Continuous Treatment Difference-in-Differences with Unknown Controls: A Data-Driven Approach

Elird Haxhiu [*1] and Thomas Helgerman [*2]

[1]Department of Economics, University of Michigan
[2]Carlson School of Management, University of Minnesota

June 1, 2026

Abstract

This paper studies difference-in-differences (DiD) research designs where all observations receive a continuous treatment (or dose) in response to an aggregate policy, so there is no group that is *ex post* unexposed. We evaluate the common practice of dichotomizing the dose space into "treated" and "control" groups using an empirical percentile, then running a standard DiD design. We interpret this design as attempting to estimate a Binned Average Treatment Effect on the Treated (BATT), where the treatment group is specified by the researcher. We develop a framework to estimate the BATT when the dose takes effect only after a cutoff value, the Minimum Effective Dose (MED), and propose a first stage model selection procedure to determine a suitable control group. To conduct valid inference after model selection, we use a bootstrap aggregated estimator of the BATT, where a confidence interval is generated by a second-order bootstrap of this procedure, following Efron (2014). To conclude, we provide simulation evidence that suggests that this procedure delivers marked improvements in mean-squared error relative to common practice.

JEL codes: C14, C23, C24.
Key words: Difference-in-Differences, Parallel Trends, Threshold, Dose Response Function.

[*]Contact: haxhiu@umich.edu, tehelg@umn. We thank John Bound, Melvin Stephens, Florian Gunsilius, Andrew Goodman-Bacon, Emir Murathanoglu, Dean Yang, Toni Whited, Charles Brown, Luis Espinoza, and participants at the Michigan Labor Lunch seminar and the CSOM Summer Econ Seminar for their helpful discussions and suggestions. We thank Daniel Boutain for excellent research assistance.

1 Introduction

Without additional assumptions, it is generally impossible to infer the effects of a treatment by comparing participants to non-participants or comparing the participants over time. For valid inference, unit comparisons must restrict selection into treatment, which is generally difficult to justify in the absence of an instrument. On the other hand, time comparisons (e.g. interrupted time series methods) cannot allow for contemporaneous trends. Difference-in-differences (DiD) research designs combine these two estimators to infer causal effects by subtracting the change in outcomes over time for non-participants from the change for participants. This is valid whenever the average change in outcomes for participants in the absence of treatment (a counterfactual moment) is equal to the average change in outcomes for those who go untreated, known as a Parallel Trends assumption (PTA). Panel data thus enables identification of a causal effect without ruling out selection into treatment levels or contemporaneous trends.

A common extension in practice admits a continuously distributed treatment (or dose) variable, where all units are exposed to some level of the dose $D_i \geq 0$. One motivation behind these research designs is an aggregate policy where the researcher hypothesizes that units face heterogeneous exposure according to some dose variable. A prominent example is Card (1992) which measures a state's exposure to a higher federal minimum wage by the share of teenage workers who fall below the new legislated minimum. The estimating equation follows via analogy to binary DiD

$$Y_{i,t} = \alpha_i + \theta_t + \beta \cdot D_i P_t + \varepsilon_{i,t} \tag{1}$$

where $Y_{i,t}$ is some outcome of interest, α_i and θ_t are unit and time fixed-effects respectively, P_t indicates the periods that units are exposed, and ε_{it} is some error term. The main coefficient of interest in the Two-Way Fixed-Effects (TWFE) regression above is β, which represents an aggregation of the "effect" of the continuous treatment. Researchers typically interpret this as the weighted average dose response function (DRF), in line with the "average derivative" interpretation of regression coefficients (Yitzhaki, 1996; Angrist and Pischke, 2009).

Recently Callaway et al. (2024), hereafter referred to as CGS, decompose β and derive sufficient assumptions to identify well-defined causal parameters. Whenever researchers have access to "pure control" units, or observations unexposed to the treatment with dose $D_i = 0$, only minor modifications to the parallel trends assumption allows identification of the Average Treatment Effect on the Treated (ATT) parameter, defined for each positive dose level. But what if researchers

do not have access to a pure control group? We focus on these cases and study inference in the absence of well-defined controls.

To understand the data constraints that researchers face, we conduct a metastudy of all papers published in the *American Economic Review* (*AER*) between 2000 and 2018. Out of the 44 papers estimating a continuous treatment DiD model, 31 estimate a full dose regression like (1). Without pure control units, this estimator relies exclusively on comparisons across dose levels, aggregating them into an estimate of the overall dose response. However, the standard parallel trends assumption does not identify the ATT nor the average effect of marginal increases in the dose (Average Causal Response). This is because comparisons across treated units at different doses must additionally restrict selection into treatment levels, which is not typical in standard DiD. In the case of a continuous dose, the least squares estimator of β converges to a weighted average of differences between adjacent ATT estimates across the dose distribution, which only reveals a causal response under restrictions on counterfactual treatment effect heterogeneity (strong parallel trends).

If we are not willing to make strong assumptions, what is left to do? The remaining 13 papers dichotomize the dose variable (at some researcher-specified cutoff value) and estimate a traditional difference-in-differences model. This approach bins all "intensely" exposed units together and compares them with everyone else; the hope is that even if some control units receive a small dose, as long as the effect is monotonic this will reveal an attenuated version of the average treatment effect for units classified as treated. We show that this estimator does not escape the problems inherent in comparing treated units at different doses and, without stronger assumptions, estimates a sum of treatment and selection effects. At its core, this paper is concerned with identifying a set of plausible and testable assumptions that allow for inference using an estimator of this type.

Inference in these settings is particularly difficult, beyond the usual Fundamental Problem of Causal Inference (Imbens and Rubin, 2015) that treated and untreated outcomes cannot be observed for the same unit at the same time. In this case, in every period we observe *either* a treated or an untreated outcome for *all* units, so that contemporaneous treatment-control comparisons cannot be made. This issue is well-known in macroeconomics (Rambachan and Shephard, 2019), which exploits the timing of exogenous shocks for identification. Instead, we focus on the use of contemporaneous comparisons, which remains the dominant paradigm in applied microeconometrics.

We begin by constructing a potential outcomes framework for applications where all units

are treated at a certain time period and consider a counterfactual where all units are instead untreated, which allows us to formalize precise definitions of treatment effect parameters of interest. In particular, we introduce the *Binned Average Treatment Effect on the Treated* (BATT) parameter, which gives the ATT averaged across all units receiving a researcher-specified dose or higher. We view this as the most natural estimand that a dichotomized design (discussed above) would be attempting to estimate. To estimate the BATT, the treatment group is obvious, as it is specified by the researcher, but it is unclear which units will make suitable controls.

The central assumption in our approach is that certain groups experience outcomes identical to a world in which the policy is not passed. To be precise, borrowing from the pharmacological literature, we assume that there exists a "Minimum Effective Dose" (MED), where units with a dose below the MED experience an outcome indistinguishable from one in which they were untreated (Ruberg (1989) defines this formally). As a result, any unit receiving a dose below the MED can be used as a valid control in a standard difference-in-differences design.

If we knew *ex ante* what the MED was, a design in which the researcher used only units receiving this dose or lower as controls would allow for straightforward estimation. In practice, however, this is generally unknown, and we propose a two-step estimator to estimate the BATT when the MED exists. The first step is a model selection step. Under the assumption that the MED is at least as high as the lowest dose, we can write down an expression for the mean-squared error (MSE) of any estimator without knowledge of the value of the MED. Using this expression, we propose choosing the set of control units that minimize the empirical MSE. This procedure encapsulates an intuitive tradeoff: when we are unsure which units are actually untreated, adding another set of units as controls could potentially raise MSE by increasing bias if these units are actually treated, but even if this case, the resulting drop in variance could lead to lower MSE. In the second step, we utilize this set of controls to estimate the BATT with a standard difference-in-differences estimator.

Our estimator can be thought of as a data-adaptive procedure that is "sandwiched" between two common estimators. Consider a researcher who would like to estimate the BATT for all units with treatment above the median. One option is to use all of the units below the median as controls - this corresponds to the dichotomized estimator that is common in the literature. Another is to only use the lowest dose (which might be 0) as controls - this corresponds to the estimator proposed by Callaway et al. (2024). We introduce a model selection step that will include anywhere from just the lowest dose to all possible controls, depending on the estimated impact on MSE of

each choice. Importantly, this procedure is consistent in the case where only the lowest dose is a suitable control, choosing only this dose in the limit. Our primary application of interest is a setting with a discrete number of doses with many observations at each dose, though we briefly consider the case of a continuous dose space.

1.1 Connections to Literature

This paper contributes to three literatures. Recent technical advances to difference-in-differences dealing with staggered adoption (Goodman-Bacon, 2021; Sun and Abraham, 2021; Wooldridge, 2021; de Chaisemartin and D'Haultfoeuille, 2020), heterogeneous treatment effects (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021), pre-testing (Roth, 2022), and functional form specification (Roth and Sant'Anna, 2021) are increasingly well understood. However, identification issues related to continuous treatments are actively being litigated among econometricians and applied researchers (de Chaisemartin et al., 2026; Callaway et al., 2024; de Chaisemartin et al., 2023; Sun and Shapiro, 2022; Butts, 2022). Our contribution is to show that researchers can still rely on traditional parallel trends assumptions with continuous treatments lacking pure controls, and to clarify the assumptions this type of inference requires. This is not a free lunch as we need to assume the MED exists, which is not always guaranteed in practice.

