as above, except that the additional *no essential heterogeneity* assumption implies that distribution of heterogeneous treatment effects $\tilde{\beta}_{Ti}$ is independent of treatment status. As a result, the same difference can now be interpreted as the *ATE*.

Proposition 2: *Assumptions 1–5 imply that the naïve difference in the average outcome between the control group without the antidote* $(D_i = 0, S_i = 1, W_i = 1)$ *and the control group with the antidote* $(D_i = 0, S_i = 1, W_i = 0)$ *identifies the ASEU, i.e.,*

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

The formal proof of this proposition is provided in *Appendix B*. The intuition behind this result is that the difference within the control group is free from selectivity bias, and as per *assumption 5*, v_i is independent of W_i. The only effect experienced by this group is treatment spillover. As a result, this difference identifies the *ASEU*.

Proposition 3: *Assumptions 1–6 imply that the naïve difference in the average outcome between the treatment group with the antidote and the control group with the antidote identifies the SELB, i.e., with*

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

The formal proof of this proposition is provided in *Appendix B*. The intuition behind this result is that the difference between antidotal groups in the treatment and control groups does not reflect any treatment or spillover effects. These groups essentially mimic the no-treatment outcomes for both the treatment and control groups. Thus, the difference in their averages measures the spillover effects or *SELB*.

3.7 Identification in the Presence of Concurrent Treatments

One advantage of the *AV* approach is that neither the treatment effect (*ATT* or *ATE*) nor the *ASEU* (SUTVA bias) are confounded by concomitant treatments as long as the *AV* is unrelated to those other concomitant treatments. See *Appendix C* for the proof.

3.8 The Imperfect Antidote Case

An antidote may sometimes be imperfect. This can arise either when the antidote fails to completely nullify the effects of the treatment or when there is imperfect compliance with regard to antidote uptake. In our prior example, a defective earplug fails to provide complete noise protection.

Define τ to depict the antidote's efficacy such that $0 \leq \tau \leq 1$. When $\tau = 1$, the antidote $W_i = 0$ fully nullifies the effects of the treatment and spillover. Conversely, when $\tau = 0$, the antidote has no effect in neutralizing these effects.

Given an imperfect antidote, point estimates for the *ATE* (or *ATT*), *ASEU* and *SELB* cannot be identified. However, one can bound these estimates based on additional widely accepted assumptions such as *Monotone Treatment Response* (*MTR*), *Monotone Treatment Selection* (*MTS*) and *Optimal Treatment Selection* (*OTS*) (Manski 1997; Manski and Pepper 2000). *MTR* assumes the treatment never harms. *MTS* implies the treated group's average outcome always differ from that of the control group. *OTS* suggests the treated group benefits on average, while the control group is harmed by the treatment. *Appendix D* formally defines each of these assumptions. Figure 2 (top panel and bottom panel) presents two sets of bounds (derived in *Appendix D*) with two sets of additional assumptions.

3.9 Immunity: When antidotes are immutable

Immutable antidotes are traits assigned by nature. They are innate characteristics that entities do not choose or develop, yet these traits shield them from the effect of treatment. Because of this

inherent randomness of this type of antidote, the immutability assumption implies that $\eta_i \perp\!\!\!\perp W_i$, $v_i \perp\!\!\!\perp W_i$ and $\phi_i \perp\!\!\!\perp W_i$ are generally satisfied, that is $E[\tilde{\beta}_{Ti}|W_i = 1] = E[\tilde{\beta}_{Ti}|W_i = 0]$,

Assumptions	Bounds
More Y is preferred than less Y; and $Y_i \geq 0$; positive monotone treatment response (MTR); no essential heterogeneity.	$E[Y_i\|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] - E[Y_i\|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$ $\leq \beta_T \leq$ $E[Y_i\|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1]$
	$E[Y_i\|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] - E[Y_i\|D_i = 0, S_i = 1, W_i = 0, \tau < 1]$ $\leq \delta \leq$ $E[Y_i\|D_i = 0,\ S_i = 1, W_i = 1, \tau < 1]$
	$-E[Y_i\|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1]$ $\leq \theta \leq$ $E[Y_i\|D_i = 1, S_i = 0,\ W_i = 0, \tau < 1]$
More Y is preferred than less Y; and $Y_i \geq 0$; Monotone Treatment Selection (MTS) and Optimum Treatment Selection (OTS); no essential heterogeneity.	$E[Y_i\|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] - E[Y_i\|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$ $\leq \beta_T \leq$ $E[Y_i\|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] - E[Y_i\|D_i = 0,\ S_i = 0, W_i = 0, \tau < 1]$
	$E[Y_i\|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] - E[Y_i\|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$ $\leq \delta \leq$ $E[Y_i\|D_i = 0, \quad S_i = 1, W_i = 1, \tau = 1] - E[Y_i\|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1]$
	$-E[Y_i\|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$ $\leq \theta \leq$ $E[Y_i\|D_i = 1, S_i = 0, W_i = 0, \tau < 1] - E[Y_i\|D_i = 0, S_i = 1, W_i = 0,\ \tau < 1]$

Figure 2. Assumptions and identified bounds: Here τ $(0 \leq \tau \leq 1)$ represents the efficacy of the treatment. $\tau = 1$ means the antidote is fully effective and $\tau = 0$ means antidote is fully ineffective. See *Appendix D* for details. As these expressions suggest, the length of the bound is dependent on the efficacy of the antidote.

$E[\widetilde{\delta_\iota}\,|W_i = 1] = E[\widetilde{\delta_\iota}|W_i = 0]$, and $E[\theta_i\,|W_i = 1] = E[\theta_i|W_i = 0]$. In other words, immutability means *ATE* or *ATT* is the same for the treated units with or without antidote, and the *ASEU* is also the same for the untreated units across both groups.

In our *AV* application middle aged women are typically unaffected by paid family leave because of demographic considerations. Women in that age group tend not to have young children or sick parents. For them, immunity W_i is likely independent of η_i and v_i because there is no selection on the gain.

Factors such as age, gender, and institutional preconditions are examples of potentially immutable variables that could act as antidotes which in many cases are unrelated to the effectiveness of either the treatment or potential spillovers. Parameters β_T, δ and θ are all identified when using an immutable antidotal variable.

3.10 Nonrandom Antidote Adoption

As of now, we assume that W_i is administered randomly and hence is independent of η_i, ϕ_i, v_i, and ω_i^0. However, this assumption is violated when respondents select whether to take the antidote based on the gain. (Immutable antidotes are innate so that selection based on the gain is not relevant.) In *Appendix E* we examine the implications of relaxing each part of this assumption. To summarize, it is still possible to identify the *ASEU* and *SELB*, but one cannot identify the *ATE* or *ATT* when the treatment effect heterogeneity η_i is mean dependent on W_i. It is, still possible to identify both the treatment effect and the selectivity bias, though not the *ASEU* when the spillover effect heterogeneity v_i is mean dependent on W_i. Finally, the selectivity bias cannot be identified if ϕ_i is mean dependent on W_i, but the treatment effect and the bias that is caused by a violation of the *SUTVA can still be identified*. However, as illustrated in the next section, one is able to test mean independence of W_i and u_i.

3.11 Testing Whether the Antidote Assignment (W) is Random

One generally cannot identify all three parameters of interest (the treatment effect (β_T), the bias arising from a *SUTVA* violation (δ), and the selectivity bias (θ) from a single cross-section when W_i is mean dependent of u_i. However, as shown in *Appendix F*, an advantage of the *AV* method is that it allows testing whether W_i is correlated with u_i using a cross-section of data where the

treatment and control groups can be observed without treatment. A natural source of such data is a cross-section from the pre-treatment period. If available, one can run the following simple regressions to implement the test:

$$\text{Treated group: } Y_i = \gamma_0 + \gamma_1 W_i + \zeta_i \tag{14a}$$

$$\text{Untreated group: } Y_i = \gamma_0' + \gamma_1' W_i + \zeta_i' \tag{14b}$$

within the treated and untreated groups. When $\gamma_1 = 0$ and $\gamma'_1 = 0$, and remain the same post-treatment, one can infer that non-randomness of W_i does not affect the parameter identification, as it should not, given there is no treatment to nullify provided the insignificance remains valid in post-treatment period. If $\gamma_1 \neq 0$ and $\gamma'_1 = 0$, then one identifies *ASEU* (δ), but the *ATE* or *ATT* and *SELB* (β_T and θ) cannot be identified. Conversely, if $\gamma_1 = 0$ and $\gamma'_1 \neq 0$, then one identifies β_T, but cannot identify *ASEU* (δ) and *SELB* (θ). If $\gamma_1 \neq 0$ and $\gamma'_1 \neq 0$, then none of the parameters of interest are identified. However, even in this pre-treatment test, for causal inference in post-treatment period one must assume that the insignificance of γ_1 and γ'_1 holds in the post-treatment period. Unlike this approach, standard IV validity cannot be tested using pre-treatment data, as any variation in the IV without variation in treatment (which remains zero in this case) implies weak correlation and violates the relevance condition.

3.12: When W_i nonrandom:

When the antidote is likely not randomly assigned, or the test above fails, one can invoke conditional ignorability—a weaker assumption than unconditional ignorability.

Assumption 7 (conditional unconfoundedness or ignorability): *Given a set of covariates X_i, the antidote assignment W_i is independent of potential outcomes $Y_i(1)$, $Y_i(1^S)$, $Y_i(0)$, i.e., $[Y_i(1), Y_i(1^S), Y_i(0)] \perp W_i|X_i$ where X_i is the vector of covariates.*

When *Assumption 5* is not satisfied, but *Assumption 7* is, one can still retrieve all three effects. The corresponding regression equation would be modified to

$$Y_i = \mu_0 + \beta_T W_i D_i + \delta W_i(1 - D_i) + \theta_i D_i + X'_i\rho + u_i \tag{15}$$

where ρ is the vector of corresponding coefficients.

4. SIMULATIONS

To validate the approach and test for consistency of the estimators, we conduct several simulation exercises. First, we simulate data based on a process where W_i is randomly assigned, meaning it is independent of D_i and the error term u_i. Second, we relax the independence of W_i and D_i while generating the data. We use various parameter values and sample sizes (100, 1000, 10000, 100000, 1000000). See *Appendix G* for the details. The results (*Appendix Table G.1*) show that the method reproduces the estimates which approach their true values as sample size grows, implying the consistency of the estimators.

