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

November 18, 2024

Link to most recent version

Abstract

This paper studies difference-in-differences (DD) 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. This setting stands in contrast to the recent literature re-examining DD estimators which typically requires that a subset of observations never receive the treatment to identify the Average Treatment Effect on the Treated (ATT). We develop a framework to estimate the treatment effect when the dose takes effect only after a cutoff value, the Minimum Effective Dose (MED), and introduce the average treatment effect on the *effectively* treated (ATET) as our target estimand. We propose a sample splitting estimator of the ATET and MED under non-parametric assumptions on the dose response function. First, in a hold-out sample, we borrow methods from the pharmacological literature to estimate the MED in a model selection step. This is then used to estimate the ATET with the remaining observations in a second step. This estimator is asymptotically conservative: it does not erroneously identify any treated units as untreated in the limit even if the MED is on the boundary of the parameter space, but as a result, it provides an attenuated estimate of the ATET. We use the bootstrap procedure in Efron (2014) to construct standard errors for the ATET estimate that reflect uncertainty over the value of the MED. Our simulations suggest that this estimator performs well in finite samples.

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 (DD) 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 DD

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

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 DD 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 DD. 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.

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. 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.

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. A natural treatment group is this setting is all units that were impacted by the policy; under the assumption of an MED, this group consists of all units receiving a dose above the MED. We define the average treatment effect across this group as the average treatment effect on the effectively treated (ATET), which is our target estimand.

If we knew *ex ante* what the MED was, a design in which the researcher dichotomized at this value would recover the ATET by standard arguments. In practice, however, this is generally unknown, and we propose a two-step estimator to estimate the ATET when the MED exists. The first step is a model selection step where we estimate the MED. With a discrete dose space, this is a standard sequential hypothesis test problem, with standard methods leveraging pairwise comparisons to identify the highest dose that is statistically different from the placebo (Dunnett and Tamhane, 1992). We propose a non-parametric method that uses the entire p-value distribution, generated by pairwise dose comparisons, developed by Mallik et al. (2011) and adapted by Sales (2024) to a sequential hypothesis testing problem. In the second step, we utilize this parameter to estimate the ATET with a standard difference-in-differences estimator.

This procedure has many desirable qualities. In the limit, the MED estimator will never mistakenly identify a treated unit as a control. As a result, it will be asymptotically conservative, as the only statistically error will be choosing an MED below the true value, which still identifies a group of valid control units. Our simulations show that this property is achieved at a relatively low number of observations per dose, suggesting excellent finite sample performance. Further, this procedure will estimate an MED at the boundary, which provides a check on our identifying assumption. Put differently, if there are no untreated units, the MED estimator will choose only the lowest dose as the control group.

However, since the chosen MED can lie below the true value even in the limit, our second stage

estimator will provide in expectation an attenuated ATET. This is due to the reliance on the p-value distribution: even asymptotically, a test of a true null hypothesis can deliver an arbitrarily low p-value, lowering the estimated MED. We are not aware of any consistent estimator of the MED with a fixed number of doses, though the limiting distribution of our estimator suggests that it will often correctly identify the ATET.

To minimize the bias from joint estimation of the model and estimand, we randomly split the dataset into two parts to tackle MED and ATT estimation separately. Specifically, we implement K-fold cross-fitting as a generalization of this procedure, given potential improvements in finite-sample performance (see Chernozhukov et al. (2018)). To estimate the standard error of the ATET estimator, we use the smoothed bootstrap estimator in Efron (2014) that explicitly deals with discontinuities at the boundary of model regimes in estimators defined by a model selection step. We show it achieves proper coverage of the ATET in simulations.

Our primary application of interest is a setting with a finite number of doses with many observations at each dose. As an extension, we consider what can be done with a continuous dose space with one observation per dose. While there are similar methods available in this setting, we argue that they will likely perform very poorly. Instead, we suggest discretizing the dose space and utilizing our MED estimator to identify a group of untreated units. With this pool, the researcher can readily identify the dose response curve using methods introduced in Callaway et al. (2024) that require a pool of untreated units. We conclude by showing that a similar procedure can also be used to produce a consistent estimator of the ATT at any specific dose value in the finite case, or range of dose values in the continuous case.

