Difference-in-differences Designs for Controlled Direct Effects: An Application to Reducing Intergroup Prejudice*

Matthew Blackwell† Adam Glynn‡ Hanno Hilbig§
Connor Halloran Phillips¶

May 26, 2022

Abstract

Recent experimental studies in the social sciences have demonstrated that short, perspective-taking conversations are effective at reducing support for discriminatory public policies, but it remains unclear if these effects occur even if subjective feelings about the minority group are unchanged. Unfortunately, the identification and estimation of the controlled direct effect—the natural causal quantity of interest for this question—has required strong selection-on-observables assumptions for any mediator. Given that this assumption is too strong for many social science settings, in this paper we show how to identify and estimate controlled direct effects under a difference-in-differences design where we have measurements of the outcome and the mediator before and after treatment is assigned. This design allows us to weaken the identification assumptions to allow for linear and time-constant unmeasured confounding between the mediator and the outcome. Furthermore, we develop a semiparametric efficient and multiply robust estimator for these quantities. We find that there is a robust controlled direct effect of perspective-taking conversations when subjective feelings are neutral but not positive or negative. An open-source software package implements the approach with a variety of flexible, machine-learning algorithms for nuisance function estimation.

*Thanks to David Broockman, Kosuke Imai, Josh Kalla, and Soichiro Yamauchi for comments and suggestions. Working paper, comments welcome. Software to implement the methods in this paper can be found in the `DirectEffects` R package.

†Department of Government and Institute for Quantitative Social Science, Harvard University. web: http://www.mattblackwell.org email: mblackwell@gov.harvard.edu

‡Department of Political Science and Institute for Quantitative Theory and Methods, Emory University. email: aglynn@emory.edu

§Department of Government and Institute for Quantitative Social Science, Harvard University. email: hhilbig@g.harvard.edu

¶Department of Government, Harvard University. email: connorphillips@g.harvard.edu

1 Introduction

Can behavioral interventions change voter attitudes toward anti-minority policies even when prejudice against the minority group is unchanged? A recent experimental literature has found that short conversations involving perspective-taking can reduce prejudicial attitudes toward minority groups like transgender people and increase support for public policies such as nondiscrimination laws against these groups (Broockman and Kalla, 2016). Experiments such as these are an important tool for making principled causal inferences about complex political phenomena. While these experiments are well-suited to providing evidence for the existence of a causal effect, there has been considerable disagreement about their usefulness for assessing the direct effect of an intervention holding fixed, for example, prejudicial attitudes, which is important for these questions since subjective feelings about outgroups have been shown to be durable over a person's life (Sears and Funk, 1999; Tesler, 2015). Methods for estimating these types of direct effects have required strong and sometimes implausible assumptions beyond the randomization of the experimental design. In particular, common methods require a "no unmeasured confounders" assumption for the mediator, turning a carefully designed experiment into essentially an observational study. And while researchers have some options when the mediator is manipulable, learning about causal mechanisms in experimental studies generally is an uphill battle.

The goal of this paper is to show how to identify and estimate controlled direct effects under weaker assumptions with experimental designs that feature a multi-wave panel design. These designs measure covariates and outcomes in multiple waves, with at least one wave occurring before treatment is administered, allowing significant reductions in an experiment's implementation costs (Broockman, Kalla and Sekhon, 2017). We show that these designs have the additional benefit of allowing the identification of controlled direct effects of treatment fixing the value of a mediator under a parallel trends assumption as in a difference-in-differences design. Parallel trends, while still being a strong and untestable assumption, allows for time-constant unmeasured confounders between the mediator and the outcome, which is generally thought to be much weaker than the standard no unmeasured confounders assumption that mediation analyses generally require.

We introduce two quantities of interest, one conditional on just the baseline level of the mediator and one conditional on the observed treatment and mediator values, similar to the average treatment effect on the treated. We call the latter the path-conditional average controlled direct effect (ACDE-PC). The identification of these quantities depends on different, nonnested assumptions. Identification of the the baseline conditional ACDE requires changes in the mediator to be unrelated to trends in the outcome for both the treated and control groups, conditional on baseline and post-treatment covariates. We can identify the ACDE-PC, on the other hand, with parallel trends in the control condition alone, though such an assumption must omit any posttreatment confounders for the mediator. And unlike standard mediation analyses, neither assumption admits a decomposition of the the overall effect into a direct and indirect effect without further assumptions.

We build on these identification results to develop multiply robust, semiparametrically efficient estimators for these effects leveraging both propensity score and outcome regression modeling. The "multiply robust" property of this estimator means that it will be consistent and asymptotically normal even when only a subset of these models is correctly specified. Furthermore, we propose to fit this estimator utilizing the cross-fitting strategy of Chernozhukov et al. (2018) to allow for weaker conditions on the estimation of the nuisance parameters and a simple variance estimator. This cross-fitting procedure can also allow for flexible estimation of both the outcome regression and propensity score models through machine learning techniques (see, e.g., Bradic, Ji and Zhang, 2021), which we leverage for our empirical application.

The essential intuition behind our approach is that pretreatment measures of the outcome allow us to view the *changes* in the outcome of interest as the dependent variable rather than the levels themselves. Our main identification assumptions are then implied by standard sequential ignorability assumptions applied to changes in the outcome rather than levels. Thus, the identification will be robust to any time-constant, linear confounders for the relationship between the mediator and the outcome. While we focus on a context where the treatment is randomly assigned, it is simple to generalize our results to an observational situation with an additional parallel trends assumption for treatment.

This paper brings together two branches of the causal inference literature in the social and biomedical sciences. First, we build on the difference-in-differences (DID) framework that has been a workhorse of quantitative social sciences. In particular, our approach closely follows on similar work on DID estimators that leverage inverse probability weighting (IPW) estimators (Abadie, 2005) or combinations of IPW and regression (Sant'Anna and Zhao, 2020). Those papers generally targeted the average treatment effect (on the treated) for a single binary treatment, whereas we focus on estimation of the controlled direct effect. Second, there is a large literature on the estimation of these controlled direct effects mostly in the context of "no unmeasured confounders" assumptions (Robins and Greenland, 1992; Robins, 1994; Hernán, Brumback and Robins, 2001; Goetgeluk, Vansteelandt and Goetghebeur, 2008; Blackwell and Strezhnev, 2022). While there have been some efforts to identify these effects with instrumental variables (Robins and Hernán, 2009) or fixed-effects (Blackwell and Yamauchi, 2021), there have been very few attempts to identify these quantities leveraging changes in the outcomes over time in a difference-in-differences design as we do in this paper (see below for some exceptions). Finally, our proposed estimators build on a growing literature on "doubly robust" estimators (see, for example, Seaman and Vansteelandt, 2018, for a review of this literature) that has recently broadened to allow for adaptive "machine learning" algorithms in the estimation of various nuisance functions (Chernozhukov et al., 2018).

A handful of other studies have connected DID designs to direct effects more broadly. Deuchert, Huber and Schelker (2019) and Huber, Schelker and Strittmatter (2022) use a principal stratification approach to identification and estimation of different mediation quantities under monotonicity and parallel trends assumptions without intermediate covariates. Our work builds on their approach by not requiring monotonicity, incorporating nonbinary mediators, and allowing baseline and intermediate confounders at the expense of losing the decomposition-based interpretation of the quantities of interest. Concurrently with our work, Shahn et al. (2022) developed estimation techniques for estimating the parameters of a structural nested mean model under a DID-style assumption similar to our own. Their setting differs from our own in that they focus on estimating the effects of time-varying treatments where the outcomes are measured between each treatment. In ours, we only observe the

outcome after the treatment and mediator have been realized.

The paper proceeds as follows. In Section 2 we describe the experimental setting for the empirical application. Section 3 introduces the main quantities of interest and establishes the core identification results. We introduce the main estimation strategy in Section 4 along with how we leverage cross-fitting. In Section 5 we present simulation evidence for our approach, showing how it can avoid some of the pitfalls of the usual approaches to this problem and how flexible approaches can reduce bias even under misspecification. Section 6 presents the results of our empirical application and Section 7 concludes with a discussion.

2 Motivating application

Broockman and Kalla (2016) report the results of a door-to-door canvassing intervention designed to reduce discrimination toward transgender people (those who identify with a gender different from their sex assigned at birth). The intervention consisted of a 10-minute conversation that encouraged active "perspective taking," where the respondent is encouraged to think about a time when they themselves were judged negatively for being different and asked to reflect on if and how the conversation changed their mind. The respondents were recruited from a list of registered voters with a mailer for a baseline survey. Respondents to this survey were then randomly assigned to either the above intervention or a placebo conversation about recycling. The researchers followed up with online surveys to measure outcomes at various times after these conversations: 3 days, 3 weeks, 6 weeks, and 3 months.

The original analysis of the experiment found that the conversations about transphobia increased support for nondiscrimination laws and feelings of acceptance toward transgender people relative to the recycling conversation at 6 weeks and 3 months. The subjective feelings were measured via a "feeling thermometer" which asks respondents to rate their feelings of warmth toward a group on a 0–100 scale. Our present goal is to determine if the intervention has a *direct* effect on policy views for fixed feelings of warmth toward transgender people. This effect can help us understand if the total effects on nondiscrimination policies described above are entirely due to increased positive feelings

toward transgender people or if it is possible to affect policy views without changing "hearts and minds" about a particular minority group. This is crucial for understanding persuasion in diverse democracies since it shows whether or not personal tolerance of outgroups is required to increased support for *legal* tolerance of those same groups. But unmeasured confounding between subjective feelings and policy preferences make standard appraoches to estimating direct effects implausible, and so we leverage the experiment's multiwave design and rely on parallel trends-style assumptions to identify effects.

3 Setting and assumptions

Let D_{it} be a binary indicator of unit i receiving treatment in period t and let $M_{it} \in \mathcal{M}$ be a discrete mediator. In the simplest case, we have a binary mediator $\mathcal{M} = \{0, 1\}$, but we allow for arbitrary discrete mediators. We follow the canonical differences-in-differences framework and assume that $D_{i1} = 0$, and define $D_i = D_{i2}$. The mediator may take on any value in the first and second period. We also have a set of pretreatment, $\mathbf{X}_i$, and posttreatment, $\mathbf{Z}_i$, covariates, where $\mathbf{Z}_i$ are causally prior to M_{i2}. We assume the observed data $\mathbf{O}_i = (\mathbf{X}_i, \mathbf{Z}_i, D_i, M_{i1}, M_{i2}, Y_{i1}, Y_{i2})$ is independent and identically distributed across i.

Let $Y_{it}(d_t, m_t)$ be the potential outcome for a unit with treatment set to $D_{it} = d$ and mediator set to $M_{it} = m$. We assume the usual consistency assumption that $Y_{it} = Y_{it}(D_{it}, M_{it})$. There are potential versions of the intermediate covariates and posttreatment mediator as well, $\mathbf{Z}_i(d)$ and $M_{i2}(d)$, with similar consistency assumptions. Given the DID setup, we have $Y_{i1} = Y_{i1}(0, M_{i1})$. A key feature of differences-in-differences designs is the use of analyzing changes in the outcome over time to adjust for time-constant confounding. To that end, let $\Delta Y_i(d, m) = Y_{i2}(d, m) - Y_{i1}(0, M_{i1})$, where we connect this to the observed changes over time as $\Delta Y_i = \Delta Y_i(D_i, M_{i2})$.

Our goal is estimate the direct effect of treatment fixing the value of the (posttreatment) mediator to a particular value. We introduce a few different estimands to this end. The first is this controlled direct effect conditional on the mediator taking that value at baseline:

$$\tau_m = \mathbb{E}\{Y_{i2}(1, m) - Y_{i2}(0, m) \mid M_{i1} = m\},$$

for some $m \in \mathcal{M}$. We refer to this as the baseline-conditional average controlled direct effect or ACDE-BC. In the context of our application, this is the effect of the perspective-taking intervention for a fixed level of subjective feelings about transgender people for units that had that same level of subjective feelings at baseline. This effect is useful when assessing the effect for a particular value of the mediator, but it is also useful to have a summary measure of the direct effect at differing levels of the mediator. Let $\pi_{1m} = \mathbb{P}(M_{i1} = m)$ and the define the marginalized direct effect as

$$\tau = \sum_{m \in \mathcal{M}} \tau_m \pi_{1m} = \sum_{m \in \mathcal{M}} \mathbb{E}[Y_{i2}(1, m) - Y_{i2}(0, m) \mid M_{i1} = m] \pi_{1m},$$

which we call the marginalized ACDE-BC. This estimand treats each level of the baseline mediator as a separate DID study and aggregates them based on their size. In this way, it is similar to a conditional version of the average factorial effect in a factorial experiment or the average marginalized component effect in conjoint studies.

We also investigate the controlled direct effect on those who were treated *and* hold their value of the mediator constant over time,

$$\gamma_m = \mathbb{E}\{Y_{i2}(1, m) - Y_{i2}(0, m) \mid D_i = 1, M_{i1} = m, M_{i2} = m\},$$

which is similar to the average treatment effect on the treated in point exposures settings. We call this the path-conditional average controlled direct effect or ACDE-PC and we can marginalize it similarly to τ_m and τ. In the context of our application, this has the same interpretation as the ACDE-BC except that the effect is only among those units who were treated and who held the same subjective feelings about transgender people before and after treatment. Below, we show that this quantity can be identified under an alternative set of assumptions that may be more plausible in some empirical settings.

3.1 Assumptions

We build our identification from two key features of the experimental design under question: randomization and panel data. Randomization allows us to identify the effects of treatment broadly,

while the panel nature of the data allows us to leverage the key identifying assumption of a difference-in-differences design that there are parallel trends in certain potential outcomes over time. Let $\mathbf{Y}(\bullet) = \{Y_{it}(d,m) : t = 1,2 \ d = 0,1 \ m \in \mathcal{M}\}$ be the set of all potential outcomes, with similar notation defined for $M_{i2}(d)$ and $\mathbf{Z}_i(d)$.

Assumption 1 (Treatment Randomization)**.** $\{\mathbf{Y}(\bullet), M_{i2}(\bullet), \mathbf{Z}_i(\bullet), M_{i1}\} \perp\!\!\!\perp D_i$.

Assumption 2 (Mediator Parallel Trends)**.** *For* $d \in \{0,1\}$*, and* $m, m', m'' \in \mathcal{M}$

$$\mathbb{E}\{\Delta Y_i(d,m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, \mathbf{Z}_i, M_{i2} = m'\} = \mathbb{E}\{\Delta Y_i(d,m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, \mathbf{Z}_i, M_{i2} = m''\}.$$

Assumption 1 comes from the design of the experiment, though it is possible to generalize this assumption to a selection-on-observables or parallel trends assumption for an observational study. Assumption 2 states that the over-time trends in the potential outcomes are mean-independent of the mediator value in period 2, conditional on some covariates that might be pretreatment ($\mathbf{X}_i$) or posttreatment ($\mathbf{Z}_i$). For example, suppose a unit switches their subjective feelings about transgender people from neutral to positive (say, $M_{i1} = m$ to $M_{i2} = m'$) before and after treatment. Parallel trends states that their potential outcomes if they had not switched, $\Delta Y_i(d,m)$, has the same expectation as those units who in fact remain neutral, conditional on $\mathbf{X}_i$ and $\mathbf{Z}_i$. Note that this places no restrictions on the baseline mediator, so we allow for unmeasured confounding between the outcome and the baseline mediator. Thus, we allow for pretreatment subjective feelings to be arbitrarily related to baseline attitudes about laws relating to transgender people.

