Learning From Imperfect Identification Strategies: Automating Causal Inference When Classic Assumptions Fail[*]

Guilherme Duarte
gjduarte@upenn.edu

Luke Keele
luke.keele@uphs.upenn.edu

Dean Knox
dcknox@upenn.edu

Jonathan Mummolo
jmummolo@princeton.edu

August 17, 2023

Abstract

Social science has developed an expansive design-based toolkit for applied causal inference, but the identifying assumptions that undergird standard approaches often fail in applied settings. In response, researchers often present unreliable results, narrow research questions post-hoc, or abandon projects altogether. In this paper, we demonstrate an alternative approach—automated partial identification—that allows researchers to learn as much as possible in these imperfect settings, while transparently acknowledging the limitations of their data and design. Using our Autobounds algorithm, analysts declare an estimand, state assumptions, and supply discrete data. The program then returns sharp bounds on the estimand, or a point-identified solution if one exists. Replicating numerous published studies across subfields of empirical political science, we show how our approach accommodates and extends classic designs including selection on observables, instrumental variables, difference in differences, and mediation analysis. Autobounds allows analysts to easily relax key assumptions and confront challenges—such as selection, mismeasurement, or missingness—to credibly study meaningful social questions.

Keywords: causal inference, partial identification, constrained optimization, linear programming, polynomial programming

[*] Guilherme Jardim Duarte is a Ph.D. candidate in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Luke Keele is a Research Associate Professor of Statistics in Surgery, Perleman School of Medicine, University of Pennsylvania. Dean Knox is an Andrew Carnegie Fellow and an assistant professor in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Jonathan Mummolo is an assistant professor of Politics and Public Affairs, Princeton University. Authors listed in alphabetical order. We gratefully acknowledge financial support from AI for Business and the Analytics at Wharton Data Science and Business Analytics Fund. This research was made possible in part by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the authors. Authors listed in alphabetical order.

Contents

1 Introduction

Social science has developed an expansive design-based toolkit for applied causal inference, greatly improving researchers' ability to draw credible inferences from observational data. But as all researchers know, applied work is messy: particular scenarios rarely conform perfectly to the ideal conditions under which standard approaches return reliable estimates of causal quantities. For example, (i) so-called selection-on-observables approaches function well in theory, but rely on the assumption that all relevant causal factors are measured and modeled correctly (Imbens and Rubin, 2015); (ii) instrumental variables approaches rely on the well-known "exclusion restriction" (Angrist et al., 1996a), meaning no direct causal path exists between the instrument and outcome; (iii) difference-in-differences designs rely on a "parallel trends" assumption (Abadie et al., 2010; Xu, 2017) that in the absence of treatment, treatment and control groups would trend in parallel over time; (iv) mediation designs rely on the "sequential ignorability" assumption (Imai et al., 2011) that no unobserved factors affect any combination of treatment, mediator, and outcome; and (v) analyses of selected data (Heckman, 1979) rely on the assumption that selection is not driven by any unobserved factor that also affects outcomes (Elwert and Winship, 2014). These assumptions, discussed in more detail below, are often difficult to defend outside the sterile examples that appear in statistics textbooks. Applied researchers have limited options when confronting thorny, real-world scenarios where there is reason to doubt key assumptions. Typically, researchers either ignore the problem and present unreliable results, narrow their focus to questions of lesser importance; or abandon projects altogether.

In this paper, we argue that an alternative approach, *partial identification*, represents a far more fruitful path for applied work when ideal conditions are unachievable. Rather than ignoring violations of key assumptions and making precise, unreliable claims, the goal of partial identification is to explicitly account for these violations and seek to *bound* the quantity of interest—that is, to report best- and worst-case values for the causal question, given the imperfect information at hand. In other words, partial identification acknowledges the difficulties of empirical research and asks, "Given all available information, what is the most that can be learned *in spite of these obstacles*?" We demonstrate how our proposed techniques allow analysts to relax all of the above assumptions.

The concept of partial identification is not new; a decades-long literature (Robins, 1989; Manski, 1990a; Heckman and Vytlacil, 2001; Zhang and Rubin, 2003; Cai et al., 2008; Swanson et al., 2018; Gabriel et al., 2020; Molinari, 2020) has developed and refined techniques for bounding causal quantities in particular scenarios (Lee, 2009; Gabriel et al., 2020; Kennedy et al., 2019; Knox et al., 2020; Li and Pearl, 2021; Sjölander et al., 2014). However, depending on the problem, the process of deriving sharp bounds—the narrowest possible range of answers to a causal question—can be at best tedious, potentially involving dozens of pages of algebra, and at worst entirely intractable. Scholars have at various points attempted to automate this process in specific settings (Balke and Pearl, 1994, 1997) or derive bounds for several variants of a problem (Swanson et al., 2018; Gabriel et al., 2020). However, a general and computationally feasible solution was only recently presented in Duarte et al. (2023). With this algorithm, Autobounds, users declare a causal quantity of interest (i.e. an estimand, such as an average treatment effect), state assumptions, and provide discrete data, i.e. data in which all variables take a finite and countable number of values—however incomplete or mismeasured. Using a dual-relaxation branch-and-bound optimization technique (Vigerske and Gleixner, 2018; Gamrath et al., 2020; Belotti et al., 2009), the algorithm then efficiently searches over possible data generating processes (DGPs) and locates ones which satisfy the constraints implied by the stated assumptions and observed data. When complete, it outputs *sharp bounds* on the estimand, the most precise possible solution given the information at hand, or a point-identified solution if one exists. As Duarte et al. (2023) states, "This approach can accommodate scenarios involving any classic threat to inference, including but not limited to missing data, selection, measurement error, and noncompliance" (1).

In short, this new tool allows researchers to flexibly confront idiosyncratic applied research environments without relying on canned, often implausible assumptions that come bundled with off-the-shelf causal-inference techniques. Instead, analysts can selectively invoke only the assumptions that are plausible in their applied settings, acknowledging the design and data limitations that they face. Researchers can transparently narrow the range of possible answers as much as possible until more data is obtained or additional information comes to light that justifies new assumptions. This approach can do more than simply add or remove assumptions; it can also partially relax assumptions, for example by stipulating that the treatment group may deviate from parallel trends by no more than some amount, or that a monotonicity assumption may be violated in no more than some percentage of cases. This allows researchers to conduct sensitivity analyses on virtually any aspect of an analysis. Our approach allows researchers to precisely characterize the empirical consequences of a potential assumption by quantifying how much it narrows the bounds, helping direct future research where it can advance scientific progress the most. As we show below, this tool can also, in certain cases, test every possible observable implication of a set of assumptions, flagging them as collectively *false* if observed data could not have possibly been generated by the theorized process.

Perhaps most importantly, this approach allows for *question-driven* research. The flexibility of this tool means that the statistical quantity of interest—the research question, formally stated—need not be retroactively adjusted to suit a particular applied setting. For example, researchers can use this tool to easily bound the average treatment effect in an entire population in a setting where an instrumental variables approach could only target a local average treatment effect among "compliers" (Angrist et al., 1996a). In other words, while the bounds will vary from case to case, the goal of a study need not change—a desirable feature as scholars seek to accumulate knowledge on a particular research question across independent studies.

To demonstrate the benefits of this approach, we replicate and extend published findings across empirical subfields of political science to show how standard assumptions can be relaxed in applied settings. These studies include examinations of support for the “Brexit” referendum (Hobolt, 2016); the consequences of bombing campaigns in Vietnam (Kocher et al., 2011); the determinants of counterinsurgency in the Peruvian Civil War (Schubiger, 2021); voter mobilization (Davenport et al., 2010); and racial bias in policing (Knox et al., 2020). These examples cover many classic approaches in the causal inference toolkit, including selection-on-observables, instrumental-variables, difference-in-differences, and mediation designs. In some cases, we show that relaxing standard assumptions shows that no firm conclusion can be drawn given available data (i.e., the bounds on the causal effect of interest include zero). In other cases, we show that informative bounds can be recovered even after abandoning assumptions previously thought to be pivotal. Regardless of the informativeness of the bounds, we show in each case how automated partial identification allows researchers to confront difficult challenges in a principled manner, all while pursuing the same meaningful social question that motivated a study to begin with.