The approach in Butts (2022) to estimate treatment effects at specific locations with geo-coded data is most similar to ours. In studying treatment explicitly as continuous distance, he notes that researchers must know the threshold distance beyond (or below) which treatment effects begin, and proposes a non-parametric method to estimate the treatment effect curve (what we call the dose response function) with large data. Similarly, existing approaches to identifying treatment effects in a difference-in-differences setting rely on additional variation not present in our set-up. de Chaisemartin et al. (2026) take a related but markedly different approach to leverage continuity of the expected change in potential outcomes as the dose approaches 0. Our framework is inspired in large part by Callaway et al. (2024), but they assume that *ex ante* untreated units are available and observed by the researcher. de Chaisemartin et al. (2023) identify causal responses with additional variation in the treatment variable: If the dose changes over time and there exists a group of stayers or a group of quasi-stayers, then the average derivative among the treated is identified. Sun and Shapiro (2022) provide impossibility results on identifying causal responses in TWFE regressions absent such additional information. They show that with the existence of a pure control, a modified instrumented difference-in-differences approach is sufficient to target

an average of causal responses among treated units. We contribute to this work by providing a consistent framework to estimate a new causal parameter, the BATT, without pure controls, a case that is common in research designs estimating continuous treatment DD.

Finally, we contribute to a literature spanning several disciplines which highlights the problems inherent to dichotomizing a continuous variable. The motivation for this specification varies by application: the psychology literature dichotomizes either one or two continuous variables to use a more familiar one-way or two-way ANOVA design (MacCallum et al., 2002); the clinical literature targets a threshold for biomarkers that maximizes the predictive use of flagging patients who fall above this threshold (Altman et al., 1994); and the epidemiology literature categorizes continuous controls into several groups to model a more flexible linear specification (Brenner, 1997). For example, number of cigarettes smoked per week might be dichotomized into "non-smoker," "light smoker" and "heavy smoker."

Some of this work may not be relevant to economists who often explicitly model continuous variation in data. In fact, well-known results on measurement error would immediately identify this as problematic (see Bound et al. (2001) for a summary). However, we suspect that dichotomization persists due to the lure of a difference-in-differences research design, in part because of the ability to falsify its main identifying assumption (parallel trends) with a visual and statistical pre-trends test given the availability of at least one additional period before treatment. We show that similar problems arise as in earlier work, which warrants a clear statement of necessary assumptions that must hold to identify meaningful estimands. We aim to do this in the remainder of the paper.

Section 2 describes our potential outcomes framework, formalizing the baseline assumptions needed to identify and estimate the BATT. Additionally, we use the framework to characterize the state of current practice where a continuous variable is dichotomized to allow for a standard DiD design and provide a simple test for researcher manipulation of this threshold. In Section 3, we introduce our core MED assumption and consider inference when this is known. Section 4 describes our estimator when the MED is unknown, and Section 5 provide simulation evidence for its performance relative to other estimators. Section 6 concludes.

2 Framework

2.1 Setup

Following Callaway et al. (2024), we consider an environment with two time periods $t \in \{\tau-1, \tau\}$ and N units $i \in \{1, ..., N\}$, where some outcome $Y_{i,t}$ is observed. Units are assigned a treatment dose D_i, which is also observed by the researcher. Data are independent and identically distributed across units.

A1 **Random sampling:** $\{Y_{i,\tau}, Y_{i,\tau-1}, D_i\}_{i=1}^N$ is independent and identically distributed (iid)

The dose distribution is given by the cumulative distribution function $F_D(d)$ with bounded and compact support on $\mathbb{R}_{++}$ and a well-defined probability distribution function. Formally,

A2 **Dose distribution:** $D_i \sim F_D(d)$ over compact supp$\{D_i\} := \mathcal{D} \subset \mathbb{R}_+$, which admits a Radon-Nikodym derivative $f(d)$ such that $f(d) > 0 \, \forall d \in \mathcal{D}$. Let $d_l = \inf \mathcal{D}$ and $d_u = \sup \mathcal{D}$.

In most applications, the support $\mathcal{D}$ is either discrete or continuous over a closed interval. To facilitate discussion, we clarify assumptions for each of these cases

A2.1 **Discrete dose distribution:** The support of the dose distribution is given by $\mathcal{D} = \{d_1, ..., d_J\}$ for some finite J where $0 < d_1 < d_2 < ... < d_J$. The CDF $F_D(d)$ admits a probability mass function $f(d)$ such that $f(d) > 0 \, \forall d \in \mathcal{D}$.

A2.2 **Continuous dose distribution:** The support of the dose distribution is given by $\mathcal{D} = [d_l, d_u]$ for some $0 < d_l < d_u$. The CDF $F_D(d)$ admits a probability density function $f(d)$ such that $f(d) > 0 \, \forall d \in \mathcal{D}$.

In Section 4, when we introduce our core estimation strategy, we assume a finite number of doses as in A2.1, which is the empirical setting it is designed for. In Section 4.3, we briefly consider how this design would extend to a continuous dose distribution as in A2.2. All results in the current section apply to both and so we only specify A2.

2.2 Potential Outcomes

In each period, let $Y_{i,t}(D_i, L_t)$ denote potential outcomes, which take two arguments. The first argument is the dose assigned to each unit. In addition, we assume there is some policy of interest to the researcher that is implemented at time τ. The second argument, L_t, is an indicator for the

policy of interest being implemented in time τ. In applications of interest, $L_{\tau-1} = 0$ and $L_\tau = 1$, which we formalize by writing observed outcomes as a function of potential outcomes:

A3 **Observed Outcomes:** In period $\tau - 1$, observed outcomes are given by $Y_{i,\tau-1} = Y_{i,\tau-1}(D_i, 0)$. In period τ, observed outcomes are given by $Y_{i,\tau} = Y_{i,\tau}(D_i, 1)$.

Even though its value is fixed, we introduce L_t to define a counterfactual where the policy was not implemented ($L_\tau = 0$) for all units in period τ. To define outcomes in period $\tau - 1$, Callaway et al. (2024) assume any unit that receives a positive dose experiences an outcome equivalent to its untreated outcome. In our setting, no units receive a 0 dose, but are unaffected by their dose value since the policy is not in place until period τ. Introducing L_t allows us to formalize this intuition.

2.3 Target Causal Estimands

With $P_t = 1\{t = \tau\}$ indicating the time when the policy is implemented, observed outcomes are

$$Y_{i,t} = (1 - P_t)Y_{i,t}(D_i, 0) + P_t Y_{i,t}(D_i, 1) \tag{2}$$

The policy indicator L_t enables us to consider a useful thought experiment, which we use to define our target estimands. Our building block is an *individual treatment effect* (ITE), or the difference between the observed (treated) outcome and a counterfactual one in a world where the policy was not implemented

$$\mu_i(D_i) \equiv Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0) \qquad \text{ITE}$$

By considering how this changes as the dose is varied we trace out the dose response function (DRF), which might vary across units.

We target a variety of averages of these parameters. Following Callaway et al. (2024), the *average treatment effect at dose* d (ATT$(\cdot|d)$) is given by

$$\mathbb{E}[\mu_i(a)|D_i = d] = \mathbb{E}[Y_{i,\tau}(a, 1) - Y_{i,\tau}(a, 0) \mid D_i = d] \qquad \text{ATT}(a|d)$$

Intuitively, the individual treatment effect of receiving dose a is averaged over all units receiving the dose d. Note that this is defined as a function of a for every dose level d. Of particular interest is the average effect at the actual dose received, ATT$(d|d)$. This will be the fundamental building

block of every causal parameter we target in this paper.

It will be useful here to consider several different ways in which the ATT$(d|d)$ can be aggregated to generate estimands that might be of interest to the practitioner. Letting $T_i = 1$ denote all units that receive treatment allows us to define the *average treatment effect on the treated* (ATT):

$$\mathbb{E}[\mu_i(D_i)|T_i = 1] = \mathbb{E}[Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0)|T_i = 1] \qquad \text{ATT}$$

Note that even if all units are treated, this is still well-defined, but inference becomes more complicated. Without a comparison group, estimation of the ATT must necessarily arise from extrapolating an estimated counterfactual in the pre-period and comparing this to the observed trend.[1]

In light of these difficulties, practitioners have generally opted to leverage differences in intensity of treatment instead of treatment status. In a dose response setting, as Callaway et al. (2024) point out, these designs estimate a different causal parameter, the average causal response. This parameter is not entirely appropriate in our setting, though, since in a binned designed, a large portion of dose variation is not utilized. This suggests an aggregation across several dose values, which we call the *Binned Average Treatment Effect on the Treated* (BATT)

$$\text{BATT}[\mathcal{D}_r] = \mathbb{E}[Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0)|D_i \in \mathcal{D}_r]$$

where $\mathcal{D}_r$ is a set of dose values specified by the researcher. Many estimators used in practice can be thought of as targeting a particular BATT. Consider, for example, a difference-in-differences estimator binarized at the median. The group considered "treated" in this design is given by all $D_i \in [\text{median}(D_i), \infty]$. If all units below the median were untreated, this design would recover BATT$[\text{median}(D_i), \infty]$. It is, of course, a strong assumption that all units below a researcher-specified cutoff are untreated, and we consider how to choose a set of comparison units in a more rigorous manner in Section 4.

2.4 Current Practice

We use the framework described above to characterize current practice, focusing on specifications which dichotomize the continuous treatment at some researcher-specified threshold. One reason

[1]Perhaps the simplest implementation of this idea is the interrupted time series (ITS), though more sophisticated approaches exist as well (Botosaru et al., 2024). However, ITS methods are not widely used in microeconometrics, perhaps because they depend on estimating a secular trend in the pre-period (Turner et al., 2021), violating the core assumption of a TWFE design with an unrestricted time fixed effect.

for this approach is to identify causal effects under minimal assumptions on the data-generating process, which we also pursue by eschewing parametric assumptions.