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 treatement 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 (**?**Callaway et al., 2024; de Chaisemartin et al., 2023; de Chaisemartin and D'Haultfoeuille, 2022; Sun and Shapiro, 2022; Butts, 2022). Our contribution is to show that re-

searchers 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. **?** 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 ATET, without pure controls, a case that is common in research designs estimating continuous treatment DD.

We use numerous results from the literature on threshold and change-point estimation. In a simple model of a minimum effective dose with a homogeneous linear dose response, the MED constitutes a change-point in an underlying linear model. Inference is complicated because the change-point is not identified under the null hypothesis of no dose response (Hansen, 1996). We show how these methods can be utilized in a continuous difference-in-differences setting whenever relevant parametric assumptions hold. Hansen (1999) develops this theory in a panel data setting, which now even has a dedicated Stata command (Wang, 2015). Recent work has developed theory to accommodate dynamic panels (Seo and Shin, 2016), interactive fixed effects (Miao et al., 2020), and heterogeneous change-point and slope coefficients (Miao et al., 2020).

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, dichotomization persists due to the lure of a difference-in-differences research design, in part due to 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 identifying 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 ATT. Additionally, we use the framework to characterize the state of current practice in continuous DD estimation. In Section 3, we propose our non-parametric estimator and simulation evidence of its performance in finite samples. Section 4 describes some extensions to our main results, and Section 5 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 $\text{supp}\{D_i\} := \mathcal{D} \subset \mathbb{R}_+$, 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 the next section, when we introduce our core estimation strategy, we assume a finite number of doses as in A2.1. In the section after, we consider 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$(d|d)$. We can generalize this by averaging over all treated units at the dose at which they were treated. 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. 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.

In light of these difficulties, practitioners have generally opted to leverage differences in intensity of treatment instead of treatment status. However, as Callaway et al. (2024) point out, as a result, these designs estimate a different causal parameter, the average causal response. Intuitively, comparison of "more" and "less" treated units prevents reliance on extrapolation by making within period comparisons, but necessitates focus on a different estimand.

We cannot escape this trade-off but opt for an estimand similar to the ATT and likely of interest to the researcher in its own right. Let T_i^E indicate units which are *effectively* treated, where a unit is effectively treated if it's potential outcome is changed by the policy of interest; that is, if $Y_{i,\tau}(d,0) \neq Y_{i,\tau}(d,1)$. Then, we can define the *average treatment effect on the effectively treated* (ATET)

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

In our view, this aligns much more closely with practitioner intentions in a discretized design. The ATET is our primary estimand of interest, framing our discussion of identification going forward.

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 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 conduct 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 DD 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 DD 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 that 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 $\text{ATT}(d|d)$'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 $\text{ATT} > 0$ but the binned estimator is negative. To avoid this outcome, we would need to impose the further assumption that $\mu(d|d)$ 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. We also explore the relationship between this estimator and the ATET under the assumption of a minimum effective dose in Appendix Section B.1.

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.[1] 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.6 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

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

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).

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 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 $[\epsilon, 1-\epsilon]$. Using the z-score that would be used to define the rejection region of this test, we can calculate the actual Type 1 error probability would result. We reproduce part of Table 1 from Miller and Siegmund (1982) to give a sense of how these levels change.

Table 1: Inflation Rejection Rates of Null Hypothesis

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

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.

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 DD estimates do not account for uncertainty over what this cutoff is. In the next section, we introduce a non-parametric method to estimate this cutoff, in effect systematizing the process of choosing d_r, as well as a bootstrapping approach that accounts for the uncertainty over the value of the cutoff.

3 Minimum Effective Dose (MED)

We first describe minimal assumptions that identify the ATET. To do so, we propose splitting the sample so one part of the dataset is used to estimate a suitable control group, and the other utilizes this group to estimate the ATET. We then use the smoothed bootstrap method to calculate standard errors and conduct simulations to understand the finite sample properties of our estimator.