This paper proceeds as follows. We first review the basic concepts of partial identification in causal inference. We then discuss the mechanics of the Autobounds algorithm at a high level to convey intuition on how partial identification can be automated. Having established the basic approach, we then apply it to several common research designs and inferential obstacles, replicating and extending published work to show how our algorithm performs in practice. These exercises illustrate how Autobounds allows analysts to relax or abandon key assumptions when their validity is suspect, investigate more meaningful estimands, and direct future research in efficient ways. We conclude with a discussion of the potential for automated partial identification to promote more credible, productive and transparent empirical social science.

2 What is Partial Identification?

First, we provide a review of the basic concepts on which our methods depend. Most of what we review are standard concepts from the causal inference literature. See Keele (2015) for a more detailed review. The first critical concept we introduce is that of a target or causal estimand. A target estimand is a contrast of counterfactuals that define the scientific question of interest. That is, the estimand defines the *exact* form of the causal effect of interest. It is separate from a statistical estimator or a specific point estimate that would be calculated from observed data. One way to define estimands is using the potential outcomes framework (Rubin, 1974). Next, we outline notation based on potential outcomes to formalize some example estimands.

Under this framework, potential outcomes represent unit level behaviors in the presence or absence of an intervention or treatment, and the actual outcome depends on the actual treatment received. We denote a binary treatment with $D_i \in \{0, 1\}$. The potential outcomes are $Y_i(D)$. The actual outcome is a function of treatment assignment and potential outcomes such that $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. In this framework, estimands are defined as contrasts of potential outcomes. For example, one possible counterfactual contrast or estimand is the average treatment effect (ATE):

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

which is the average difference in the pair of potential outcomes averaged over the entire population of interest. The equation for the ATE is referred to as a causal estimand, since it based on a contrast of potential outcomes – that is, it represents a contrast of counterfactuals. The target estimand does not represent a specific point estimate estimated from observed data. Often causal estimands are defined as averages over specific subpopulations. For example, we might average over subpopulations defined by pretreatment covariates such as sex and estimate the ATE for females only. When the estimand is defined for a specific subpopulation, it is said to be more local. Frequently, the average treatment effect is defined for the subpopulation exposed to the treatment or the average treatment effect on the treated (ATT).

$$ATT = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] \tag{2}$$

The estimand is a formal statement of the causal effect is of interest in a study. Moreover, the choice of the estimand is a critical part of scientific study. The ATE and ATT estimands are more standard estimands, since the counterfactual contrast only depends on the treatment. There are a number of estimands where the contrast of potential outcomes depend on quantities that are affected by the treatment. Estimands of this type are of particular interest, since conditioning on quantities affected by the treatment is widely known to cause bias (Rosenbaum, 1984a). For example, the method of instrumental variables focuses on a causal estimand, the complier average causal effect, that depends on the treatment (Angrist et al., 1996b).

For any estimand, we face what is known as an identification problem, since there are terms in the estimand that are unobservable. The key issue in causal identification problems is that even if we had samples of infinite size, we still could not estimate the average causal effect without observing both potential outcomes. In sum, no amount of data will allow us to observe the counterfactual quantities represented by the potential outcomes. As an example, we briefly outline the identification problem in the ATE estimand. First, we define π as the proportion of the sample assigned to the treated condition. Based on π, we can decompose the true ATE into the following set of terms:

$$\mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \pi \left\{\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1]\right\} + (1-\pi)\left\{\mathbb{E}[Y_i(1)|D_i = 0] - \mathbb{E}[Y_i(0)|D_i = 0]\right\}$$

Here, the ATE is a function of five quantities. Without additional assumptions we can estimate only three of those quantities directly from observed data. We can estimate π using $\mathbb{E}[D_i]$, and $\mathbb{E}[Y_i(1)|D_i = 1]$ and $\mathbb{E}[Y_i(0)|D_i = 0]$ using $\mathbb{E}[Y_i|D_i = 1]$ and $\mathbb{E}[Y_i|D_i = 0]$. However, we cannot estimate $\mathbb{E}[Y_i(1)|D_i = 0]$ and $E[Y_i(0)|D_i = 1]$ from the data without assumptions. One is the average outcome under treatment for those units in the control condition, and the other is the average outcome under control for those in the treatment condition. That is, we face an identification problem, since these two quantities are unobserved counterfactuals, and no additional amount of data will allow us to estimate these quantities. Therefore, we must find a set of assumptions that allow for identification. In an *identification analysis*, the analyst provides both a formal statement of the set of assumptions needed to identify a particular causal estimand and a proof that those assumptions lead to an identified causal effect. Often the identification analysis seeks to formally prove which non-model based (functional form) assumptions are needed for identification. This type of identification analysis is said to be nonparametric, since it is a formal statement of the assumptions that must hold to compute a causal effect with some hypothetical set of data that is infinite in size and without reference to any specific statistical model. Typically, investigators invoke an existing identification analysis and stipulates identification under that set of assumptions. A set of identification assumptions are often referred to as a "identification strategy." In sum, an identification strategy is simply a research design intended to solve the causal inference identification problem (Angrist and Pischke, 2010).

One tool that can be utilized for identification analysis is causal graphs or directed acyclic graphs (DAGs) (Pearl, 1995a). DAGs allow investigators to formalize identification concepts in a graphical manner. With DAGs, based on a given graph, we can derive nonparametric identification results and identify which variable or sets of variables are necessary for identification. Once the DAG is written down, it can be defended as a causal representation of a theory. Based on that structure one can then derive whether a causal effect is nonparametrically identified or not. DAGs are an essential part of the methodology we review next.

Most identification strategies are designed to result in point identification—identification of a single parameter that describes the causal effect of D_i. One alternative is the method of partial identification Manski (1990b, 1995). In a partial identification analysis, analysts seek to identify sharp bounds on the effect of interest. Sharp bounds are the upper and lower bounds that form an interval of possible values for the causal effect of interest, which provably cannot be narrowed without additional assumptions. The advantage of identifying bounds instead of single parameters is that, in most cases, one can invoke weaker identification assumptions. While point identification analyses use stronger assumptions to uniquely recover a single, that precise estimate that will generally be biased if these stronger assumptions do not hold. Partial identification allows analysts to navigate the trade-off between informativeness of bounds and strength of assumptions by using a nested series of models that add assumptions one at a time, obtaining successively tighter bounds on the quantity of interest. This approach clarifies the relationship between the strength of causal modeling assumptions and the amount of information about the quantity of interest that can be extracted from the available data. See Mebane and Poast (2013) and Keele and Minozzi (2012) for examples in political science.

3 A Brief Primer on Autobounds

Partial identification offers a principled approach to characterizing causal effects in the presence of incomplete information. However, deriving sharp bounds on effects manually is often analytically intractable. To address this, Duarte et al. (2023) provides an easy-to-use algorithm to automatically compute sharp bounds given the specific parameters of a causal question. For details on how this algorithm functions, we refer reads to Duarte et al. (2023); here, we seek to convey high-level intuition on how Autobounds works.

Autobounds takes four inputs: a causal theory, represented in a Directed Acyclic Graph (DAG); any additional assumptions not captured by the DAG (e.g. “no defiers”); a causal estimand (the target quantity, such as an Average Treatment Effect); and discrete data (i.e. all variables take a countable and finite number of values). The user also specifies two tolerance parameters corresponding to the desired level of provable sharpness and the desired width of the bounds.[1] The algorithm then searches over possible solutions—values of the causal estimand—that are consistent with all these parameters.

At the heart of this procedure is the concept of principal stratification (Frangakis and Rubin, 2002), the process of characterizing units in a study into their essential types based on how they would respond, counterfactually, when variables in the model take different values. Perhaps the most well-known example of principal stratification appears in the instrumental variables approach outlined in Angrist et al. (1996a). In this framework, an exogenous instrument, Z, is thought to encourage treatment, D, which in turn affects an outcome Y. Given a dichotomous instrument and treatment, Angrist et al. (1996a) describes four principal strata: "always takers," units which would accept treatment regardless of the value of Z; "never takers," units which would never accept treatment; "compliers," units which accept treatment if encouraged by the instrument Z but not otherwise; and "defiers," units which accept treatment in the absence of encouragement by Z, and reject treatment if encouraged.