Assumption 2 is implied by the following sequential ignorability assumption with changes in the outcome as the dependent variable,

$$\Delta Y_i(d,m) \perp\!\!\!\perp M_{i2} \mid M_{i1} = m, D_i = d, \mathbf{X}_i, \mathbf{Z}_i. \tag{1}$$

The parallel trends assumptions are weaker since (a) they only place restrictions on the averages of the potential outcomes rather than their entire distributions; and (b) they only place restrictions on the potential outcomes for the same mediator status as the baseline mediator, m. This sequential ignorability version of the assumption does retain the core benefit of a differences-in-differences

design: both D_i and M_i can still be correlated with time-constant factors that affect both Y_{i1} and Y_{i2} in the same way. That is, they still allow for time-constant unmeasured confounding, albeit in a restricted, linear way.

As an example of how this unmeasured confounding might manifest, suppose that there is a time-constant unmeasured confounder, U_i, that is correlated with M_{i2}. Further, suppose we have the following models for our potential outcomes:

$$Y_{i1}(d,m) = f_{1dm}(\mathbf{X}_i) + g(U_i, \mathbf{X}_i) + \varepsilon_{i1}, \qquad Y_{i2}(d,m) \quad = f_{2dm}(\mathbf{X}_i, \mathbf{Z}_i(d)) + g(U_i, \mathbf{X}_i) + \varepsilon_{i2},$$

where we assume that ε_{it} are i.i.d. and independent of all variables $\mathbf{O}_i$. Under this model, the usual sequential ignorability assumption for $Y_{i2}(d,m)$ and M_{i2} conditional on just $\mathbf{X}_i$ and $\mathbf{Z}_i$ would not hold because of the unmeasured confounder, U_i. But because that confounder enters into the model for potential outcomes in a linear, additive, and time-constant manner, it will be unrelated to the *changes* in the potential outcomes over time.

One downside to the parallel trends in Assumption 2 is that the condition must hold for $d = 1$, where $\Delta Y_i(d,m)$ combines two different sources of trends. There is the secular trend in the outcome over time plus the effect of switching from untreated in $t = 1$ to treated in $t = 2$. Thus, for $d = 1$, this assumption implies that the over-time effect of treatment on the mediator is mean-independent of the over-time effect of treatment on the outcome. In the above example, this holds because the unmeasured confounding between M_{i2} and Y_{i2}, captured by $g(U_i, \mathbf{X}_i)$, is constant across treatment conditions. This speaks to the *time-constant* nature of the confounding we can adjust for in this setting: it ceases to be time-constant if the confounding is affected by treatment which obviously can change over time. This mean-independence of the treatment effect is still significantly weaker than other no-interactions assumptions used to identify mediation effects that require no interaction between D_i and M_{i2} at the individual level (Robins, 2003).

In settings where the posttreatment mediator might be correlated with the controlled direct effect, we propose an alternative identifying assumption based on parallel trends among controls only. This assumption will help identify the ACDE-PC, γ_m.

Assumption 3 (Control Parallel Trends)**.** *For all* $m, m', m'' \in \mathcal{M}$ *and* $d \in \{0, 1\}$

$$\mathbb{E}\{\Delta Y_i(0, m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, M_{i2} = m'\} = \mathbb{E}\{\Delta Y_i(0, m) \mid D_i = d, \mathbf{X}_i, M_{i1} = m, M_{i2} = m''\}.$$

This assumption states that parallel trends holds for the mediator in the control group conditioning just on the pretreatment covariates. This says that, conditional on pretreatment covariates, the value of the second period mediator is unrelated to the trends in the control group—or in the context of our application, that changes in subjective feelings are unrelated to changes in support for laws. Combined with randomization of D_i, this implies that every group defined by their values of D_i and M_{i2} would have followed the same average trend if, possibly contrary to fact, they had remained at $M_{i2} = m$ and stayed in the control condition. Given the lack of intermediate covariates, this is similar to a standard difference-in-differences design with a multileveled treatment (combining D_i and M_{i2}).

The exclusion of posttreatment covariates in this identifying assumption is a major limitation, so it is important to consider why they cannot be included. To identify the ACDE, we will need to impute the trends for $(D_i = 0, M_{i2} = m)$ group among those with, say, $(D_i = 1, M_{i2} = m)$. We typically accomplish this by adjusting for covariates through weighting or regression and those methods would require assumptions on both the treated and control potential outcome trends as in Assumption 2. If we include posttreatment confounders, however, our adjustment would require information about the joint distribution of the potential outcomes of the posttreatment covariates, $\mathbf{Z}_i(1)$ and the potential outcomes $\Delta Y_i(0, m)$. Unfortunately, this joint distribution is never identified due to the fundamental problem of causal inference. We could assume $\mathbf{Z}_i$ is unaffected by D_i, but then it ceases to be posttreatment. We could alternatively assume parallel trends holds for $\mathbf{Z}_i(1, m)$ with respect to $\Delta Y_i(0, m)$, conditional on $\mathbf{X}_i$, $M_{i1} = m$ and $D_i = 1$, but this seems to call into question why it would be needed to block confounding for M_{i2}. Thus, it appears that in this setting we either can either restrict our parallel trends assumption to the control treatment of D_i or allow for posttreatment confounders, but not both simultaneously.

3.2 Identification

Let $\pi_{2m}(m_1, d, \mathbf{x}, \mathbf{z}) = \mathbb{P}(M_{i2} = m \mid M_{i1} = m_1, D_i = d, X_i = \mathbf{x}, Z_i = \mathbf{z})$ be the generalized propensity score for M_{i2}, and let $\pi_d = \mathbb{P}(D_i = 1)$ be the probability of treatment. We define W_{i1m} to be an indicator for the baseline mediator being equal to m, so that $W_{i1m} = 1$ when $M_{i1} = m$ and 0 otherwise, with W_{i2m} being similarly defined for M_{i2}. We use the convention that if $\mathbf{Z}_i$ is omitted from $\pi_2(\cdot)$, it represents the propensity score just as a function of $\mathbf{X}_i$. We also define the marginal probability of the mediator in the second period as $\overline{\pi}_{m2}(d, m_1) = \mathbb{P}(M_{i2} = m \mid M_{i1} = m_1, D_i = d)$. We now present the main identification result.

Theorem 1. *Under Assumptions 1 and 2, we have*

$$\tau_m = \mathbb{E}\left[\frac{W_{i1m}W_{i2m}}{\pi_{1m}\pi_{2m}(m, D_i, \mathbf{X}_i, \mathbf{Z}_i)}\left(\frac{D_i}{\pi_d} - \frac{(1 - D_i)}{(1 - \pi_d)}\right)\Delta Y_{i2}\right],$$

and under Assumptions 1 and 3,

$$\gamma_m = \mathbb{E}\left[\frac{W_{i1m}W_{i2m}}{\pi_{1m}\overline{\pi}_{2m}(m, 1)}\left(\frac{D_i}{\pi_d} - \frac{(1 - D_i)\pi_{2m}(m, 1, \mathbf{X}_i)}{(1 - \pi_d)\pi_{2m}(m, 0, \mathbf{X}_i)}\right)\Delta Y_i\right].$$

Theorem 1 shows that the conditional controlled direct effects can be identified from weighted averages of the changes in the outcome over time, where the weights depend on the propensity score for the mediator in posttreatment period. The proof strategy and result generalize the approach of Abadie (2005) to the setting of controlled direct effects. Furthermore, identification of the effects conditional on the baseline mediator imply that the marginalized effects, τ and γ, are also identified.

While we have focused so far on the direct effects of treatment, mediation analyses often target indirect effects as well. The presence of posttreatment confounders, $\mathbf{Z}_i$, usually precludes the possibility of identifying mediation quantities like the natural indirect effect (Robins, 2003; Avin, Shpitser and Pearl, 2005) and this is true for our controlled direct effect parameter τ_m. While the controlled direct effect on the treated, γ_m, requires no posttreatment confounders for identification like the mediation quantities, it also focuses on units who do not change their mediator status before and after treatment ($M_{i1} = M_{i2} = m$). It would be possible to identify and estimate mediation quantities if we were to combine different aspects of our assumptions and assume that Assumption 2 holds without

conditioning on posttreatment confounders and we were willing to make parallel trends assumptions with regard to sequences such as $Y_{i2}(0,m) - Y_{i1}(0,m')$. Under these assumptions, the standard techniques of mediation analysis can be used to estimate quantities like the average natural direct effect (Imai, Keele and Yamamoto, 2010). Here we focus on our less restrictive assumption and leave mediation analyses to future research.

We can also compare the identification result to how the same data might be identified under a standard sequential ignorability assumption for levels rather than changes. This assumption maintains that $Y_{i2}(d,m) \perp\!\!\!\perp M_{i2} \mid D_i = d, \mathbf{X}_i, \mathbf{Z}_i$ and we can identify τ_m as $\mathbb{E}[\omega_{im} Y_{i2}]$, where

$$\omega_{im} = \frac{W_{i1m} W_{i2m}}{\pi_{1m}\pi_{2m}(m, D_i, \mathbf{X}_i, \mathbf{Z}_i)} \left(\frac{D_i}{\pi_d} - \frac{(1-D_i)}{(1-\pi_d)} \right),$$

meaning that the difference between our DID identification result and the sequential ignorability identification result is $\mathbb{E}[\omega_{im} Y_{i1}]$. This would be the estimand of attempting to estimate the "controlled direct effect" of treatment on a pretreatment measurement of the outcome, which under sequential ignorability should be zero. Thus, one way to view the DID approach we present is as a bias correction to standard sequential ignorability approaches using the lagged outcomes as a negative control (Lipsitch, Tchetgen Tchetgen and Cohen, 2010; Sofer et al., 2016).

This analysis assumes we use the same conditioning set under a parallel trends approach and a sequential ignorability approach, but what if we condition on the lagged dependent variable (LDV) in the latter? Several authors have shown there is a bracketing relationship between the DID approach and this LDV approach in the case of a single treatment variable Angrist and Pischke (2009); Ding and Li (2019). In Supplemental Materials A, we derive the difference between the DID target of inference and LDV target of inference for the ACDE-BC and for the ACDE-PC. In the latter case, we show that when either parallel trends or sequential ignorability with an LDV holds, the two approaches should bound the true ACDE-PC in the limit, as in Ding and Li (2019).

4 Estimation

We now turn to estimation of the controlled direct effects. It is straightforward to build a standard inverse probability weighting (IPW) estimator based on the identification result as in Abadie (2005),

but IPW estimators can be unstable when the propensity scores models are incorrectly specified. To create efficient and stable estimators, we focus on developing a set of multiply robust estimators based on semiparametric efficiency theory. These estimators are multiply robust in the sense that we will specify a number of models—for the propensity scores and for various outcome regressions—and the resulting estimator will be consistent and asymptotically normal when some, but not necessarily all, of the models are correctly specified. Furthermore, even if all models are misspecified, multiply robust estimators can improve performance over estimators that rely on any of the individual misspecified models in isolation. Finally, we integrate cross-fitting into our estimation approach so that data-adaptive machine learning models can be leveraged to make estimates less sensitive to particular functional form assumptions.

We define the following outcome regression functions:

$$\mu_{dm_2}(m_1, \mathbf{x}, \mathbf{z}) = \mathbb{E}[\Delta Y_i(d, m_2) \mid M_{i2} = m_2, M_{i1} = m_1, D_i = d, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}]$$

$$\nu_{dm_2}(m_1, \mathbf{x}) = \mathbb{E}[\mu_{dm_2}(m_1, \mathbf{x}, \mathbf{Z}_i) \mid M_{i1} = m_1, D_i = d, \mathbf{X}_i = \mathbf{x}]$$

Since our focus is on estimands where the mediator is the same before and after treatment, we define $\mu_{i,dm} = \mu_{dm}(m, \mathbf{X}_i, \mathbf{Z}_i)$ and $\nu_{i,dm} = \nu_{dm}(m, \mathbf{X}_i)$. The first of these functions is the regression of the potential outcomes on all the mediators, treatments, and covariates, which we sometimes call the "long" regression. By consistency, this is equivalent to the regression of the observed changes in the outcome over time on those same variables. The second function is the average of the first regression over the distribution of the intermediate covariates, $\mathbf{Z}_i$, as a function of treatment and the pretreatment covariates, which we sometimes call the "short" regression.

Building a multiply robust estimator requires the specification of working models for three nuisance functions. These are functions that we need to estimate but are not of direct interest. First, we must specify a model for the propensity scores, which we refer to as $\widehat{\pi}_{2m}(m_1, d, \mathbf{x}, \mathbf{z})$. In many cases this might be a logistic regression for a binary mediator or a multinomial logistic regression for more general discrete mediators, though our setup allows for more flexible machine learning models. We assume another working model for the "long" regression, $\widehat{\mu}_{dm_2}(m_1, \mathbf{x}, \mathbf{z})$. These are working models because we do not necessarily assume that either one of them is correctly specified. While the propen-

sity score and long regression models are fairly straightforward, the short regression is a bit more complicated. One simple way to estimate ν_{dm_2} would be to regress fitted values $\widehat{\mu}_{D_i m_2}(M_{i1}, \mathbf{X}_i, \mathbf{Z}_i)$ on D_i, M_{i1} and X_i or some functions of these variables, but this depends on the correct specification of the long regression working model. We can instead build a doubly robust estimator of the short regression based on the two working models above,

$$\widehat{\nu}_{dm_2}(m_1, \mathbf{x}) = \widehat{\mathbb{E}}\left[\widehat{\mu}_{dm_2}(m_1, \mathbf{x}, \mathbf{Z}_i) + \frac{W_{i2m}(\Delta Y_i - \widehat{\mu}_{dm_2}(m_1, \mathbf{x}, \mathbf{Z}_i))}{\widehat{\pi}_{2m_2}(m_1, d, \mathbf{x}, \mathbf{Z}_i)} \mid M_{i1} = m_1, D_i = d, \mathbf{X}_i = \mathbf{x}\right],$$

where $\widehat{\mathbb{E}}[\cdot \mid A = a]$ is the working regression of a variable as a function $A = a$. This estimator will be corrected specified when the functional form of the dependence on D_i, M_{i1} and $\mathbf{X}_i$ is correct and either $\widehat{\pi}_2$ or $\widehat{\mu}_{dm_2}$ are correctly specified.