3.1 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 DD 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 Assumption A5, potential outcomes in (2) can be simplified to

$$Y_{i,t} = (1 - P_t)Y_{i,t}(D_i, 0) + P_t\{Y_{i,t}(D_i, 1)T_i + Y_{i,t}(D_i, 0)(1 - T_i)\} \tag{8}$$

where $T_i := 1\{D_i \geq d_c\}$ denotes the (effectively) treated group. As this expression illustrates, the assumption of an MED allows us to recover a standard binary difference-in-differences setting. Under the standard parallel trends assumption, a difference-in-differences estimator will recover the average treatment effect for all (effectively) treated units above the MED

Proposition 2. *Suppose that assumptions A1-A3, A4 (Parallel Trends), and A5 (Minimum Effective Dose) hold. Then, the binned difference-in-differences estimator using the MED d_c recovers the ATET.*

In our estimation, this is the most straightforward justification of the binarization approach that 30% of the papers in our metastudy pursue. However, as we point out here, it relies on the assumption of a minimum effective dose, as well as a parallel trends assumption. Further, the DD estimator studied above is generally infeasible, as it relies on knowledge of the MED d_c.

3.2 Recasting as a Threshold Model

To overcome this, we propose a model selection step to choose d_c by leveraging advancements in the threshold estimation literature. The standard approach involves assuming a homogenous, parametric dose response function and estimating treatment effects and the threshold d_c simultaneously. For example, we could specify the dose response function as $\mu_i(D_i; \beta) = \beta_1 + \beta_2(D_i - d_c)$ reflecting a linear "partial linear dose" approach that allows for a jump (β_1) in treatment effects beyond the cutoff (d_c) as well as a constant multiple effect given their dose (β_2). Then, the true parameters $(\beta_0, \beta_1, \beta_2, d_c)$ will solve

$$\underset{\beta,d}{\operatorname{argmin}} \sum_{i=1}^{N} [\Delta Y_{it} - \beta_0 - \mu_i(D_i; \beta_1, \beta_2) \cdot 1\{D_i \geq d_c\}]^2 \tag{9}$$

For every choice of d_c, $(\beta_0, \beta_1, \beta_2)$ can be estimated via least squares. Searching over all d_c and choosing the parameter set minimizing the expression in (9) will yield the solution (Hansen, 2000). This method will work when the dose response function is homogeneous, specified correctly, and meets certain regularity conditions. We suspect that these restrictions will be unpalatable to practitioners and instead propose a nonparametric approach that relaxes many of these restrictions.

Using (8), we can write observed outcomes in each time period as

$$\begin{aligned} Y_{i,\tau-1} &= Y_{i,\tau-1}(D_i, 0) \\ Y_{i,\tau} &= Y_{i,\tau}(D_i, 0) + [Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0)]T_i \end{aligned}$$

Taking the first difference of observed outcomes yields

$$\begin{aligned} \triangle Y_{i,\tau} &= \triangle Y_{i,\tau}(D_i, 0) + [Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0)]T_i \\ &= \triangle Y_{i,\tau}(D_i, 0) + \mu_i(D_i)\mathbb{1}(D_i \geq d_c) \end{aligned}$$

Finally, we can rewrite the right hand side using conditional expectations

$$\begin{aligned} \triangle Y_{i,\tau} &= \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|D_i] + \mathbb{E}[\mu_i(D_i)|D_i]\mathbb{1}(D_i \geq d_c) + \varepsilon_i \\ &= \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)] + \mathbb{E}[\mu_i(D_i)|D_i]\mathbb{1}(D_i \geq d_c) + \varepsilon_i \end{aligned}$$

Where the second line follows from parallel trends (A4). We've written the first difference of observed outcomes as a standard threshold response model with an aggregate dose response function of the form $\mathbb{E}[\mu_i(D_i)|D_i]$. For estimation to become feasible, we need to impose some restrictions on what this functions looks like. We opt for 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 monontonicity; we need only that the aggregate dose response does not change sign. Inference in this setting can be broken down into two steps. The first step involves model selection: we need to identify which dose is equal to d_c. The second is estimation, where we estimate the ATET. To conduct proper inference, standard errors should account for uncertainty in both steps.

3.3 First Stage: Model Selection and Choosing d_c

We restrict our focus to Assumption A2.1 that the dose distribution is discrete, with doses taking the values $\{d_1, d_2, ..., d_J\}$. This implies a simple expression for the conditional expectation function

$$\mathbb{E}[\triangle Y_{i,t}|D_i = d_j] = \begin{cases} \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)], & \text{for } d_i \leq d_c \\ \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)] + \mathbb{E}[\mu_i(D_i)|D_i = d_j], & \text{for } d_i > d_c \end{cases}$$

We add the assumption that there are both treatment and control units, which makes this estimation procedure worthwhile:

A7 **Interior MED (Discrete):** $d_1 < d_c < d_J$

However, it is also worth emphasizing that if this assumption fails we will also be able to identify an MED in the boundary. If $d_1 = d_c$, then functionally no MED exists, as we find evidence of a dose response after just the first dose. If $d_J = d_c$, then all units are untreated, and the implied ATET estimate would be 0.