While this classic setup identifies four principal strata based on how the treatment responds to the instrument, it is possible to represent any discrete causal model in terms of principal strata based on how every relevant variable in the system could possibly respond under various scenarios. As causal systems grow in complexity, the number of principal strata rapidly multiplies. However, so long as all variables in the system are discrete, the number of principal strata will be countable and finite, since there are only so many ways discrete variables can respond to other elements of the system. And if there are only a finite number of ways variables can respond to other variables, including unmeasured confounders, then there are only so many values the causal estimand can possibly take.[2]

Building on this intuition, Autobounds efficiently enumerates all the principal strata implied by a causal model and crucially, eliminates strata from consideration that cannot possibly

[1] See discussion of ϵ^{thresh} and θ^{thresh} in Duarte et al. (2023).

[2] This is true even in the case where unobserved confounders are continuous.

exist given the stated assumptions and observed data, or which provide redundant information. The program then expresses the causal estimand in terms of these principal strata, specifically, as a polynomial expression which can be computationally optimized, subject to the constraints implied by the causal theory, assumptions and data.

Once expressed in this fashion, discovering sharp bounds becomes a computational task. The main obstacle is obtaining solutions in a feasible amount of time, especially since, the number of principal strata explodes rapidly. To address this obstacle, Duarte et al. (2023) uses spatial branch and bound techniques that iteratively eliminate whole swaths of the model space as new extreme values (valid bounds) are discovered. The process continues until sharp bounds are obtained, or a point-identified solution is disovered (i.e. the bounds collapse on a point). In addition, in cases where computation time becomes prohibitive, Autobounds has the desirable feature of being "anytime" (Dean and Boddy, 1988), meaning bounds are always valid (but potentially non-sharp) no matter when the program is terminated.

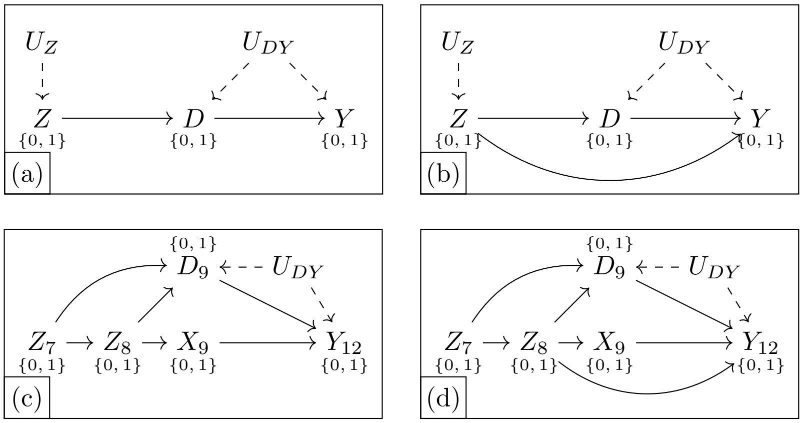
4 Applications

4.1 Instrumental Variables

When selection on observables cannot be credibly assumed—i.e., when treatment is not as-if randomly assigned, even conditional on observed covariates X—a popular identification strategy is instrumental variables. With this approach, treatment D and outcome Y still share a common unobserved confounder, U, but analysts locate an instrumental variable, Z, that "encourages" treatment to occur as-if randomly, which allows for the identification of a local average treatment effect of X on Y (Angrist et al., 1996a). This DGP is represented in the left panel of Figure 1.

As shown in Angrist et al. (1996a), the traditional instrumental variables strategy will recover the local average treatment effect among "compliers" (LATE)—units which respond to treatment status as instructed—if several assumptions hold. These conditions include (i)

Figure 1: **IV DAGs.** Panel (a) depicts the simple IV setting of Angrist et al. (1996a). In this setting, if monotonicity of Z to D is assumed (no defiers), the local ATE on the outcome Y, among those that "comply" with encouragement, is point identified. Panel (b) represents the same DAG with an additional arrow from Z to D, which indicates a violation of exclusion restriction. Panel (c) depicts the assumed DGP of Kocher et al. (2011). Instruments Z_7 and Z_8 represent Viet Cong control of hamlets in July and August and are assumed to encourage aerial bombing in September, D_9. Blocking control in September, X_9, is assumed to eliminate the direct effect of the instruments on hamlet control in December, Y_{12}. We probe the observable implications of this assumed model and find that it is inconsistent with data. Panel (d) is identical to panel (c) with the additional structural assumption that Z_8 directly causes Y_{12}.

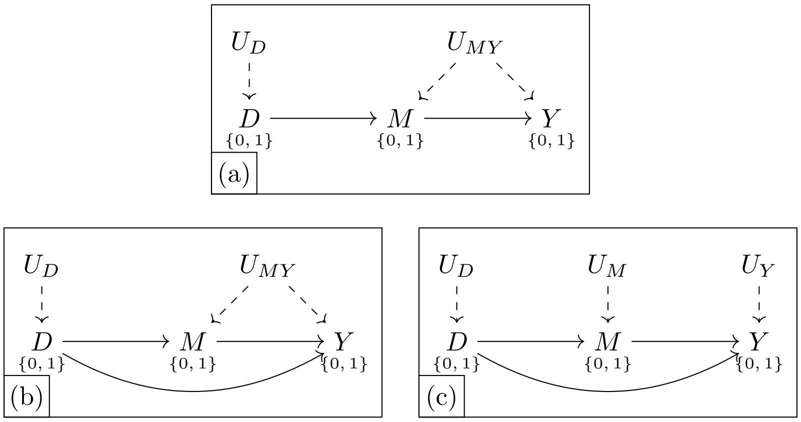


ignorability of Z_i, satisfied if the instrument is as-if randomly assigned; (ii) a non-null effect of Z_i on X_i, also known as "relevance"; (iii) an exclusion restriction, or the absence of a direct effect of Z_i on Y_i; and (iv) monotonicity, or the absence of "defiers" that behave inversely to instructions.[3] As Balke and Pearl (1997) shows, even when monotonicity is not assumed, it is possible to calculate sharp bounds for the ATE using a linear-programming approach. Autobounds generalizes this approach allowing for the calculation of sharp bounds not only for the ATE, but also for nonlinear quantities such as the LATE, and indeed for essentially any estimand. Moreover, our estimator will produce valid bounds both with or without monotonicity assumptions; in cases where the estimand is point identifiable, as in Angrist et al. (1996a), we obtain convergent bounds in which the best-case scenario is exactly equal to the worst.

Besides calculating bounds for estimands, Autobounds can also test for the well-known

[3] This result also assumes a stable unit treatment value assumption (SUTVA), which we employ throughout this paper.

Figure 2: **Mediation DAG.** Panel (a) depicts the assumed IV model of Green and Gerber (2002) and Gerber et al. (2003), in which randomized encouragement to vote in one election (D) is assumed to influence turnout in that election only (M) and not subsequent elections (Y). Panel (b) depicts the likely scenario of Davenport et al. (2010), in which D was delivered immediately before the November 2007 election (M) and analysts are interested in a subsequent election (Y) that occurred in January 2008. In this case, a direct $D \to Y$ effect is difficult to rule out, though analysts may assume that encouragement similarly has a monotonic effect on turnout in the second election. Finally, panel (c) depicts a strong additional assumption that analysts may wish to invoke, that turnout in both elections are not confounded.



instrumental inequalities (Pearl, 1995b; Bonet, 2001). In their general form, these state that for a valid instrument Z, discrete treatment D, and discrete outcome Y,

$$\max_d \sum_y \left[\max_z \ \Pr(Y = y, D = d | Z = z) \right] \leq 1$$

As this inequality is an implication of the IV graph, if data fails to satisfy it, then analysts may conclude there is a violation of the assumption set—potentially a direct $Z \to Y$ effect, or confounding of Z and some other variable.