With these intermediate estimators in hand, we now construct multiply robust estimators for our controlled direct effects. We will build these estimators from estimating equations with influence function as $\psi_m(\mathbf{O}_i; \boldsymbol{\eta})$ for τ_m and $\phi_m(\mathbf{O}_i; \boldsymbol{\eta})$ for γ_m, where $\boldsymbol{\eta} = (\pi_{2m}, \mu_{dm}, \nu_m, \pi_{1m}, \pi_d)$ is the vector of nuisance parameters for τ_m, $\boldsymbol{\eta} = (\pi_{2m}, \mu_{dm}, \overline{\pi}_{md})$ is the vector of nuisance parameters for γ_m, and $\overline{\pi}_{md} = \mathbb{P}(M_{i1} = m, D_i = d, M_{i2} = m)$. These estimating equations have the form

$$\begin{aligned}\psi_m(\mathbf{O}_i; \boldsymbol{\eta}) = &\left(\frac{W_{i1m} D_i W_{i2m}}{\pi_{1m}\pi_d\pi_{i,2m}}\right)(\Delta Y_i - \mu_{i,1m}) - \left(\frac{W_{i1m}(1 - D_i)W_{i2m}}{\pi_{1m}(1 - \pi_d)\pi_{i,2m}}\right)(\Delta Y_i - \mu_{i,0m}) \\ &+ \frac{W_{i1m} D_i}{\pi_{1m}\pi_d}(\mu_{i,1m} - \nu_{i,1m}) - \frac{W_{i1m}(1 - D_i)}{\pi_{1m}(1 - \pi_d)}(\mu_{i,0m} - \nu_{i,0m}) + \frac{W_{i1m}}{\pi_{1m}}(\nu_{i,1m} - \nu_{i,0m}),\end{aligned}$$

for the ACDE-BC, and

$$\begin{aligned}\phi_m(\mathbf{O}_i; \boldsymbol{\eta}) = &\left(\frac{W_{i1m} D_i W_{i2m}}{\overline{\pi}_{m1}}\right)(\Delta Y_i - \mu_{i,1m}) - \left(\frac{W_{i1m}(1 - D_i)W_{i2m}}{\overline{\pi}_{m0}}\right)\left(\frac{\pi_{2m}(1, m, \mathbf{X}_i)}{\pi_{2m}(0, m, \mathbf{X}_i)}\right)(\Delta Y_i - \mu_{i,0m}) \\ &+ \frac{W_{i1m} D_i W_{i2m}}{\overline{\pi}_{m1}}(\mu_{i,1m} - \mu_{i,0m})\end{aligned}$$

for the ACDE-PC. This estimator combines imputation (via the outcome regressions) with weighting (via the propensity scores) to estimate τ. These estimating equations are generalizations of similar doubly robust approaches to estimating ACDEs under sequential ignorability (Murphy et al., 2001; Orellana, Rotnitzky and Robins, 2010; van der Laan and Gruber, 2012) and are similar to doubly robust approaches to estimation of the effect of point exposures under difference-in-differences designs (Sant'Anna and Zhao, 2020).

Given a set of estimators for the propensity scores and the regression functions, $\widehat{\boldsymbol{\eta}}$, we can use these estimating equations to derive the following estimators:

$$\widehat{\tau}_m = \frac{1}{N}\sum_{i=1}^{N}\psi_m(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}) \qquad \widehat{\gamma}_m = \frac{1}{N}\sum_{i=1}^{N}\phi_m(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}).$$

We first establish a multiply robust consistency result for this estimator.

Theorem 2. *(a) Under Assumptions 1 and 2 and suitable regularity conditions, $\widehat{\tau}_m$, is consistent for τ_m when either $\widehat{\pi}_{2m} \xrightarrow{p} \pi_{2m}$ or both $\widehat{\mu}_{dm_2} \xrightarrow{p} \mu_{dm_2}$ and $\widehat{\nu}_{dm_2} \xrightarrow{p} \nu_{dm_2}$. (b) Under Assumptions 1 and 3, $\widehat{\gamma}_m$ is consistent for γ_m when either $\widehat{\pi}_{2m} \xrightarrow{p} \pi_{2m}$ or $\widehat{\mu}_{dm} \xrightarrow{p} \mu_{dm}$.*

Theorem 2 ensures that our estimators will be consistent for their intended estimands when either the propensity score model for the posttreatment mediator or the outcome regressions are correctly specified. While we focus on the case where treatment is randomized, this result could easily be expanded to handle a treatment that satisfies selection on observables or a parallel trends assumption. In that case, we would require an additional propensity score model for D_i and the number of correctly specified model combinations that would ensure consistency would expand.

4.1 Semiparametric efficiency

In addition to being multiply robust, our estimator $\widehat{\tau}_m$ has the benefit of being locally semiparametrically efficient when the working models are all correctly specified. This means $\widehat{\tau}_m$ will have the lowest variance among all semiparametric estimators—that is, estimators that make no assumptions about the working models.

Theorem 3. *(a) Under Assumptions 1 and 2 and suitable regularity conditions, the efficient influence function for τ_m is $\widetilde{\psi}_m(\mathbf{O}_i;\boldsymbol{\eta}) = \psi_m(\mathbf{O}_i;\boldsymbol{\eta}) - W_{i1m}\pi_{1m}^{-1}\tau_m$,, and the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i;\boldsymbol{\eta})^2]$. (b) Under Assumptions 1 and 3 and suitable regularity conditions, the efficient influence function for γ_m is $\widetilde{\phi}_m(\mathbf{O}_i;\boldsymbol{\eta}) = \phi_m(\mathbf{O}_i;\boldsymbol{\eta}) - W_{i1m}D_iW_{i2m}\overline{\pi}_{m1}^{-1}\gamma_m$, and the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\phi}_m(\mathbf{O}_i;\boldsymbol{\eta})^2]$.*

The regularity conditions here involve technical requirements to ensure pathwise differentiability of the efficient influence function. See, for example, Bickel et al. (1998, Chapter 3 3) for more details on these conditions.

4.2 Variance estimation and crossfitting

As we have shown, to obtain estimators from the doubly robust estimator $\widehat{\tau}_m$ we first need to fit a series of working models, $\widehat{\boldsymbol{\eta}}$. When we use the same observations to fit these models as we use in the sample mean $N^{-1}\sum_{i=1}^{N}\psi_m(\mathbf{O}_i;\widehat{\boldsymbol{\eta}})$, our estimates can become less stable and we must account for using the data twice in our uncertainty estimates.

We can avoid both of these issues by relying on a estimation framework called *cross-fitting*, a generalization of the long-used method of sample splitting (Chernozhukov et al., 2018). Sample splitting is a simple way to make the estimates of the nuisance parameters (the working models for the propensity scores and outcome regressions here) independent of the final estimates of the quantities of interest. We first randomly split the sample into two groups, the main and auxiliary samples. We use the auxiliary sample to fit the working models, then use those fitting models to obtain predicted values for the main sample to plug into the main estimation of the $\widehat{\tau}_m$. The downside of sample splitting is that we only use half of our sample for the estimation of the main quantities of interest, which motivated the development of cross-fitting. With cross-fitting, we simply swap the roles of the main and auxiliary samples and obtain a second estimate of the quantity of interest. We then take the average of the two "sample split" estimators as our final estimate. Cross-fitting retains the benefits of sample-splitting in terms of making inference more stable and straightforward without the drawback of reduced efficiency. We can further generalize this approach to more than a single split by creating K roughly equally-sized folds, where $K \geq 2$. Finally, to account for the variability of this splitting process, we repeat this process several times and take the median of the estimates across the different splits as recommmended by Chernozhukov et al. (2018).

To be specific, we randomly partition the data into K groups by drawing $(B_1, \ldots, B_n)$ independently of the data, where B_i is distributed uniformly over $\{1, \ldots, K\}$. We take $B_i = b$ to mean that

unit i is split into group b. Let $\psi_m(\mathbf{O}_i, \widehat{\boldsymbol{\eta}}_{-b})$ be the value of the influence function when the nuisance parameters are estimated without the group $B_i = b$. We also let $\mathbb{P}_n^b$ denote the conditional empirical distribution for the group $B_i = b$,

$$\mathbb{P}_n^b\{f(\mathbf{O}_i)\} = \frac{\sum_{i=1} f(\mathbf{O}_i)\mathbb{I}(B_i = b)}{\sum_{i=1} \mathbb{I}(B_i = b)}$$

Then, we can define the crossfitting estimator as

$$\widehat{\tau}_m = \sum_{b=1}^{K} \left\{ \frac{1}{n} \mathbb{I}(B_i = b) \right\} \mathbb{P}_n^b\{\psi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_{-b})\} = \mathbb{P}_n \left\{ \psi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_{-B_i}) \right\},$$

with $\widehat{\gamma}_m$ defined similarly. Let $\|f\| = \left(\mathbb{E}[(f(\mathbf{O}_i))^2]\right)^{1/2}$ for any function f.

Theorem 4. *(a) Let Assumptions 1 and 2 hold and suppose that (i) $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\| = o_p(1)$, (ii) $\|\widehat{\mu}_{dm} - \mu_{dm}\| \times \|\widehat{\pi}_{2m} - \pi_{2m}\| = o_p(n^{-1/2})$. Then, $\sqrt{N}(\widehat{\tau}_m - \tau_m)$ will converges in distribution to $N(0, \mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta})^2])$. (b) Under the same assumptions with Assumption 3 replacing Assumption 2, $\sqrt{N}(\widehat{\gamma}_m - \gamma_m)$ will converge in distribution to $N(0, \mathbb{E}[\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta})^2])$.*

We can easily obtain consistent variance estimators for this procedure with

$$\widehat{\mathbb{V}}[\widetilde{\tau}_m] = \frac{1}{N^2} \sum_{i=1}^{N} \left\{ \psi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_{-B_i}) - \widehat{\tau} \right\}^2 \qquad \widehat{\mathbb{V}}[\widetilde{\gamma}_m] = \frac{1}{N^2} \sum_{i=1}^{N} \left\{ \phi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_{-B_i}) - \widehat{\gamma}_m \right\}^2$$

An additional benefit of this crossfitting procedure is that it allows for "plug-and-play" integration with machine learning algorithms so that we can replace, say, a standard logistic regression model for a propensity score with an data-adaptive algorithm such as the logistic lasso. Bradic, Ji and Zhang (2021) has derived the rate conditions on these types of algorithms needed to ensure consistent and asymptotic normality of dynamic treatment effects like the ones we study here. In our own empirical application, we leverage a version of the lasso for outcome regressions and a random forest approach for estimation of the generalized propensity score for a three-level discrete mediator.

4.3 Extension to Time-varying Treatments

Given our empirical setting, we have focused on a framework where the two causal variables of interest—the treatment and the mediator—are distinct. But our approach can also be used in situations where the treatment and the mediator are two measurements of the same variable over time.

This type of time-varying treatment is very common in the social and biological sciences. There is, in fact, a large literature that studies difference-in-differences designs with time-varying treatments (Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021), though the vast majority of these studies tend to assume that (a) there is only one switch from a control condition to a treatment condition (often referred to as a *staggered adoption* design) and/or (b) only contemporaneous, not lagged, treatment affects the outcome (a so-called *no carryover* assumption). Our approach would allow for the estimation of direct effects of lagged treatment in cases where a unit can switch back from treatment to control.

We consider a three-period setting where we map our original causal variables, (M_{i1}, D_i, M_{i2}), to three measurements of the treatment variable, (D_{i0}, D_{i1}, D_{i2}). To match the standard DID setting, we assume that $D_{i0} = 0$ for all i, but that (D_{i1}, D_{i2}) can take any value in $\{0, 1\}^2$. Let $Y_{i0} = Y_{i0}(0)$ be the baseline measure of the outcome and let $Y_{i2}(d_1, d_2)$ be the potential outcome after both treatments are administered. Then our main identifying assumption becomes

$$\mathbb{E}[Y_{i2}(d_1, d_2) - Y_{i0}(0) \mid D_i = (0, d_1, 1), \mathbf{X}_i, \mathbf{Z}_i] =$$
$$\mathbb{E}[Y_{i2}(d_1, d_2) - Y_{i0}(0) \mid D_i = (0, d_1, 0), \mathbf{X}_i, \mathbf{Z}_i],$$

which would allow us to identify the ACDE of D_{i1}, $\mathbb{E}[Y_{i2}(1, 0) - Y_{i2}(0, 0)]$ under randomization of D_{i1}. This identification is exactly the same as the identification of τ_m above. If we make a parallel trends assumption for D_{i1} instead of a randomization assumption, then we would be able to identify the ACDE on the treated, $\mathbb{E}[Y_{i2}(1, 0) - Y_{i2}(0, 0) \mid D_{i1} = 1]$. These results imply that our estimators can be used to estimate the effects of treatment histories with time-varying confounding and without a strict exogeneity assumption. To accomplish this, one simply uses the same multiply robust estimators as above using (D_{i0}, D_{i1}, D_{i2}) in place of (M_{i1}, D_i, M_{i2}). It should be possible to extend our approach to an arbitrary number of time periods, though this is beyond the scope of the current paper.

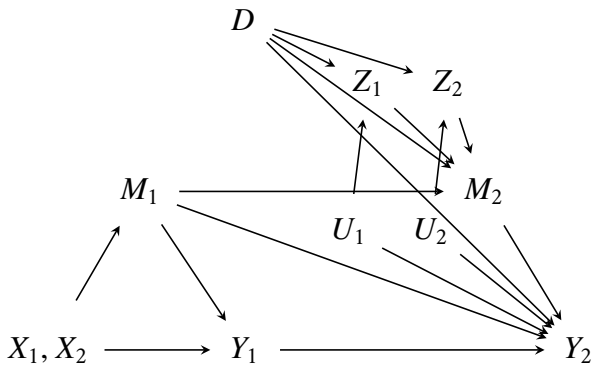

Figure 1: Directed acyclic graph showing the simulation setup.

5 Simulation Results

We now evaluate the finite-sample performance of our estimator with a simulation experiment. We are interested in how the multiply robust estimation techniques compare to traditional difference-in-differences approaches, but also how different machine learning techniques in the multiply robust approach are able to handle misspecification. We evaluate the performance of our estimator against two alternative approaches for computing direct effects—traditional regression DID controlling for baseline covariates $\mathbf{X}_i$ and the mediator, and the same specification also controlling for intermediate covariates $\mathbf{Z}_i$. As the results show, our method performs well against these alternatives even when the working models are misspecified, particularly at larger sample sizes.