Under this assumption, there are $J - 2$ possible specifications to choose between, as d_c can take any value in the set $\{d_2, ..., d_{J-1}\}$. Denote each potential specification by the dose d that is asserted to equal d_c. Each specification is associated with a falsifiable hypothesis $H_0(d)$, defined by

$$H_0(d) : \text{For all } d', d'' < d, \mathbb{E}[\triangle Y_{i,t}|D_i = d'] = \mathbb{E}[\triangle Y_{i,t}|D_i = d'']$$

These hypotheses have two important properties. First, they are nested: if $H_0(d)$ is true, then $H_0(d')$ is true for all $d' < d$. Second, they demarcate the parameter of interest, d_c, as follows: $H_0(d)$ is true if and only if $d \leq d_c$, under the no crossing assumption. These properties allow us to utilize the procedure developed in Sales (2024) to choose d_c.

Consider an equality of means test for each hypothesis, letting $P(d)$ denote the p-value from each test. There are a large number of composite hypotheses that could be tested; our preferred method will be to compare the mean at every dose with the mean at d_1. If the null is true, $P(d) \sim \mathcal{U}(0, 1)$, and so $\mathbb{E}[P(d)] = 1/2$. If it is false, $P(d) \to_n 0$. By the plug-in principle, in the limit $\hat{d}_c$ solves

$$\arg\min_{d \in \mathcal{D}} \sum_{i:D_i \leq d} \left[P(D_i) - \frac{1}{2}\right]^2 + \sum_{i:D_i > d} [P(D_i)]^2 \tag{10}$$

Mallik et al. (2011) first proposed this method to identify d_c for a threshold dose response model with continuously distributed dose variable. Sales (2024) extends the method to test nested hypotheses, suggesting it is well suited for threshold estimation.

Figure 1 illustrates how this estimator functions. We consider a dose distribution with 100 values evenly spaced between 0 and 1 with $d_c = 0.5$. Figure 1a plots the limiting distribution of p-values: all differences to the right of d_c produce a p-value of 0, while all differences to the left of d_c constitute a draw from a $U[0,1]$ distribution. The lines consist of a "well estimated" stump function given by equation 10 that arrives at d_c. Unfortunately, this estimator will not be consistent for d_c. We are not aware of any estimator that is consistent for d_c when there are a finite number of doses and no parametric assumptions, though we do not have a formal impossibility result.

The figure illustrates that if we have an "unlucky" draw right to the left of the cutoff, with a p-value close to 0, we will estimate a threshold below the true d_c. Regardless of the number of observations around the threshold, since $P(d) \sim \mathcal{U}(0,1)$ in the limit, this uncertainty must persist. We nevertheless recommend this estimator because of two desirable properties. First, it always identifies a valid control group asymptotically. In the limit, $\hat{d}_c$ exceeds d_c with probability zero:

Proposition 3. *Suppose that assumptions A1, A2.1, A3, A4 (Parallel Trends), A5 (Minimum Effective Dose), and A6 (No Crossing) hold. Then,* $\mathbb{P}(\hat{d}_c > d_c) \to_n 0$

To see this, we can rearrange the problem in Equation 10 and verify it is equivalent to

$$\hat{d}_c := \arg\max_{d \in \mathcal{D}} \sum_{i:D_i \leq d} \left[P(D_i) - \frac{1}{4} \right] \tag{11}$$

In the limit, any dose above the threshold will have a p-value of 0, so the value of the objective function will decrease by $1/4$ for every such dose included.

Second, our simulations indicate that this estimator performs very well in the limit, even if it is not consistent. This asymptotic distribution of this estimator is difficult to derive analytically, but, since the limiting distribution of the p-values are known, simple to simulate. We run 100,000 simulations with 100 dose values where d_c is equal to 50. In each simulation, we draw a p-value from a $U[0,1]$ distribution for all doses at or below d_c and set this p-value equal to 0 for all doses above d_c. The estimated asymptotic distribution of $\hat{d}_c$ is plotted in Figure 1b. Our estimator identifies d_c correctly in around 70% of all simulations, is no more than one dose below d_c in

around 90% of simulations, and is no more than two doses below d_c in around 95% of simulations. We also verify that it never takes a value above d_c. So, while it is not consistent, it does take the true value of the parameter of interest for the large majority of simulations.