It is not necessary that W_i is independent of D_i. This is because each parameter is identified based on subsamples in which one of either W_i or D_i remain fixed, while the other varies. See *Appendix G* for details on the reasoning and simulation results (*Appendix Table G.2*).

5. 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 of analysis estimated the difference in take-up from before to after the law in California, relative to a control. We replicate this type of 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 potentially inherent in DID. Moreover, unlike the *AV* approach, previous DID studies did not estimate *ASEU* and *SELB*.

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.

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.[1] Table 1 presents the summary statistics for leave taking hours before and after CPFL became effective. Between 2001-2003 and 2004-2006, women in California increased leave taking, whereas in other states women's leave taking decreased (from 1.21 hours in 2001-03 to 1.59 hours in 2004-2006 in California compared to a decrease from 2.06 hours in 2001-2003 to 1.91 hours in 2004-2006 in other states). 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 1.35 versus 1.06 for the old in California and 1.98 versus 1.88 for the old in other states).

Table 1: Hours of Leave Taking

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

Source: Nonmilitary respondents in the IPUMS-CPS (AESC rounds); authors' computations. Hours of leave taking are defined as the difference in a worker's usual weekly work hours and actual work hours.

Table 2 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 1. California's incidence of leave taking rises to 3.3 percent in 2004-06 from 2.7 percent in 2001-03. Other states

[1] 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.

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 1 and 2 we denote California as receiving treatment D_i = 1 and the other states as untreated D_i = 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 because they typically do not have young children and hence are unlikely to take family leave.[2] We assign W_i = 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_i = 1 for those women 25-40. The W_i = 0 group (45-55 year old) 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.

Table 2: Incidence of leave (proportion)

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

Source: Nonmilitary respondents in the IPUMS-CPS (AESC rounds); authors' computations. For each respondent leave taking is a binary variable which takes value 0 if the respondent is at work, and 1 if the respondent has a job but is on leave at the time of the interview.

DID estimates CPFL's effect by computing the before and after differences between California and the rest of the US. This *ATT* (or *ATE* if California has the same potential distribution of treatment effects as other states) amount to 0.53 (computed as (1.59-1.21) - (1.91-2.06) = 0.53) for hours and a 0.009 (computed as (0.033-0.027) - (0.038-0.041) = 0.009) increase in the incidence of taking a leave.

[2] 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 for this reason we drop from the sample. Further, the incidence and amount of parent-motivated leave taking for 45-55 and 25-40 years 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. Nevertheless, we also bound our estimates using the approach outlined for imperfect antidotes.

The antidotal variable approach entails four groups. Group 1 (D_i = 1 and W_i = 1) are those who receive treatment and do not have the antidote (age 25-40 in California). Group 2 (D_i = 0 and W_i = 1) constitute those not receiving the antidote (age 25-40) in the other (control) states. Group 3 (D_i = 1 and W_i = 0) comprise those receiving treatment but getting the antidote (age 45-55 in California). Finally, group 4 (D_i = 0 and W_i = 0) are those in the control group who get the antidote (are 45 – 55 in other states). 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 1 (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 2 (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 (*ATE*) as $\bar{Y}_1$ - $\bar{Y}_3$, *SELB* as $\bar{Y}_3$ - $\bar{Y}_4$, and *ASEU* bias as $\bar{Y}_2$ - $\bar{Y}_4$. Thus, the *ATE* is 0.91 hours, the *SELB* is -0.82 hours, and the *ASEU* is 0.05 hours. For incidence, the values are 0.014, -0.012, and 0.002 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. 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 (0.53) is about half the antidotal variable estimate (0.91) for hours

leave, and 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 if other factors change in California and the control states. 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% (computed as (-.009-(-.012))/(.009-.014)=0.60) of the discrepancy when considering incidence. As we will show shortly, we observe no statistically significant treatment or *SUTVA* effects when examining 2001-2003, an expected placebo test.

Table 3: Testing mean independence of the antidotal variable W (2001-03 subsample)

VARIABLES	(1) HRABSNT D=1	(2) HRABSNT 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

Source: IPUMS-CPS (AESC rounds); authors' computations.
HRABSNT=hours of leave; DLEAVE=incidence of leave. Robust standard errors in parentheses.

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 (10). However, to do so, we first test whether the antidotal variable W_i is mean independent of u_i. Mean independence implies we can obtain unbiased and consistent estimates. We utilize hours and incidence of leave

as dependent variables in an OLS regression on W_i using 2001-03 data, the period prior to the policy implementation. For each dependent variable we run an OLS regression on W_i, once for California and once for the other states. Table 3 presents the results. An insignificant coefficient implies mean independence between the antidotal variable W_i and the error term u_i assuming the results carry over in the post-treatment period. While one might expect average leave-taking to differ between young and old during 2001–03 in both California and non-California states, the data in Table 3 show no significant differences. As illustrated, the coefficients for the antidotal variable W_i are insignificant in each of these regressions. Thus, the results suggest that W_i is mean independent of u_i. Based on this finding we proceed to estimate the *ATE*, *SELB*, and *ASEU* using the antidotal variable method.

Table 4 (columns 2 and 4) presents the causal effect of CPFL on weekly leave taking hours obtained from regressions based on (10) and (15). Young women in California take 0.91 hours more leave, which is the same as previously computed (based on Table 1) because this regression with no covariates simply reports differences in mean values between the four groups. 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 a significant effect on the rest of the country's labor market. Of course, in the PFL example, if SUTVA is violated within California—say, through changes in older workers' relative wages—the ATE estimate may be biased. However, this is unlikely, as older workers typically hold different types of jobs, despite similar leave-taking rates.

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 taking exceed zero, and paid family leave does not lower leave taking when implemented. Accordingly, based on formulas given in Figure 2, we construct the following bounds on the ATE, and the

SUTVA and selectivity biases 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.

Table 4: Regression results for the United States

VARIABLES	(1) HRABSNT 2001-03	(2) HRABSNT 2004-06	(3) HRABSNT 2001-03	(4) HRABSNT 2004-06
DW: β_T	-0.229	0.908	-0.274	0.888
	(0.284)	(0.287)	(0.285)	(0.288)
D: θ	-0.630	-0.821	-0.644	-0.819
	(0.243)	(0.223)	(0.244)	(0.222)
W(1-D): δ	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

Source: IPUMS-CPS (AESC rounds); authors' computations.
HRABSNT=hours of leave during last week.
Robust standard errors in parentheses; coefficient of $DW = \beta_T$ (the treatment effect); coefficient of $D = \theta$ (selectivity bias); coefficient of $W(1 - D) = \delta$ (spillover bias).

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 (10) and (15) 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) of Tables 4 and 5, 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 5 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 55 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. The bounds for incidence of leave taking are $0.014 \leq \beta_T \leq 0.039$, $0.002 \leq \delta \leq 0.039$, and $-0.037 \leq \theta \leq 0.025$. 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.

Table 5: Regression Results for the United States

	(1)	(2)	(3)	(4)
VARIABLES	DLEAVE 2001-03	DLEAVE 2004-06	DLEAVE 2001-03	DLEAVE 2004-06
DW: β_T	-0.00590	0.0139	-0.00732	0.0131
	(0.00555)	(0.00577)	(0.00556)	(0.00579)
D: θ	-0.00899	-0.0120	-0.00949	-0.0125
	(0.00479)	(0.00426)	(0.00480)	(0.00427)
W(1-D): δ	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

Source: IPUMS-CPS (AESC rounds); authors' computations.
DLEAVE=incidence of leave.
Robust standard errors in parentheses; coefficient of $DW = \beta_T$ (the treatment effect); coefficient of $D = \theta$ (selectivity bias); coefficient of $W(1 - D) = \delta$ (spillover bias).

One reason for a virtually zero *SUTVA* bias is we compare 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 react to the introduction of California's paid family leave policy, but this reaction is likely muted for those in states distant to California. Utilizing all states but California as the controls possibly lead 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 6 and 7 present these results. As before, we observed a positive treatment effect, essentially the same magnitude as before (0.91 for hours and 0.013 for incidence). Selectivity is larger (-1.41 for hours and -0.038 for incidence) meaning California differs more from its neighboring states than from the whole US. Here, there is a negative *SUTVA* effect (-0.84 for hours and -0.03 for incidence) 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 pre-treatment zero coefficient, compared to the non-zero post-treatment coefficient, implies the possibility of other unobserved confounders, thus exacerbating the difference between California and its neighbors in 2004-2006.

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 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.

6. 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

Table 6: Regression Results for California and Neighboring States

VARIABLES	(1) HRABSNT 2001-03	(2) HRABSNT 2004-06	(3) HRABSNT 2001-03	(4) HRABSNT 2004-06
DW: β_T	-0.229	0.908	-0.243	0.903
	(0.284)	(0.287)	(0.290)	(0.294)
D: θ	-0.0563	-1.403	-0.0576	-1.408
	(0.394)	(0.437)	(0.397)	(0.437)
W(1-D): δ	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

Source: IPUMS-CPS (AESC rounds); authors' computations.
HRABSNT=hours of leave during last week.
Robust standard errors in parentheses; coefficient of $DW = \beta_T$ (the treatment effect); coefficient of $D = \theta$ (selectivity bias); coefficient of $W(1 - D) = \delta$ (spillover bias).

Table 7: 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: β_T	-0.00590	0.0139	-0.00644	0.0133
	(0.00555)	(0.00578)	(0.00557)	(0.00585)
D: θ	-0.000968	-0.0379	-0.00139	-0.0383
	(0.00784)	(0.00967)	(0.00793)	(0.00970)
W(1-D): δ	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

Source: IPUMS-CPS (AESC rounds); authors' computations.
DLEAVE=incidence of leave.
Robust standard errors in parentheses; coefficient of $DW = \beta_T$ (the treatment effect); coefficient of $D = \theta$ (selectivity bias); coefficient of $W(1 - D) = \delta$ (spillover bias).

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 mean independent of the error term, which can be tested using pre-treatment data. This is a weaker assumption than standard IV approaches, which requires a variable related to the treatment but unrelated to the dependent variable directly, 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 who do not receive the antidote. A violation of this assumption prevents identification of the treatment effect.

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 percentage points, 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.