The DGP follows the DAG in Figure 1. Treatment has independent probability $\pi_d = 0.5$, and we generate two observed baseline variables, $\mathbf{X}_i = (X_{i1}, X_{i2})' \sim \mathcal{N}_2(0, \sigma_x^2 \mathbf{I}_2)$, where $\sigma_x^2 = 0.01$, and two unobserved independent baseline variables $U_{i1}, U_{i2} \sim \mathcal{N}(0, 0.01)$. We draw the baseline mediator as $M_{i1} = \mathbb{I}(X_{i1} + X_{i2} + \varepsilon_{im1} \geq 0)$. The baseline outcome follows $Y_{i1} = 1 + 0.4M_{i1} + \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_{iy1}$, where $\boldsymbol{\beta} = (0.5, 0.5)'$, and then we generate the intermediate confounders with heterogeneous treatment effects, $Z_{ij} = \delta_{ij} D_i + 5U_{ij} + \varepsilon_{izj}$, where $\delta_{ij} \sim \mathcal{N}(0.25, 0.0025)$ for $j \in \{1, 2\}$. The posttreatment mediator follows $M_{i2} = \mathbb{I}(-1 + 1.5D_i + 0.4M_{i1} + \mathbf{Z}_i'\boldsymbol{\gamma} + \varepsilon_{im2} \geq 0)$, where $\boldsymbol{\gamma} = (0.75, 0.75)'$. The

second period outcome is

$$Y_{i2} = Y_{i1} + 0.4M_{i1} + 0.2D_i + 0.3M_{i2} + 0.1D_iM_{i2} + 5U_{i1} + 5U_{i2} + \varepsilon_{iy2},$$

where $(\varepsilon_{im1}, \varepsilon_{iy1}, \varepsilon_{iz1}, \varepsilon_{iz2}, \varepsilon_{im2}, \varepsilon_{iy2}) \sim \mathcal{N}_6(0, \Sigma_\varepsilon)$ and Σ_ε is a diagonal matrix with $\text{diag}(\Sigma_\varepsilon) = (0.01, 0.01, 0.04, 0.04, 1, 0.01)'$. In order to test how these methods perform when the relevant models are misspecified, we also construct transformations of the covariates $X_{i1}, X_{i2}, Z_{i1}, Z_{i2}$ as follows, employing a similar setup to Kang and Schafer (2007):

$$X^*_{i1} = (\exp(X_{i1}/2) - 1)^2 \quad X^*_{i2} = X_{i2}/(1 + \exp(X_{i1})) + 10$$
$$Z^*_{i1} = (X_{i1}Z_{i1}/25 + 0.6)^3 \quad Z^*_{i2} = (X_{i2} + Z_{i2} + 20)^2.$$

For each simulated dataset, we construct five estimates for the marginalized ACDE-BC, τ. First, we simply regress $\Delta Y_i = Y_{i2} - Y_{i1}$ on D_i, controlling for M_{i1}, M_{i2}, X_{i1} and X_{i2} ("DID + Mediator"). Second, we add the intermediate covariates to this specficiation ("DID + Mediator + Covariates"). Finally, we use our multiply robust ACDE estimator with the same outcome regression the DID + Mediator + Covariate estimator with three different propensity scores estimators for M_{i2}: logistic regression ("MR ACDE (Logit)"), the lasso ("MR ACDE (Lasso)"), and random forests ("MR ACDE (RF)"). For the lasso and random forest approaches, we include all squared terms and two-way interactions of the covariates.

We ran 1000 replications of this DGP and computed the average bias, the root mean square error (RMSE), and the coverage of nominal 95% confidence intervals for sample sizes of 250, 500, and 1000 and using either the "correctly specified" covariates $(X_{i1}, X_{i2}, Z_{i1}, Z_{i2})$ or the "incorrectly specified" transformed versions $(X^*_{i1}, X^*_{i2}, Z^*_{i1}, Z^*_{i2})$. We calculated the true values of τ_m and τ as part of the Monte Carlo simulation.

Figure 2 presents the results of this simulation. Under both correctly and incorrectly specified models, we can see that the DID estimators exhibit large biases at all sample sizes and have correspondingly high RMSEs and low coverage. This performance is being driven, as expected, by confounding bias when excluding the intermediate covariates and posttreatment bias when the covariates are included. Under this DGP, these biases can be made larger or smaller by manipluating the strength of the relationships on those paths.

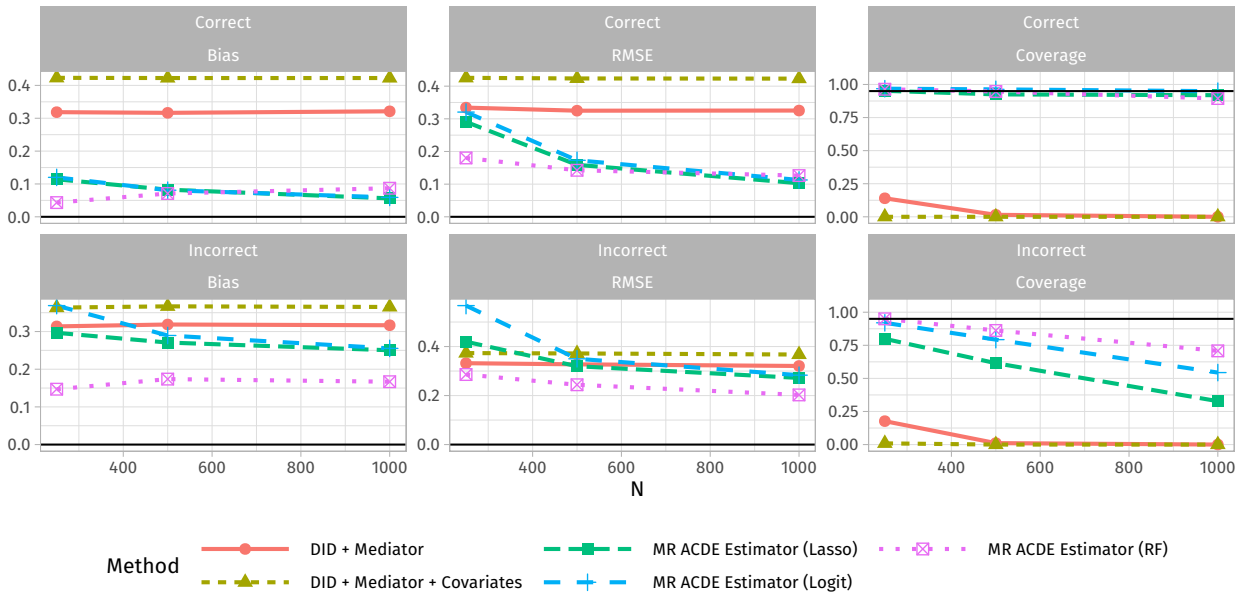

Figure 2: Performance of our multiply robust estimator as compared with difference-in-differences controlling for mediator and baseline covariates and difference-in-differences controlling for the mediator, baseline covariates, and intermediate covariates.

The performance of our MR ACDE estimator varies more across the correct and incorrect specification and across the estimation engines used. When the correctly specified variables are used, all of the multiply robust methods are similar in having low bias, low RMSE, and roughtly correct coverage. However, when using the incorrectly specified covariates, there is a much larger gap between the three MR approaches. All of these have higher bias in the misspecified setting, but the increase is more muted for the random forest compared to the others. The least flexible approach, the logit, shows the largest biases. The lasso approach here is at somewhat of a disadvantage because we only include squares and first-order interactions, but the true specification has cubic transformations. A richer set of basis functions could improve its performance. These results show that even when hampered by misspecified covariates, the flexible approaches can reduce bias.

6 Empirical Application

We now apply these methods to estimate whether a pro-transgender intervention changes support for nondiscrimination laws, holding constant feelings of warmth towards transgender people. To do so, we rely on data from a study by Broockman and Kalla (2016). The authors implemented a canvassing intervention that consisted of a brief conversation, encouraging respondents to engage in "perspective taking." Respondents were recruited from a list of registered voters, and were first asked to complete a baseline survey. Households were then assigned to receive either the perspective taking intervention (treatment) or information about recycling (control).

In addition to the baseline survey prior to the intervention, respondents completed four post-intervention surveys, which were conducted three days, three weeks, six weeks and three months after the intervention. The main outcome of interest is a seven-point scale of support for transgender nondiscrimination laws. The main finding in Broockman and Kalla (2016) is that the canvassing intervention increased support for nondiscrimination laws in the third and fourth post-intervention periods. (The authors speculate that the absence of treatment effects in the first two periods could be due to respondents' lack of knowledge about the meaning of the term 'transgender' and so included a definition in the subsequent waves.) While Broockman and Kalla (2016) report treatment effects based on cross-sectional differences between treatment and control groups after the intervention, we instead use changes in the outcome ΔY_i.

Our goal is to understand how subjective feelings toward transgender people may be a part of the effect of these interventions. To this end, we define our mediator to be survey items that measure feelings towards transgender people, which are observed in the baseline period and all posttreatment periods. For the purposes of the empirical application, we define Y_{i2} such that it is measured six weeks after the intervention, while M_{i2} is measured three weeks after the intervention. To construct the mediator, we rely on the transgender feeling thermometer, which is measured on a scale of 0-100, where higher values indicate "warmer" feelings toward the group. As we show in figure 3, thermometer scores often show a significant amount of clumping at "even" numbers such as 0, 50, and 100, and much of the informational content could be summarized as a person feeling cooly, warmly, or neutral

about a group. Thus, we transform these scores into a three-level discrete variable such that $M_{i1} = 1$ for participants who score below 50 on the thermometer (cooler feelings), $M_{i1} = 2$ for participants who score exactly 50 (neutral feelings), and $M_{i1} = 3$ for participants who score above 50 points (warm feelings). We use the same transformation for M_{i2}.

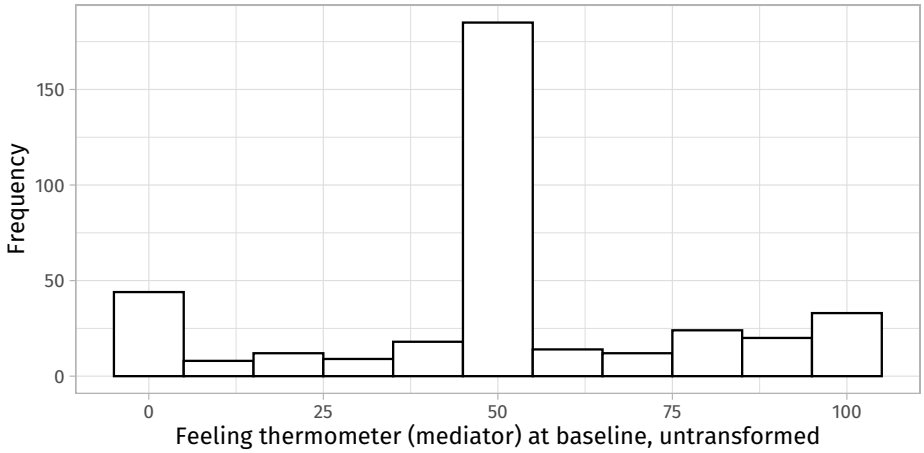

Figure 3: Distribution of the mediator at baseline, prior to discretizing

In addition, we include several pre- and posttreatment covariates, which mainly measure basic demographics (age, race, gender, etc), political leanings, and gender-related attitudes. We present a full list of these covariates is given in Table SM.1 in the Supplemental Materials. All pretreatment covariates are measured at baseline, while we obtain posttreatment covariates by differencing the first posttreatment period (three days after the intervention) relative to the baseline. For the ACDE-BC and ACDE-PC estimators, we use adaptive estimation for the nuisance functions with the lasso approach of Belloni and Chernozhukov (2013) from the `hdm` R package for the outcome regression and random forests from the `ranger` R package for the propensity scores for M_{i2}. For these, we pass all the covariates plus first-order interactions and squared terms for continuous variables (though these flexible terms are not included for the standard DID estimates). As dictated by our identifying assumptions, we omit intermediate covariates for the estimation of the ACDE-PCs. We restrict our sample to individuals for which all covariates are observed ($N = 369$).

We present the results in Figure 4, which includes the estimates $\widehat{\tau}_m$ (circles) and $\widehat{\gamma}_m$ (triangles), as well as standard difference-in-differences estimates. The DID results just conditioning on baseline

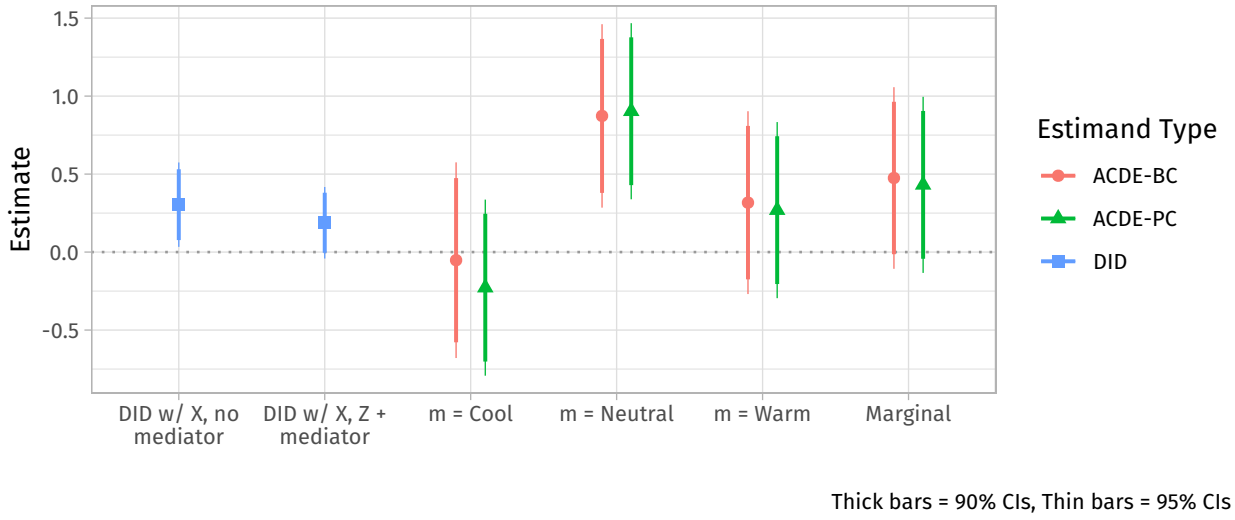

Figure 4: Controlled direct effect estimates using different estimation strategies.

covariates ("DID w/ X no mediator") replicate the main results of the original study: the perspective-taking intervention increased support for nondiscrimination laws. The magnitude of the DID estimate (0.304) is very similar to the cross-sectional estimate of the effects from the original study (0.36). Once we add intermediate covariates and the mediator into our DID analysis, however, the effect attenuates by almost 40% (0.188). Such a change might led an analyst to conclude that feelings about transgender people mediate the effect of the intervention. Of course, this ignores the potential for posttreatment bias in condition on the intermediate confounders.

When we look at our approach to estimating the controlled direct effects of the treatment, we see a slightly different and more nuanced set of results. For both of the marginal ACDEs, we see that the direct effect estimates are larger in magnitude than the DID baseline, by around 56% in the case the ACDE-BC. These effects also have much larger standard errors due in part to the estimation of the nuisance functions. The uncertainty in these results makes it difficult to compare to the overall DID estimates, but in terms of the point estimates, we would reach the opposite conclusion as the naive DID approach of simply conditioning on the intermediate covariates and mediator.