2.4.1 Metastudy

To get a sense of the use and justification of a binned approach, we conducted a small metastudy. Using Google Scholar, we query all papers published in the *American Economic Review* between 2000 and 2018 which contain the keywords "difference-in-difference" and "continuous" in the manuscript. We find 178 total papers that satisfy these requirements and after checking each one, retain only those that estimate a continuous treatment difference-in-difference model. Of these 44 papers, 31 estimate a full dose regression like (1). The remaining 13 dichotomize the dose at some value or percentile, comparing units above and below the researcher-defined cutoff. In our framework, this is equivalent to the bivariate regression

$$\Delta Y_{i,t} = \alpha + \beta \cdot \mathbb{1}(D_i \geq d_r) + \varepsilon_{i,t} \tag{3}$$

Researchers choose some threshold d_r and use more exposed units ($d_i \geq d_r$) as a "treatment" group to compare to less exposed units ($d_i < d_r$) in a "control" group, to recover a binary DiD structure. Our metastudy suggests that one prominent motivation to dichotomize is to recover a traditional difference-in-differences setup that relies on comparisons between treated and control units. In the "2" $\times$ 2 setting here, regression recovers a classic difference-in-differences of means

$$\hat{\beta} = \hat{\mathbb{E}}[\Delta Y_{it} | D_i \geq d_r] - \hat{\mathbb{E}}[\Delta Y_{it} | D_i < d_r] \tag{4}$$

Where $\hat{\mathbb{E}}$ denotes the sample average equivalent of the population expectation. While this design has intuitive appeal, it is unclear what the causal estimand is in this approach, let alone if the binned estimator is unbiased or consistent.

2.4.2 What Parameter Does TWFE Recover?

Consistent with this design mimicking a standard binary difference-in-differences design, practitioners often invoke a standard parallel trends assumption. We introduce this assumption formally in our setting:

A4 **Parallel Trends:** $\mathbb{E}[\triangle Y_{i\tau}(D_i, 0) | D_i = d] = \mathbb{E}[\triangle Y_{i\tau}(D_i, 0) | D_i = d'] \; \forall d, d' \in \mathcal{D}$

In a DiD setting with an untreated group, this assumption would state that the trend of any treated group in the absence of treatment would equal that of the control group in expectation. In our setting, since there is no zero dose, we assume that if no policy was passed and no units were treated, units at each dose level would have the same trend in outcome between period $\tau - 1$ and τ in expectation. In a setting with a zero dose group, this assumption is satisfied by requiring that units at each dose have parallel trends to the untreated group.

Now we can consider what the estimator in (4) will deliver under assumptions A1-A4. Intuitively, since the "control" group in this setting receives a treatment, we do not recover the ATT, but rather the difference in average treatment effects for both groups. More precisely, we are averaging the dose-specific ATT$(d|d)$ estimands, weighted by the observed dose distribution:

Proposition 1. *If assumptions A1-A4 hold, the binned difference-in-differences estimator recovers*

$$\int_{d_r}^{d_u} \text{ATT}(d|d) \frac{f(l)}{1 - F(d_r)} \text{d}l - \int_{d_l}^{d_r} \text{ATT}(d|d) \frac{f(l)}{1 - F(d_r)} \text{d}l$$

Note that, even though untreated outcomes are not observed in period τ, we arrive at an expression involving causal estimands as untreated potential outcomes are differenced out.

There is no clear relationship between the term in Proposition 1 and the ATT. Using the law of iterated expectations, we can write the ATT as

$$\text{ATT} = \int_{d_l}^{d_u} \text{ATT}(d|d) \frac{f(l)}{1 - F(d_r)} \text{d}l$$

In fact, they are in tension; while the binned estimator recovers the difference of averaged ATT$(d|d)$'s for the "treated" and "control" group, the ATT is the weighted sum of these objects:

$$\text{ATT} = (1 - F(d_r)) \int_{d_r}^{d_u} \text{ATT}(d|d) \frac{f(l)}{1 - F(d_r)} \text{d}l + F(d_r) \int_{d_l}^{d_r} \text{ATT}(d|d) \frac{f(l)}{1 - F(d_r)} \text{d}l$$

Perhaps the clearest way this can cause problems is if the dose response function is not monotonic, in which case it can both be true that ATT > 0 but the binned estimator is negative. To avoid this outcome, we would need to impose the further assumption that ATT$(d|d)$ is monotonic - This would ensure that $\mathbb{E}[\hat{\beta}]$ has the same sign as the ATT, but the bias from such an estimator would still have unknown sign.

In general, it is not possible to "rescue" this approach by making parametric assumptions. To see this, consider the case of a homogenous, linear dose response function, given by $\mu_i(d_i) = \beta d_i$,

with a dose distribution that is approximately normally distributed with mean μ and standard deviation σ.[2] The ATT in this case will simply be equal to $\beta\mu$. However, using well-known results from selection models, we show in Appendix Section A.8 that the binned estimator will yield

$$\beta\sigma \frac{\phi(\frac{d_r-\mu}{\sigma})}{\Phi(\frac{d_r-\mu}{\sigma})(1-\Phi(\frac{d_r-\mu}{\sigma}))}$$

Which is approximately equal to $1.5\beta\sigma$ when d_r is chosen to be the median. In this case, the binned estimator is a function of the dose distribution's standard deviation, while the ATT is a function of the dose distribution's mean, even under the restrictive assumption of a linear dose response.

Instead, we might want to make the weaker claim that, while we can't estimate the ATT, this above/below comparison still delivers an estimand of significance. Unfortunately, the expression in Proposition 1 does not have a causal interpretation without further assumptions. We can see this by decomposing the difference in Proposition 1 into a causal component and a selection component. Instead of comparison across doses as in a continuous design, we are comparing across distributions of doses, as illustrated in the following lemma:

Lemma 1. *If assumptions A1-A4 hold, the binned difference-in-differences estimator can be written as*

$$\int_{d_r}^{d_u} \mu(d|d)\frac{f(l)}{1-F(d_r)}dl - \int_{d_l}^{d_r} \mu(d|F^{-1}_{d>d_r}(F_{d<d_r}(l)))\frac{f(l)}{F(d_r)}dl$$
$$+ \int_{d_l}^{d_r} \{\mu(d|F^{-1}_{d>d_r}(F_{d<d_r}(l))) - \mu(d|d)\}\frac{f(l)}{F(d_r)}dl$$

The first line is a causal estimand, measuring the causal response resulting from moving from the dose distribution below d_r to the dose distribution above d_r. The second line is a selection term, representing the fact that units that receive different doses might not have the same dose response at dose d. Fundamentally, this lemma reinforces that by discretizing we canot avoid the issues inherent in dose comparisons emphasized in Callaway et al. (2024).[3]

2.4.3 Inflated Type 1 Error Rate

The analysis thus far implicitly assumed that the researcher cutoff d_r was chosen without regard to what the data look like. However, since there is not much consensus on how to choose a cutoff

[2] This is an approximation as even for a large, positive μ this distribution can take negative values, violating A2.

[3] The same decomposition would hold for any bijective map between $\{D_i|D_i < d_r\}$ and $\{D_i|D_i \geq d_r\}$, changing both the causal and selection terms. We choose the quantile transformation as it preserves dose rank.

for empirical analyses, this is largely left to researcher discretion. This is worrying as the strategic choice of where this cutoff is can lead to substantial inflation of the Type I error rate. This is well known in the clinical literature, where it was common practice to choose a cutoff for some biomarker by minimizing the p-value, arguing that such a procedure would lead to the most predictive cutoff point (Altman et al., 1994). While we are not aware of any work promoting this approach, concerns about p-hacking and publication bias in difference-in-differences (Brodeur et al., 2020) suggest that we should be cautious about the sensitivity of results to researcher choice of d_r.

It will come as no surprise that choosing a threshold "optimally" - that is, with the lowest p-value - will lead to an inflated Type I error rate. Rather, it is the degree of inflation that occurs that is of concern, which can be calculated (approximately) from a known asymptotic distribution. Recall the model in (3) where the researcher regresses an outcome $\triangle Y_i$ on a dichotomized dose variable. The coefficient estimate $\hat{\beta}$ converges to $\mathbb{E}[\triangle Y_i | D_i \geq d_r] - \mathbb{E}[\triangle Y_i | D_i < d_r]$ as this is simply a t-test of the difference in means between the "treatment" and "control" groups. Under the null of no difference in means between groups, let $T_n(d)$ denote the value of this t-statistic for sample size n and cutoff d. We consider the maximum of these statistics from some lower bound d_1 to some upper bound d_2, denoted by

$$\underset{d \in [d_1, d_2]}{\text{maximize}} \quad |T_n(d)| \tag{5}$$

Lausen and Schumacher (1992) show that this object converges to the supremum of the absolute value of a standardized Brownian bridge, given by

$$\sup_{t \in [\epsilon_1, \epsilon_2]} \frac{|B_0(t)|}{(t(1-t))^{1/2}} \tag{6}$$

Where $\epsilon_1 = F(d_1)$ and $\epsilon_2 = F(d_2)$. Miller and Siegmund (1982) provide the following asymptotic approximation for the Type I error of this distribution

$$\mathbb{P}\left[\sup_{t \in [\epsilon, 1-\epsilon]} \frac{|B_0(t)|}{(t(1-t))^{1/2}} \geq z\right] = \frac{4\phi(z)}{z} + \phi(z)\left(z - \frac{1}{z}\right) \ln\left(\frac{(1-\epsilon)^2}{\epsilon^2}\right) + o\left(\frac{\phi(z)}{z}\right) \tag{7}$$

Where we have simplified the set of cutoffs to search over to the symmetric range $[\epsilon, 1-\epsilon]$. Plugging in the cutoff z_α that would be used to define the rejection region of a test with Type I error of size α, we can calculate the actual Type 1 error probability would result. Table 1 reproduces part of Table 1 from Miller and Siegmund (1982) to give a sense of how these levels change.

Consider a test with the standard 5% Type I error rate. Searching between the 10$^{\text{th}}$ and 90$^{\text{th}}$

percentile would result in a realized Type I error rate of 49%, an inflation of 10 times the presumed level. Compressing the interval of search ameliorates the problem only slightly. Searching in the interquartile range results in an error rate of 31%, and even highly significant results (p = 0.01) would exhibit almost 10 times the assumed level of error.