[IN PROGRESS. originally written for mediation example and needs to be converted]

4.1.1 Problem Formulation

Case 2(a): Gerber et al. (2003), Simple IV

Structural Eq.	Response Func.	Response Form
$D = f_D(U_D)$	$f_D^{(U_D=u_D)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$M = f_M(D, U_{MY})$	$f_M^{(U_{MY}=u_{MY})}(d)$	$\{0,1\} \mapsto \{0,1\}$
$Y = f_Y(M, U_{MY})$	$f_Y^{(U_{MY}=u_{MY})}(m)$	$\{0,1\} \mapsto \{0,1\}$

Case 2(b–c): Mediation

$D = f_D(U_D)$	$f_D^{(U_D=u_D)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$M = f_M(D, U_{MY})$	$f_M^{(U_{MY}=u_{MY})}(d)$	$\{0,1\} \mapsto \{0,1\}$
$Y = f_Y(D, M, U_{MY})$	$f_Y^{(U_{MY}=u_{MY})}(d, m)$	$\{0,1\}^2 \mapsto \{0,1\}$

Estimand

Gerber et al. (2003) presumably target the LATE, given their use of 2SLS. However, it is implausible that

4.1.2 Replication and Extension of Davenport et al. (2010)

We are interested in three different estimands. The first estimand is related to post-treatment effect heterogeneity, or a difference between conditional ATEs (CATE) in which the grouping variable is a mediator (Stephens et al., 2016; Ertefaie et al., 2018)

$$\mathbb{E}[Y(D=1) - Y(D=0)|M=1] - E[Y(D=1) - Y(D=0)|M=0]$$

The remaining estimands are nested counterfactual mediation quantities. The second is the natural direct effect (NDE),

$$\mathbb{E}[Y(D=1, M(D=0)) - Y(D=0, M(D=0))]$$

and the third is the natural indirect effect (NIE)

$$\mathbb{E}[Y(D=0, M(D=1)) - Y(D=0, M(D=0))]$$

We bound the three estimands in the following scenarios with DAGs and functional assumptions:

1. Figure 2(b)

2. Figure 2(b) with weak (on-average) monotonicity from D to Y,

$$\mathbb{E}[Y(D=1, M=m) - Y(D=0, M=m)] \geq 0, \forall m$$

3. Figure 2(b) with strong (unit-level) monotonicity from D to Y,

$$\Pr[Y(D=0)=1, Y(D=1)=0] = 0$$

4. Figure 2(b) with strong monotonicity from D and M to Y,

$$\Pr[Y(D=0, M=m)=1, Y(D=1, M=m)=0] = 0 \ \forall \ m$$

5. Figure 2(b) with no interaction,

$$\mathbb{E}[Y(D=1, M=0) - Y(D=0, M=0) - Y(D=1, M=1) + Y(D=0, M=1)] = 0$$
$$\mathbb{E}[Y(D=0, M=1) - Y(D=0, M=0) - Y(D=1, M=1) + Y(D=1, M=0)] = 0$$

6. Figure 2(a)

7. Figure 2(c)

Table 1: Results for the re-analysis of (Davenport et al., 2010) in Autobounds. Columns indicate each different estimand: a) Difference of Heterogeneous ATEs; b) Natural Direct Effect (NDE); c) Natural Indirect Effect (NIE). Rows indicate which sets of assumptions (DAGs) are being used. (1–5) use Figure 2(b), with the following additional assumptions: (1) has no additional assumptions; (2) assumes weak monotonicity of D to Y; (3) assumes strong monotonicity of D to Y; (4) assumes monotonicity of D and M to Y; and (5) assumes no interactions. Finally, (6) uses 2(a) and (7) uses 2(c)

.

	Difference of CATEs	NDE	NIE
Scenario 1	[-0.54, 0.145]	[-0.649, 0.145]	[-0.661, 0.155]
Scenario 2	[-0.54, 0.145]	[-0.58, 0.145]	[-0.661, 0.155]
Scenario 3	[-0.015, 0.033]	[-0.645, 0.039]	[-0.645, 0.039]
Scenario 4	[-0.52, 0.045]	[0, 0.155]	[-0.661, 0.01]
Scenario 5	[-0.54, 0.145]	[-0.649, 0.145]	[-0.661, 0.155]
Scenario 6	[-0.156, -0.001]	[-0.003, -0.003]	[0.013, 0.013]
Scenario 7	[-0.41, 0.035]	[0,0]	[0.01, 0.001]

4.2 Difference in Differences

Difference-in-differences (DID) is another widely used identification strategy designed to neutralize the influence of unobserved confounding. In the simplest case, the DID strategy involves comparing the outcomes of two groups of observations (treatment and control) in two time periods (before and after some intervention is applied to the treatment group, but not to the control group). The average pre-post change in the treatment group is then contrasted with the average pre-post change in the control group and under certain assumptions, this comparison identifies the average treatment effect on the treated (ATT; formalized below).

To demonstrate the usefulness of our technique for probing these assumptions, we replicate and extend Schubiger (2021), which relies in part on a DID strategy to estimate the effect of exposure to state violence on counterinsurgent mobilization in the Peruvian Civil War. As the study states, "The core challenge to answering this question lies in the fact that even though state violence was highly unpredictable during the counterinsurgency campaign of 1983–85, targeting did not occur at random, thus being potentially related to other important determinants of communities' propensity for counterinsurgent collective action," (1388).

In this study, the units of analyis are *centro poblados*, "settlements of various sizes and types, such as villages and towns," some of which targets of state violence. This analysis examines two time periods: 1983–1985, during which time state violence—including human rights violations—was imposed on various towns and villages in response to a counterinsurgency, (the pre period); and 1986–1988, during which time some localities employed "self-defense committees" which engaged in violent clashes with the state (the post period). The analysis examines two groups of localities: those that experienced state violence in the pre-period (the treatment group) and those that did not (the control group). The outcome is a binary indicator of "whether a given *centro poblado* was affected by violence against or perpetrated by self-defense committees in the period after the counterinsurgency campaign (1986–88)" (1390).

A key identifying assumption necessary for DID is parallel trends—in the absence of any intervention, the average outcomes in the treatment and control groups would have continued in parallel in the post-intervention period. This assumption is generally regarded as untestable, since by definition the treatment group's counterfactual trend in the post-period cannot be observed. To probe this assumption, researchers typically examine pre-trends to see if outcomes were trending in parallel prior to the intervention. However, as Schubiger (2021) notes, "As there is only one pretreatment period, pretreatment trends cannot be explored in detail. . . " (1395).

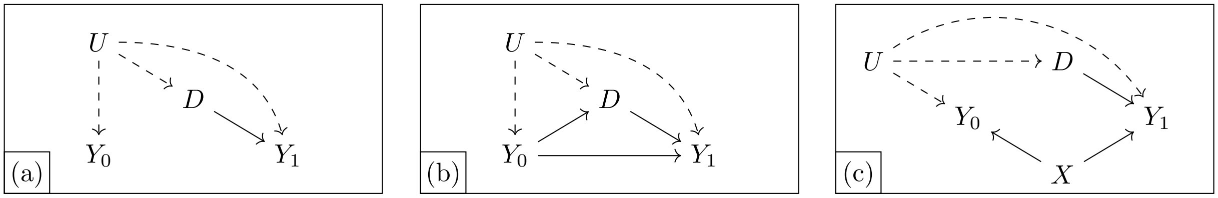
In what follows, we use Autobounds to demonstrate what can be learned from these data when the parallel trends assumption is relaxed.

4.2.1 Problem Formulation

Difference-in-differences is a strategy that involves at least three variables: pre-treatment outcome Y_0, treatment D, and post-treatment outcome Y_1. Our estimand is the average treatment effect on the treated (ATT)[4] in the post period:

[4] By the axiom of consistency, $E[Y_1(D=1)|D=1] = E[Y_1|D=1]$, so the first element of the estimand is immediately identifiable. However, the second is not.

Figure 3: **xxx.** xxx.



$$\mathbb{E}[Y_1(D=1)|D=1] - \mathbb{E}[Y_1(D=0)|D=1]. \tag{3}$$

For simplicity of exposition, we will first work with the minimal model in Figure 3(a), in which the initial outcome Y_1 is confounded with both treatment D and the subsequent outcome Y_1, but does not have a direct causal effect. More complex causal models in which the pre-treatment outcome has various effects on subsequent variables, such as those depicted in Figure 3(b), are straightforward to accommodate with Autobounds. We will further demonstrate how relaxations of the DID framework that use bracketing trends based on some auxiliary variable—a framework that includes results such as monotone trends (discussed below; Hasegawa et al., 2019; Ye et al., 2020)—are easily accommodated by Autobounds. An example DGP for this approach is given in Figure 3(c).