3.4 Second Stage: Estimating the ATET and Constructing Honest Standard Errors

From Proposition 2, a standard difference-in-differences estimator will identify the ATET if d_c is known. Following the plug-in principle, we utilize the same estimator using an estimator of d_c:

$$\widehat{\text{ATET}} = \hat{\mathbb{E}}[\Delta Y_{it} | d_i > \hat{d}_c] - \hat{\mathbb{E}}[\Delta Y_{it} | d_i \leq \hat{d}_c] \tag{12}$$

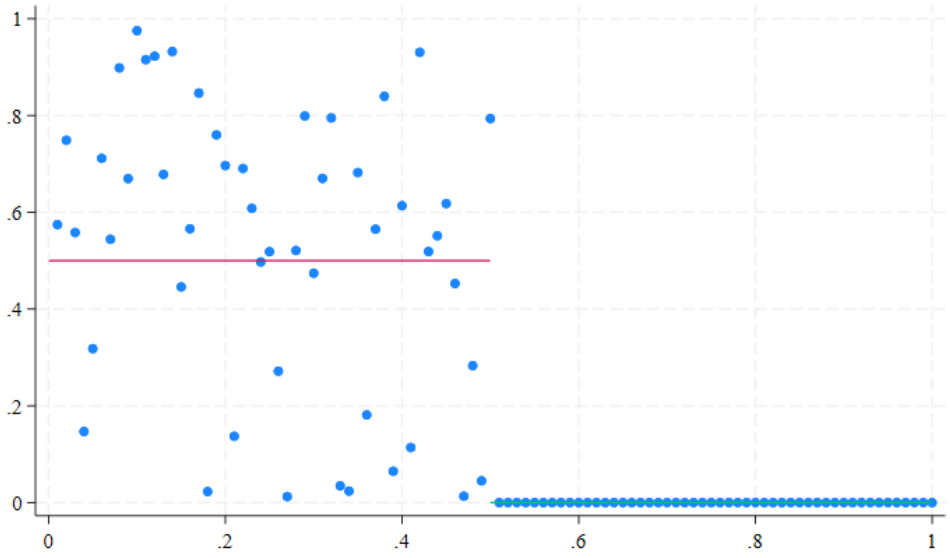
Where $\hat{d}_c$ is selected using the procedure outlined in the previous section. Since $\hat{d}_c$ is not a consistent estimator of d_c, we cannot use standard consistency arguments. However, since $\hat{d}_c$ is asymptotically conservative, we will always estimate the difference in means in the control group accurately. Occasionally, control units will be misclassified as treated, but the no crossing property will guarantee that this will only attenuate the estimator, which we establish formally:

Proposition 4. *Suppose assumptions A1, A2.1, A3, A4 (Parallel Trends), A5 (Minimum Effective Dose), A6 (No Crossing), and A7 (Interior MED) hold. Then the standard difference-in-differences estimator in* (12) *using the first-step estimator* $\hat{d}_c$ *from* (11) *is an attenuated estimator of the ATET asymptotically.*