2.4.4 Testing for Research Manipulation

To guard against researcher manipulation, when a discretized DiD design is presented, the addition of a simple plot would provide a check of the sensitivity of results to the researcher choice of d_r. For any given choice of d_r, write the p-value of the $\hat{\beta}$ as $p_n(d_r)$. By choosing a sufficiently fine grid over the dose space $\mathcal{D}$, we can generate an empirical estimate of the function $p_n(\cdot)$. We recommend researchers generate this plot with horizontal lines denoting typical levels of significance and a vertical line denoting the researcher's cutoff choice, as pictured in Figure 1.

These plots give a visual depiction of how sensitive the reported results are to the researcher choice of d_r. Figure 1a plots an example when the p-value varies markedly with d_r. Though results at the indicated threshold of 0.875 are statistically significant moving in either direction quickly produces a much higher p-value.[4] It is often tempting to conclude from results of this form that the dose only impacts the outcome for "highly treated" units, but it is important to be cautious about drawing this conclusion from a binarized design. To illustrate this, we run a simple simulation where a $\mathcal{N}(0,1)$ outcome variable is compared to a placebo dose distributed $\mathcal{U}[0,1]$.[5] In each simulation, we conduct a grid search for the minimum p-value and record which dose it is achieved at. Across 1,000 replications, we find that for 60% of simulations, this dose is either above the 90th percentile or below the 10th percentile of the dose distribution (that is, concentrated in the tails).[6] In contrast, Figure 1b plots an example where the p-value is robust to the choice of d_r, though this in and of itself is not evidence that the regression is properly specified.

In any case, these calculations suggest that it would be advantageous to "tie the hands" of the researcher to restrict their choice of d_r and limit the potential inflation of Type I error. However, a heuristic approach to this would also likely result in incorrect standard errors, as standard DiD

[4]The data generating process for this exercise involves regressing a noise term ($\mathcal{N}(0,1)$) on various placebo indicators, so in the true relationship $\beta = 0$.

[5]Since the dose distribution is binarized, these can be thought of as draws across the percentile distribution.

[6]Intuitively, this occurs because the regressor in equation (3) is a Bernoulli random variable with variance $F(d_r)(1-F(d_r))$. As d_r approaches the tails of the dose distribution, this goes to 0. The empirical variance appears in the denominator of $\hat{\beta}$ and the square root of the empirical variance appears in the denominator of se$(\hat{\beta})$; so, for fixed n, as d_r approaches the tails, it inflates $\hat{\beta}$ at a faster rate than se$(\hat{\beta})$, leading to a higher likelihood of a lower p-value.

estimates also do not account for uncertainty over what this cutoff is. We now turn to developing a principled way of conducting inference of this type.

3 Minimum Effective Dose (MED)

3.1 Choosing a Target Estimand and Treatment Group

As we discussed above, current practice involves researchers choosing a threshold d_r that defines a treatment and control group for a difference-in-differences estimator. This approach, though common, has many undesirable qualities. First, it is unclear what causal estimand this design is targeting. A standard difference-in-differences design would estimate the ATT, but it is unlikely that the true treated group is identified by a researcher-specified d_r. Further, the comparison group is defined as the complement of the treatment group. This has the advantage of using all available data, but leads to inconsistencies when treatment heterogeneity is considered. Sometimes, after estimating a design split at the median, a practitioner might instead set d_r as the 90th percentile of the dose distribution to estimate treatment effects for "highly treated" units. It does not seem consistent that units between the median and 90th percentile, designated as treated in the first design, are now used as control units in the second.

To avoid these issues, we propose separating these steps into separate decisions. To begin, the researcher should specify what the target causal estimand is. In general, this will usually be BATT$[\mathcal{D}]$, ATT$(d|d)$, or some combination of the two. Once this has been established, the researcher should choose an appropriate treatment and comparison group to construct an estimator. The causal estimands introduced in this paper all have a very natural choice of treatment group: any ATT$(d|d)$ should use all units receiving dose d as the treated group, while BATT$[\mathcal{D}]$ should use all units receiving between doses in set $\mathcal{D}$.[7] Choosing a comparison group is much more difficult. In our general set-up all units are treated, and even if certain units were not treated, we would need a way to identify them. We first outline the assumption needed for an appropriate comparison group to be available in the data, then turn to the task of choosing which of these units should be inluded.

[7] In the case of BATT[median$(d), \infty$], since the median is estimated, any estimator would need to account for uncertainty over the value of median(d). We do not consider such a design, restricting the researcher to choosing some known set of doses.

3.2 Existence Assumption

All empirical settings suffer from the fundamental problem of causal inference, as we are unable to observe treated and untreated observations in the same time period for any unit. In a standard difference-in-differences setting, assumptions on the progression of counterfactual untreated outcomes for a treated group identify causal effects by comparing outcomes for treated and untreated groups before and after a policy change. In our setting, inference is further constrained as these comparisons are not possible because *all* units are *either* treated or untreated in each time period.

One way to restore the standard DiD setting is to assume that untreated outcomes are observed for some units, even if all of them were treated. We operationalize this by assuming that a *Minimum Effective Dose* exists - intuitively, this restricts the individual treatment effect function $\mu_i(d_i)$ to equal zero for units receiving a dose below some threshold d_c. Formally,

A5 **Minimum Effective Dose (MED) exists:** $\exists\, d_c \in \mathcal{D}$ s.t. $\forall D_i < d_c$, $Y_{i\tau}(D_i, 1) = Y_{i\tau}(D_i, 0)$

Note that we do not assume units that are untreated exist. Rather, we assume that low dose units do not exhibit a treatment response so that their treated and untreated outcomes are identical at the dose they experienced. Under A5, we now have a suitable set of comparison units available to use to estimate our target estimand.

3.3 Institutional Knowledge of d_c

In certain applications, institutional knowledge implies that some d_c known *a priori* can function as a minimum effective dose and enable estimation. Under the standard parallel trends assumption, a difference-in-differences estimator using effectively untreated units as a comparion group will recover the average treatment effect for all units of interest:

Proposition 2. *Suppose that assumptions A1-A3, A4 (Parallel Trends), and A5 (Minimum Effective Dose) hold. Then, an estimator of the form:*

$$\triangle Y_{i,\tau} = \alpha + \beta \mathbb{1}(D_i \in \mathcal{D}_r) + \varepsilon_{i,\tau} \tag{8}$$

estimated over units in the set $(-\infty, d_c] \cup \mathcal{D}_r$ *will identify BATT*$[\mathcal{D}_r]$

This is a standard 2x2 difference-in-differences estimator where units not of interest to the researcher ($d_i \notin \mathcal{D}_r$) nor of use as suitable controls ($d_i > d_c$) are not used in estimation. Often in

practice, when the object of interested is the BATT for all units above a researcher specified cutoff (BATT$[d_r, \infty]$), this estimator has a donut shape, as units in the range (d_c, d_r) are thrown away to enable a precise comparison.[8]

As Proposition 2 states, this depends on the parallel trends assumption (A4), which states that $\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|D_i = d]$ does not depend on d. With the additional information that units below d_c are effectively untreated, we can test this assumption directly for these units. Essentially, we are making the assumption that the conditional expectation function of $Y_{i,\tau}(D_i, 0)$ is constant with respect to dose, providing the null hypothesis

$$H_0 : \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|D_i, D_i < d_c] = \alpha \tag{9}$$

If the dose distribution is discrete (Assumption 2.1), then we can test this restriction directly by regressing $\triangle Y_{i,\tau}(D_i, 0)$ on a saturated set of dose indicators:

$$\triangle Y_{i,\tau}(D_i, 0) = \alpha + \sum_{d \leq d_c, d \neq d_1} \mathbb{1}(D_i = d)\beta_d + \varepsilon_i \tag{10}$$

And testing the joint restriction that $\beta_d = 0 \, \forall d \leq d_c, d \neq d_1$. For continuous dose distributions (Assumption 2.2), tests for parametric restrictions on the shape of the conditional expectation function have been well-developed in the non-parametric econometrics literature (see Li and Racine (2007) for a textbook treatment), which rely on comparing data fit for a parametric regression against a non-parametric alternative. The test developed by Hsiao et al. (2007) is implemented in the R `np` package and can be readily utilized by practitioners.

If knowledge of d_c is going to be asserted, we recommend the addition of this statistical test, which functions as the equivalent of a pre-trends test in the dose space. In a standard binary DiD setting, we can't test directly whether or not untreated outcomes would have progressed similarly in the absence of treatment, but we can test whether or not this assumption was true in the lead up to treatment. Similarly, in our continuous DiD setting, we can't test directly whether or not untreated outcomes would have progressed similarly in the absence of dose assignment for effectively treated units, but we can test whether or not this was true for effectively untreated units.

[8]There is no econometric issue if $d_c > d_r$, though the set of comparison units will need to be changed to $d_i < \min\{d_c, d_r\}$ There is, however, perhaps a conceptual issue as to why we would include in our treatment group units that are assumed to be untreated.

However, we suspect that in many settings practitioners will not have *a priori* knowledge of the value of d_c In this case, inference can be broken down into two steps. The first step involves model selection: we need to identify which units will work as comparison units. The second is estimation of the target estimand. To conduct proper inference, standard errors should account for uncertainty in both steps. We now move to presenting a framework for a two-step procedure in a discrete dose setting.