Whereas we apply the AV technique to analyze the California Paid Family Leave program, it potentially has applications beyond this example. To conclude, we provide two examples, one of a perfect antidote related to virus immunity, and the other an imperfect antidote related to lead poisoning, but there are also many other possible examples. First, suppose one wishes to measure a causal effect of exposure to a virus, say hours of sleep, ability to work, or some other outcome. Individuals previously infected by the virus develop antibodies. They are immune to exposure (the treatment), whereas others are susceptible to exposure. The antibodies make no difference in the outcome (sleep or the ability to work) whether a person previously had been infected. As such, past infection with the same virus acts as a perfect antidote to new exposure. Second, suppose one wishes to estimate the causal effect of lead on adult IQ. According to epidemiology literature (Thomas et al. 2011), the effects of lead differ based on the aminolevulinate delta-dehydratase (ALAD) genotype so that individuals with different types of ALAD genotypes experience different effects of lead exposure. At the same time, in the absence of lead exposure, ALAD genotypes variations are not related to IQ. Thus, the different ALAD types act as an antidote to lead exposure. Here ALAD variations provide an imperfect antidote because they do not completely nullifying lead's adverse effects.

Acknowledgements

We thank Joshua Angrist, Robert Basmann, Alfonso Flores-Lagunes, Ivan Korolev, David Slichter and participants in the 2021 Camp Econometrics for extremely valuable comments on an earlier version. We thank three anonymous referees, the editor and associate editor of this journal, Alfonso Flores-Lagunes, Jan Ondrich, Isaac Opper, David Slichter, Marlon Tracey and Tymon Slocznski, as well as participants of the 2023 Camp Econometrics for insightful comments on this version.

Disclosure Statement

The authors report no competing interests to declare.

References

Adamopoulou, Effrosyni (2012) "Peer Effects in Young Adults' Marital Decisions," Working Paper 12-28 Departamento de Economía, Economic Series Universidad Carlos III de Madrid.

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.

Ashenfelter, Orley (1978) "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics* 60(1): 47-57.

Baird, Mathew, John Engberg and Isaac Opper (2023) "Optimal Allocation of Seats in the Presence of Peer Effects: Evidence from Job Training Programs," *Journal of Labor Economics* 41(2):479-509.

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.

Benjamin-Chung, Jade, Benjamin F Arnold, David Berger, Stephen P Luby, Edward Miguel, John M Colford Jr and Alan E Hubbard (2018) "Spillover effects in epidemiology: parameters, study designs and methodological considerations," *International Journal of Epidemiology* 2018, 332–347.

Christafore, David and Susane Leguizamon (2019) "Neighbourhood inequality spillover effects of gentrification," *Papers in Regional Science* 98(3): 1469-1485.

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

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

DiMaggio, Paul and Filiz (2012) "Network Effects and Social Inequality." *Annual Review of Sociology* (38): 93-118.

DiTraglia, Francis J., Camilo García-Jimeno, Rossa O'Keeffe-O'Donovan, and Alejandro Sánchez-Becerra (2023) "Identifying Causal Effects in Experiments with Spillovers and Non-Compliance." *Journal of Econometrics* 235(2): 1589–1624.

Forastiere, Laura, Edoardo M. Airoldiy, and Fabrizia Mealliz (2021) "Identification and estimation of treatment and interference effects in observational studies on networks," *Journal of the American Statistical Association*, 116(534), 901–918.

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.

Hudgens, Michael G., and M. Elizabeth Halloran. (2008). "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103(482): 832–42.

Huber, Martin and Andreas Steinmayr (2021) "A Framework for Separating Individual-Level Treatment Effects from Spillover Effects," *Journal of Business and Economic Statistics* 39(2): 422-436.

Imai, Kosuke, and Zhichao Jiang. (2020) "Identification and sensitivity analysis of contagion effects in randomized placebo-controlled trials." *Journal of the Royal Statistical Society Series A: Statistics in Society* 183, no. 4: 1637-1657.

Imai, Kosuke, Zhichao Jiang, and Anup Malani (2021) "Causal Inference With Interference and Noncompliance in Two-Stage Randomized Experiments." *Journal of the American Statistical Association* 116(534):632–44.
https://www.tandfonline.com/doi/epdf/10.1080/01621459.2020.1775612?needAccess=true

Kang, Ji Young, Areum Lee, Eunsun Kwon and Sojung Park (2022) "The Effects of California Paid Family Leave on Labor Force Participation Among Low-income Mothers One Year after Childbirth" *Journal of Social Policy51*(4): 707 – 727.

Lipsitch, M., Tchetgen Tchetgen, E. J., and Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* 21(3):383.

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

Liu, Lan, Michael G. Hudgens, Bradley Saul, John D. Clemens, Mohammad Ali, and Michael E. Emch (2019a) "Doubly Robust Estimation in Observational Studies with Partial Interference" *Stat* 8(1): e214.

——— (2019b) "Doubly Robust Estimation in Observational Studies with Partial Interference" *Stat* 8(1): e214.

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.

Sävje, Fredrik, Peter M. Aronow, and Michael G. Hudgens (2021) “Average Treatment Effects in the Presence of Unknown Interference.” *The Annals of Statistics* 49(2): 673-701.

Sherman, Lawrence W. and David Weisburd (1995) “General Deterrent Effects of Police Patrol In Crime "Hot Spots": A Randomized, Controlled Trial,” *Justice Quarterly* 12:625-48.

Sobel, Michael E. (2006) “What Do Randomized Studies of Housing Mobility Demonstrate?” *Journal of the American Statistical Association* 101(476): 1398–1407.

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.

Thomas, Deena, Ananya Roy, Howard Hu, Bhramar Mukherjee, Rama Modali, Kavitha Palaniappan, and Kalpana Balakrishnan (2011) "IQ and Blood Lead Levels: Effect Modification by ALAD Amongst Children in Chennai, India." *Epidemiology* 22(1): S135-S136.

VanderWeele, Tyler J., and James M. Robins (2007) "Four types of effect modification: a classification based on directed acyclic graphs." *Epidemiology*: 561-568.

VanderWeele, Tylere J. (2009) “On the Distinction Between Interaction and Effect Modification,” *Epidemiology*: 863-871.

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.

Wilke, Anna M., Donald P. Green, and Jasper Cooper. (2020). "A placebo design to detect spillovers from an education–entertainment experiment in Uganda." *Journal of the Royal Statistical Society Series A: Statistics in Society* 183, no. 3: 1075-1096.

Supplementary Appendices

Causal Inference Using Antidotal Variables

Appendix A: The Identification Problem

Theorem A.1: *If* Δ *represents the naïve difference in the average outcome between the treated and the untreated group with spillover, then*

$$\Delta = ATE - ASEU + (1 - \pi)(ATT - ATU) + SELB$$

Proof:

Section 3.1.1 suggests that for unit i, the potential naïve difference in outcomes between the treated and untreated states is $\Delta_i = Y_i(1) - Y_i(1^S)$. However, this difference is unobservable since either $Y_i(1)$ or $Y_1(1^S)$ is always missing. Moreover, because $Y_i(1^S) \neq Y_i(0)$, Δ_i does not capture individual i's treatment effect. In our setting, in the absence of an antidote, we observe the naïve difference in the average outcome of the treatment group and control group. This difference can be written as

$$\Delta = E[Y_i(1)|D = 1, S_i = 0, W = 1] - E[Y_i(1^S)|D = 0, S_i = 1, W = 1] \qquad (A.1)$$

where, $W_i = 1$ indicates absence of the antidote, the default case as in standard treatment effect analysis. The averages $E[Y_i(1)|D = 1, S_i = 0, W = 1]$ represents the average outcome Y for the treated group, and $E[Y_i(1^S)|D = 0, S_i = 0, W = 1]$ represents the average outcome Y for the untreated group exposed to treatment spillover.

The definition of *ATE* is

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \qquad (A.2)$$

Neither $E[Y_i(1)]$ nor $E[Y_i(0)]$ is observed for the entire population. However, if π represents the fraction of the population receiving treatment, $E[Y_i(1)]$ can be rewritten as:

$$E[Y_i(1)] = \pi E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] + (1 - \pi)E[Y_i(1)|D_i = 0, S_i = 1, W_i = 1] \quad (A.3)$$

Rearranging (*A.3*) and substituting it into (*A.2*) yields:

$$E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 1]$$

$$= ATE + (1-\pi)\{E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(1)|D_i = 0, S_i = 1,\ W_i = 1]\} + E[Y_i(0)] \quad (A.4)$$

In a similar way, the *ASEU* can be expressed as

$$ASEU = E[Y_i(1^S) - Y_i(0)|D = 0, W = 1]$$
$$= E[Y_i(1^S)|D = 0, S_i = 1, W = 1)] - E[Y_i(0)|D = 0, S_i = 1, W = 1)] \quad (A.5)$$

Rearranging (*A.5*) yields

$$E[Y_i(1^S)|D = 0, S_i = 1, W = 1)] = ASEU + E[Y_i(0)|D = 0, S_i = 1, W = 1)] \quad (A.6)$$

Substituting $E[Y_i(1)|D = 1, S_i = 0, W = 1]$ and $E[Y_i(1^S)|D = 0, S_i = 1, W = 1)]$ from (*A.4*) and (*A.6*) into (*A.1*) yields

$$\Delta = ATE + (1-\pi)\{E[Y_i(1)|D = 1, S_i = 0, W = 1] - E[Y_i(1)|D = 0, S_i = 1, W = 1]\} + E[Y_i(0)]$$
$$-ASEU - E[Y_i(0)|D = 0, S_i = 1, W = 1)] \quad (A.7)$$

The following component of (*A.7*) can be rewritten as

$$(1-\pi)\{E[Y_i(1)|D = 1, S_i = 0, W = 1] - E[Y_i(1)|D = 0, S_i = 1, W = 1]\}$$
$$= (1-\pi)\begin{Bmatrix} E[Y_i(1)|D = 1, S_i = 0, W = 1] - E[Y_i(0)|D = 1, S_i = 0, W = 1] \\ -(E[Y_i(1)|D = 0, S_i = 1, W = 1] + E[Y_i(0)|D = 0, S_i = 1, W = 1]) \\ +E[Y_i(0)|D = 1, S_i = 0, W = 1] - E[Y_i(0)|D = 0, S_i = 1, W = 1] \end{Bmatrix} \quad (A.8)$$

Note that

$$ATT = E[Y_i(1)|D = 1, S_i = 0, W = 1] - E[Y_i(0)|D = 1, S_i = 0, W = 1]$$

$$ATU = (E[Y_i(1)|D = 0, S_i = 1, W = 1] - E[Y_i(0)|D = 0, S_i = 1, W = 1])$$

$$SELB = E[Y_i(0)|D = 1, S_i = 0, W = 1] - E[Y_i(0)|D = 0, S_i = 1, W = 1]$$

where *ATT* is the average treatment effect on the treated, *ATU* is the average treatment effect on the untreated, and *SELB* represents the selectivity bias.