Figure 4 also shows the variation in the ACDE-BC and ACDE-PC across the "controlled" level of the mediator. The ACDEs for remaining feeling negatively toward transgender people ($m = 1$)

are actually slightly negative, though these effects are not statistically significant. The ACDEs for remaining feeling positively ($m = 3$) are almost identical in the point estimate to the baseline DID estimate with larger standard errors. Finally, the ACDEs for remaining neutral are both much larger than the baseline estimates but also statistically significant. Note that this is driven both by the direct effect nature of the estimand and how the overall effect of the intervention for those that felt neutrally at baseline is also higher.

These results are substantively important for the study of political behavior since they show that perspective-taking conversations can have political effects even when subjective feelings are unchanged. This points to the ability for political campaigns to persuade citizens about legal discrimination without necessarily altering their own personal feelings about the group, though we acknowledge that our effects are concentrated among those with neutral feelings toward transgender people at baseline. This is vital since a number of studies in political science have shown that subjective feelings toward outgroups are often formed in childhood and very difficult to change durably (Sears and Funk, 1999; Tesler, 2015). This is a positive sign for the health of democracies and how they can increase tolerant public policies without necessarily increasing interpersonal tolerance.

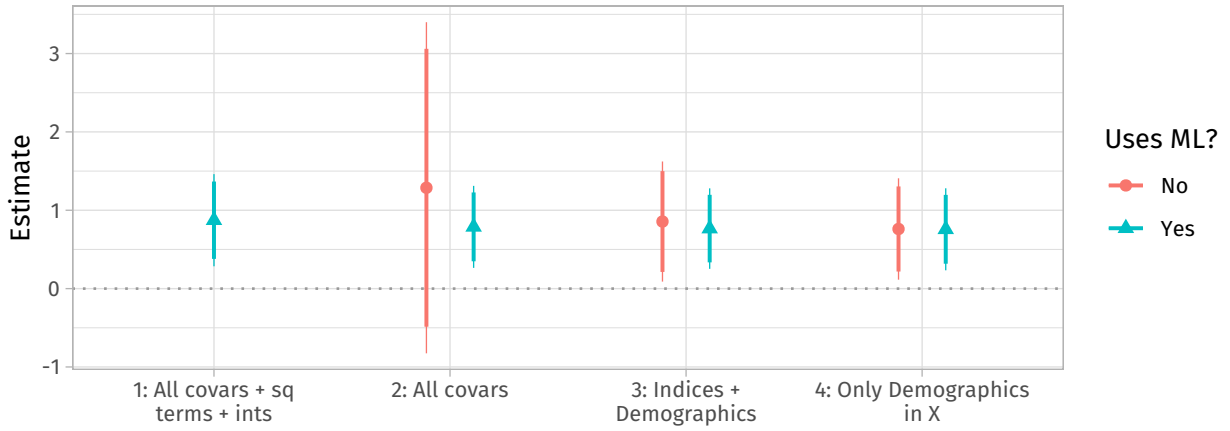

Figure 5: ACDE estimates for $m = 2$ across different covariate specifications for the adaptive/ML (blue triangles) and standard (red circles) estimation of the nuisance functions.

Finally, we also investigate how the use of adaptive estimation techniques for the nuisance func-

tions impacts the stability of our estimates across different specification choices. In particular, we varied the choice of variables to pass to either a standard set of models (OLS for the outcome regression and a multinomial logistic regression for the propensity scores) or the adaptive estimators described above. The sets of variables are (a) only demographics in the baseline covariates, (b) demographics and LGBT opinion indices in the baseline covariates, (c) the full set of baseline covariates, and (d) the full set of covariates plus squared terms for all continuous variables and all first-order interactions. In this last specification, the number of covariates is far larger than the number of units, so we only used the adaptive design for this specification. Figure 5 presents the results, which show that the adaptive design has a massive impact on the stability of estimates and their uncertainty across these specifications. The increase in uncertainty when adding additional controls is overwhelming for the standard estimators, but has almost no impact on the adaptive approach. Thus, the combination of cross-fitting and adaptive nuisance estimation perhaps provides a path toward much less model-dependent estimates and less opportunities for intentional or unintentional p-hacking.

7 Conclusion

In this paper, we have introduced a novel identification strategy for controlled direct effects under a difference-in-difference design embedded in a multiwave experimental study. Our key identifying assumptions allow for the mediator to be related to the baseline levels of the potential outcomes, which is far weaker than the selection-on-observables assumption traditionally used to identify the controlled direct effects. Our assumptions do require so-called parallel trends assumptions, meaning that the mediator must be unrelated to the *changes* in the potential outcomes over time. This approach highlights how having access to baseline measures of the outcome can allow researchers to weaken key assumptions in the pursuit of evaluating causal mechanisms. We have also built on recent work on doubly and multiply robust estimators to propose a multiply robust, semiparametrically efficient estimator for our proposed quantities. These estimators allow researchers to take full advantage of adaptive machine learning algorithms for estimating nuisance functions like propensity scores and outcome regressions.

Through both simulations and the empirical application, we have shown that proper adjustment for intermediate covariates can lead to different substantive conclusions. Our simulations show that naive adjustment can lead to severe bias and undercoverage of confidence intervals. In the empirical application, we saw the estimated direct effect of a perspective-taking intervention fixing feeling thermometer scores move in different directions compared to the overall average effects of the intervention along with large differences in the uncertainty of the estimates. Untangling the mechanisms in this case is somewhat difficult due to the estimation uncertainty, but we can detect large differences between the estimated direct effects for different levels of the mediator.

There are several avenues for future research in this area. We have focused here on a situation with effectively two time periods and two causal variables (a treatment and a mediator), but it should be possible to generalize this approach to handle treatment history of arbitrary length. This might allow for identification and inference for causal effects in marginal structural models with weaker assumptions on confounding between the outcome and the treatment history. In addition, in situations with more pretreatment measurements it may be possible to use those past measurements to measure and correct for deviations from the parallel trends assumptions. These are the key identification assumptions for our approach and attention to them is crucially important.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." *The Review of Economic Studies* 72(1):1–19.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Avin, Chen, Ilya Shpitser and Judea Pearl. 2005. Identifiability of Path-specific Effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05 San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. pp. 357–363.

Belloni, Alexandre and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19(2):521–547.

Bickel, P.J., C.A. Klaassen, Y. Ritov and J.A. Wellner. 1998. *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

Blackwell, Matthew and Anton Strezhnev. 2022. "Telescope matching for reducing model dependence in the estimation of the effects of time-varying treatments: An application to negative advertising." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 185(1):377–399.

Blackwell, Matthew and Soichiro Yamauchi. 2021. "Adjusting for Unmeasured Confounding in Marginal Structural Models with Propensity-Score Fixed Effects.".
URL: *https://arxiv.org/abs/2105.03478*

Bradic, Jelena, Weijie Ji and Yuqian Zhang. 2021. "High-dimensional Inference for Dynamic Treatment Effects." *arXiv:2110.04924 [cs, econ, math, stat]* .

Broockman, David E., Joshua L. Kalla and Jasjeet S. Sekhon. 2017. "The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs." *Political Analysis* 25(4):435–464.

Broockman, David and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Callaway, Brantly and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225(2):200–230.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21(1):C1–C68.

Deuchert, Eva, Martin Huber and Mark Schelker. 2019. "Direct and Indirect Effects Based on Difference-in-Differences With an Application to Political Preferences Following the Vietnam Draft Lottery." *Journal of Business & Economic Statistics* 37(4):710–720.

Ding, Peng and Fan Li. 2019. "A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment." *Political Analysis* 27(4):605–615.

Goetgeluk, Sylvie, Sijn Vansteelandt and Els Goetghebeur. 2008. "Estimation of controlled direct effects." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70(5):1049–1066.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225(2):254–277.

Hernán, Miguel A., Babette A. Brumback and James M. Robins. 2001. "Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments." *Journal of the American Statistical Association* 96(454):440–448.

Huber, Martin, Mark Schelker and Anthony Strittmatter. 2022. "Direct and Indirect Effects based on Changes-in-Changes." *Journal of Business & Economic Statistics* 40(1):432–443.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25(1):51–71.

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22(4):523–539.

Kennedy, Edward H., Sivaraman Balakrishnan and Max G'Sell. 2020. "Sharp instruments for classifying compliers and generalizing causal effects." *The Annals of Statistics* 48(4).

Lipsitch, Marc, Eric Tchetgen Tchetgen and Ted Cohen. 2010. "Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies." *Epidemiology* 21(3):383–388.

Murphy, S A, M J van der Laan, J M Robins and Conduct Problems Prevention Research Group. 2001. "Marginal Mean Models for Dynamic Regimes." *Journal of the American Statistical Association* 96(456):1410–1423.

Newey, Whitney K. 1990. "Semiparametric efficiency bounds." *Journal of Applied Econometrics* 5(2):99–135.

Orellana, Liliana, Andrea Rotnitzky and James M. Robins. 2010. "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content." *International Journal of Biostatistics* 6(2):–.

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics* 23(8):2379–2412.

Robins, James M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems*, ed. P. J. Green, N. L. Hjort and S. Richardson. Oxford University Press pp. 70–81.

Robins, James M. and Miguel A Hernán. 2009. Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, ed. Garrett Fitzmaurice, Marie Davidian, Geert Verbeke and Geert Molenberghs. Chapman & Hall/CRC pp. 553–599.

Robins, James M. and Sander Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3(2):143–155.

Sant'Anna, Pedro H.C. and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of Econometrics* 219(1):101–122.

Seaman, Shaun R. and Stijn Vansteelandt. 2018. "Introduction to Double Robust Methods for Incomplete Data." *Statistical Science* 33(2):184 – 197.

Sears, David O. and Carolyn L. Funk. 1999. "Evidence of the Long-Term Persistence of Adults' Political Predispositions." *The Journal of Politics* 61(1):1–28.

Shahn, Zach, Oliver Dukes, David Richardson, Eric Tchetgen Tchetgen and James Robins. 2022. "Structural Nested Mean Models Under Parallel Trends Assumptions.".
URL: *https://arxiv.org/abs/2204.10291*

Sofer, Tamar, David B. Richardson, Elena Colicino, Joel Schwartz and Eric J. Tchetgen Tchetgen. 2016. "On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences." *Statistical Science* 31(3).

Sun, Liyang and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225(2):175–199.

Tesler, Michael. 2015. "Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed." *American Journal of Political Science* 59(4):806–824.

van der Laan, Mark J. and Susan Gruber. 2012. "Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions." *The International Journal of Biostatistics* 8(1).

Supplemental Materials (to appear online)

A Comparison to Sequential Ignorability with a Lagged Dependent Variable

In this section we contrast the targets of inference under the difference-in-differences framework and the sequential ignorability with lagged dependent variable framework. For simplicity, we assume a binary mediator and that $M_{i1} = 0$ throughout and suppress any such conditioning statement. Let $F_{Y_1}(y \mid d, m, \mathbf{x}, \mathbf{z})$ be the cumulative density function of Y_{i1} given $D_i = d$, $M_{i2} = m$, $\mathbf{X}_i = \mathbf{x}$, and $\mathbf{Z} = \mathbf{z}$, $G_{Y_1}(y \mid d, \mathbf{x}, \mathbf{z})$ be the same distribution function without conditioning on M_{i2}, and let $\overline{\mu}(d, m, \mathbf{x}, \mathbf{z}, y) = \mathbb{E}[Y_{i2} \mid D_i = d, M_{i2} = m, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}, Y_{i1} = y]$.

Next, we describe the targets of inference for both the DID and LDV approaches. These are quantities that are, under each set of assumptions, identify the ACDE but remain valid observational quantities even when those assumptions do not hold. Our bracketing result will order these quantities and so is valid regardless of whether or not the identification assumptions actually hold. First, we write the quantity that, under parallel trends, would identify $\mathbb{E}[Y_{i2}(0,0)]$:

$$\widetilde{\mu}_{0,\text{DID}} = \mathbb{E}[Y_{i1} \mid D_i = 0, M_{i2} = 0] + \int_{\mathbf{x},\mathbf{z}} \mathbb{E}[\Delta Y_i \mid D_i = 0, M_{i2} = 0, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}] dP(\mathbf{x}, \mathbf{z} \mid D_i = 0),$$

with $\widetilde{\mu}_{1,\text{DID}}$ being defined similarly. Under Assumption 1 and 2, $\widetilde{\tau}_{\text{DID}} = \widetilde{\mu}_{1,\text{DID}} - \widetilde{\mu}_{0,\text{DID}}$ would identify the ACDE, τ. Under a lagged dependent variable, the g-computational formula gives the following identification formula for $\mathbb{E}[Y_{i2}(0,0)]$:

$$\widetilde{\mu}_{0,\text{LDV}} = \int_{\mathbf{x},\mathbf{z},y} \mathbb{E}[Y_{i2} \mid D_i = 0, M_{i2} = 0, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}, Y_{i1} = y] dG_{Y_1}(y \mid \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0),$$

with $\widetilde{\mu}_{1,\text{LDV}}$ defined similarly. If LDV sequential ignorability holds, $Y_{i2}(d, m) \perp\!\!\!\perp M_{i2} \mid D_i = d, \mathbf{X}_i, \mathbf{Z}_i, Y_{i1}$, then $\widetilde{\tau}_{\text{LDV}} = \widetilde{\mu}_{1,\text{LDV}} - \widetilde{\mu}_{0,\text{LDV}}$ would identify the ACDE.

Theorem 5. *The difference between $\widetilde{\mu}_{0,DID}$ and $\widetilde{\mu}_{0,LDV}$ is*

$$\begin{aligned} \widetilde{\tau}_{DID} - \widetilde{\tau}_{LDV} = &\int_{\mathbf{x},\mathbf{z},y} \Delta_1(y) \left(dF_{Y_1}(y \mid 1, 0, \mathbf{x}, \mathbf{z}) - dG_{Y_1}(y \mid 1, \mathbf{x}, \mathbf{z}) \right) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0) \\ &- \int_{\mathbf{x},\mathbf{z},y} \Delta_0(y) \left(dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) - dG_{Y_1}(y \mid 0, \mathbf{x}, \mathbf{z}) \right) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0), \end{aligned} \tag{2}$$

where $\Delta_d(y) = \overline{\mu}(d, 0, \mathbf{x}, \mathbf{z}, y) - y$.