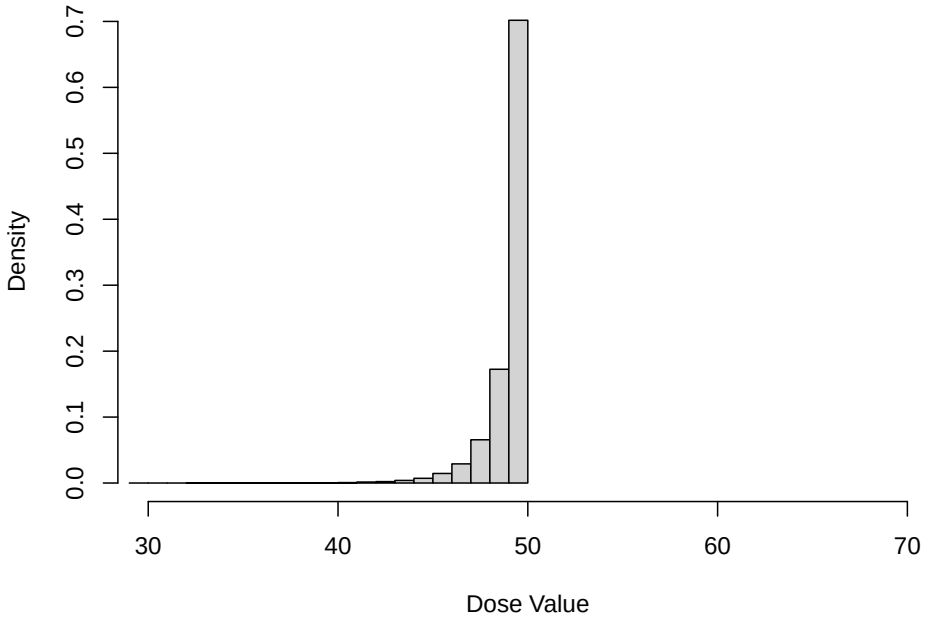
To improve finite sample performance, we use cross-fitting based on separate partitions of the data to estimate $\hat{d}_c$ and $\widehat{\text{ATET}}$. Following Chernozhukov et al. (2018), we separate the data into K equally sized groups $\{G_1, ..., G_K\}$. For each group G_k, we use the observations in its complement $G_k^C := G \setminus G_k$ for the model selection step to choose $\hat{d}_c$. In the second stage we estimate the ATET in G_k using $\hat{d}_c$ in the plug-in estimator given by (12). We repeat this for every group, obtaining a set of K estimates $\{\widehat{\text{ATET}}_k\}_{k=1,...,K}$ and average them to obtain the final estimate $\widehat{\text{ATET}}_{CF}$.

We expect this estimator to exhibit significant finite sample improvements over the two-sample procedure since it safeguards against spurious inference due to outliers in small samples by implementing the estimator across a broader range of the original data (Chernozhukov et al., 2018). Whenever $K = 2$, this constitutes a simple sample-splitting estimator, which uses each half of the original sample to estimate $\hat{d}_c$ and $\widehat{\text{ATET}}$.

Figure 1: Visualizing the Asymptotic Distribution



(a) p-value Distribution in the Limit



(b) Estimator distribution in the limit

Figure 1a plots an example of what the distribution of p-values should look like in the limit. The dose distribution is discrete uniform over $[0, 1]$ taking 100 values, where the minimum effective dose d_c is at 0.5. We plot a straight line at $1/2$ before the cutoff and at 0 after the cutoff to illustrate where the best fit line estimated in (10) and (11) should lie. Using the asymptotic distribution of this estimator, Figure 1b plots the proportion of times our estimator chooses each dose value as d_c in a simulation exercise.

3.4.1 Accounting for Uncertainty Over the Choice of d_c

To account for the uncertainty resulting from model selection as well as estimation, we use the bootstrap aggregation method and the corresponding standard errors proposed in Efron (2014). We first generate B bootstrap samples from our dataset, use each to produce an estimate of our parameter of interest $\widehat{\text{ATET}}_b$, then take the average of these to generate a point estimate:

$$\widehat{\text{ATET}}_B = \frac{1}{B}\sum_{b=1}^{B}\widehat{\text{ATT}}_b$$

As $\hat{d}_c$ changes, $\widehat{\text{ATET}}_b$ 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 $\widehat{\text{ATET}}_b$, generating a "lumpy" bootstrap distribution. Taking the average over the set of bootstrap samples corrects this by generating a smooth estimator in the presence of these discrete changes. 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

$$\hat{\sigma}_B = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{B}(W_{ij} - W_j)(\widehat{\text{ATET}}_b - \widehat{\text{ATT}}_B)/B}$$

And the standard interval $\widehat{\text{ATET}}_B \pm 1.96 \times \hat{\sigma}_B$ delivers a finite sample approximation to a 95% confidence interval for the ATET, which we test in the next section.

3.5 Simulation Evidence

To understand the finite sample properties of our proposed estimator, we run a Monte Carlo simulation with 100 repetitions. Our sample consists of $i = \{1, ..., N\}$ units observed across $j = \{1, ..., M\}$ dose values. The data generating process is given by

$$\triangle Y_{i,\tau} = \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)] + \mathbb{E}[\mu_i(D_i)|D_i]\mathbb{1}(D_i \geq d_c) + \varepsilon_i$$

For simplicity, we set $\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)] = 0$ and $\mathbb{E}[\mu_i(D_i)|D_i] = 1$. The dose distribution is bounded by $d_L = 0$ and $d_U = 1$, and the true threshold is set at $d_c = 0.5$. For each simulation, we draw

M doses, where the dose is spaced evenly across $[0, 1]$. For each unit i, we assign their outcome according to the DGP, where $\varepsilon_i \sim \mathrm{N}(0, 1)$.