Substituting *ATT, ATU*, and *SELB* into (*A.8*) yields:

$$= (1-\pi)(ATT - ATU) + (1-\pi)SELB \qquad (A.9)$$

Similarly, the other component of (*A.7*) can be written as

$$\begin{aligned} -E[Y_i(0)|D=0, S_i=1, W=1)] + E[Y_i(0)] \\ = -E[Y_i(0)|D=0, S_i=1, W=1)] + \pi E[Y_i(0)|D=1, S_i=0, W=1] \\ + (1-\pi)E[Y_i(0)|D=0, S_i=1, W=1] \end{aligned}$$

$$= \pi E[Y_i(0)|D=1, S_i=0, W=1] - \pi E[Y_i(0)|D=0, S_i=1, W=1]$$

$$= \pi\{E[Y_i(0)|D=1, S_i=0, W=1] - E[Y_i(0)|D=0, S_i=1, W=1]\}$$

$$= \pi SELB \qquad (A.10)$$

Substituting (*A.9*) and (*A.10*) into (*A.7*) yields

$$\Delta = ATE - ASEU + (1-\pi)(ATT - ATU) + (1-\pi)SELB + \pi SELB$$

or,

$$\Delta = ATE - ASEU + (1-\pi)(ATT - ATU) + SELB \qquad (QED)$$

Lemma A1: *In the absence of selectivity ($\theta_i = 0$), spillover ($S_i = 0$), antidote ($W_i = 1$), and essential heterogeneity, ATE is represented by β_T, i.e.*

$$ATE = E[Y_i(1)] - E[Y_i(0)] = \beta_T$$

Proof: In the absence of selectivity ($\theta_i = 0$), spillover ($S_i = 0$), the antidote ($W_i = 1$), equation (*9*) simplifies to:

$$Y_i = \mu_0 + \beta_T D_i + \eta_i D_i + \psi_i$$

Given the consistency assumption ($Y_i = Y_i(D_i)$), we can express this in its potential outcomes form as:

$$Y_i(D_i) = \mu_0 + \beta_T D_i + \eta_i D_i + \psi_i \qquad (AL1.1)$$

The unconditional expectations, with respect to treated and untreated units, are given by:

$$E[Y_i(1)] = \mu_0 + \beta_T + E[\eta_i] \qquad (AL1.2a)$$

$$E[Y_i(0)] = \mu_0 \qquad (AL1.2b)$$

In absence of any essential heterogeneity that is $E[\eta_i|D_i = 1] = E[\eta_i] = 0$ and $E[\psi_i] = 0$, (*AL1.2a*) further reduces to

$$E[Y_i(1)] = \mu_0 + \beta_T$$

Thus, by definition of *ATE*

$$ATE = E[Y_i(1)] - E[Y_i(0)] = \beta_T W_i \qquad (AL1.3)$$

Without the 'no essential heterogeneity' assumption, the conditional expectations of (*AL1.1*) with respect to the treatment group would be

$$E[Y_i(1)|D_i = 1] = \mu_0 + \beta_T + E[\eta_i|D_i = 1] \qquad (AL1.4a)$$

$$E[Y_i(0)|D_i = 1] = \mu_0 \qquad (AL1.4b)$$

The difference of (*AL1.4a*) and (*AL1.4b*) gives the *ATT*, that is

$$ATT = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] = \beta_T + E[\eta_i|D_i = 1]$$

Incorporating all group identifiers yields:

$$\begin{aligned} ATT = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1] \\ = \beta_T W_i + E[\eta_i|D_i = 1, S_i = 0, W_i = 1] \end{aligned} \qquad (AL1.5)$$

Lemma A2: *In the absence of selectivity ($\theta_i = 0$), the antidote ($W_i = 1$), and essential heterogeneity in the spillover effect, the ASEU is represented by δ, i.e.*

$$\delta = ASEU = E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1] - E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1]$$

Proof: Given the consistency property, the potential outcomes for the untreated group are

$$E[Y_i(1)|D_i = 0, S_i = 1, W_i = 1] = \mu_0 + \delta \qquad (AL2.1a)$$

$$E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1] = \mu_0 \qquad (AL2.1b)$$

With no essential heterogeneity in ϕ_i, the difference of these (*AL2.1a*) and (*AL2.1b*) gives the *ASEU*, that is

$$\begin{aligned} ASEU &= E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1] - E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1] \\ &= \delta \end{aligned} \qquad (AL2.2)$$

Lemma A3: *In the absence of the treatment and antidote, the difference in average outcomes between the treated and untreated groups reflects selectivity bias (SELB), represented by* θ*, i.e.,*

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

Proof: In the absence of the treatment, spillover, and the antidote, the average outcome for the treated group as per (*10*) is:

$$\begin{aligned} E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1] &= \mu_0 + E[\theta D_i + \psi_i|D_i = 1, S_i = 0, W_i = 1] \\ &= \mu_0 + \theta + E[\psi_i|D_i = 1, S_i = 0, W_i = 1] = \mu_0 + \theta \end{aligned} \qquad (AL3.1)$$

In the absence of the treatment, spillover, and the antidote, the average outcome for the treated group under equation (10) is:

$$\begin{aligned} E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1] &= \mu_0 + E[\theta D_i + \psi_i|D_i = 0, S_i = 1, W_i = 1] \\ &= \mu_0 + E[\psi_i|D_i = 1, S_i = 0, W_i = 1] = \mu_0 \end{aligned} \qquad (Al3.2)$$

Subtracting (*AL3.2*) from (*AL3.1*) yields

$$SELB = E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1] = \theta \qquad (AL3.3)$$

Appendix B: Nonparametric Identification of *ATT or ATE, ASEU, SELB*

Proposition 1a: *Assumptions 1–5 imply that the naïve difference in the average outcome between the treatment group without the antidote* ($D_i = 1, S_i = 0, W_i = 1$) *and the treatment group with the antidote* ($D_i = 1, S_i = 0, W_i = 0$) *identifies the ATT, i.e.,*

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

Proposition 1b: *Assumptions 1–6 imply that the observed difference in the average outcome between the treatment group without the antidote* ($D_i = 1, S_i = 0, W_i = 1$) *and the treatment group with the antidote* ($D_i = 1, S_i = 0, W_i = 0$) *identifies the ATE, i.e.,*

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

Proof:

In the presence of an antidote, the sample is divided into four groups:

(i) Treatment group without the antidote ($D_i = 1, S_i = 0, W_i = 1$),

(ii) Treatment group with the antidote ($D_i = 1, S_i = 0, W_i = 0$),

(iii) Control group without the antidote ($D_i = 0, S_i = 1, W_i = 1$),

(iv) Control group with the antidote ($D_i = 0, S_i = 1, W_i = 0$).

As shown in Figure 1, these correspond to n_1, n_3, n_2 and n_4 respectively.

At the same time, *Assumptions 1, 3 and 5 (no treated-to-treated spillover, fully effective AV and unconditional unconfoundedness)* suggest that

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

The left-hand sides of equations (*B.1*) represent the average outcomes of the treatment groups with the antidote. Essentially, *Assumption 1, 3 and 5* say that this group's outcome would be the same as if they had not received treatment or experienced any spillover. As such, this mean $E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0]$ serve as the counterfactuals for $E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1]$.

When the antidote is randomly assigned (*Assumption 5*), $E[\omega_i^0|D_i = 1, S_i = 0, W_i = 1] = E[\omega_i^0|D_i = 1, S_i = 0, W_i = 0]$. This assumption implies that:

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] = E[Y_i(1)|D_i = 1, S_i = 0] \qquad (B.2a)$$

Moreover, the same random assignment of W_i also implies that:

$$E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1] = E[Y_i(0)|D_i = 1, S_i = 0] \qquad (B.2b)$$

Combining (*B.1*) and (*B.2b*) one can write

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0] = E[Y_i(0)|D_i = 1, S_i = 0] \qquad (B.3)$$

Now consider the definition of *ATT*

$$ATT = E[Y_i(1)|D_i = 1, S_i = 0] - E[Y_i(0)|D_i = 1, S_i = 0]$$

Substituting values of $E[Y_i(1)|D_i = 1, S_i = 0]$ and $E[Y_i(0)|D_i = 1, S_i = 0]$ from (*B.2a*) and (*B.3*) yields

$$ATT = \beta_T = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0]$$

Applying consistency assumption (*Assumption 2)*

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

Thus, in presence of essential heterogeneity, *ATT* is identified

To go from *Proposition 1a* to *Proposition 1b* one requires *Assumption 6*, that is the absence of *essential heterogeneity in* $\tilde{\beta}_{Ti}$. This assumption implies

$$E[\eta_i|D_i] = E[\eta_i] = 0$$

That is, *ATT=ATU=ATE.* Thus, with *Assumption 1-5* and *Assumption 6* (no essential heterogeneity), the same difference identifies *ATE*, i.e.,

$$ATE = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1] - E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0]$$

Again, applying consistency property in *Assumption 2,*

$$ATE = E[Y_i|D_i = 1, S_i = 0, W_i = 1] - E[Y_i|D_i = 1, S_i = 0, W_i = 0] \qquad (B.5)$$

Proposition 2: *Assumptions 1–6 imply that the naïve difference in the average outcome between the control group without the antidote (*$D_i = 0, S_i = 1, W_i = 1$) and control group with the antidote *(*$D_i = 0, S_i = 1, W_i = 0$) identifies *ASEU, i.e.,*

$$ASEU = \delta = E[Y_i|D_i = 0, S_i = 1, W_i = 1] - E[Y_i|D_i = 0, S_i = 1, W_i = 0]$$

Proof:

Following the same line of proof, we now consider two subsamples: the control group without the antidote ($D_i = 0, S_i = 1, W_i = 1$) and the control group with the antidote ($D_i = 0, S_i = 1, W_i = 0$). As shown in Figure 1, these correspond to n_2 and n_4 subsamples, respectively.

No *essential heterogeneity in* $\tilde{\delta}_i$ implies

$$E[v_i|D] = E[v_i] = 0$$

Again, *Assumption 3* suggests that the antidote is fully effective in nullifying the spillover effect. This means that:

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0] = E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1] \qquad (B.6)$$

The left-hand side of equations (*B.6*) represents the average outcomes of the control groups with the antidote. Essentially, this indicates that this group's average outcome would be the same as if it had not experienced any spillover. As such, this mean serves as the counterfactual for $E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1]$.