4 MSE-Optimal Estimation

We restrict our focus to Assumption A2.1 that the dose distribution is discrete, with doses taking the values $\{d_1, d_2, ..., d_J\}$. For a researcher interested in estimating BATT$[\mathcal{D}_r]$, units receiving any doses $d_j \in \mathcal{D}_r$ will be used to estimate the evolution of outcomes for the treatment group. Let $\bar{d}_r$ denotes the lowest dose in $\mathcal{D}_r$. We consider units receiving doses in the set $\{d_j : d_j < \bar{d}_r\}$ as potential controls.[9] Under the assumption of a minimum effective dose, the safest choice would be the lowest dose, d_1, which must be untreated. In this case, a DiD estimator of the form

$$\hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i \geq d_r] - \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i = d_1] \tag{11}$$

would be an unbiased estimator for BATT$[d_r, \infty]$. However, unless $d_c = d_1$, this would not be the efficient estimator. At minimum, we would want to include all units with dose assignments up to and including d_c, and for any fixed n, it might be optimal to include higher doses as long as the variance reduction outweighs the increase in bias. The intuition here is that, especially in the case when there are few effectively untreated units, we might be off including "just barely" treated units as comparisons as well. In light of this, we propose a first-step procedure that estimates the MSE-optimal set of controls, then uses these control units in a second-step DiD estimator.

4.1 MSE-Optimal Controls

Let $\hat{\theta}(\mathcal{D}_C)$ denote the estimator of the form

$$\hat{\theta}(\mathcal{D}_C) = \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i \in \mathcal{D}_r] - \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i \in \mathcal{D}_C] \tag{12}$$

[9] There is no conceptual issue considering units receiving some dose $d_j \notin \mathcal{D}_r, d_j > \bar{d}_r$ as a potential control, but this would imply a very strange shape for the dose response function and accordingly we make the simplifying assumption to only consider units receiving a dose below $\bar{d}_r$.

Where $\theta = \text{BATT}[\mathcal{D}_r]$ denotes the estimand of interest. It is simple to derive the expression for the MSE as the sum of the squared bias and variance terms:

Proposition 3. *Under assumptions A1-A5, if $d_1 < d_c$, then the bias and variance of $\hat{\theta}(\mathcal{D}_C)$ can be expressed as*

$$\begin{aligned} \textit{Bias}(\hat{\theta}(\mathcal{D}_C)) &= \sum_{d \in \mathcal{D}_c, d \neq d_1} \frac{n_d}{n_c} (\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0) | d_i = d_1] - \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0) | d_i = d]) \\ \mathbb{V}(\hat{\theta}(\mathcal{D}_C)) &= \frac{\mathbb{V}(\triangle Y_{i,\tau}(D_i, 0) | d_i \in \mathcal{D}_r)}{n_t} + \frac{\mathbb{V}(\triangle Y_{i,\tau}(D_i, 0) | d_i \in \mathcal{D}_c)}{n_c} \end{aligned}$$

This derivation leverages the fact that the progression of outcomes for units receiving d_1 are assumed to be untreated. Then, the bias that results from the inclusion of each additional dose d_j is the given by the DiD estimator comparing d_j with d_1; the contribution of each dose to the bias term is given by the fraction of total control units that receive that dose. And the variance is given simply by the sum of the variance of the treatment and control means.

We propose selecting as controls the units that minimize the empirical estimate of mean-squared error. By replacing population averages with sample means in Proposition 4, we can construct an estimator of the MSE that would result from any proposed set of control units $\widehat{\mathcal{D}}_C$, given by $\widehat{\text{MSE}} = \widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))^2 + \hat{\mathbb{V}}(\hat{\theta}(\mathcal{D}_C))$. Then, we can define our estimator as the subset of $\mathcal{D}_C$ that minimizes the empirical estimator of mean-squared error, that is, $\widehat{\mathcal{D}}_C = \text{argmin}_{\mathcal{S} \subseteq \mathcal{D}_C} \widehat{\text{MSE}}(\mathcal{S})$. Of course, due to Jensen's Inequality, the expected value of the square of our empirical estimate of bias will not equal the population quantity squared; instead, $\mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))^2] = \text{Bias}(\hat{\theta}(\mathcal{D}_C))^2 + \mathbb{V}[\widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))]$. Accordingly, we develop a finite sample bias-corrected estimator of $\mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))^2]$:

Proposition 4. *Under assumptions A1-A5, if $d_1 < d_c$, the following are unbiased estimators for $\textit{Bias}(\hat{\theta}(\mathcal{D}_C))^2$ and $\mathbb{V}(\hat{\theta}(\mathcal{D}_C))$:*

$$\begin{aligned} \widehat{\textit{Bias}}^2_{bc}(\hat{\theta}(\mathcal{D}_C)) &= \widehat{\textit{Bias}}(\hat{\theta}(\mathcal{D}_C))^2 - \left[\frac{(n_c - n_{d_1})^2}{n_c^2 n_{d_1}} \hat{\sigma}^2_{d_1} + \sum_{d \neq d_1} \frac{n_d}{n_c^2} \hat{\sigma}^2_d \right] \\ \hat{\mathbb{V}}(\hat{\theta}(\mathcal{D}_C)) &= \frac{\hat{\mathbb{V}}(\triangle Y_{i,\tau}(D_i, 0) | d_i \in \mathcal{D}_r)}{n_t} + \frac{\hat{\mathbb{V}}(\triangle Y_{i,\tau}(D_i, 0) | d_i \in \mathcal{D}_c)}{n_c} \end{aligned}$$

Where $\hat{\sigma}^2_d$ is the standard variance estimator of $\triangle Y_{i,\tau}(D_i, 0)$ for units receiving dose d.

Since $\mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))^2] > \text{Bias}(\hat{\theta}(\mathcal{D}_C))^2$, $\widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))^2]$ represents a more conservative estimator, overestimating the bias from additional units and leading toward less control units overall.

In finite samples, unless we are willing to make restrictions on the shape of the dose response function and heteroskedasticity, there is no reason that the optimal set of control units will be given by the interval $\{d \in \mathcal{D}_C : d \leq d_0\}$ for some choice of d_0. For example, consider a pathological example where $\mathbb{V}(\triangle Y_{i,\tau}(D_i, 0)|d_i) = \sigma^2 * \mathbb{1}(i \text{ is odd})$ and the dose response function is flat on $\{d_1, d_2, d_3, d_4, d_5\}$. Then, the optimal choice of $\hat{\mathcal{D}}_C$ would include d_2 and d_4 for $\sigma > 0$. Unfortunately, finding the optimum involves searching over $2^{|\mathcal{D}_C|}$ possible subsets, which becomes intractable with even moderate amounts of doses. Accordingly, for tractability, we search only for sets of the form $\{d \in \mathcal{D}_C : d \leq d_0\}$, noting that since this must contain a consistent estimator ($\{d_1\}$), it will not impact any of the results below.

Intuitively, this estimator sits between a fully discretized design that uses all units in $\mathcal{D}_C$, and the most conservative estimator that uses only d_1. The latter design is conservative insofar as it will estimate BATT($\mathcal{D}_r$) under the lowest feasible minimum effective dose, and as Callaway et al. (2024) point out, will identify BATT($\mathcal{D}_r$) − ATT($d_1|d_1$) in the event that units receiving d_1 are not untreated. The former design is very common in practice and will identify BATT($\mathcal{D}_r$) − BATT($\mathcal{D}_C$) in the absence of a minimum effective dose. Our estimator will choose a maximum dose somewhere between d_1 and d_r, informed by an estimate of the MSE induced by each choice.

Asymptotically, in the event that only units receiving d_1 are effectively untreated, our procedure will collapse to the more conservative estimator; alternatively, in the event that all $d < d_c$ are effectively untreated for some $d_c > d_1$, the set of controls $\widehat{\mathcal{D}}_C$ will asymptotically exclude all units receiving a dose about d_c. Formally, we also need to assume a no-crossing property:

A6 **No crossing:** Exactly one of the following is true:

A6.1 $\mathbb{E}[\mu_i(D_i)|D_i] > 0 \; \forall D_i > d_c$

A6.2 $\mathbb{E}[\mu_i(D_i)|D_i] < 0 \; \forall D_i > d_c$

Note that this is far weaker than assuming monotonicity; we need only that the aggregate dose response does not change sign. Functionally, this rules out a situation where the DRF returns to zero above d_c.[10] Now, we can establish this formally:

Proposition 5. *Under assumptions A1-A6, if $d_1 < d_c$, $\widehat{\mathcal{D}}_C$ will not include any treated units in the limit. If the minimum effective dose is d_1, then $\widehat{\mathcal{D}}_C$ will include only d_1 in the limit.*

[10] Note that the MED assumption does not rule this out; it only restricts units below d_c to exhibit no dose response.

4.2 Second Stage: Estimating BATTs and Constructing Honest Standard Errors

Following the plug-in principle, we utilize the a standard DiD estimator in the second step:

$$\hat{\theta}(\widehat{\mathcal{D}}_C) = \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i \geq d_r] - \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i \in \widehat{\mathcal{D}}_C] \tag{13}$$

Where $\mathcal{D}_C$ is selected using the procedure outlined in the previous section. Since $\widehat{\mathcal{D}}_C$ will not include treated units in the limit, it is straightforward to show that $\hat{\theta}(\widehat{\mathcal{D}}_C)$ is consistent for BATT$(\mathcal{D}_r)$

Proposition 6. *Under assumptions A1-A6, if $d_1 < d_c$, then the standard difference-in-differences estimator in* (13) *using the first-step estimator $\widehat{\mathcal{D}}_C$ is consistent for BATT$[\mathcal{D}_r]$.*

However, this result hides the instability inherent in a model selection step. Even in the best case that every dose to the left of d_r is untreated, the model will oscillate between different choices of $\mathcal{D}_C$, introducing variance into $\hat{\theta}(\widehat{\mathcal{D}}_C)$. To limit the increased variance from model selection, we opt instead for a bootstrap aggregating (bagging) estimator (Breiman, 1996). To calculate this, we generate B bootstrap samples from our dataset, using each to produce an estimate of our parameter of interest $\hat{\theta}_b(\widehat{\mathcal{D}}_C)$, then take the average of these to generate a point estimate:

$$\hat{\theta}_B(\widehat{\mathcal{D}}_C) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b(\widehat{\mathcal{D}}_C)$$

As $\widehat{\mathcal{D}}_C$ changes, $\hat{\theta}_b(\widehat{\mathcal{D}}_C)$ changes in a discrete manner; since there are a discrete number of doses, moving from $\hat{d}_c = d_j$ to $\hat{d}_c = d_{j+1}$ entails a discrete jump in $\hat{\theta}_b(\widehat{\mathcal{D}}_C)$, generating a "lumpy" distribution. Taking the average over the set of bootstrap samples corrects this by generating a smooth estimator in the presence of these discrete changes.