Proof. Using iterated expectations, we can write $\widetilde{\mu}_{0,\text{DID}}$ as

$$\begin{aligned}
\widetilde{\mu}_{0,\text{DID}} &= \int_{\mathbf{x},\mathbf{z},y} y dG_{Y_1}(y \mid \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0) \\
&\quad + \int_{\mathbf{x},\mathbf{z},y} \mathbb{E}[\Delta Y_i \mid D_i = 0, M_{i2} = 0, \mathbf{X}_i = \mathbf{x}, \mathbf{Z}_i = \mathbf{z}, Y_{i1} = y] dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0) \\
&= \int_{\mathbf{x},\mathbf{z},y} y dG_{Y_1}(y \mid \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0) \\
&\quad + \int_{\mathbf{x},\mathbf{z},y} \Delta_0(y) dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0)
\end{aligned}$$

Combining this with the definition of $\widetilde{\mu}_{0,\text{LDV}}$, we obtain

$$\begin{aligned}
\widetilde{\mu}_{0,\text{DID}} - \widetilde{\mu}_{0,\text{LDV}} &= \int_{\mathbf{x},\mathbf{z},y} \Delta(y) dF_{Y_1}(y \mid 0, 0, \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 1) \\
&\quad - \int_{\mathbf{x},\mathbf{z},y} \Delta_0(y) dG_{Y_1}(y \mid \mathbf{x}, \mathbf{z}) dP(\mathbf{x}, \mathbf{z} \mid D_i = 0).
\end{aligned}$$

Applying the same logic to $\widetilde{\mu}_{1,\text{DID}} - \widetilde{\mu}_{1,\text{LDV}}$ yields the result. □

Our ACDE-TP estimand, on the other hand, has a more specific relationship with the sequential ignorability approach. In fact, because the identification assumptions for that estimand are simply parallel trends for a four-category outcome, we can apply the results of Ding and Li (2019) to obtain a bracketing result between the DID estimand and the LDV estimand. Let $\widetilde{\gamma}_{\text{DID}}$ and $\widetilde{\gamma}_{\text{LDV}}$ be the targets of inference for these two settings, identified in a similar manner to the two above. Following Ding and Li (2019), we first invoke conditions on the data generating process:

Condition 1 (Stationarity)**.** $\partial \overline{\mu}(d, m, \mathbf{x}, \mathbf{z}, y)/\partial y < 1$ *for all* y.

Condition 2 (Stochastic Monotonicity)**.** *Either (a)* $F_{Y_1}(y \mid d, 1, \mathbf{x}, \mathbf{z}) \geq F_{Y_1}(y \mid d, 0, \mathbf{x}, \mathbf{z})$ *for all* y*; or (b)* $F_{Y_1}(y \mid d, 0, \mathbf{x}, \mathbf{z}) \geq F_{Y_1}(y \mid d, 1, \mathbf{x}, \mathbf{z})$.

Condition 1 is a limit on the growth of the time series of the outcome and with a linear model, it would require that the coefficient on the lagged dependent variable be less than one. This is a commonly invoked assumption with panel and time-series data. Condition 2 characterizes the relationship between the lagged dependent variable and the mediator, with Condition 2(a) meaning that

the group with $M_{i2} = 1$ has higher baseline outcomes across the entire distribution compared to the $M_{i2} = 0$ group and vice versa for Condition 2(b). We say Condition 1 and 2 are conditions rather than assumptions because they are both empirically testable Ding and Li (2019).

Ding and Li (2019) have shown that under Conditions 1 and 2(a) we have $\widetilde{\gamma}_{\text{DID}} \geq \widetilde{\gamma}_{\text{LDV}}$, and under Conditions 1 and 2(b), we have $\widetilde{\gamma}_{\text{DID}} \leq \widetilde{\gamma}_{\text{LDV}}$. Thus, if one of these two sets of conditions holds and one of the two sets of identifying assumptions holds, then the two estimands will bracket the true value of the ACDE-TP.

B Proofs

Proof of Theorem 1. We prove part (a) for τ_m. The proof for γ_m is very similar and so we omit it. Below we combine $\mathbf{X}_i$ and $\mathbf{Z}_i$ into a single vector $\mathbf{X}_i$ since their role in the proof is the same. We begin with the first term of τ_m. By randomization and the law of total probability we have:

$$\begin{aligned}
\mathbb{E}\{Y_{i2}(1,m) \mid M_{i1} = m\} &= \mathbb{E}\{Y_{i2}(1,m) \mid D_i = 1, M_{i1} = m\} \\
&= \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(1,m) \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m) \\
&= \mathbb{E}(Y_{i1}(0,m) \mid D_i = 1, M_{i1} = m) \\
&\quad + \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(1,m) - Y_{i1}(0,m) \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m)
\end{aligned}$$

The first term is identified and, using Assumption 2 we can write the second term as:

$$\int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i(1,m) \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m)$$

Recall that $\overline{\pi}_{2m_2}(d,m) = \mathbb{P}(M_{i2} = m_2 \mid D_i = d, M_{i1} = m)$. By consistency and then Bayes' rule, this becomes,

$$\begin{aligned}
&\int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} dP(\mathbf{x} \mid D_i = 1, M_{i1} = 0) \\
&= \int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i = 1, M_{i1} = m, \mathbf{X}_i = \mathbf{x}, M_{i2} = m\} \frac{\overline{\pi}_{2m}(1,m)}{\pi_{2m}(1,m,\mathbf{x})} dP(\mathbf{x} \mid D_i = 1, M_{i1} = m, M_{i2} = m)
\end{aligned}$$

Once again applying the law of total probability and then using the definition of conditional probability, we can simplify this to:

$$\mathbb{E}\left\{ \frac{\Delta Y_i}{\pi_{2m}(1,m,\mathbf{X}_i)} \,\middle|\, D_i = 1, M_{i1} = m, M_{i2} = m \right\} (\overline{\pi}_{2m}(1,m)) = \mathbb{E}\left\{ \frac{W_{i1m} D_i W_{i2m} \Delta Y_{i2}}{\pi_{1m} \pi_d \pi_{2m}(1,m,\mathbf{X}_i)} \right\}$$

Thus, we can write the first term in the τ_m (using randomization on the first term):

$$\mathbb{E}\{Y_{i2}(1,m) \mid M_{i1}=m\} = \mathbb{E}(Y_{i1}(0,m) \mid M_{i1}=m) + \mathbb{E}\left\{\frac{W_{i1m}D_iW_{i2m}}{\pi_{1m}\pi_d\pi_{2m}(1,m,\mathbf{X}_i)}\Delta Y_{i2}\right\}$$

We now turn to the second term of the τ_m. Again using the law of total probability and Assumption 2, we have:

$$\begin{aligned}
&\mathbb{E}\{Y_{i2}(0,m) \mid M_{i1}=m\} = \mathbb{E}\{Y_{i2}(0,m) \mid D_i=0, M_{i1}=m\} \\
&= \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(0,m) \mid D_i=0, M_{i1}=m, \mathbf{X}_i=\mathbf{x}\} dP(\mathbf{x} \mid D_i=1, M_{i1}=0) \\
&= \mathbb{E}(Y_{i1}(0,m) \mid M_{i1}=m) \\
&\quad + \int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(0,m) - Y_{i1}(0,m) \mid D_i=0, M_{i1}=m, \mathbf{X}_i=\mathbf{x}\} dP(\mathbf{x} \mid D_i=0, M_{i1}=m)
\end{aligned}$$

Once again, using the law of total probability and Assumption 2, this term becomes:

$$\begin{aligned}
&\int_{\mathbf{x}} \mathbb{E}\{Y_{i2}(0,m) - Y_{i1}(0,m) \mid D_i=0, M_{i1}=m, \mathbf{X}_i=\mathbf{x}, M_{i2}=m\} dP(\mathbf{x} \mid D_i=0, M_{i1}=m) \\
&\int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i=0, M_{i1}=m, \mathbf{X}_i=\mathbf{x}, M_{i2}=m\} dP(\mathbf{x} \mid D_i=0, M_{i1}=m) \\
&= \int_{\mathbf{x}} \mathbb{E}\{\Delta Y_i \mid D_i=0, M_{i1}=m, \mathbf{X}_i=\mathbf{x}, M_{i2}=m\} \frac{\overline{\pi}_{2m}(0,m)}{\pi_{2m}(0,m,\mathbf{x})} dP(\mathbf{x} \mid D_i=0, M_{i1}=m, M_{i2}=m)
\end{aligned}$$

Finally, using the law of total probability and the definition of conditional expectation, we can write this term as:

$$\begin{aligned}
&\mathbb{E}\left\{\frac{\Delta Y_i}{\pi_{1m}(1-\pi_d)\pi_{2m}(0,m,\mathbf{X}_i)} \,\middle|\, D_i=0, M_{i1}=m, M_{i2}=m\right\}\overline{\pi}_{2m}(0,m)) \\
&\qquad = \mathbb{E}\left\{\frac{W_{i1m}(1-D_i)W_{i2m}}{\pi_{1m}(1-\pi_d)\pi_{2m}(0,m,\mathbf{X}_i)}\Delta Y_{i2}\right\}
\end{aligned}$$

Combining this with the results on the first term gives the desired result for τ_m. □

Proof of Theorem 2. We write $\psi_{i,\tau}(m) = \psi_{i,0}(m) - \psi_{i,0}(m)$, where

$$\begin{aligned}
\psi_{1m}(\mathbf{O}_i;\boldsymbol{\eta}) &= \left(\frac{W_{i1m}D_iW_{i2m}}{\pi_{1m}\pi_d\pi_{i2m}}\right)(\Delta Y_i - \mu_{i,1m}) + \frac{W_{i1m}D_i}{\pi_{1m}\pi_d}(\mu_{i,1m} - \nu_{i,1m}) + \frac{W_{i1m}}{\pi_{1m}}\nu_{i,1m} \\
\psi_{0m}(\mathbf{O}_i;\boldsymbol{\eta}) &= \left(\frac{W_{i1m}(1-D_i)W_{i2m}}{\pi_{1m}(1-\pi_d)\pi_{i2m}}\right)(\Delta Y_i - \mu_{i,0m}) + \frac{W_{i1m}(1-D_i)}{\pi_{1m}(1-\pi_d)}(\mu_{i,0m} - \nu_{i,0m}) + \frac{W_{i1m}}{\pi_{1m}}\nu_{i,0m}.
\end{aligned}$$

Based on Theorem 1, we can express the estimation error for a given estimator for the nuisance function $\widehat{\boldsymbol{\eta}}$ as

$$\begin{aligned}\widehat{\tau}_m - \tau_m = &\frac{1}{N}\sum_{i=1}^{N}(\psi_{1m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}) - \mathbb{E}[\Delta Y_i(1,m) \mid M_i = m]) \\ &- \frac{1}{N}\sum_{i=1}^{N}(\psi_{0m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}) - \mathbb{E}[\Delta Y_i(0,m) \mid M_i = m]).\end{aligned}$$

We demonstrate the double robustness result on the first expression $\psi_{1m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}})$ and the corresponding result for $\psi_{0m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}})$ following similarly. We first consider the case where the propensity score model is correctly specified, so that

$$\begin{aligned}\widehat{\pi}_{2m_2}(m_1,d,\mathbf{x},\mathbf{z}) &\xrightarrow{p} \pi_{2m_2}(m_1,d,\mathbf{x},\mathbf{z}),\\ \widehat{\mu}_{1,m_2}(m_1,\mathbf{x},\mathbf{z}) &\xrightarrow{p} \mu^*_{1,m_2}(m_1,\mathbf{x},\mathbf{z}),\\ \widehat{\nu}_{1,m_2}(m_1,\mathbf{x}) &\xrightarrow{p} \nu^*_{1,m_2}(m_1,\mathbf{x}),\end{aligned}$$

where $\mu^*_{1,m}$ and $\nu^*_{1,m}$ are functions that do not necessarily correspond to $\mu_{1,m}$ and $\nu_{1,m}$. Then by Slutsky's Theorem, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\psi_{1m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}) &= \mathbb{E}\left\{\left(\frac{W_{i1m}D_iW_{i2m}}{\pi_{1m}\pi_d\pi_{i2m}}\right)\left(\Delta Y_i - \mu^*_{i,1m}\right) + \frac{W_{i1m}D_i}{\pi_{1m}\pi_d}\left(\mu^*_{i,1m} - \nu^*_{i,1m}\right) + \frac{W_{i1m}}{\pi_{1m}}\nu^*_{i,1m}\right\} + o_p(1)\\ &= \mathbb{E}\left\{\left(\frac{W_{i1m}D_i}{\pi_{1m}\pi_d}\right)\left(\mu_{i,1m} - \mu^*_{i,1m}\right) + \frac{W_{i1m}D_i}{\pi_{1m}\pi_d}\left(\mu^*_{i,1m} - \nu^*_{i,1m}\right) + \frac{W_{i1m}}{\pi_{1m}}\nu^*_{i,1m}\right\} + o_p(1)\\ &= \mathbb{E}\left\{\frac{W_{i1m}}{\pi_{1m}}\mu_{i,1m}\right\} + o_p(1)\\ &= \mathbb{E}[\Delta Y_i(1,m) \mid M_{i1} = m] + o_p(1)\end{aligned}$$

The second equality follows from iterated expectations and the definition of π_{i2m}, the third by randomization of D_i and the last by the definition of conditional expectation. This, combined with the equivalent result for $\psi_{0m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}})$, establishes consistency when the propensity score model is correct.

Now we turn to the setting where the outcome regressions are correctly specified so that

$$\begin{aligned}\widehat{\pi}_{2m_2}(m_1,d,\mathbf{x},\mathbf{z}) &\xrightarrow{p} \pi^*_{2m_2}(m_1,d,\mathbf{x},\mathbf{z}),\\ \widehat{\mu}_{1,m_2}(m_1,\mathbf{x},\mathbf{z}) &\xrightarrow{p} \mu_{1,m_2}(m_1,\mathbf{x},\mathbf{z}),\\ \widehat{\nu}_{1,m_2}(m_1,\mathbf{x}) &\xrightarrow{p} \nu_{1,m_2}(m_1,\mathbf{x}).\end{aligned}$$

With these, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\psi_{1m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}}) &= \mathbb{E}\left\{\left(\frac{W_{i1m}D_iW_{i2m}}{\pi_{1m}\pi_d\pi^*_{i2m}}\right)\left(\Delta Y_i-\mu_{i,1m}\right)+\frac{W_{i1m}D_i}{\pi_{1m}\pi_d}\left(\mu_{i,1m}-\nu_{i,1m}\right)+\frac{W_{i1m}}{\pi_{1m}}\nu_{i,1m}\right\}+o_p(1)\\
&= \mathbb{E}\left\{\left(\frac{W_{i1m}D_i\pi_{i2m}}{\pi_{1m}\pi_d\pi^*_{i2m}}\right)\left(\mu_{i,1m}-\mu_{i,1m}\right)+\frac{W_{i1m}D_i}{\pi_{1m}\pi_d}\left(\mu_{i,1m}-\nu_{i,1m}\right)+\frac{W_{i1m}}{\pi_{1m}}\nu_{i,1m}\right\}+o_p(1)\\
&= \mathbb{E}\left\{\left(\frac{W_{i1m}D_i}{\pi_{1m}\pi_d}\right)\left(\mu_{i,1m}-\nu_{i,1m}\right)+\frac{W_{i1m}}{\pi_{1m}}\nu_{i,1m}\right\}+o_p(1)\\
&= \mathbb{E}\left\{\frac{W_{i1m}}{\pi_{1m}}\mu_{i,1m}\right\}+o_p(1)\\
&= \mathbb{E}[\Delta Y_i(1,m)\mid M_{i1}=m]+o_p(1)
\end{aligned}$$

This, combined with the equivalent result for $\psi_{0m}(\mathbf{O}_i;\widehat{\boldsymbol{\eta}})$, establishes consistency when the outcome regressions are correct. The result for γ_m also follows similarly.