Again, when the antidote is randomly assigned, the average outcome of the control with the antidote will equal the average outcome of the full control group under the different treatment status.

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1] = E[Y_i(1^S)|D_i = 0, S_i = 1] \qquad (B.7a)$$

$$E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1] = E[Y_i(0)|D_i = 0, S_i = 1] \qquad (B.7b)$$

Again, by *Assumption 3*,

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0] = E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1]$$

Substituting value of $E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1]$ from (*B.7b*)

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0] = E[Y_i(0)|D_i = 0, S_i = 1] \qquad (B.7b')$$

Now consider the definition of *ASEU*

$$ASEU = E[Y_i(1^S)|D_i = 0, S_i = 1] - E[Y_i(0)|D_i = 0, S_i = 1]$$

Substituting values of $E[Y_i(1^S)|D_i = 0, S_i = 1]$ and $E[Y_i(0)|D_i = 0, S_i = 1]$ from (*B.7a*) and (*B.7b'*)

$$ASEU = \delta = E[Y_i(1^S)|D_i = 0, S_i = 1, W = 1] - E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0]$$

With *Assumption 2* (consistency assumption),

$$ASEU = \delta = E[Y_i|D_i = 0, S_i = 1, W_i = 1] - E[Y_i|D_i = 0, S_i = 1, W_i = 0] \qquad (B.8)$$

Proposition 3: *Assumptions 1–4 imply that the difference in the average outcome between the treatment group with the antidote ($D_i = 1, S_i = 0, W_i = 0$) and the control group with the antidote ($D_i = 0, S_i = 1, W_i = 0$) identifies SELB, i.e., with*

$$\theta = E[Y_i|D_i = 1, S_i = 0, W_i = 0] - E[Y_i|D_i = 0, S_i = 1, W_i = 0].$$

Proof:

The selectivity bias is defined as

$$\theta = E[Y_i(0)|D_i = 1, S_i = 0] - E[Y_i(0)|D_i = 0, S_i = 1] \qquad (B.10)$$

Equations $(B.3)$ and $(B.7b')$ show that

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0] = E[Y_i(0)|D_i = 1, S_i = 0] \qquad (B.11a)$$

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0] = E[Y_i(0)|D_i = 1, S_i = 0] \qquad (B.11b)$$

Thus, directly substituting $(B.11a)$ and $(B.11b)$ into$(B.10)$ yields

$$SELB = \theta = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0] - E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0]$$

Again, with *Assumption 2* (consistency assumption),

$$SELB = \theta = E[Y_i|D_i = 1, S_i = 0, W_i = 0] - E[Y_i|D_i = 0, S_i = 1, W_i = 0] \qquad (B.12)$$

Appendix C: Concurrent Treatment Does Not Confound Either β_T or δ

The advantage of the antidotal approach is that neither the estimates of the treatment effect nor the *SUTVA* bias will be confounded by any concomitant treatments if the antidote nullifies the effect of the treatment of interest and is unrelated to the concomitant treatments that causes confounding. To see this, consider (10) again.

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

Suppose T_i is another treatment administered alongside the main treatment. Treated units (D_i=1) also receive this additional treatment ($T_i = 1$), while untreated units ($D_i = 0$) do not ($T_i = 0$).

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

The error term u_i now includes any heterogeneity in α_T. As long as this heterogeneity is independent of all other factors, and unrelated to W_i, it does not affect either the treatment effect or the *SUTVA* bias estimates. Accordingly, the observed averages for groups n_1, n_2, n_3, and n_4 are

$$n_1\colon E[Y_i(1)|D_i = 1, S_i = 0, T_i = 1,\ W_i = 1]$$
$$= \mu_0 + \beta_T + \theta + \alpha_T + E[u_i|D_i = 1, S_i = 0, T_i = 1,\ W_i = 1]$$
$$n_2\colon E[Y_i(1^S)|D_i = 0, S_i = 1, T_i = 1, W_i = 1] = \mu_0 + \delta + E[u_i|D_i = 0, S_i = 1, T_i = 0,\ W_i = 1]$$
$$n_3\colon E[Y_i(1)|D_i = 1, S_i = 0, T_i = 1, W_i = 0]$$
$$= \mu_0 + \theta + \alpha_T + E[u_i|D_i = 1, S_i = 0, T_i = 1, W_i = 0]$$
$$n_4\colon E[Y_i(1^S)|D_i = 0, S_i = 1, T_i = 0, W_i = 0] = \mu_0 + E[u_i|D_i = 0, S_i = 1, T_i = 0, W_i = 0]$$

Given that u_i is mean independent of W_i by *Assumption 5*, and if all other assumptions are satisfied, the treatment effect is calculated by differencing the averages of n_1 and n_3, i.e.,

$$E[Y_i(1)|D_i = 1, S_i = 0, T_i = 1,\ W_i = 1] - E[Y_i(1)|D_i = 1, S_i = 0, T_i = 1, W_i = 0] = \beta_T$$

$$= ATE$$

By the consistency assumption (*Assumption 2*)

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

Similarly, with *Assumption 1-6*, the *ASEU* is the difference between n_2 and n_4

$$E[Y_i(1^S)|D_i = 0, S_i = 1, T_i = 0,\ W_i = 1] - E[Y_i(1^S)|D_i = 0, S_i = 1, T_i = 0, W_i = 0] = \delta$$

$$= ASEU$$

Again, with the consistency assumption, the difference in the observed counterparts of these means yield the estimate of $ASEU$, i.e.,

$$E[Y_i|D_i = 0, S_i = 1, T_i = 0,\ W_i = 1] - E[Y_i|D_i = 0, S_i = 1, T_i = 0, W_i = 0] = \delta = ASEU$$

However, with concomitant treatments as in this case, *SELB* is not identified. This is evident from the difference between n_3 and n_4

$$E[Y_i|D_i = 1, S_i = 0, T_i = 1, W_i = 0] - E[Y_i|D_i = 0, S_i = 1, T_i = 0, W_i = 0] = \theta + \alpha_T$$

which shows the estimate combines the effects of the concomitant treatment and the selectivity component without affecting the *ATT* or *ATE* or *ASEU*.

Appendix D: Bounds When the Antidote is Imperfect

An imperfect antidote does not fully negate treatment and its spillover effects. Consequently, β_T, θ and δ are not generally identified. Nevertheless, these parameters can be set-identified with additional assumptions.

For illustration, consider regression equation (10) in the case of perfect antidotes.

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

Let τ represent an unknown fraction ($0 \leq \tau \leq 1$) indicating the extent the antidote neutralizes the treatment's effect. With this modification, equation (D.1) can be rewritten as:

$$Y_i = \mu_0 + [1 - \tau(1 - W_i)]\beta_T D_i + \theta D_i + \delta[1 - \tau(1 - W_i)]S_i + u_i \qquad (D.2)$$

When $\tau = 1$, applying the antidote ($W_i = 0$) fully nullifies the effects of the treatment and spillover. Conversely, when $\tau = 0$, the antidote is completely ineffective at neutralizing the effects.

When τ is between 0 and 1, and $W_i = 1$, the treatment and spillover effects (i.e., β_T, δ) remain intact. However, when $W_i = 0$, unlike in the case of a fully effective antidote, the treatment and spillover effects do not drop to zero but instead remains at $(1 - \tau)\beta_T$ and $(1 - \tau)\delta$, which is a fraction of the original effect β_T and δ.

Before addressing identification, we revisit potential outcomes and assess their observability. Consider the three potential outcomes defined in (1a), (1b), and (1c). The first corresponds to receiving treatment, denoted as 1. The second applies when untreated but exposed to treatment spillover, denoted as 1^S. The third occurs when neither treatment nor spillover is received, denoted as 0. Thus, $Y_i(1)$ represents the potential outcome when individual i is treated, $Y_i(1^S)$ applies when untreated but exposed to spillover, and $Y_i(0)$ applies when neither treatment nor spillover occurs.

Thus, in the case of a perfect antidote ($\tau = 1$), we express the definition of the effects of interest as follows:

$$\beta_T = E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 1, \tau = 1] - E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \ (D.3a)$$

$$\delta = E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1, \tau = 1] - E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] \ (D.3b)$$

$$\theta = E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1, \tau = 1] - E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] \quad (D.3c)$$

Here, β_T denotes the Average Treatment Effect on the Treated (*ATT*) or the Average Treatment Effect (*ATE*) in the absence of essential heterogeneity, δ represents the Average Spillover Effect on the Untreated (*ASEU*), and θ captures the Average Selectivity Bias (*SELB*). Here, an identification issue arises because $E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1]$, $E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1]$ are not observed. However, *Appendix B* (with a fully effective antidote, i.e., $\tau = 1$) shows that $E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] = E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1]$ and $[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] = E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1]$, where $E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1]$ and $E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1]$ are observed, i.e., the average outcomes of the subsample n_3 and n_4. Thus, as per *Propositions 1a-3*, the effects are identified as

$$\beta_T = E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 1, \tau = 1] - E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] \quad (D.4a)$$

$$\delta = E\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1\right] - E\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1\right] \quad (D.4b)$$

$$\theta = E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] - E\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1\right] \quad (D.4c)$$

The identification issue with an imperfect antidote ($\tau < 1$) arises because the two subsamples (n_3 and n_4) with a perfect antidote are unobserved. Even if

$$E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 1, \tau = 1] = E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 1, \tau < 1]$$

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1, \tau = 1] = E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 1, \tau < 1]$$

the other averages are not equal, i.e.,

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] \neq E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$$

$$E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1] \neq E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1]$$

This means, unlike the perfect antidote case, $E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 0,\ \tau = 1]$ and $E\left[Y_i\left((1^S)\right)\middle|D_i = 0, S_i = 1, W_i = 0,\ \tau = 1\right]$ are no longer observed. As such, this means

$$E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \neq E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1] \quad (D.5a)$$

$$E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] \neq E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1] \quad (D.5b)$$

With just the observed means of these four groups (n_1, n_2, n_3, n_4), one cannot identify the *ATT* or *ATE* , *ASEU* and *SELB*.

To enhance clarity, Table *D1* outlines the observed and unobserved group averages in the context of an imperfect antidote.