4.2.1 Accounting for Uncertainty Over the Choice of d_c

In the limit, both the bias and variance terms go to 0 for any set $\mathcal{D}_C$ that is comprised only of untreated units. As a result, the doses selected will oscillate between these potential models for any finite n, and inference that does not account for the model selection step will not provide a reliable finite sample approximation. This is not an issue that is unique to our problem; Leeb and Pötscher (2005) show that this is generically true of many estimators that embed a model selection step and arises because the sequence of finite sample distributions of the estimator do not converge uniformly to its asymptotic distribution.

Intuitively, fixing any set of untreated doses, the standard difference-in-differences estimator would have its usual asymptotic properties. For any fixed n, our estimator with a model selection step will take each possible set of untreated doses with some positive probability, resulting in a mixture distribution. The other key advantage of a bagging estimator in this setting, apart from variance reduction, is that it smooths across the discontinuities introduced by selecting among a finite set of models. As a result, we can construct valid standard errors and confidence intervals for the bagging estimator using the method proposed by Efron (2014).

The simplest way to proceed is to conduct a second bootstrap: for each draw of the bagging estimator, draw B random samples with replacement and calculate the corresponding bagging estimator. This will leave is with B bagging estimators, allowing us to construct standard errors using usual approaches. However, this requires B^2 replications and is computationally intensive; Efron (2014) suggests an adjustment to the standard bootstrap standard error calculation to obtain a better approximation of uncertainty. Let W_{ij} denote the number of times that observation j was drawn in bootstrap replication i, and let W_j denote the average of W_{ij} across all bootstrap replications B for each observation j. Then, the standard error estimate is defined as

$$\hat{\sigma}_B = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{B}(W_{ij} - W_j)(\hat{\theta}_b(\widehat{\mathcal{D}}_C) - \hat{\theta}_B(\widehat{\mathcal{D}}_C))/B}$$

And the standard interval $\hat{\theta}_B(\widehat{\mathcal{D}}_C) \pm 1.96 \times \hat{\sigma}_B$ delivers a finite sample approximation to a 95% confidence interval for ATT(d_r).

4.3 Continuous Dose Case

Under the structure we have imposed here, estimation with a continuous dose distribution is somewhat trivial, but we discuss it briefly for the sake of completeness. Recall Assumption A2.2 that doses have some continuous distribution over (d_l, d_u). As before, the researcher would like to estimate the ATT for units receiving dose d_r or higher, leaving (d_l, d_r) as potential controls. If we assume that some minimum effective dose $d_c > d_l$ exists, we would want to choose an estimator $\hat{d}_c$ to define a local average of the form

$$\frac{1}{|\{d_i : d_i < \hat{d}_c\}|}\sum_{d_i < \hat{d}_c} \triangle Y_{i\tau}(d_i) \tag{14}$$

This estimator is trivial insofar as, for any $\hat{d}_c$ such that $\hat{d}_c \to_n d_l$, in the limit this average will converge to $\mathbb{E}[\triangle Y_{i\tau}(0)]$, since eventually $\mathbb{P}(\hat{d}_c < d_c) \to_n 1$. Ideally, we would choose $\hat{d}_c$ to minimize mean-squared error: moving $\hat{d}_c$ away from d_l would lower variance but potentially raise bias. This is, functionally, a local non-parametric regression estimate at d_l, and any such estimator would do.[11]

5 Simulation Evidence

To assess the relative efficiency of our estimator, we conduct a simulation exercise to explore mean-squared error across a variety of data generating processes. We consider a simple design where the difference in outcomes, $\triangle Y_{i\tau}$, is drawn directly:

$$\triangle Y_{i\tau} = \mu(d_i) + \triangle \varepsilon_{i\tau} \tag{15}$$

The dose distribution is discrete and takes 20 equally spaced values within (0,1). Our baseline dose distribution is discrete uniform, where there are 30 observations per dose, resulting in total observations of 600. We consider two alternative distributions that are more concentrated around the mean to reflect what we expect is more commonly found in practice. These distributions are derived using weights given by the expression $[d(1-d)]^{1-\beta}$: these baseline weights are normalized to sum to 1, then observations at each dose are given by distributing (600) observations across doses proportionally to their weight. We consider $\beta \in \{1, 1.5, 2\}$; $\beta = 1$ results in the discrete uniform distribution, and as β increases, mass moves from the tails to the center of the distribution.

We consider three common shapes for the dose response function, μ:

$$\mu(d) = \begin{cases} 1 \cdot \mathbb{1}(d > d_c) & \text{(jump)} \\ (d - 0.25) \cdot \mathbb{1}(d > d_c) & \text{(linear)} \\ (d - 0.25)^2 \cdot \mathbf{1}(d > d_c) & \text{(quadratic)} \end{cases} \tag{16}$$

For each DRF, we impose a minimum effective dose of d_c. In the jump design, there is a constant treatment effect of 1 after passing this threshold. In the linear design, there is an increasing treatment effect in the dose with a linear shape, with slope equal to 1. In the quadratic design, this

[11] The weakest assumption that we could make here would be that $d_c = d_l$, and we would likely want to use a local linear regression to estimate $\mathbb{E}[\triangle Y_{i\tau}(d_l)]$ (Fan et al., 1996). This is the estimator proposed by de Chaisemartin et al. (2026), who develop the optimal procedure in detail.

is an increasing treatment effect in the dose with a quadratic shape, with slope equal to $2d$. The designs are intended to represent situations where the DRF is discontinuous at d_c (jump), the DRF is continuous at d_c but the derivative of the DRF is discontinuous at d_c (linear), and where both the DRF and its derivative are discontinouous at d_c (quadratic).

Errors are all drawn initially from a $\mathcal{N}(0, \text{ATT})$ distribution, where ATT is calculated using $\mu(\cdot)$. Since the treatment effect itself depends on the shape of the DRF, we scale the error variance to ensure that the ratio of the ATT to the error variance remains constant. We introduce two different types of heterogeneity: one where the standard deviation σ is positively correlated with dose (and consequently $\triangle Y_{i\tau}$) and one where σ is negatively correlated with dose. We implement this by multiplying the initial error draws by $d/0.5$ or $(1-d)/0.5$, respectively.

5.1 Results

The results of our simulation exercise are presented in Table 2. To make comparison simpler, we report mean-squared error (MSE) relative to using only the untreated first dose, as proposed in Callaway et al. (2024), as well as a design that binarizes the dose space at the researcher specified dose value, as is common in practice. We find that our estimator provides sizable efficiency gains over this baseline estimator. These gains are robust across the shape of the dose response function and the centrality of the dose distribution but depend on the heteroskedasticity of the errors. Specifically, we see MSE gains only when the errors are homoskedastic or heterkskedastic and negatively correlated with dose. This matches econometric intuition: heteroskedasticity that is positively correlated with dose reduces the gains in variance reduction that can be obtained by adding additional control units at the bottom of the dose distribution. Importantly, even with heteroskedastic errors that are positively correlated with dose, our estimator displays almost no loss of efficiency relative to the first dose estimator, suggesting that it presents a uniform improvement over this baseline.

MSE comparisons with the binarized estimator are not useful, since this estimator is not unbiased, but since this is almost always the design preferred in the empirical literature, it is presented for comparison. Relative to this estimator, our procedure reduces bias substantially and is broadly comparable to using only the first dose. Of course, our estimator is not unbiased in finite samples by design, as it trades off bias for variance reduction, but it is clear that this only adds minimal bias, at least at the modest sample size chosen for this exercise.

References

Altman, D. G., B. Lausen, W. Sauerbrei, and M. Schumacher (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute 86*(11), 829–835.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Botosaru, I., R. Giacomini, and M. Weidner (2024). Forecasted treatment effects. Working Paper.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. *Handbook of Econometrics 5*, 3705–3843.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*(2), 123–140.

Brenner, H. (1997). A potential pitfall in control of covariates in epidemiologic studies. *Epidemiology 9*(1), 68–71.

Brodeur, A., N. Cook, and A. Heyes (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review 110*(11), 3634–60.

Butts, K. (2022). Difference-in-differences with geocoded microdata. *Journal of Urban Economics*.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2024). Difference-in-differences with a continuous treatment. Working Paper.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Card, D. (1992). Using regional variation in wages to measure the effects of the federal minimum wage. *Industrial and Labor Relations Review*, 22–37.

de Chaisemartin, C., D. Ciccia, X. D'Haultfoeuille, and F. Knau (2026). Difference-in-differences estimators when no unit remains untreated. Working Paper.

de Chaisemartin, C. and X. D'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *Journal of Econometrics*.

de Chaisemartin, C., X. D'Haultfoeuille, F. Pasquier, and G. Vazquez-Bare (2023). Difference-in-differences estimators for treatments continuously distributed at every period. *Working Paper*.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association 109*(507), 991–1007.

Fan, J., I. Gijbels, T.-C. Hu, and L.-S. Huang (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 113–127.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Hsiao, C., Q. Li, and J. S. Racine (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics 140*(2), 802–826.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Lausen, B. and M. Schumacher (1992). Maximally selected rank statistics. *Biometrics 48*(1), 73–85.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

MacCallum, R. C., S. Zhang, K. J. Preacher, and D. D. Rucker (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods 7*(1), 19–40.