□

B.1 Efficient influence function

Proof of Theorem 3. Define the collection of potential outcomes in each period as $\mathbf{Y}_{i2}(\bullet)=\{Y_{i2}(0,m),Y_{i2}(1,m)\}_{m\in\mathcal{M}}$ and $\mathbf{Y}_{i1}(\bullet)=\{Y_{i1}(0,m)\}_{m\in\mathcal{M}}$ with representative values $\mathbf{y}_2(\bullet)$ and $\mathbf{y}_1(\bullet)$, respectively. Then the full data is given by

$$\mathbf{H}_i=(\mathbf{Y}_{i2}(\bullet),\mathbf{Y}_{i1}(\bullet),M_{i2},\mathbf{Z}_i,D_i,\mathbf{X}_i,M_{i1}),$$

and let $\mathbf{h}$ be a possible value of $\mathbf{H}$. Then the density of $\mathbf{H}$ for some sigma-finite measure is

$$\begin{aligned}
\overline{q}(\mathbf{h}) = \prod_{m_2\in\mathcal{M}}\prod_{m_1\in\mathcal{M}} &\overline{f}(\mathbf{y}_2(\cdot),\mathbf{y}_1(\cdot)\mid m_2,D_i=1,m_1,\mathbf{z},\mathbf{x})^{w_{m_2}dw_{m_1}}\\
&\times\overline{f}(\mathbf{y}_2(\cdot),\mathbf{y}_1(\cdot)\mid m_2,D_i=0,m_1,\mathbf{z},\mathbf{x})^{w_{m_2}(1-d)w_{m_1}}\\
&\times\pi_{2m_2}(1,m_1,\mathbf{z},\mathbf{x})^{w_{m_2}dw_{m_1}}\pi_{2m_2}(0,m_1,\mathbf{z},\mathbf{x})^{w_{m_2}(1-d)w_{m_1}},\\
&\times f(\mathbf{z}\mid D_i=1,m_1,\mathbf{x})^{dw_{m_1}}f(\mathbf{z}\mid D_i=0,m_1,\mathbf{x})^{(1-d)w_{m_1}}\\
&\times f(\mathbf{x}\mid m_1)^{w_{m_1}}\pi_d^d(1-\pi_d)^{(1-d)}\pi_{1m_1}^{w_{m_1}}
\end{aligned}$$

where w_{m_1} is 1 when $M_{i1}=m_1$ and 0 otherwise, with w_{m_2} defined similarly. In addition to the propensity scores that have already been defined, this density contains the following:

- $\overline{f}(\mathbf{y}_2(\cdot), \mathbf{y}_1(\cdot) \mid m_2, D_i = d, m_1, \mathbf{z}, \mathbf{x})$ is the density of the potential outcomes conditional on $M_{i2} = m_2$, $D_i = d$, $M_{i1} = m_1$, $\mathbf{Z}_i = \mathbf{z}$, and $\mathbf{X}_i = \mathbf{x}$, where $m_1, m_2 \in \mathcal{M}$, $d \in \{0, 1\}$, $\mathbf{z} \in \mathbb{R}^{k_z}$, and $\mathbf{x} \in \mathbb{R}^{k_x}$.

- $f(\mathbf{z} \mid D_i = d, m_1, \mathbf{x})$ is the density of $\mathbf{Z}_i$ conditional on $D_i = d$, $\mathbf{X}_i = \mathbf{x}$, and $M_{i1} = m_1$.

- $f(\mathbf{x} \mid m_1)$ is the density of $\mathbf{X}_i$ conditional on $M_{i1} = m_1$.

We now the turn to the density of the observed data, $\mathbf{O}_i = (Y_{i2}, Y_{i1}, M_{i2}, \mathbf{Z}_i, D_i, \mathbf{X}_i, M_{i1})$. We write the density of the observed outcomes as

$$f(y_2, y_1 \mid m_2, 1, m_1, \mathbf{z}, \mathbf{x}),$$

which marginalizes the $\overline{f}(\cdot)$ over the potential outcomes where $D_i \neq 1$, $M_{i2} \neq m_2$, or $M_{i1} \neq m_1$. Consider a possible value of the observed data

$$\mathbf{o} = (y_2, y_1, j_2, d, j_1, z, x)'$$

The density of the observed data $\mathbf{O}_i$ can be written as

$$\begin{aligned} q(\mathbf{o}; \theta) = \prod_{m_2 \in \mathcal{M}} \prod_{m_1 \in \mathcal{M}} & \left[f(y_2, y_1 \mid m_2, 1, m_1, \mathbf{z}, \mathbf{x}) \pi_{2m_2}(1, m_1, \mathbf{z}, \mathbf{x}) \right]^{d\mathbf{1}(m_2 = j_2, m_1 = j_1)} \\ & \times \left[f(y_2, y_1 \mid m_2, 0, m_1, \mathbf{z}, \mathbf{x}) \pi_{2m_2}(0, m_1, \mathbf{z}, \mathbf{x}) \right]^{(1-d)\mathbf{1}(m_2 = j_2, m_1 = j_1)} \\ & \times \left[f(\mathbf{z} \mid D_i = 1, m_1, \mathbf{x})^d f(\mathbf{z} \mid D_i = 0, m_1, \mathbf{x})^{(1-d)} \right]^{\mathbf{1}(m_1 = j_1)} \\ & \times f(\mathbf{x} \mid m_1)^{\mathbf{1}(m_1 = j_1)} \pi_d^d (1 - \pi_d)^{(1-d)} \pi_{1m_1}^{\mathbf{1}(m_1 = j_1)}. \end{aligned}$$

We consider a regular parametric submodel for the joint distribution of $\mathbf{O}_i$, with log likelihood

$$\begin{aligned} &\log q(\mathbf{o}; \theta) = \\ &\sum_{m_2 \in \mathcal{M}} \sum_{m_1 \in \mathcal{M}} \Big[d\mathbf{1}(m_2 = j_2, m_1 = j_1) \left(\log f(y_2, y_1 \mid m_2, 1, m_1, \mathbf{z}, \mathbf{x}; \theta) + \log \pi_{2m_2}(1, m_1, \mathbf{z}, \mathbf{x}; \theta) \right) \\ &\qquad + (1 - d)\mathbf{1}(m_2 = j_2, m_1 = j_1) \left(\log f(y_2, y_1 \mid m_2, 0, m_1, \mathbf{z}, \mathbf{x}; \theta) + \log \pi_{2m_2}(0, m_1, \mathbf{z}, \mathbf{x}; \theta) \right) \Big] \\ &+ \sum_{m_1 \in \mathcal{M}} \mathbf{1}(m_1 = j_1) \left(d \log f(\mathbf{z} \mid 1, m_1, \mathbf{x}; \theta) + (1 - d) \log f(\mathbf{z} \mid 0, m_1, \mathbf{x}; \theta) + \log f(x \mid m_1; \theta) \right) \end{aligned}$$

where, $q(\cdot;\theta_0) = q(\cdot)$ so that θ_0 is the true value of the parameters. This parametric submodel yields the following score:

$$S(\mathbf{o};\theta) = S_y(y_2, y_1, j_2, s, j_1, \mathbf{z}, \mathbf{x};\theta) + S_m(j_2, s, j_1, \mathbf{z}, \mathbf{x};\theta) + S_z(\mathbf{z}, j_1, s, \mathbf{x};\theta) + S_x(\mathbf{x}, j_1;\theta)$$

where,

$$S_y(y_2, y_1, j_2, s, j_1, \mathbf{z}, \mathbf{x};\theta) = \sum_{m_1\in\mathcal{M}} \sum_{d\in\{0,1\}} \sum_{m_2\in\mathcal{M}} \mathbf{1}\left(m_1 = j_1, d = s, m_2 = j_2\right) \frac{d}{d\theta} \log f(y_2, y_1 \mid m_2, d, m_1, \mathbf{z}, \mathbf{x};\theta)$$

$$S_m(j_2, s, j_1, \mathbf{z}, \mathbf{x};\theta) = \sum_{m_1\in\mathcal{M}} \sum_{d\in\{0,1\}} \sum_{m_2\in\mathcal{M}} \mathbf{1}\left(m_1 = j_1, d = s, m_2 = j_2\right) \frac{\dot{\pi}_{2m_2}(d, m_1, \mathbf{z}, \mathbf{x};\theta)}{\pi_{2m_2}(d, m_1, \mathbf{z}, \mathbf{x};\theta)}$$

$$S_z(\mathbf{z}, s, j_1, \mathbf{x};\theta) = \sum_{m_1\in\mathcal{M}} \sum_{d\in\{0,1\}} \mathbf{1}\left(m_1 = j_1, d = s\right) \frac{d}{d\theta} \log f(\mathbf{z} \mid d, m_1, \mathbf{x};\theta)$$

$$S_x(\mathbf{x}, j_1;\theta) = \sum_{m_1\in\mathcal{M}} \mathbf{1}(m_1 = j_1) \frac{d}{d\theta} \log f(\mathbf{x} \mid m_1;\theta)$$

Let $L_0^2(F_W)$ be the usual Hilbert space of zero-mean, square-integrable functions with respect to the distribution F_W. The tangent space of the model is $\mathcal{H} = \mathcal{H}_y + \mathcal{H}_m + \mathcal{H}_z + \mathcal{H}_x$, where

$$\mathcal{H}_y = \{S_y(Y_{i2}, Y_{i1}, M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) : S_y(Y_{i2}, Y_{i1}, M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) \in L_0^2(F_{Y_2,Y_1|M_2,D,M_1,\mathbf{Z},\mathbf{X}})\}$$

$$\mathcal{H}_m = \{S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) : S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) \in L_0^2(F_{M_2|D,M_1,\mathbf{Z},\mathbf{X}})\}$$

$$\mathcal{H}_z = \{S_z(\mathbf{Z}_i, D_i, M_{i1}, \mathbf{X}_i) : S_z(\mathbf{Z}_i, D_i, M_{i1}, \mathbf{X}_i) \in L_0^2(F_{\mathbf{Z}|D,M_1,\mathbf{X}})\}$$

$$\mathcal{H}_x = \{S_x(\mathbf{X}_i, M_{i1}) : S_x(\mathbf{X}_i, M_{i1}) \in L_0^2(F_{\mathbf{X}|M_1})\},$$

The further restrictions on the tangent space are that we have $\mathbb{E}[S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i) \mid D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i] = \sum_{m_2\in\mathcal{M}} \dot{\pi}_{2m_2}(D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)$ and

$$\mathbb{E}[S_m(M_{i2}, D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)^2 \mid D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i] = \sum_{m_2\in\mathcal{M}} \dot{\pi}_{2m_2}(D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i)^2 / \pi_{2m_2}(D_i, M_{i1}, \mathbf{Z}_i, \mathbf{X}_i).$$

We can write the ACDE as a function of the regular parametric submodel as

$$\begin{aligned}\tau_m(\theta) = &\int_{\mathbf{x}}\int_{\mathbf{z}}\int_{y_1,y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x};\theta) f(\mathbf{z} \mid 1, m, \mathbf{x};\theta) f(\mathbf{x} \mid m;\theta) dy_2 dy_2 d\mathbf{z} d\mathbf{x} \\ &- \int_{\mathbf{x}}\int_{\mathbf{z}}\int_{y_1,y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 0, m, \mathbf{z}, \mathbf{x};\theta) f(\mathbf{z} \mid 0, m, \mathbf{x};\theta) f(\mathbf{x} \mid m;\theta) dy_2 dy_2 d\mathbf{z} d\mathbf{x},\end{aligned}$$

where $\tau_m(\theta_0) = \tau_m$.

Our proposed influence function will be the efficient influence function if it is in the tangent space $\mathcal{H}$ and meets the following condition:

$$\frac{\partial \tau_m(\theta_0)}{\partial \theta} = \mathbb{E}\left[\psi_{\tau_m}(\mathbf{O}_i) S(\mathbf{O}_i; \theta_0)\right].$$

We can derive the pathwise derivative as

$$\begin{aligned}
\frac{\partial \tau_m(\theta_0)}{\partial \theta} = &\int_{\mathbf{x}} \int_{\mathbf{z}} \int_{y_1, y_2} \Big[(y_2 - y_1) S(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}) f(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}) f(\mathbf{z} \mid 1, m, \mathbf{x}) \\
&\qquad \times f(\mathbf{x} \mid m) dy_2 dy_2 d\mathbf{z} d\mathbf{x}\Big] \\
&+ \int_{\mathbf{x}} \int_{\mathbf{z}} \int_{y_1, y_2} \Big[(y_2 - y_1) S(y_2, y_1 \mid m, 0, m, \mathbf{z}, \mathbf{x}) f(y_2, y_1 \mid m, 1, m, \mathbf{z}, \mathbf{x}) f(\mathbf{z} \mid 0, m, \mathbf{x}) \\
&\qquad \times f(\mathbf{x} \mid m) dy_2 dy_2 d\mathbf{z} d\mathbf{x}\Big] \\
&+ \int_{\mathbf{x}} \int_{\mathbf{z}} (\mu_{1m}(m, 1, m, \mathbf{z}, \mathbf{x}) - \nu_{1m}(m, \mathbf{x})) S(\mathbf{z} \mid 1, m, \mathbf{x}) f(\mathbf{z} \mid 1, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{z} d\mathbf{x} \\
&+ \int_{\mathbf{x}} \int_{\mathbf{z}} (\mu_{0m}(m, 0, m, \mathbf{z}, \mathbf{x}) - \nu_{0m}(m, \mathbf{x})) S(\mathbf{z} \mid 0, m, \mathbf{x}) f(\mathbf{z} \mid 0, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{z} d\mathbf{x} \\
&+ \int_{\mathbf{x}} (\nu_{1m}(m, \mathbf{x}) - \nu_{0m}(m, \mathbf{x}) - \tau_m) S(\mathbf{x} \mid m) f(\mathbf{x} \mid m) d\mathbf{x},
\end{aligned}$$

Upon inspection, $\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_0)$ satisfies the condition and is in $\mathcal{H}$. Thus, by Theorem 3.1 of Newey (1990), $\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_0)$ is the efficient influence function for τ_m and the latter is a pathwise differentiable parameter. This also implies that the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_0)^2]$.