Group averages with imperfect antidote

	Observed		Unobserved
W=1	n_1: $E[Y_i(1)\|D=1, S_i=0, W_i=1, \tau=1]$		
W=1	n_2: $E[Y_i(1^S)\|D=0, S_i=1, W_i=1, \tau=1]$		
W=0	n_3: $E[Y_i(1)\|D=1, S_i=0, W_i=0, \tau<1]$	$\neq$	n_3: $E[Y_i(1)\|D=1, S_i=0, W_i=0, \tau=1]$
W=0	n_4: $E[Y_i(1^S)\|D=0, S_i=1, W_i=0, \tau<1]$	$\neq$	n_4: $E[Y_i(1^S)\|D=1, S_i=0, W_i=0, \tau=1]$

Bounds: A Summary of Identification Strategy

Consider *Propositions 1a to 3*. In *Proposition 1a*, $E[Y_i(1)|D_i=1, S_i=0,\ W_i=1, \tau=1]$ is observed but $E[Y_i(1)|D_i=1,\ S_i=0, W_i=0, \tau=1]$ is not observed. Hence, β_T is not identified. However, the upper bound and lower bounds of $E[Y_i(1)|D_i=1,\ S_i=0, W_i=0, \tau=1]$ would yield the lower bound and upper bound of β_T.

Similarly, in *Proposition 2*, $E\left[Y_i\left((1^S)\right)\middle|D_i=0,\ S_i=1, W_i=1, \tau=1\right]$ is observed but $E\left[Y_i\left((1^S)\right)\middle|D_i=0,\ S_i=1, W_i=0, \tau=1\right]$ is not observed. Hence, δ is not identified. However, the upper bound and lower bounds of $E\left[Y_i\left((1^S)\right)\middle|D_i=0,\ S_i=1, W_i=0, \tau=1\right]$ would yield the lower bound and upper bound of δ.

In case of *Proposition 3*, neither $E[Y_i(1)|D_i=1,\ S_i=0, W_i=0, \tau=1]$ nor $E\left[Y_i\left((1^S)\right)\middle|D_i=0,\ S_i=1, W_i=0, \tau=1\right]$ is observed. Hence, θ is not identified. However, the upper bound of $E[Y_i(1)|D_i=1,\ S_i=0, W_i=0, \tau=1]$ and lower bound of

$\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1\right]$ yield the upper bound estimate of θ. In the same way, the lower bound of $E[Y_i(1)|D_i = 1,\ \ S_i = 0, W_i = 0, \tau = 1]$ and upper bound of $\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1\right]$ yield the lower bound estimate of θ.

Below, we derive bounds on these unobserved terms under various assumptions, allowing us to establish bounds on the β_T, δ and θ.

Additional Assumptions:

As mentioned earlier, when $\tau < 1$, the effects can no longer be point-identified. However, with appropriate additional assumptions, we can construct bounds around the effect. In doing so, we maintain all previous assumptions, except *Assumption 3,* which implies the perfect antidote case. In addition, we adopt a few other assumptions commonly used in studies that construct bounds. Specifically, we introduce three additional sets of assumptions.

Assumption Set 1*: More Y is preferred than less Y; and $Y_i \geq 0$.*

Assumption Set 2*: Positive monotone treatment response (MTR), which states that for any i,*

$$Y_i(1) \geq Y_i(0)$$

Assumption Set 3*: Optimal treatment selection (OTS) and monotone treatment selection (MTS).*

OTS assumption: The treatment group ($D = 1$) prefers treatment over no treatment, while the control group ($D = 0$) prefers no treatment over treatment.

$$\text{Member of treatment group: } Y_i(1) \geq Y_i(0)$$

$$\text{Member of control group: } \quad Y_i(1) \leq Y_i(0)$$

Hence, *OTS Assumption* implies

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \geq E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \qquad (D.6a)$$

$$E[Y_i(1)|D_i = 0, S_i = 1, W_i = 1, \tau = 1] \leq E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1, \tau = 1] \quad (D.6b)$$

MTS Assumption: The average outcome of the treatment group is always greater than or equal to the average outcome of the control group, both with and without treatment.

This means

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1, \tau = 1] \geq E[Y_i(1)|D_i = 0, S_i = 1, W_i = 1, \ \tau = 1] \quad (D.7a)$$

$$E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1, \ \tau = 1] \geq E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1, \ \tau = 1] \quad (D.7b)$$

Theorems and lemmas

Lemma D.1**:** *When the antidote is perfect, the average outcome of the treatment group without treatment is equal to that of the treatment group with treatment and a perfect antidote. Similarly, the average outcome of the control group without the treatment spillover is equal to that of the control group with treatment spillover and a perfect antidote, implying*

$$E[Y_i(0)|D_i = 1, \ S_i = 0, W_i = 1, \tau = 1] = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0, \tau = 1]$$

$$E[Y_i(0)|D_i = 0, \ S_i = 1, W_i = 1, \tau = 1] = E[Y_i(1^S)|D_i = 0, \ S_i = 1, W_i = 0, \tau = 1]$$

Proof:

For the treatment group, when treatment is combined with a perfect antidote, the treatment becomes ineffective. Consequently, the outcome with treatment and a perfect antidote is equal to the outcome without treatment. Hence, for the treatment group, the average outcome with a perfect antidote and no treatment (irrespective of the antidote) will be the same, i.e.,

$$E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1, \tau = 1] = E[Y_i(1)|D_i = 1, S_i = 0, \ W_i = 0, \tau = 1] \quad (D.8a)$$

Similarly, for the control group, when treatment spillovers occur but are accompanied by a perfect antidote, the spillover effect disappears, as it is inherently the same as the treatment. In other words, the spillover becomes ineffective, making the average outcomes equal to those without the spillover (irrespective of the antidote), i.e.,

$$E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1, \tau = 1] = E\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1\right] \quad (D8b)$$

Q.E.D

Theorem D.1: If *Assumption set 1 and 2* are true, the following inequalities hold.

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1] \geq\ E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1, \tau = 1]$$

$$E\left[Y_i\left((1^S)\right)\middle|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1\right] \geq\ E[Y_i(0)|D_i = 0, S_i = 1, W_i = 1, \tau = 1]$$

Proof:

The key point is that the *MTR*, combined with *Assumption Set 1*, implies that the treatment always increases the outcome. Consequently, treatment with an imperfect antidote, which retains some of the treatment's effect, results in higher outcomes than those with the treatment or its spillover and a perfect antidote, regardless of whether the treatment is received directly or through a spillover. Based on this, one can write:

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1] \geq\ E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 0, \tau = 1] \quad (D.9a)$$

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0, \tau < 1] \geq\ E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0, \tau = 1] \quad (D.9b)$$

Since a perfect antidote nullifies the effects of a treatment and spillover, one can write:

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] = E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \quad (D.10a)$$

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0,\ \tau = 1] = E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1,\ \tau = 1] \quad (D.10b)$$

Substituting (*D.10a*) and (*D.10b*) into (*D.9a*) and (*D.9b*), respectively, gives:

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1] \geq E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \quad (D11a)$$

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0, \tau < 1] \geq E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1,\ \tau = 1] \quad (D11b)$$

Q.E.D.

Theorem D.2a: *Under Assumption Sets 1 and 3 (i.e., OTS and MTS), the average outcome for the treatment group with a perfect antidote is always greater than or equal to the average outcome for the control group with the treatment spillover and an imperfect antidote, i.e.,*

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] \geq E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0,\ \tau < 1]$$

Proof: As per *Lemma D1*,

$$E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0, \tau = 1] \quad (D.12a)$$

$$E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1,\ \tau = 1] = E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0,\ \tau = 1] \quad (D.12b)$$

The OTS assumption along with random assignment of W_i state that treatment improves the outcomes for the treatment group. Since an imperfect antidote retains some of the treatment effect, the average outcome with an imperfect antidote is higher than the average outcome with a perfect antidote for the treatment group.

$$E[Y_i(1)|D_i = 1, S_i = 0,\ W_i = 0, \tau < 1] \geq E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] \quad (D.13a)$$

The OTS Assumption along with random assignment of W_i also state that treatment lowers the outcomes for the control group. Since an imperfect antidote retains some of the treatment effect,

the average outcome with an imperfect antidote is lower than the average outcome with a perfect antidote for the control group, i.e.,

$$E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1] \leq E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1] \quad (D.13b)$$

The MTS Assumption, on the other hand, implies that, in the absence of treatment, the average outcome of the treatment group is higher than that of the control group.

$$E[Y_i(0)|D_i = 1, S_i = 0,\ W_i = 1, \tau = 1] \geq E[Y_i(0)|D_i = 0, S_i = 1,\ W_i = 1, \tau = 1] \quad (D.14)$$

By combining *Lemma D.1* and (*D.14*) one can write

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0, \tau = 1] \geq E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] \quad (D15)$$

Further combining (D13b) with (D15) suggests

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] \geq E[Y_i(1_s)|D_i = 0, S_i = 1,\ W_i = 0, \tau < 1] \quad (D16)$$

Q.E.D.

Theorem D.2b: *Under the OTS assumption, the average Y for the treatment group with an imperfect antidote is always greater than or equal to the average Y for the treatment group with treatment and a perfect antidote, i.e.,*

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0, \tau < 1] \geq E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1]$$

Proof:

Assumption 1 and the OTS along with random assignment of W_i indicate that treatment results in a higher average outcome for the treatment group compared to no treatment.

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 1, \tau = 1] \geq E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \quad (D.17)$$

Because with a perfect antidote $E[Y_i(0)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] = E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1]$, one can rewrite (D17) as

$$E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \geq E[Y_i(1)|D_i = 1,\ S_i = 0, W_i = 0, \tau = 1] \quad (D.18)$$

Because an imperfect antidote retains some of the treatment effect, which is always greater than no treatment under these assumptions, one can write

$$[Y_i(1)|D_i = 1, S_i = 0, W_i = 0, \tau < 1] \geq E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1] \quad (D.19)$$

Q.E.D.

Theorem D.3a: *Under the OTS assumption, the average outcome for the control group with a treatment spillover and an imperfect antidote is always less than or equal to the average outcome for the control group with a treatment spillover and a perfect antidote, i.e.,*

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0, \tau < 1] \leq E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1]$$

Proof:

Assumption 1 and the OTS along with random assignment of W_i, imply that the treatment would result in a lower average outcome for the control group compared to no treatment.

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0,\ \tau < 1] \leq E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] \quad (D.20)$$

because with a perfect antidote $E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0, \tau = 1] = E[Y_i(0)|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1]$, one can rewrite the above equation as

$$E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1] \leq E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1] \quad (D.21)$$

In this case, treatment spillover, being somewhat similar to the treatment, would cause the control group's average outcome to decline compared to what it would be without the treatment. Hence, under these assumptions, one can write:

$$E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0, \tau < 1] \leq E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau = 1] \quad (D.22)$$

Q.E.D.