Case 3(a): Simple DID

Structural Eq.	Response Func.	Response Form
$Y_0 = f_{Y_0}(U)$	$f_{Y_0}^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(U)$	$f_D^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Y_1 = f_{Y_1}(D,U)$	$f_{Y_1}^{(U=u)}(d)$	$\{0,1\} \mapsto \{0,1\}$

Case 3(b): DID with Effects of Pre-treatment Outcome

Structural Eq.	Response Func.	Response Form
$Y_0 = f_{Y_0}(U)$	$f_{Y_0}^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$D = f_D(Y_0, U)$	$f_D^{(U=u)}(y_0)$	$\{0,1\} \mapsto \{0,1\}$
$Y_1 = f_{Y_1}(Y_0, D, U)$	$f_{Y_1}^{(U=u)}(y_0, d)$	$\{0,1\}^2 \mapsto \{0,1\}$

Case 3(c): Bracketing Trends in X

Structural Eq.	Response Func.	Response Form
$X = f_X(U)$	$f_X^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Y_0 = f_{Y_0}(U, X)$	$f_{Y_0}^{(U=u)}(x)$	$\{0,1\} \mapsto \{0,1\}$
$D = f_D(Y_0, U)$	$f_D^{(U=u)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$Y_1 = f_{Y_1}(X, D, U)$	$f_{Y_1}^{(U=u)}(x, d)$	$\{0,1\}^2 \mapsto \{0,1\}$

Our estimand,

$$\begin{aligned}
\mathcal{T}_{\text{ATT}} &= \mathbb{E}[Y_1(D=1)|D=1] - \mathbb{E}[Y_1(D=0)|D=1] \\
&= \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_1}^{(U=u)}(d=1) = 1, f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u)} \\
&\quad - \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_1}^{(U=u)}(d=0) = 1, f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u)}
\end{aligned}$$

is a fractional expression and thus cannot be inserted directly into the optimization problem. To handle this, we redefine our estimand using an auxiliary variable, $\mathcal{T}' = s$, which then becomes the objective function of the polynomial programming problem; we then create a new polynomial constraint of the form

$$\begin{aligned}
& s \cdot \textstyle\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u) \\
&\quad = \textstyle\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_1}^{(U=u)}(d=1) = 1, f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u) \\
&\qquad - \textstyle\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_1}^{(U=u)}(d=0) = 1, f_D^{(U=u)}(\varnothing) = 1 \right\} \cdot \Pr(U=u)
\end{aligned}$$

We now turn to the key *parallel trends* assumption, which states that on average, the post-treatment potential outcome under control, $Y_1(d=0)$, will differ from the pre-treatment outcome, Y_0, by exactly the same amount for both treated and control groups. That is,

$$\mathbb{E}\left[Y_1(D=0) - Y_0 | D=1\right] = \mathbb{E}\left[Y_1(D=0) - Y_0 | D=0\right],$$

which translates to

$$\begin{aligned}
0 &= \Pr[Y_1(D=0)=1|D=1] - \Pr[Y_0=1|D=1] \\
&\quad - \Pr[Y_1(D=0)=1|D=0] + \Pr[Y_0=1|D=0] \\
&= \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_1}^{(U=u)}(d=0)=1, f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)} \\
&\quad - \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_0}^{(U=u)}(\varnothing)=1, f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing)=1 \right\} \cdot \Pr(U=u)} \\
&\quad + \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_1}^{(U=u)}(d=0)=1, f_D^{(U=u)}(\varnothing)=0 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing)=0 \right\} \cdot \Pr(U=u)} \\
&\quad - \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_{Y_0}^{(U=u)}(\varnothing)=1, f_D^{(U=u)}(\varnothing)=0 \right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{ f_D^{(U=u)}(\varnothing)=0 \right\} \cdot \Pr(U=u)}
\end{aligned}$$

which creates a polynomial-fractional constraint with differing denominators, in which the fraction can be cleared by further algebraic manipulation.

An alternative strategy is to divide the control group into two groups on some variable, $X=0$ and $X=1$, creating two observed trends in their outcomes. The treatment group's trend is then assumed to be *bracketed* by the trends in the two control groups (Ye et al., 2020). That is,

$$\mathbb{E}[Y_1(d=0) - Y_0|D=0, X=0] \;\leq\; \mathbb{E}[Y_1(d=0) - Y_0|D=1] \;\leq\; \mathbb{E}[Y_1(d=0) - Y_0|D=0, X=1]$$

The monotone trends approach of Hasegawa et al. (2019) is a special case of the bracketing approach in which $X = \mathbb{1}\{Y_0 \geq \theta\}$ for some threshold θ.[5] In this case, the key bracketing assumption may be justified if the outcome follows a "rich-get-richer" pattern, such as an exponential growth model.

These inequalities,

$$\mathbb{E}\left[Y_1(D=0) - Y_0 | D=0, X=x\right] \bigstar \mathbb{E}\left[Y_1(D=0) - Y_0 | D=1\right],$$

(where $\star$ is $\leq$ for $x=0$ and $\geq$ for $x=1$) then translate to

$$\begin{aligned}
0 \bigstar\ & \Pr[Y_1(D=0, X=x) = 1 | D=1] - \Pr[Y_0 = 1 | D=1] \\
& - \Pr[Y_1(D=0, X=x) = 1 | D=0, X=x] + \Pr[Y_0 = 1 | D=0, X=x] \\
\bigstar\ & \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_{Y_1}^{(U=u)}(d=0, x) = 1, f_D^{(U=u)}(\varnothing) = 1\right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_D^{(U=u)}(\varnothing) = 1\right\} \cdot \Pr(U=u)} \\
& - \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_{Y_0}^{(U=u)}(\varnothing) = 1, f_D^{(U=u)}(\varnothing) = 1\right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_D^{(U=u)}(\varnothing) = 1\right\} \cdot \Pr(U=u)} \\
& + \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_{Y_1}^{(U=u)}(d=0, x) = 1, f_D^{(U=u)}(\varnothing) = 0, f_X^{(U=u)}(\varnothing) = x\right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_D^{(U=u)}(\varnothing) = 0, f_X^{(U=u)}(\varnothing) = x\right\} \cdot \Pr(U=u)} \\
& - \frac{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_{Y_0}^{(U=u)}(\varnothing) = 1, f_D^{(U=u)}(\varnothing) = 0, f_X^{(U=u)}(\varnothing) = x\right\} \cdot \Pr(U=u)}{\sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{f_D^{(U=u)}(\varnothing) = 0, f_X^{(U=u)}(\varnothing) = x\right\} \cdot \Pr(U=u)}
\end{aligned}$$

As before, constraints arising from the axioms of probability are $\Pr(U=u) \geq 0$ for all u and $\sum_{u \in \mathcal{S}(U)} \Pr(U=u) = 1$; constraints arising from empirical evidence in the simple DID

[5] We note that this approach behaves poorly with binary data, as X then perfectly separates units with $Y_0 = 0$ and $Y_0 = 1$ for $\theta \in (0, 1)$.

case of Figure 3(a) are

$$\begin{aligned}&\Pr(Y_0 = y_0, D = d, Y_1 = y_1)\\&= \sum_{u \in \mathcal{S}(U)} \mathbb{1}\left\{\begin{array}{r} f_{Y_0}^{(U=u)}(\varnothing) = y_0, \\ f_{D}^{(U=u)}(\varnothing) = d, \\ f_{Y_1}^{(U=u)}(d) = y_1 \end{array}\right\} \cdot \Pr(U = u).\end{aligned}$$

We omit the extensions to Figure 3(b–c) cases, which are largely identical.

4.2.2 Replication and Extension of Schubiger (2021)

We demonstrate how Autobound can be used to relax the parallel trends assumption by replicating and extending (Schubiger, 2021). Initially, we introduced the DAG of Figure 3a. Faithfully to the original paper, we assumed parallel trends. We then introduced data with respect to all the variables, D, Y_0, and Y_1. Two estimands were analyzed. For the ATT, we obtained convergent bounds of 0.047. This is exactly the result obtained by Schubiger (2021). However, Autobounds also allows us to calculate bounds for the ATE: $[0.001, 0.956]$. In other words, we can be sure that the result will be positive. This last result is an improvement over the bounds if we had assumed pure confounding with no parallel trends, where the ATE would have been $[-0.044, 0.956]$ and the ATT, $[-0.940, 0.052]$. The results for the ATE and ATT from assuming parallel trends do not change from DAG 1 to DAG 2.

Finally, we tested the model in 3c. Here, rather than assuming parallel trends, we assumed a type of bracket trends, where for all control variables X in Schubiger (2021), we calculated trend values $g(x) = E[Y|D = 0, X = x] - E[Z|D = 0, X = x]$. We then assume that $\min_x g(x) \leq E[Y|D = 1] - E[Z|D = 1] \leq \max_x g(x)$. With this assumption, we find that the bounds for the ATT are $[0.032, 0.052[$ and the bounds for the ATE are $[0.001, 0.956]$.