Miller, R. and D. Siegmund (1982). Maximally selected chi square statistics. *Biometrics 38*(4), 1011–1016.

Rambachan, A. and N. Shephard (2019). A nonparametric dynamic causal model for macroeconometrics. Working Paper.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*.

Roth, J. and P. H. Sant'Anna (2021). When is parallel trends sensitive to functional form? *Working Paper*.

Ruberg, S. J. (1989). Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association 84*(407), 816–822.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

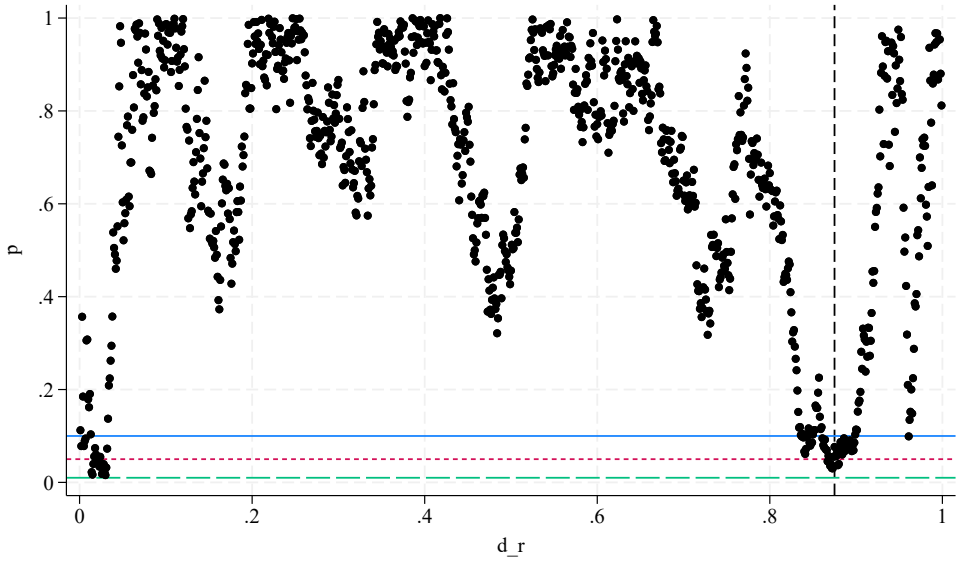
Sun, L. and J. M. Shapiro (2022). A linear panel model with heterogeneous coefficients and variation in exposure. *Journal of Economic Perspectives*.

Turner, S. L., A. Karahalios, A. B. Forbes, M. Taljaard, J. M. Grimshaw, and J. E. McKenzie (2021). Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series. *BMC Medical Research Methodology 21*(134), 1–19.

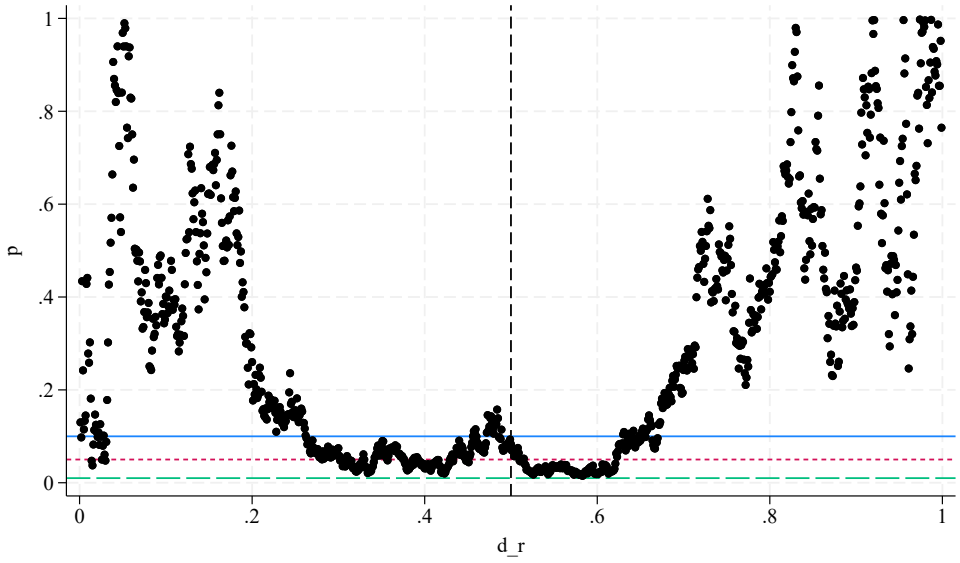
Wooldridge, J. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Working Paper*.

Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business & Economic Statistics 14*(4), 478–486.

Figure 1: Testing for Researcher Manipulation



(a) p-value Sensitive to Researcher Choice



(b) p-value Robust to Researcher Choice

We draw 10,000 observations and estimate equation (3) over a grid of d_r values equally spaced by 0.001. Each figure plots the p-value of $\hat{\beta}$ against the corresponding cutoff choice d_r. The horizontal lines represent conventional significance levels $(0.10, 0.05, 0.01)$. The dose distribution is continuous uniform over $[0, 1]$. In Figure 1a, the outcome is equal to a random noise $\mathcal{N}(0, 1)$ term, where a vertical line denoting a choice of $d_r = 0.875$ is plotted. In Figure 1a, the outcome is equal to one if the dose is higher than 0.5 plus a random noise $\mathcal{N}(0, 1)$ term, where a vertical line denoting a choice of $d_r = 0.5$ is plotted.

Table 1: Inflated Rejection Rates of Null Hypothesis

	Search Region (ϵ)		
Significance Level	$1/3$	$1/4$	$1/5$
$\alpha = .10$	.40	.49	.55
$\alpha = .05$	.24	.31	.35
$\alpha = .01$	.07	.09	.11

This is a partial reproduction of Table 1 from Miller and Siegmund (1982). Each cell denotes the approximate Type I error, given by (7), resulting from a statistical test with significance level α of $\hat{\beta}$ in (4), where the cutoff d_r is chosen to minimize the p-value over the search region $[\epsilon, 1-\epsilon]$.

Table 2: Simulation Results (Bias Corrected)

DGP	β	ε	Binarized			First Dose			MSE-Optimal		
			Bias	Var	Relative MSE	Bias	Var	Relative MSE	Bias	Var	Relative MSE
jump	1	0	-0.64	0.01	10.74	-0.00	0.04	1.00	-0.03	0.04	0.93
jump	1	+	-0.64	0.02	24.25	0.00	0.02	1.00	0.00	0.02	0.99
jump	1	-	-0.64	0.00	3.05	0.01	0.13	1.00	-0.05	0.10	0.79
jump	1.5	0	-0.74	0.01	5.52	0.00	0.10	1.00	-0.04	0.08	0.86
jump	1.5	+	-0.73	0.02	27.57	0.01	0.02	1.00	0.01	0.02	1.01
jump	1.5	-	-0.74	0.00	1.69	0.02	0.33	1.00	-0.05	0.25	0.76
jump	2	0	-0.80	0.01	2.87	-0.01	0.23	1.00	-0.07	0.18	0.82
jump	2	+	-0.80	0.02	31.18	0.00	0.02	1.00	0.00	0.02	1.03
jump	2	-	-0.80	0.00	0.70	0.00	0.92	1.00	-0.08	0.65	0.71
linear	1	0	-0.14	0.00	1.61	0.00	0.01	1.00	-0.01	0.01	0.78
linear	1	+	-0.15	0.01	4.97	-0.00	0.01	1.00	-0.00	0.01	1.01
linear	1	-	-0.15	0.00	0.53	0.00	0.04	1.00	-0.01	0.03	0.67
linear	1.5	0	-0.17	0.00	1.17	0.00	0.03	1.00	-0.02	0.02	0.71
linear	1.5	+	-0.17	0.01	5.28	0.00	0.01	1.00	0.00	0.01	1.03
linear	1.5	-	-0.17	0.00	0.27	0.01	0.11	1.00	-0.01	0.07	0.68
linear	2	0	-0.18	0.00	0.46	0.01	0.08	1.00	0.00	0.06	0.72
linear	2	+	-0.19	0.01	6.54	-0.01	0.01	1.00	-0.01	0.01	1.03
linear	2	-	-0.18	0.00	0.11	0.02	0.31	1.00	-0.00	0.23	0.73
quadratic	1	0	-0.04	0.00	0.61	-0.00	0.01	1.00	-0.01	0.00	0.74
quadratic	1	+	-0.04	0.00	1.84	0.00	0.00	1.00	0.00	0.00	1.02
quadratic	1	-	-0.04	0.00	0.13	-0.00	0.02	1.00	-0.00	0.01	0.69
quadratic	1.5	0	-0.05	0.00	0.29	0.01	0.01	1.00	0.00	0.01	0.71
quadratic	1,5	+	-0.05	0.00	2.34	-0.00	0.00	1.00	-0.00	0.00	1.01
quadratic	1.5	-	-0.05	0.00	0.09	-0.01	0.03	1.00	-0.01	0.02	0.67
quadratic	2	0	-0.05	0.00	0.14	-0.01	0.03	1.00	-0.01	0.02	0.73
quadratic	2	+	-0.05	0.00	2.34	0.00	0.00	1.00	0.00	0.00	1.00
quadratic	2	-	-0.05	0.00	0.03	0.00	0.10	1.00	-0.00	0.07	0.73

A Proofs

A.1 Proposition 1

Proof. Consider the probability limit of $\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0]$, where $T_i = \mathbb{1}(d_i \geq d_r)$.