Theorem D.3b: *Under Assumptions 1 and 3, the average outcome for the treatment group with the treatment and an imperfect antidote is always greater than or equal to the average outcome for the control group with the treatment spillover and a perfect antidote, i.e.,*

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \geq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$$

Proof:

By the *MTS* assumption again,

$$E[Y_i(0)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] \geq E[Y_i(0)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] \quad (D.23)$$

However, since the outcome for the treatment group in the absence of the treatment is equal to the outcome with treatment and a perfect antidote, i.e., $E[Y_i(0)|D_i = 1, S_i = 0, W_i = 1,\ \tau = 1] = E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0,\ \tau = 1]$, one can write

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0,\ \tau = 1] \geq E[Y_i(0)|D_i = 0, S_i = 1, W_i = 0,\ \tau = 1] \quad (D.24)$$

Similarly, if an imperfect antidote is administered to the treatment group with treatment, some effect of the treatment would persist. Based on the same OTS assumption, now one can write

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0,\ \tau < 1] \geq E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0,\ \tau = 1] \quad (D.25)$$

Combining (*D.24*) and (*D.25*) and since $E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0,\ \tau = 1] = E[Y_i(0)|D_i = 0, S_i = 1, W_i = 0,\ \tau = 1]$, one can write

$$E[Y_i(1)|D_i = 1, S_i = 0, W_i = 0,\ \tau < 1] \geq E[Y_i(1^S)|D_i = 0, S_i = 1, W_i = 0,\ \tau = 1] \quad (D.26)$$

Q.E.D.

Bounding Average treatment effect *(β_T)*

As per Proposition 1a and *1b* in the main text ATE (i.e., β_T) can be written as

$$\beta_T = E[Y(1)_i|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(0)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$$

The known component in this expression is $E[Y(1)_i|D_i = 1, W_i = 1, S_i = 0, \tau = 1$, and the unknown component is $E[Y(1)_i|D_i = 1, W_i = 0, S_i = 0, \tau = 1$. Instead, $E[Y(1)_i|D_i = 1, W_i = 1, S_i = 0, \tau < 1$ is observed when the antidote is imperfect. Bounding the ATE requires determining the bounds for the latter term, $E[Y(1)_i|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$.

We provide two possible bounds for β_T with different sets of assumptions. The first one uses *Assumptions 1 and 2*. The second one uses *Assumptions 1 and 3*. We explain these one by one.

(1) Constructing bounds on β_T under Assumptions 1 & 2 when the antidote is imperfect and there is no essential heterogeneity

***Result 1a**: The lower bound of the ATE is*

$$\beta_T \geq E[Y(1)_i|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$

This directly follows from Theorem D.2b which states

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \geq E[Y(1)_i|D_i = 1, W_i = 0, S_i = 0, \tau = 1] \qquad (27)$$

Substituting this maximum value of this unobserved entity $E[Y(1)_i|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$ yields the lower bound of β_T, i.e.,

As per *Theorem D.1*, and equation $(D.3)$, one can now construct the lower bound of the ATE

$$\beta_T \geq E[Y(1)_i|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$

Result 1b: The *upper bound of* β_T

$$\beta_T \leq E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1]$$

The assumption $Y_i \geq 0$ implies that $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1] \geq 0$, providing a lower bound estimate for $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$. Consequently, the upper bound estimate of the ATE is:

$$\beta_T \leq E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] \qquad (D28)$$

Thus, *Results 1a and 1b* suggest that bounds for β_T under *Assumption 1 and 2* are

$$E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$
$$\leq \beta_T \leq$$
$$E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1]$$

By the consistency assumption, the equation above can be rewritten as:

$$E[Y_i|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] - E[Y_i|D_i = 1, S_i = 0, W_i = 0,\ \tau < 1]$$
$$\leq \beta_T \leq$$
$$E[Y_i|D_i = 1,\ S_i = 0, W_i = 1, \tau = 1] \qquad (D29)$$

(2) Constructing bounds on β_T *under Assumptions 1 & 3 when the antidote is imperfect and there is no essential heterogeneity*

Result 2a: The lower bound of β_T is

$$\beta_T \geq E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$

As per *Theorem D.2b*,

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \geq E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$$

Given this inequality one can substitute $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$ by its upper bound, which is observed, and obtain the lower bound of β_T, *i.e.*,

$$\beta_T \leq E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 0, W_i = 0, S_i = 0, \tau < 1] \quad (D30)$$

Result 2b: The upper bound of β_T is

$$E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 0, W_i = 0, S_i = 0, \tau < 1]$$

As per *Theorem D.2a*,

$$E[Y_i(0)|D_i = 1, W_i = 0, S_i = 0, \tau = 1] \geq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 0, \tau < 1]$$

Substituting $E[Y_i(0)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$ with its lower bound of constitutes the upper bound of β_T, that is,

$$\beta_T \leq E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(0)|D_i = 1, W_i = 0, S_i = 0, \tau = 1] \quad .(D31)$$

Thus, the bounds for β_T under *Assumption 1 and 3* are

$$[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$
$$\leq \beta_T \leq$$
$$E[Y_i(1)|D_i = 1, W_i = 1, S_i = 0, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 0, \tau < 1] \quad (D32)$$

By the consistency assumption, the equation above can be rewritten as:

$$[Y_i|D_i = 1, S_i = 0, W_i = 1, \ \tau = 1] - E[Y_i|D_i = 1, S_i = 0, W_i = 0, \tau < 1]$$
$$\leq \beta_T \leq$$
$$E[Y_i|D_i = 1, \ S_i = 0, W_i = 1, \tau = 1] - E[Y_i|D_i = 0, \ S_i = 1, W_i = 0, \tau < 1] \qquad (D32)$$

Proposition 2 states that one can show that the $ASEU$ (i.e., β_T) is

$$\delta = E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$$

Due to the imperfect antidote, $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$ is no longer observed. Instead, $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$ is observed, such that $E[Y_i(0)|D_i = 0, W_i = 0, S_i = 1, \tau = 1] \neq E[Y_i(1_s)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$.

Consequently, the *ASEU* is not point-identified.

Constructing bounds on δ

Here we provide two possible bounds for δ with two different set of assumptions. The first one uses *Assumptions 1 and 2*. The second one uses *Assumptions 1 and 3*. We explain these one by one.

(1) Constructing bounds on δ under Assumptions 1 & 2 when the antidote is imperfect and there is no essential heterogeneity

Result 3a: *The Lower bound of $\underline{\delta}$*

The lower bound of $\underline{\delta}$ follows from *Theorem D.1*, which shows that

$$E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \geq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1] \quad (D33)$$

Substituting $E[Y_i(0)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$ by its upper bound $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$ yields the lower bound of δ, i.e.,

$$\delta \geq E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$$

Result 3b: *The Upper bound of* $\underline{\delta}$

The assumption that $Y_i \geq 0$ also implies that $E[Y_i(1_s)|D_i = 0, W_i = 0, S_i = 1, \tau = 1] \geq 0$. Substituting $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$ by its lower bound zero yields the upper bound estimate of δ, i.e.,

$$\delta \leq E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau < 1] \quad (D34)$$

Thus, the bounds for δ are

$$E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \leq \delta \leq E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau < 1] \quad (D35)$$

By the consistency assumption, the equation above can be rewritten as:

$$E[Y_i|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] - E[Y_i(1^S)|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1] \leq \delta \leq E[Y_i|D_i = 0,\ S_i = 1, W_i = 1, \tau < 1] \quad (D35)$$

essential heterogeneity

Result 4a: The *Lower bound of* δ

Theorem D.3b states that

$$E[Y_i(1^S)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \geq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$$

This provides an upper bound for $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$. Thus, the lower bound of the *ASE* can be expressed as

$$\delta \geq E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \quad (D36)$$

Result 4b: The *Upper bound of* δ

Theorem D.3a shows that

$$E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \leq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$$

This provides a lower bound estimate of $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$. Hence, the upper bound estimate of *ASEU*

$$\delta \leq E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \quad (D37)$$

As such, the bounds under these assumptions are

$$E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$
$$\leq \delta \leq$$
$$E[Y_i(1^S)|D_i = 0, W_i = 1, S_i = 1, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \quad (D38)$$

By the consistency assumption, the equation above can be rewritten as:

$$E[Y_i|D_i = 0,\ S_i = 1, W_i = 1, \tau = 1] - E[Y_i|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$$
$$\leq \delta \leq$$
$$E[Y_i|D_i = 0, \quad S_i = 1, W_i = 1, \tau = 1] - E[Y_i|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1] \quad (D38)$$

Bounding the average selectivity bias (θ)

With a perfect antidote, the following difference in averages captures the selectivity bias.

$$\theta = E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$$

When the antidote is imperfect, neither $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$, nor $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$ are observed. Hence, this difference does not identify the *ASB*.

(1) Constructing bounds on θ under Assumptions 1 & 2 when the antidote is imperfect and there is no essential heterogeneity

Theorem D.2a implies

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \geq E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$$

$$E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 0, \tau < 1] \geq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$$

Because as per Assumption 1, i.e., $Y \geq 0$, the following two conditions can be constructed.

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1] \geq 0$$

$$E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1] \geq 0$$

These two sets of equations represent the lower and upper bounds of each unobserved term.