As this final example shows, replacing the relatively restricting parallel trends assumption to the more lenient bracketing trends assumption, analysts are still able to sign the ATT and ATE as positive. This shows that in situations where parallel trends are thought to be violated, it is still possible to recover substantively informative results in a difference-in-differences setting using automated partial identification.

4.3 Mediation Analysis

4.3.1 Problem Formulation

4.3.2 Replication and Extension of Grewal et al. (2019)

Grewal et al. (2019) features a series of laboratory experiments conducted in Tunisia to study the relationship between feelings of economic strain and support for Tunisia's Islamist Party. More specifically, these experiments test the hypothesis that the effect of perceived economic strain on Islamist support is mediated by a desire for divine rewards in the afterlife.

In these experiments,

encouragement:

"In the second experiment, random economic strain was induced by exposing respondents to hypothetical financial scenarios employed by Mani et al. (2013)."

"Second, even if such a manipulation is possible, the use of these designs requires the consistency assumption that the manipulation of the mediator should not affect the outcome through any pathway other than the mediator." (Imai et al., 2013, 6) [WE SHOULD DROP AND RELAX THIS ASSUMPTION, which they call the Consistency Assumption]

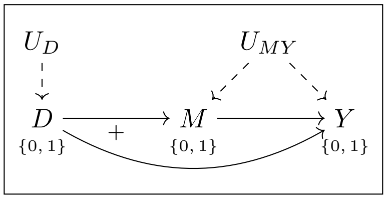
XXX i believe there is also a "no interaction effect" assumotion we can explore

4.4 Selection Bias

Selection bias is often inherent to the data sources used by social scientists. A particularly difficult problem arises when treatment status determines whether units are observed or not, and a large literature has grappled with the statistical bias that can result from this form of post-treatment conditioning (Rosenbaum, 1984b; Acharya et al., 2016; Nyhan et al., 2017; Blackwell, 2013).

In one recent example, Knox et al. (2020) examines the problem of estimating racial bias

Figure 4: **Racial Discrimination in Police Use of Force.** D represents the minority status of an encountered civilian. M is the mediating decision of whether an officer chooses to detain the civilian. Y is use of force.



in the use of force by police using only data on encounters in which police choose to detain individuals. As the paper shows, because the race of civilians involved in police encounters (the treatment, D) very likely affects whether individuals are detained in the first place (an indicator for whether a person is stopped by police, M), then comparing the rates of force (Y) used against white and nonwhite civilians among the subset of encounters that involve a detainment leads to underestimates of racial bias in the use of force, absent implausible assumptions.

The source of the confounding that results from this form of sample selection can be seen in Figure 4. As the DAG shows, even if the analyst is able to adjust for all common causes of the treatment and outcome (D and Y), and of the treatment and mediator (D and M)—equivalent to rendering encounters comparable before the officer makes the decision to initiate a stop—conditioning on the mediator induces "collider bias" (Pearl, 2009), allowing common causes of stopping (M) and the use of force (Y) to confound comparisons. Theoretically, these unobserved confounders, represented collectively by U, could be factors never recorded in police administrative data such as the officer's mood at the time of the encounter.

To address this obstacle, Knox et al. (2020) derives nonparametric sharp bounds on several causal estimands corresponding to racial bias. In general, these estimands consider the counterfactual substitution of a different individual of differing racial identity into an otherwise similar police-civilian encounter, and compare the average counterfactual rates of force between two encounters involving two racial/ethnic groups of civilians. To sharply bound these estimands given available administrative data, Knox et al. (2020) appeals to four assumptions.

The first assumption is *Mandatory Reporting*, stated formally as:

Assumption 1 (Mandatory Reporting)**.** $Y(d, 0) = 0$ *for* $d \in \{0, 1\}$*.*

This assumption assumes that all uses of force during the police-civilian interactions under study are recorded in data. In other words, if a stop does not occur ($M = 0$), then force does not occur ($Y = 0$).

The second assumption is mediator monotonicity, stated as:

Assumption 2 (Mediator Monotonicity)**.** $M(1) \geq M(0)$*.*

This assumption states that there are no encounters in which a stop would occur if the civilian was white ($D = 0$), but would not be stopped, counterfactually, if the civilian was nonwhite ($D = 1$). In other words, the assumption is that anti-white bias in stopping does not exist.

Third, Knox et al. (2020) make an assumption about the average levels of force applied in different types of encounters, defined by principal strata (Frangakis and Rubin, 2002):

Assumption 3 (Relative Non-severity of Racial Stops)**.**
$\mathbb{E}[Y(d, m)|D = d', M(1) = 1, M(0) = 1] \geq \mathbb{E}[Y(d, m)|D = d', M(1) = 1, M(0) = 0]$

This assumption holds that conditional on the race of civilians, D, the average rate of force used during police encounters is larger in encounters where civilians would be stopped regardless of race ($M(1) = M(0) = 1$) than in encounters where nonwhite civilians would be discriminatorily stopped ($M(1) = 1$, $M(0) = 0$). The logic behind this assumption is that the former class of encounters ("always-stop encounters") are theorized to be serious incidents in progress such as an armed robbery, where officers have little choice but to detain a civilian and thus a higher likelihood of using force, whereas the second class of encounters are more discretionary scenarios that allow officer bias to influence the stopping decision.

Finally, the approach requires treatment ignorability with respect to the mediator and outcome, i.e. encounters are all-else-equal prior to the stop. This assumption is encoded in the given DAG. In addition, the bounding approach in Knox et al. (2020) allows analysts to specify the severity of bias in the initial decision to stop civilians to obtain sharp bounds for that scenario. The paper shows the severity of stopping bias can be lower bounded using a standard outcome test (Knowles et al., 2001), and estimates that at least 32% of stops of Black civilians would not have occurred had similarly situated white civilians been encountered in the New York City case. In keeping with this result, we specify the parameter indicating racial bias in stopping at 0.32 in the replication below.

While Knox et al. (2020) argue that these assumptions are at least plausible in the empirical setting they examine (New York City in the 2000s, during which time the controversial "Stop, Question and Frisk" tactic was prevalent (Mummolo, 2018)), others may question their validity. Moreover, it would be difficult to apply this bounding solution in other settings where post-treatment selection occurs. This is especially true in the study of policing, where policies, norms and data availability vary tremendously across the roughly 18,000 state and local agencies in the U.S.

In the following section, we use Autobounds to relax and abandon these assumptions to demonstrate how analysts can flexibly compute sharp bounds on causal effects across diverse data environments suffering from selection bias where particular identifying assumptions may not be justified.

4.4.1 Problem Formulation

Structural Eq.	Response Func.	Response Form
$D = f_D(U_D)$	$f_D^{(U_D=u_D)}(\varnothing)$	$\varnothing \mapsto \{0,1\}$
$M = f_M(U_{MY})$	$f_M^{(U_{MY}=u_{MY})}(d)$	$\{0,1\} \mapsto \{0,1\}$
$Y = f_Y(U_{MY})$	$f_Y^{(U_{MY}=u_{MY})}(d,m)$	$\{0,1\}^2 \mapsto \{0,1\}$

Assumption 1 (Mandatory Reporting)

This assumption states that $\Pr[Y(d,0)=1]=0$ for all d, which translates to

$$0 = \sum_{u_{MY} \in \mathcal{S}(U_{MY})} \mathbb{1}\left\{ f_Y^{(U_{MY}=u_{MY})}(d, m=0) = 1 \right\} \cdot \Pr(U_{DY} = u_{DY})$$

where blocked disturbances are factorized and eliminated as before.