For our other estimand, note that

$$\begin{aligned}
\gamma_m = \mathbb{E}\big[&\mathbb{E}\left[\Delta Y_i \mid M_{i1} = m, D_i = 1, M_{i2} = m, X_i\right] \\
&- \mathbb{E}\left[\Delta Y_i \mid M_{i1} = m, D_i = 1, M_{i2} = m, X_i\right] \mid M_{i1} = m, D_i = 1, M_{i2} = m\big]
\end{aligned}$$

Thus, under the regular parametric submodel, we can write this as

$$\begin{aligned}
\gamma_m(\theta) = &\frac{\int_{\mathbf{x}} \int_{y_1, y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 1, m, \mathbf{x}; \theta) \pi_{2m}(m, 1, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta) dy_1 dy_2 d\mathbf{x}}{\int_{\mathbf{x}} \pi_{2m}(m, 1, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta)} \\
&- \frac{\int_{\mathbf{x}} \int_{y_1, y_2} (y_2 - y_1) f(y_2, y_1 \mid m, 0, m, \mathbf{x}; \theta) \pi_{2m}(m, 1, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta) dy_1 dy_2 d\mathbf{x}}{\int_{x} \pi_{2m}(m, 1, \mathbf{x}; \theta) f(\mathbf{x} \mid m; \theta)}
\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\partial \gamma_m(\theta_0)}{\partial \theta} &= \frac{\int_{\mathbf{x}} \int_{y_1,y_2} (y_2 - y_1) S(y_2, y_1 \mid m, 1, m, \mathbf{x}) f(y_2, y_1 \mid m, 1, m, \mathbf{x}) \pi_{2m}(m, 1, \mathbf{x}) f(\mathbf{x} \mid m) dy_1 dy_2 d\mathbf{x}}{\overline{\pi}_{2m}(1, m)} \\ &- \frac{\int_{\mathbf{x}} \int_{y_1,y_2} (y_2 - y_1) S(y_2, y_1 \mid m, 0, m, \mathbf{x}) f(y_2, y_1 \mid m, 0, m, \mathbf{x}) \pi_{2m}(m, 1, \mathbf{x}) f(\mathbf{x} \mid m) dy_1 dy_2 d\mathbf{x}}{\overline{\pi}_{2m}(1, m)} \\ &+ \frac{\int_{\mathbf{x}} \left(\mu_{1m}(m, 1, m, \mathbf{x}) - \mu_{0m}(m, 0, m, \mathbf{x}) - \gamma_m\right) \dot{\pi}_2(m \mid 1, m, \mathbf{x}) f(\mathbf{x} \mid m) d\mathbf{x}}{\overline{\pi}_{2m}(1, m)} \\ &+ \frac{\int_{\mathbf{x}} \left(\mu_{1m}(m, 1, m, \mathbf{x}) - \mu_{0m}(m, 0, m, \mathbf{x}) - \gamma_m\right) \pi_{2m}(1, m, \mathbf{x}) S(\mathbf{x} \mid m) f(\mathbf{x} \mid m) d\mathbf{x}}{\overline{\pi}_{2m}(1, m)}\end{aligned}$$

To verify that it is in $\mathcal{H}$, we can rewrite $\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta})$ as

$$\begin{aligned}\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}) &= \left(\frac{W_{i1m} D_i W_{i2m}}{\pi_{1m} \pi_d \overline{\pi}_{2m}(1, m)}\right) (\Delta Y_i - \mu_{i,1m}) - \left(\frac{W_{i1m}(1 - D_i) W_{i2m}}{\pi_{1m}(1 - \pi_d) \overline{\pi}_{2m}(1, m)}\right) \left(\frac{\pi_{2m}(1, m, \mathbf{X}_i)}{\pi_{2m}(0, m, \mathbf{X}_i)}\right) (\Delta Y_i - \mu_{i,0m}) \\ &+ \frac{W_{i1m} D_i}{\pi_{1m} \pi_d \overline{\pi}_{2m}(1, m)} \left(W_{i2m} - \pi_{i,2m}\right) \left(\mu_{i,1m} - \mu_{i,0m} - \gamma_m\right) + \frac{W_{i1m} D_i}{\pi_{1m} \pi_d \overline{\pi}_{2m}(1, m)} \widehat{\pi}_{i2} \left(\mu_{i,1m} - \mu_{i,0m} - \gamma_m\right).\end{aligned}$$

From there, it is straightforward to verify that

$$\frac{\partial \gamma_m(\theta_0)}{\partial \theta} = \mathbb{E}\left[\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_0) S(\mathbf{O}_i; \theta_0)\right].$$

Thus it is the efficient influence function for γ_m and the semiparametric efficiency bound is $\mathbb{E}[\widetilde{\phi}_m(\mathbf{O}_i; \boldsymbol{\eta}_0)^2]$.

□

The following lemma is from the Supplemental Materials for Kennedy, Balakrishnan and G'Sell (2020) and follows from an application of Chebyshev's inequality.

Lemma SM.1. *Let $\widehat{f}(\mathbf{o})$ be a function estimated from a sample $\mathbf{O}_{-b} = \{\mathbf{O}_i : B_i \neq b\}$ and let $\mathbb{P}_n^b$ be the empirical measure over $\mathbf{O}_b = \{\mathbf{O}_i : B_i = b\}$, which is independent of $\mathbf{O}_{-b}$. Then,*

$$(\mathbb{P}_n^b - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\left\|\widehat{f} - f\right\|}{\sqrt{n}}\right)$$

Here we describe the regularity conditions that are required to prove Theorem 4.

Assumption 4 (Regularity conditions)**.** *We assume that (a) $\mathbb{P}[\epsilon_1 \leq \widehat{\pi}_{1m} \leq 1 - \epsilon_1] = 1$, $\mathbb{P}[\epsilon_d \leq \widehat{\pi}_d \leq 1 - \epsilon_d] = 1$, and $\mathbb{P}[\epsilon_2 \leq \widehat{\pi}_{i,2m} \leq 1 - \epsilon_2] = 1$ for some values of $\epsilon_1, \epsilon_d, \epsilon_2 > 0$; (b) $\|Y_{it}\|_q \leq C_y$, $\left\|\mu_{i,dm}\right\|_q \leq C_\mu$, and $\left\|\nu_{i,dm}\right\|_q \leq C_\nu$ for some fixed strictly positive constants C_y, C_μ, C_ν and $q > 2$.*

Proof of Theorem 4. Let $\psi_{im} = \psi_m(\mathbf{O}_i; \boldsymbol{\eta}_0)$ and $\widehat{\psi}_{im,-b} = \psi_m(\mathbf{O}_i; \widehat{\boldsymbol{\eta}}_{-b})$. We can write $\widehat{\tau}_m$ and τ_m as

$$\widehat{\tau}_m = \sum_{b=1}^{K} \mathbb{P}_n \left(\widehat{\psi}_{im,-b} \mathbb{1}(B_i = b) \right) \quad \text{and} \quad \tau_m = \mathbb{E}[\psi_{im}] = \sum_{b=1}^{K} \mathbb{P} \left\{ \psi_{im} \mathbb{1}(B_i = b) \right\}.$$

Thus, we can write the estimation error as

$$\widehat{\tau}_m - \tau_m = \underbrace{(\mathbb{P}_n - \mathbb{P})\, \psi_{im}}_{\text{(I)}} + \sum_{b=1}^{K} \left[\underbrace{(\mathbb{P}_n - \mathbb{P}) \left\{ (\widehat{\psi}_{im,-b} - \psi_{im}) \mathbb{1}(B_i = b) \right\}}_{\text{(II)}} + \underbrace{\mathbb{P} \left\{ (\widehat{\psi}_{im,-b} - \psi_{im}) \mathbb{1}(B_i = b) \right\}}_{\text{(III)}} \right] \tag{3}$$

We take each of these in turn. First, (I) is the average of n iid mean-zero random variables with finite variance, so can employ the central limit theorem to establish that it will converge in distribution to $N(0, \mathbb{V}[\psi_{im}])$. Note that $\mathbb{V}[\psi_{im}] = \mathbb{E}[\widetilde{\psi}_m(\mathbf{O}_i; \boldsymbol{\eta}_0)^2]$.

For (II), we write $\pi_{D_i} = D_i \pi_d + (1 - D_i)(1 - \pi_d)$ with $\mu_{i,D_i m}$ and $\nu_{i,D_i m}$ defined similarly. Then we can write

$$\begin{aligned}
&\left\| \left(\widehat{\psi}_{im,-b} - \psi_{im} \right) \mathbb{1}(B_i = b) \right\| \leq \left\| \widehat{\psi}_{im,-b} - \psi_{im} \right\| \lesssim \left\| \widehat{\psi}_{im} - \psi_{im} \right\| \\
&= \Bigg\| \frac{W_{i1m}(2D_i - 1) W_{i2m} \left(\Delta Y_i - \mu_{i,D_i m} \right)}{\pi_{1m} \pi_{D_i} \pi_{i2m} \widehat{\pi}_{1m} \widehat{\pi}_{D_i} \widehat{\pi}_{i2m}} \left\{ (\pi_{1m} - \widehat{\pi}_{1m}) \pi_{D_i} \pi_{i2m} + \widehat{\pi}_{1m} \left(\pi_{D_i} - \widehat{\pi}_{D_i} \right) \pi_{i2m} + \widehat{\pi}_{1m} \widehat{\pi}_{D_i} \left(\widehat{\pi}_{i2m} - \pi_{i2m} \right) \right\} \\
&\quad + \frac{W_{i1m}(2D_i - 1) W_{i2m}}{\widehat{\pi}_{1m} \widehat{\pi}_{D_i} \widehat{\pi}_{i2m}} \left(\mu_{i,D_i m} - \widehat{\mu}_{i,D_i m} \right) + \frac{W_{i1m}(2D_i - 1)}{\widehat{\pi}_{1m} \widehat{\pi}_{D_i}} \left\{ (\widehat{\mu}_{i,D_i m} - \mu_{i,D_i m}) - (\widehat{\nu}_{i,D_i m} - \nu_{i,D_i m}) \right\} \\
&\quad + \frac{W_{i1m}(2D_i - 1) \left(\mu_{i,D_i m} - \nu_{i,D_i m} \right)}{\widehat{\pi}_{1m} \widehat{\pi}_{D_i} \pi_{1m} \pi_{D_i}} + \frac{W_{i1m}}{\widehat{\pi}_{1m}} \left\{ \left(\widehat{\nu}_{i,1m} - \nu_{i,1m} \right) - \left(\widehat{\nu}_{i,0m} - \nu_{i,0m} \right) \right\} \\
&\quad + \frac{W_{i1m}(\nu_{i,1m} - \nu_{i,0m})}{\widehat{\pi}_{1m} \pi_{1m}} \left(\pi_{1m} - \widehat{\pi}_{1m} \right) \Bigg\| \\
&\lesssim \|\widehat{\pi}_{1m} - \pi_{1m}\| + \|\widehat{\pi}_d - \pi_d\| + \|\widehat{\pi}_{i2m} - \pi_{i2m}\| + \max_d \left\| \widehat{\mu}_{i,dm} - \mu_{i,dm} \right\| + \max_d \left\| \widehat{\nu}_{i,dm} - \nu_{i,dm} \right\| = o(1),
\end{aligned}$$

where the last result on the first line follows because K is fixed so $N \lesssim N/K$. The last line follows from the triangle inequality, the fact that the propensity scores (and their estimates) are bounded away from zero (per Assumption 4), and combination of the bounded moment conditions from Assumption 4 and Hölder's inequality. Here we have also used the fact that the estimated and true propensity scores are bounded away from zero. By Lemma SM.1, the sum involving (II) must be $o_{\mathbb{P}}(1/\sqrt{N})$.

For (III), we similarly have

$$
\begin{aligned}
&|\mathbb{P}\left\{(\widehat{\psi}_{im,-b} - \psi_{im})\mathbb{1}(B_i = b)\right\}| \lesssim |\mathbb{P}(\widehat{\psi}_{im,-b} - \psi_{im})| \\
&= \left|\mathbb{P}\left\{\frac{W_{i1m}D_i}{\widehat{\pi}_{1m}\widehat{\pi_d}\widehat{\pi}_{i,2m}}\left(\widehat{\mu}_{i,1m} - \mu_{i,1m}\right)\left(\widehat{\pi}_{i,2m} - \pi_{i,2m}\right) - \frac{W_{i1m}(1-D_i)}{\widehat{\pi}_{1m}(1-\widehat{\pi_d})\widehat{\pi}_{i,2m}}\left(\widehat{\mu}_{i,0m} - \mu_{i,0m}\right)\left(\widehat{\pi}_{i,2m} - \pi_{i,2m}\right)\right.\right. \\
&\qquad \left.\left. + \frac{W_{i1m}}{\widehat{\pi}_{1m}\widehat{\pi}_d}\left(\widehat{\nu}_{i,1m} - \nu_{i,1m}\right)\left(\widehat{\pi}_d - \pi_d\right) - \frac{W_{i1m}}{\widehat{\pi}_{1m}(1-\widehat{\pi}_d)}\left(\widehat{\nu}_{i,0m} - \nu_{i,0m}\right)\left(\widehat{\pi}_d - \pi_d\right)\right\}\right| \\
&\lesssim \left\|\widehat{\pi}_{i,2m} - \pi_{i,2m}\right\|\left(\max_d \left\|\widehat{\mu}_{i,dm} - \mu_{i,dm}\right\|\right) + \|\widehat{\pi}_d - \pi_d\|\left(\max_d \left\|\widehat{\nu}_{i,dm} - \nu_{i,dm}\right\|\right)
\end{aligned}
$$

Given that $\|\widehat{\pi}_d - \pi_d\| = O_{\mathbb{P}}(1/\sqrt{n})$, then second term is $o_{\mathbb{P}}(1/\sqrt{N})$ so long as $\max_d \left\|\widehat{\nu}_{i,dm} - \nu_{i,dm}\right\| = o_{\mathbb{P}}(1)$, which holds by assumption. Furthermore, by assumption the first term is $o_{\mathbb{P}}(1/\sqrt{N})$, so the sum involving (III) is also $o_{\mathbb{P}}(1/\sqrt{n})$. Thus, we have,

$$\sqrt{N}(\widehat{\tau}_m - \tau_m) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\widetilde{\psi}_{im} + o_{\mathbb{P}}(1),$$

and combined with the CLT results about (I), the desired result obtains. □

C Additional Tables for Empirical Application

Table SM.1: List of covariates

Pre-treatment covariates	Post-treatment covariates
Trans law support	Obama feeling thermometer (Δ)
Registered Democrat	Trans tolerance (Δ)
Political ideology	Gender norms (Δ)
Religiousity	Trans law support (Δ)
Knows trans people	
Female	
Hispanic	
Af.-Am.	
Age	
Survey in Spanish	
Transgender tolerance	
Gender norms	
Obama feeling thermometer	