Result 5a: *The lower bound of θ:*

Substituting $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$ and $E[Y_i(1_s)|D_i = 0, W_i = 0, S_i = 0, \tau = 1]$ by their lower bound 0 and upper bound $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$ yield the lower bound estimates of *ASB*, i.e.,

$$\theta \geq 0 - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 0, \tau < 1] \qquad (D39)$$

Result 5b: *The upper bound of θ*

By the same logic, substituting $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau = 1]$ and $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 0, \tau = 1]$ by their upper bound $E[Y_i(1)|D_i = 0, W_i = 0, S_i = 0, \tau < 1]$ and lower bound 0 yield the upper bound estimates of *ASB*, i.e.,

$$\theta \leq E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] - 0 \qquad (D40)$$

Thus, the bounds for θ is

$$-E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$$

$$\leq \theta \leq$$

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$

By the consistency assumption, the equation above can be rewritten as:

$$-E[Y_i|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1]$$
$$\leq \theta \leq$$
$$E[Y_i|D_i = 1, S_i = 0, W_i = 0,\ \tau < 1] \qquad (D41)$$

(2) Constructing bounds on θ under Assumptions 1 & 3 when the antidote is imperfect and there is no essential heterogeneity

Result 6a: *The Lower bound of $\underline{\theta}$*

By *Assumption 1 and 3, the lower bound of* $E[Y_i(1)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$ *is zero and the upper bound of* $E[Y_i(1_s)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$ is $E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$. Hence the lower bound of θ is

$$\theta \geq -E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \qquad (D42)$$

Result 6b: *The Upper bound of θ*

Under *Assumption1 and 3* (especially the OTS),

$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] \geq E[Y_i(1)|D_i = 0, W_i = 0, S_i = 0, \tau = 1]$$
$$E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1] \geq E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1]$$

This gives an upper bound of $E[Y_i(1)|D_i = 0, W_i = 0, S_i = 0, \tau = 1]$ and a lower bound of $E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau = 1]$, identifying upper bound of θ. Hence,

$$\theta \leq E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \quad (D43)$$

Thus, the bounds for

$$-E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1]$$
$$\leq \theta \leq$$
$$E[Y_i(1)|D_i = 1, W_i = 0, S_i = 0, \tau < 1] - E[Y_i(1^S)|D_i = 0, W_i = 0, S_i = 1, \tau < 1] \quad (D44)$$

By the consistency assumption, the equation above can be rewritten as:

$$-E[Y_i|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1]$$

$$\leq \theta \leq$$

$$E[Y_i|D_i = 1,\ S_i = 0, W_i = 0, \tau < 1] - E[Y_i|D_i = 0,\ S_i = 1, W_i = 0, \tau < 1] \qquad (D44)$$

Appendix E: Nonrandom Antidote Take Up

As of now, we assume that W_i is independent of η_i, v_i, ϕ_i, and ω_i^0. However, our estimators may be biased and inconsistent if these assumptions are violated. We examine the implications of relaxing these four assumptions in the following sections.

(a) Error ω_i^0 is not mean independent of W_i

When ω_i^0 is mean dependent on W_i, the antidotal approach cannot identify the treatment effect and the bias that is caused by a violation of the *SUTVA*. However, the selectivity bias can be identified. Typically, this situation occurs when cohorts with $W_i = 0$ and $W_i = 1$ differ for both observed and unobserved reasons. In the Boombox example given in the text, such a situation may arise if the groups receiving (i.e., $W_i = 0$) and not receiving earplugs (i.e., $W_i = 1$) differ in their pre-treatment averages. Here $\mathrm{E}[\omega_i^0|W = 1] \neq E[\omega_i^0|W_i = 0]$ implying $E[u_i|W = 1] \neq E[u_i|W = 0]$ even if η_i, ϕ_i, v_i are independent of W_i. This implies

$$E[u_i|D_i = 1, S_i = 0, W_i = 1] \neq E[u_i|D_i = 1, S_i = 0, W_i = 0]$$
$$E[u_i|D_i = 0, S_i = 1, W_i = 1] \neq E[u_i|D_i = 0, S_i = 1, W_i = 0]$$

Therefore the non-zero term $E[u_i|D_i = 1, S_i = 0, W_i = 1] - E[u_i|D_i = 1, S_i = 0, W_i = 0]$ is added to β_T making it statistically inconsistent. Similarly, the non-zero term $E[u_i|D_i = 0, S_i = 1, W_i = 1] - E[u_i|D_i = 0, S_i = 1, W_i = 0]$ is added to δ, making it statistically inconsistent. However, the estimator of the selectivity bias remains consistent because the additive term that differentiates $W_i = 0$ and $W_i = 1$ is no longer relevant since both groups n_3 and n_4 are

antidotal i.e., for both groups $W_i = 0$. Hence, θ is identified even when W_i is not independent of ω_i^0.

(b) Treatment effect heterogeneity η_i is not mean independent of W_i

One cannot identify the treatment effect when η_i is mean dependent on W_i. However, it is still possible to identify *ASEU* and *SELB*. This situation occurs when the treatment effect heterogeneity is associated with antidote assignment. Here subjects choose antidotes based on their own assessment of the effects of the treatment. In this case the comparison of n_1 and n_3 involves components of the error term $E[\eta_i W_i D_i | D_i = 1, S_i = 0, W_i = 1]$ and $E[\eta_i W_i D_i | D_i = 1, S_i = 0, W_i = 0]$ that are unequal, implying that $E[\eta_i | D_i = 1, S_i = 0, W_i = 1] - E[\eta_i | D_i = 1, S_i = 0, W_i = 0] \neq 0$. As per (*B.5*) the difference between n_1 and n_3 is therefor $\beta_T + E[\eta_i | D_i = 1, S_i = 0, W_i = 1]$ which is not the treatment effect.

The parameter δ is identified because in the control group $D_i = 0$, so that $\eta_i W_i D_i = 0$. Thus, violating this assumption does not affect identification of the *ASEU*. Similarly, the selectivity parameter θ is identified because in n_3 and n_4, $W_i = 0$, so that $\eta_i W_i D_i = 0$.

(c) Spillover effect heterogeneity v_i is not mean independent of W_i

This situation occurs when members of the control group choose an antidote based on the spillover effect heterogeneity. Here, a spillover is tantamount to treating the control group, except the treatment does not lead to a further spillover. Based on the same reasoning presented in part (b), it is not possible to identify the *ASEU*. It is, however, still possible to identify the *ATT* OR *ATE* and the *SELB* if v_i is not mean independent of W_i.

(d) Selectivity effect heterogeneity ϕ_i is not mean independent of W_i

This occurs when one chooses to take the antidote based on heterogeneity in the treatment selection. Entities with and without antidotes have different average effects due to selectivity heterogeneity, i.e., $E[\phi_i|W_i = 1] \neq E[\phi_i|W_i = 0]$. If such is the case, then $E[\phi_i D_i|D_i = 1, S_i = 0, W_i = 1] \neq E[\phi_i D_i|D_i = 1, S_i = 0, W_i = 0]$. This means that the *ATT* or *ATE* is not identified. Similarly, violating mean independence results in $E[\phi_i D_i|D_i = 1, S_i = 0, W_i = 0] \neq 0 = E[\phi_i D_i|D_i = 0, S_i = 1, W_i = 0]$. As such, the selectivity bias is not identified. However, the bias resulting from a violation of *SUTVA* is identified because $E[\phi_i D_i|D_i = 0, S_i = 1, W_i = 1] = E[\phi_i D_i|D_i = 0, S_i = 1, W_i = 0] = 0$.

Appendix F: Testing Whether the Antidote Take Up (W) is Random

One generally cannot identify all three parameters of interest the treatment effect (β_T), the selectivity bias (θ), and the bias arising from a *SUTVA* violation (δ)) from a single cross-section when W_i is mean dependent on u_i. However, the advantage of the antidotal variable method is it allows one to test whether W_i is correlated with u_i.

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

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

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

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

In this context mean independence of u_i and W_i implies that $E[u|D_i = 1, S_i = 0, W_i = 1] = E[u_i|D_i = 1, S_i = 0, W_i = 0]$, 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

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

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

However, if one has a cross-section when treatment is not assigned (a natural possibility is the data 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, nothing gets added due to treatment. Thus, the difference in mean of Y_i for $W_i = 1$ and $W_i = 0$ is only due to the difference between $E[u_i|D_i = 1, S_i = 0, W_i = 1]$ and $E[u_i|D_i = 1, S_i = 0, W_i = 0]$ or $E[u_i|D_i = 0, S_i = 1, W_i = 1]$ and $E[u_i|D_i = 0, S_i = 1, W_i = 0]$. Thus, whether these means are different, or not, can be tested with the simple regressions within the treated/untreated groups

$$\text{Treated group: } Y_i = \gamma_0 + \gamma_1 W_i + \zeta_i \qquad (F.1)$$

$$\text{Untreated group: } Y_i = \gamma_0' + \gamma_1' W_i + \zeta_i' \qquad (F.2)$$

When $\gamma_1 = 0$ and $\gamma'_1 = 0$, one can infer that non-randomness of W_i does not affect the parameter identification, as it should not, given there is no treatment to nullify. If $\gamma_1 \neq 0$ and $\gamma'_1 = 0$, then one identifies *ASEU* (δ), but the *ATE* or *ATT* and *SELB* (β_T and θ) cannot be identified. Conversely, if $\gamma_1 = 0$ and $\gamma'_1 \neq 0$, then one identifies β_T, but cannot identify *ASEU* (δ) and *SELB* (θ). If $\gamma_1 \neq 0$ and $\gamma_1 \neq 0$, then none of the parameters of interest are identified.[3]

[3] 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.

Appendix G: Simulations

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

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

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 G.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 estimators (the treatment effect, the *ASEU*, and *SELB*) approach the true parameter values when the sample size increases. Thus the simulation exercise confirms the consistency property of the estimators. Similar results emerge when we alter the parameter values.

Table G.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.

When W_i and D_i, are dependent

Identification does not require independence between W_i and D_i. This is because each parameter is identified based on subsamples in which one of either W_i or D_i remain fixed, while the other varies. To illustrate, the treatment effect is identified by comparing outcomes in n_1 and n_3. Both of these subsamples receive treatment ($D_i = 1$), so that D_i remains constant, and therefore is uncorrelated with W_i, which varies. Similarly, the *ASEU* is identified by comparing outcomes in n_2 and n_4. Again D_i is a constant equal to 0, while W_i varies. Likewise, the antidotal groups (n_3 and n_4), in which W_i = 0, are used to determine the selectivity. In none of these three identification subgroups do W_i and D_i vary together. Thus, the correlation between treatment and antidote assignment is zero so that all the parameters are identified. Simulation results in Table G.2 illustrate this assertion.

The following simulation results demonstrate the consistency of the estimators when W_i and D_i are dependent.

Table G.2: Estimates from the simulated data (W and D are dependent)

VARIABLES	(1) n=100	(2) n=1000	(3) n=10,000	(4) n=100,000
β_T	0.235	0.325	0.310	0.318
θ	0.222	0.125	0.149	0.142
δ	0.202	0.227	0.241	0.237
μ_0	0.741	0.706	0.700	0.700
No. of simulation	100	1,000	10,000	100,000
Correlation (D_i, W_i)	0.333	0.435	0.452	0.455

Notes: True parameter values: $\beta_T = 0.32;\ \theta = 0.14;\ \delta = 0.24;\ \mu_0 = 0.7$. The results are similar for other correlations between D_i, and W_i.

As the number of observations increases, the estimated parameter based on the simulated data becomes closer to the true parameter value, indicating statistical consistency. In addition, this empirical consistency holds true for other correlations between D_i, and W_i.