Assumption 2 (Mediator Monotonicity)

This assumption states that $\Pr[M(d=1)=0, M(d=0)=1]=0$, which translates to

$$0 = \sum_{u_{MY} \in \mathcal{S}(U_{MY})} \mathbb{1}\left\{ \begin{array}{l} f_M^{(U_{MY}=u_{MY})}(d=1)=0, \\ f_M^{(U_{MY}=u_{MY})}(d=0)=1 \end{array} \right\} \cdot \Pr(U_{DY} = u_{DY})$$

Assumption 3 (Relative Non-severity of Racial Stops)

This assumption states that $\Pr[Y(d,m)=1|D=d', M(1)=1, M(0)=1] \geq \Pr[Y(d,m)=1|D=d', M(1)=1, M(0)=0]$, which translates to

$$0 \geq \frac{\displaystyle\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbb{1}\left\{ \begin{array}{l} f_Y^{(U_{MY}=u_{MY})}(d,m)=1, \\ f_D^{(U_D=u_D)}(\varnothing)=d' \\ f_M^{(U_{MY}=u_{MY})}(d=1)=1, \\ f_M^{(U_{MY}=u_{MY})}(d=0)=0 \end{array} \right\} \cdot \Pr(U_{DY}=u_{DY})}{\displaystyle\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbb{1}\left\{ \begin{array}{l} f_D^{(U_D=u_D)}(\varnothing)=d' \\ f_M^{(U_{MY}=u_{MY})}(d=1)=1, \\ f_M^{(U_{MY}=u_{MY})}(d=0)=0 \end{array} \right\} \cdot \Pr(U_{DY}=u_{DY})}$$

$$- \frac{\displaystyle\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbb{1}\left\{ \begin{array}{l} f_Y^{(U_{MY}=u_{MY})}(d,m)=1, \\ f_D^{(U_D=u_D)}(\varnothing)=d' \\ f_M^{(U_{MY}=u_{MY})}(d=1)=1, \\ f_M^{(U_{MY}=u_{MY})}(d=0)=1 \end{array} \right\} \cdot \Pr(U_{DY}=u_{DY})}{\displaystyle\sum_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbb{1}\left\{ \begin{array}{l} f_D^{(U_D=u_D)}(\varnothing)=d' \\ f_M^{(U_{MY}=u_{MY})}(d=1)=1, \\ f_M^{(U_{MY}=u_{MY})}(d=0)=1 \end{array} \right\} \cdot \Pr(U_{DY}=u_{DY})}$$

where the complaint is duplicated for each (d, d', m) triple.

Empirical Constraints

The observed data consists only of $\widehat{\Pr}(D = d, Y = y|M = 1)$.

$$\widehat{\Pr}(D = d, Y = y|M = 1) = \frac{\sum\limits_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbb{1}\left\{\begin{array}{l} f_Y^{(U_{MY}=u_{MY})}(d, m = 1) = y, \\ f_M^{(U_{MY}=u_{MY})}(d) = 1, \\ f_D^{(U_D=u_D)}(\varnothing) = d \end{array}\right\} \cdot \Pr(U_{DY} = u_{DY})}{\sum\limits_{\substack{u_D \in \mathcal{S}(U_D) \\ u_{MY} \in \mathcal{S}(U_{MY})}} \mathbb{1}\left\{f_M^{(U_{MY}=u_{MY})}(d) = 1\right\} \cdot \Pr(U_{DY} = u_{DY})}$$

Estimand

The primary objective function is the ATE among detained individuals, $\text{ATE}_{M=1} = \mathbb{E}[Y(d = 1) - Y(d = 0)|M = 1] = \mathbb{E}[Y(d = 1, M(d = 1)) - Y(d = 0, M(d = 0))|M = 1]$. This is automatically translated as

$$\begin{aligned}
\text{ATE}_{M=1} = {} & \mathbb{E}[Y(d = 1, M(d = 1))|D = 1, M(d = 1) = 1] \Pr(D = 1) \\
& + \mathbb{E}[Y(d = 1, M(d = 1))|D = 0, M(d = 0) = 1] \Pr(D = 0) \\
& - \mathbb{E}[Y(d = 0, M(d = 0))|D = 1, M(d = 1) = 1] \Pr(D = 1) \\
& - \mathbb{E}[Y(d = 0, M(d = 0))|D = 0, M(d = 0) = 1] \Pr(D = 0) \\
= {} & \Pr[Y(d = 1, M(d = 1) = 1) = 1|D = 1, M(d = 1) = 1] \Pr(D = 1) \\
& + \Pr[Y(d = 1, M(d = 1) = 0) = 1|D = 0, M(d = 0) = 1] \Pr[M(d = 1) = 0|D = 0, M(d = 0) = \\
& + \Pr[Y(d = 1, M(d = 1) = 1) = 1|D = 0, M(d = 0) = 1] \Pr[M(d = 1) = 1|D = 0, M(d = 0) = \\
& - \Pr[Y(d = 0, M(d = 0) = 0) = 1|D = 1, M(d = 1) = 1] \Pr[M(d = 0) = 0|D = 1, M(d = 1) = \\
& - \Pr[Y(d = 0, M(d = 0) = 1) = 1|D = 1, M(d = 1) = 1] \Pr[M(d = 0) = 1|D = 1, M(d = 1) = \\
& - \Pr[Y(d = 0, M(d = 0) = 1) = 1|D = 0, M(d = 0) = 1] \Pr(D = 0)
\end{aligned}$$

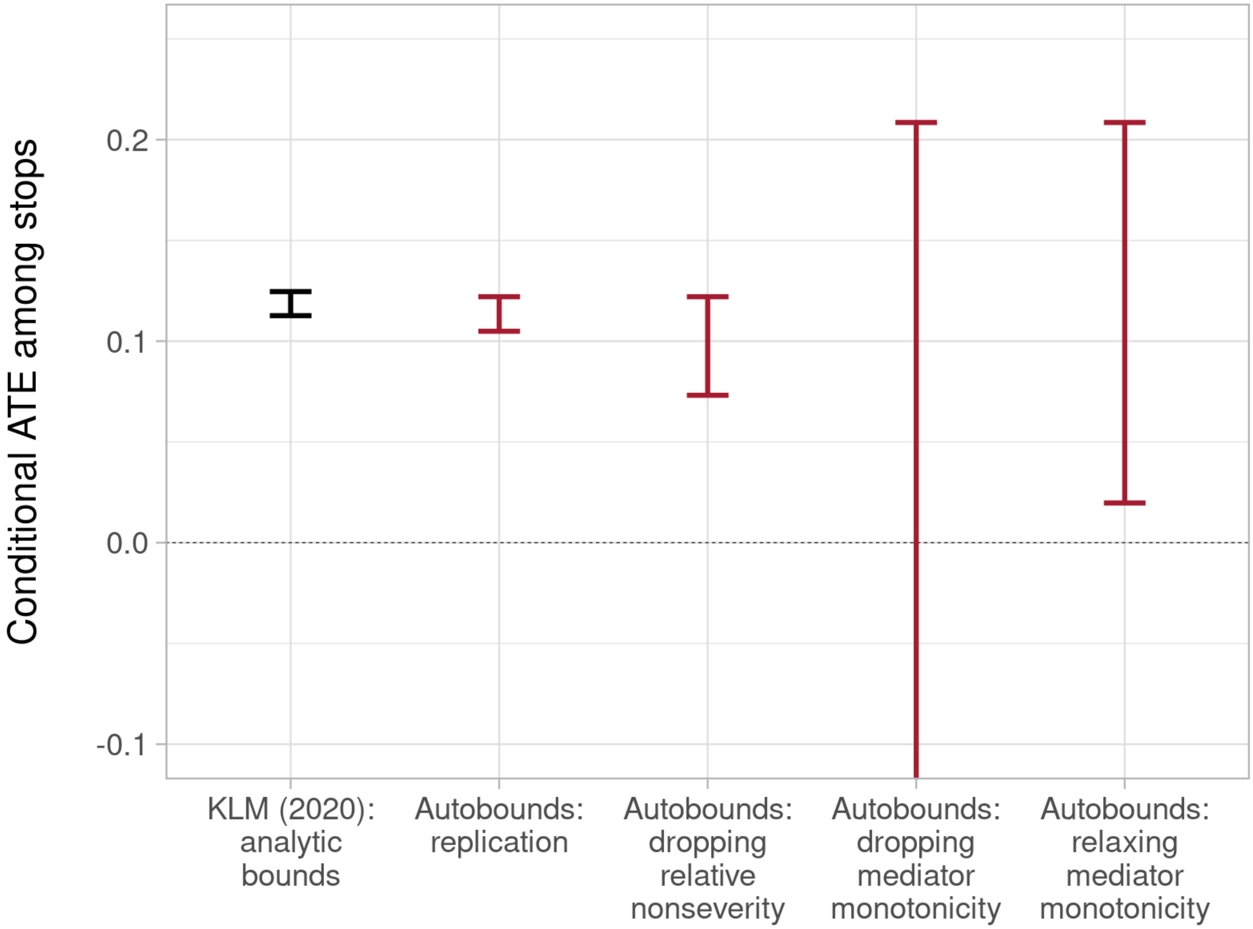
where the further translation to response functions and disturbance partitions is omitted.