$$
\begin{aligned}
&\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0] \\
&\to_p E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] \\
&= (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 1]) \\
&- (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 1]) \\
&= (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 1]) \\
&- (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 0]) \qquad \text{A4} \\
&= (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0)|T_i = 1]) - (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0]) \\
&= (E[E[Y_{i,\tau}(d, 1) - Y_{i,\tau}(d, 0)|D_i = d]|T_i = 1]) - (E[E[Y_{i,\tau}(d, 1) - Y_{i,\tau-1}(d, 1)|D_i = d]|T_i = 0]) \\
&= (E[\mu(d|d)|T_i = 1]) - (E[\mu(d|d)|T_i = 0]) \\
&= \int_{d_r}^{d_u} \mu(d|d) \frac{f(l)}{1 - F(d_r)} \mathrm{d}l - \int_{d_l}^{d_r} \mu(d|d) \frac{f(l)}{1 - F(d_r)} \mathrm{d}l
\end{aligned}
$$

□

A.2 Lemma 1

Proof. We add and subtract the following term to arrive at the decomposition in the text:

$$\int_{d_l}^{d_r} \mu(d|F^{-1}_{d>d_r}(F_{d<d_r}(l)) \frac{f(l)}{F(d_r)} \mathrm{d}l$$

□

A.3 Proposition 2

Proof. Consider the probability limit of $\hat{E}[\Delta Y_{i\tau}|D_i \in \mathcal{D}_r] - \hat{E}[\Delta Y_{i\tau}|D_i \in (-\infty, d_c]]$:

$$
\begin{aligned}
&\hat{E}[\Delta Y_{i\tau}|D_i \in \mathcal{D}_r] - \hat{E}[\Delta Y_{i\tau}|D_i \in (-\infty, d_c]] \\
&\to_p E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|D_i \in \mathcal{D}_r] - E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|D_i \in (-\infty, d_c]] \\
&= E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|D_i \in \mathcal{D}_r] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|D_i \in (-\infty, d_c]] \qquad \text{A5}
\end{aligned}
$$

$$= E[Y_{i,\tau}(D_i,1) - Y_{i,\tau-1}(D_i,0)|D_i \in \mathcal{D}_r] - E[Y_{i,\tau}(D_i,0) - Y_{i,\tau-1}(D_i,0)|D_i \in \mathcal{D}_r] \quad \text{A4}$$

$$= E[Y_{i,\tau}(D_i,1) - Y_{i,\tau-1}(D_i,0)|D_i \in \mathcal{D}_r]$$

$$= \text{BATT}[\mathcal{D}_r]$$

□

A.4 Proposition 3

Proof. The bias of $\hat{\theta}(\mathcal{D}_C)$ will be given by

$$\begin{aligned}
\text{Bias}(\hat{\theta}(\mathcal{D}_C)) &= \mathbb{E}[\hat{\theta}(\mathcal{D}_C)] - \theta \\
&= \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i \geq d_r] - \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i \in \mathcal{D}_C] - \theta \\
&= \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i \geq d_r] - \sum_{d_i \in \mathcal{D}_c} \frac{n_d}{n_c} \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = \mathcal{D}_C] \\
&\pm \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = d_1] - \theta \\
&= \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i \geq d_r] - \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = d_1] - \theta \\
&+ \sum_{d \in \mathcal{D}_c, d_i \neq d_1} \frac{n_d}{n_c} (\mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = d_1] - \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = d]) \\
&= \sum_{d \in \mathcal{D}_c, d \neq d_1} \frac{n_d}{n_c} (\mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = d_1] - \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)|d_i = d])
\end{aligned}$$

The expression for the variance follows from the standard result on the variance of the difference of two independent averages. □

A.5 Proposition 4

Proof.

$$\begin{aligned}
\mathbb{V}[\widehat{\text{Bias}}(\hat{\theta}(\mathcal{D}_C))] &= \mathbb{V}\left[\sum_{d \in \mathcal{D}_c, d \neq d_1} \frac{n_d}{n_c} (\hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i,0)|d_i = d_1] - \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i,0)|d_i = d]) \right] \\
&= \mathbb{V}\left[\frac{n_c - n_{d_1}}{n_c} \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i,0)|d_i = d_1] \right] + \mathbb{V}\left[\sum_{d \in \mathcal{D}_c, d \neq d_1} \frac{n_d}{n_c} \hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i,0)|d_i = d]) \right] \\
&= \frac{(n_c - n_{d_1})^2}{n_c^2} \frac{\mathbb{V}[\triangle Y_{i,\tau}(D_i,0)|d_i = d]}{n_{d_1}} + \sum_{d \in \mathcal{D}_c, d \neq d_1} \frac{n_d^2}{n_c^2} \frac{\mathbb{V}[\triangle Y_{i,\tau}(D_i,0)|d_i = d])}{n_d} \\
&= \frac{(n_c - n_{d_1})^2}{n_c^2 n_{d_1}} \sigma_{d_1}^2 + \sum_{d \in \mathcal{D}_c, d \neq d_1} \frac{n_d}{n_c^2} \sigma_d^2
\end{aligned}$$

It follows immediately from the definition of variance that $\widehat{Bias}^2_{bc}(\hat{\theta}(\mathcal{D}_C))$ will be unbiased. It is trivial to show that $\hat{V}(\theta(\hat{\mathcal{D}}_C))$ will also be unbiased. □

A.6 Proposition 5

Proof. Consider the expressions for Bias$(\hat{\theta}(\mathcal{D}_C))$ and $\mathbb{V}(\hat{\theta}(\mathcal{D}_C))$ as $n_d \to \infty$. Variance goes to 0, so the choice of $\mathcal{D}_C$ will depend only on bias. We know that $\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i = d_1] - \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i = d] = 0$ if and only if $d < d_c$ by A4, A5, and A6. It follows immediately that Bias$(\hat{\theta}(\mathcal{D}_C)) = 0$ requires that $\mathcal{D}_C \subseteq \{d : d \leq d_c\}$.

We now need to show that the additional terms in $\widehat{Bias}^2_{bc}(\hat{\theta}(\mathcal{D}_C))$ go to 0 in the limit. Since $n_d < n_c$ it follows immediately that $n_d/n_c^2 \to 0$. For the first term, note that

$$\frac{(n_c - n_{d_1})^2}{n_c^2 n_{d_1}} = \frac{n_c^2 - 2n_c n_{d_1} + n_{d_1}^2}{n_c^2 n_{d_1}} = \frac{1}{n_{d_1}} - \frac{2}{n_c} + \frac{n_{d_1}}{n_c}\frac{1}{n_c} \to 0$$

Finally, the last line also follows since if $d_c = d_1$, $\mathcal{D}_C \subseteq \{d_1\}$ □

A.7 Proposition 6

Proof. Let $\mathcal{D}_r = \{d : d < d_r\}$. By the law of total probability, we can write

$$\mathbb{E}[\hat{\theta}(\hat{\mathcal{D}}_C)] = \sum_{D \in 2^{\mathcal{D}_r}} (\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i \geq d_r] - \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i \in D])\mathbb{P}(\hat{\mathcal{D}}_C = D)$$

In the limit, $\mathbb{P}(\hat{\mathcal{D}}_C = D) \to 0$ if $D \subsetneq \{d : d \leq d_c\}$ by Proposition 5. So, we can write,

$$\begin{aligned}
\mathbb{E}[\hat{\theta}(\hat{\mathcal{D}}_C)] \to_p & \sum_{D \in \{d : d \leq d_c\}} (\mathbb{E}[\triangle Y_{i,\tau}(D_i, 1)|d_i \geq d_r] - \mathbb{E}[\triangle Y_{i,\tau}(D_i, 1)|d_i \in D])\mathbb{P}(\hat{\mathcal{D}}_C = D) \\
= & \sum_{D \in \{d : d \leq d_c\}} (\mathbb{E}[\triangle Y_{i,\tau}(D_i, 1)|d_i \geq d_r] - \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i \in D])\mathbb{P}(\hat{\mathcal{D}}_C = D) \qquad \text{A5} \\
= & \sum_{D \in \{d : d \leq d_c\}} \text{BATT}[\mathcal{D}_r]\mathbb{P}(\hat{\mathcal{D}}_C = D) = \text{BATT}[\mathcal{D}_r]
\end{aligned}$$

□

A.8 Linear Dose Example

From work in the proof to Proposition 4, we have that

$$
\begin{aligned}
\widehat{b}_1^{\text{BIN}} \quad \rightarrow^p \quad & E[Y_{i,\tau}(D_i,1) - Y_{i,\tau}(D_i,0) | D_i \geq d_r] - E[Y_{i,\tau}(D_i,1) - Y_{i,\tau}(D_i,0) | D_i < d_r] \\
= \quad & E[\mu_i(D_i) | D_i \geq d_r] - E[\mu_i(D_i) | D_i < d_r] \\
= \quad & \beta(E[D_i | D_i \geq d_r] - E[D_i | D_i < d_r]) \\
= \quad & \beta\left(\mu + \sigma \frac{\phi(\frac{d_r-\mu}{\sigma})}{1-\Phi(\frac{d_r-\mu}{\sigma})} - \mu + \sigma \frac{\phi(\frac{d_r-\mu}{\sigma})}{\Phi(\frac{d_r-\mu}{\sigma})}\right) \\
= \quad & \beta\sigma \frac{\phi(\frac{d_r-\mu}{\sigma})\{\Phi(\frac{d_r-\mu}{\sigma}) + (1-\Phi(\frac{d_r-\mu}{\sigma}))\}}{\Phi(\frac{d_r-\mu}{\sigma})(1-\Phi(\frac{d_r-\mu}{\sigma}))} \\
= \quad & \beta\sigma \frac{\phi(\frac{d_r-\mu}{\sigma})}{\Phi(\frac{d_r-\mu}{\sigma})(1-\Phi(\frac{d_r-\mu}{\sigma}))}
\end{aligned}
$$