4.4.2 Replication and Extension of Knox et al. (2020)

While the original bounding solution in Knox et al. (2020) relies on the aforementioned four identifying assumptions, Autobounds allows us to relax or abandon these assumptions with ease. The far left estimate in Figure 5 shows the original bounds for the $ATE_{M=1}$ for the use of any force during encounters between Black and White civilians computed in Knox et al. (2020) using an analytic derivation with no adjustment for covariates, $[0.112, 0.124]$; see Table 1 in Knox et al. (2020). The second estimate shows that when given the same parameters and data, Autobounds recovers virtually identical bounds $[0.105, 0.122]$ up to numeric stability issues that remain an area of active investigation. In the third entry in the plot, we drop Assumption 3, meaning we make no assumption about the relative severity of force used in different principal strata. Somewhat surprisingly, when we do this, the bounds on the estimand widen only slightly, $[0.073, 0.122]$. This means that analysts can still learn a great deal about the possible severity of racial bias in the use of force in scenarios where this assumption fails.

On the other hand, the following estimate in the plot shows that Assumption 2, mediator monotonicity, is much more consequential. When this assumption is dropped, the bounds on the $ATE_{M=1}$ become much less informative, $[-1.000, 0.209]$, meaning the sign of the effect can no longer be determined. However, as the final estimate shows, the effect can be signed as positive if we simply relax Assumption 2 and specify that no more than 5% of stops of White civilians occur discriminatorily, producing $[0.020, 0.209]$.

Figure 5: **Sharp Bounds on Racial Bias in the Use of Force by Police Under Various Assumptions**. The figure displays sharp bounds on racial bias in the use of force ($ATE_{M=1}$) by police in New York City using data from (Knox et al., 2020). Autobounds replicates the analytic result computed in the original paper. Dropping an assumption about the relative severity of force across principal strata of encounters still allows anaylsts to bound racial bias as positive. Dropping the assumption of no anti-White bias in stopping leads to uninformative bounds, but relaxing this assumption to allow for no more than 5% of anti-White discriminatory stops results in positive bounds on the causal estimand.



References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

Acharya, A., M. Blackwell, and M. Sen (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *Biometrics 110*(3), 512–529.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996a). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996b, June). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and J̈.-S. Pischke (2010, Spring). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Balke, A. and J. Pearl (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Belotti, P., J. Lee, L. Liberti, F. Margot, and A. Wächter (2009). Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software 24*(4-5), 597–634.

Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science 57*(2), 504–520.

Bonet, B. (2001). A calculus for causal relevance. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 40–47.

Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics 64*(3), 695–701.

Davenport, T. C., A. S. Gerber, D. P. Green, C. W. Larimer, C. D. Mann, and C. Panagopoulos (2010). The enduring effects of social pressure: Tracking campaign experiments over a series of elections. *Political Behavior 32*(3), 423–430.

Dean, T. L. and M. Boddy (1988). An analysis of time-dependent planning. pp. 49–54. American Association for Artificial Intelligence.

Duarte, G., N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser (2023). An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*.

Elwert, F. and C. Winship (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology 40*, 31.

Ertefaie, A., J. Y. Hsu, L. C. Page, D. S. Small, et al. (2018). Discovering treatment effect heterogeneity through post-treatment variables with application to the effect of class size on mathematics scores. *Journal of the Royal Statistical Society Series C 67*(4), 917–938.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Gabriel, E. E., M. C. Sachs, and A. Sjölander (2020). Causal bounds for outcome-dependent sampling in observational studies. *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2020.1832502.

Gamrath, G., D. Anderson, K. Bestuzheva, W.-K. Chen, L. Eifler, M. Gasse, P. Gemander, A. Gleixner, L. Gottwald, K. Halbig, et al. (2020). The scip optimization suite 7.0.

Gerber, A. S., D. P. Green, and R. Shachar (2003). Voting may be habit-forming: evidence from a randomized field experiment. *American journal of political science 47*(3), 540–550.

Green, D. P. and A. S. Gerber (2002). The downstream benefits of experimentation. *Political Analysis 10*(4), 394–402.

Grewal, S., A. A. Jamal, T. Masoud, and E. R. Nugent (2019). Poverty and divine rewards: The electoral advantage of islamist political parties. *American Journal of Political Science 63*(4), 859–874.

Hasegawa, R. B., D. W. Webster, and D. S. Small (2019). Evaluating missouri's handgun purchaser law: a bracketing method for addressing concerns about history interacting with group. *Epidemiology 30*(3), 371–379.

Heckman, J. and E. Vytlacil (2001). *Instrumental variables, selection models, and tight bounds on the average treatment effect*, pp. 1–15. Physica.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161.

Hobolt, S. B. (2016). The brexit vote: a divided nation, a divided continent. *Journal of European Public Policy 23*(9), 1259–1277.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*(4), 765–789.

Imai, K., D. Tingley, and T. Yamamoto (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 5–51.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Keele, L. J. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis 23*(3), 313–335.

Keele, L. J. and W. Minozzi (2012, Spring). How much is minnesota like wisconsin? assumptions and counterfactuals in causal inference with observational data. *Political Analysis 21*(2), 193–216.

Kennedy, E. H., S. Harris, and L. J. Keele (2019). Survivor-complier effects in the presence of selection on treatment, with application to a study of prompt icu admission. *Journal of the American Statistical Association 114*(525), 93–104.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–229.

Knox, D., W. Lowe, and J. Mummolo (2020). Administrative records mask racially biased policing. *American Political Science Review 114*, 619–637.

Kocher, M. A., T. B. Pepinsky, and S. N. Kalyvas (2011). Aerial bombing and counterinsurgency in the vietnam war. *American Journal of Political Science 55*(2), 201–218.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

Li, A. and J. Pearl (2021). Bounds on causal effects and application to high dimensional data. arXiv preprint arXiv:2106.12121.

Manski, C. (1990a). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Manski, C. F. (1990b). Nonparametric bounds on treatment effects. *The American Economic Review Papers and Proceedings 80*(2), 319–323.

Manski, C. F. (1995). *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

Mebane, W. R. and P. Poast (2013). Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis 22*(2), 169–182.

Molinari, F. (2020). Microeconometrics with partial identification. arXiv:2004.11751.

Mummolo, J. (2018). Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics 80*(1), 1–15.

Nyhan, B., C. Skovron, and R. Titiunik (2017). Differential registration bias in voter file data: A sensitivity analysis approach. *American Journal of Political Science 61*(3), 744–760.

Pearl, J. (1995a). Causal diagrams for empirical research. *Biometrika 82*(4), 669–710.

Pearl, J. (1995b). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435–443.

Pearl, J. (2009). *Causality.* New York: Cambridge University Press.

Robins, J. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Rosenbaum, P. R. (1984a). The consequences of adjusting for a concomitant variable that has been affected by the treatment. *Journal of The Royal Statistical Society Series A 147*(5), 656–666.

Rosenbaum, P. R. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society 147*(5), 656–666.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 6*(5), 688–701.

Schubiger, L. I. (2021). State violence and wartime civilian agency: Evidence from peru. *The Journal of Politics 83*(4), 1383–1398.

Sjölander, A., W. Lee, H. Källberg, and Y. Pawitan (2014). Bounds on causal interactions for binary outcomes. *Biometrics 70*(3), 500–505.

Stephens, A., L. J. Keele, and M. Joffe (2016). Generalized structural mean models for evaluating depression as a post-treatment effect modifier of a jobs training intervention. *Journal of Causal Inference 4*.

Swanson, S. A., M. A. Hernán, M. Miller, J. M. Robins, and T. S. Richardson (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association 113*(522), 933–947. DOI: 10.1080/01621459.2018.1434530.

Vigerske, S. and A. Gleixner (2018). Scip: Global optimization of mixed-integer nonlinear programs in a branch-and-cut framework. *Optimization Methods and Software 33*(3), 563–593.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

Ye, T., L. Keele, R. Hasegawa, and D. S. Small (2020). A negative correlation strategy for bracketing in difference-in-differences. *arXiv preprint arXiv:2006.02423*.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.