It might seem strange to use a constant dose response function, as in many empirical settings the dose response function is assumed to be monotonic. If so, the performance of our estimator hinges on the first dose past the MED, which is the most difficult dose response to distinguish from 0. We view the constant dose response as equivalent to the lowest dose response in this setting. Additionally, since our method relies on pairwise comparisons, the distribution of the dose does not impact estimation outside of the number of observations at each dose. Thus, assuming the dose is evenly spaced in this setting is without loss of generality.

3.5.1 First Stage Estimation of d_c

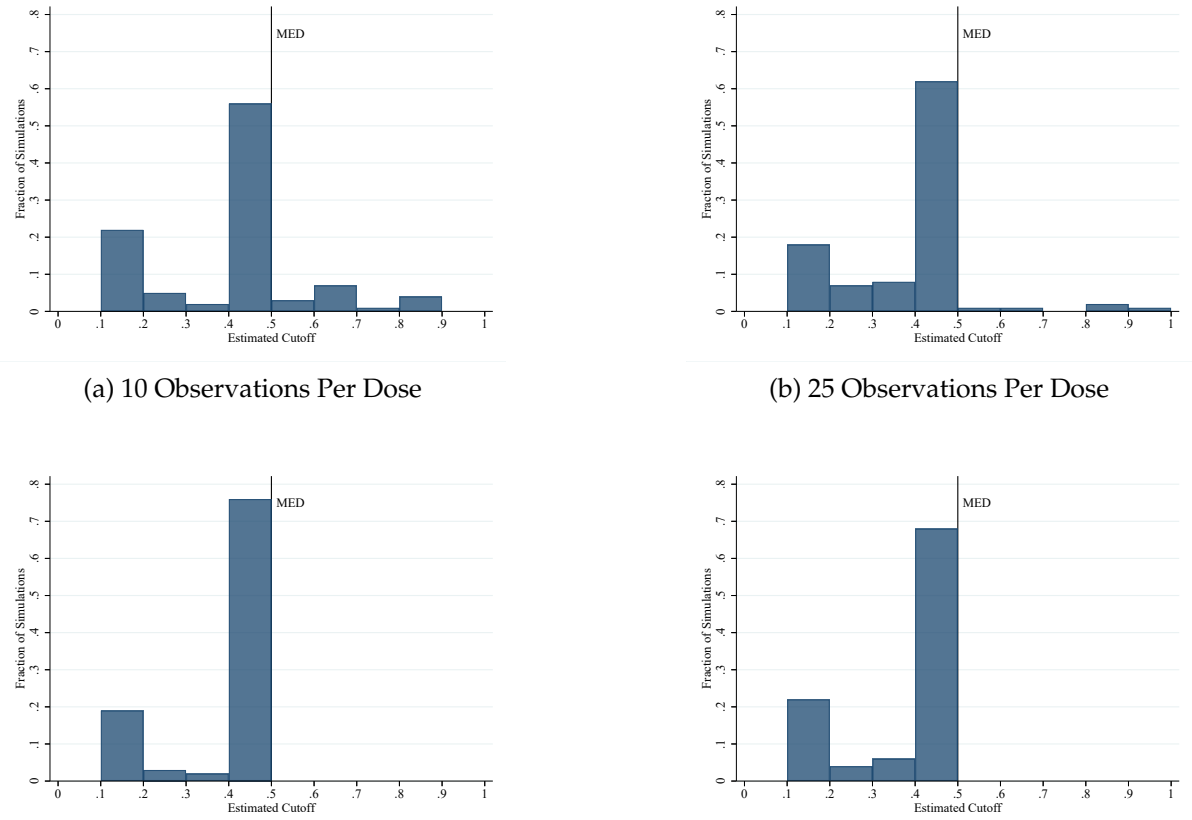
To begin, we run a series of simulations to understand the accuracy of our estimator of d_c. Our first Monte Carlo experiment considers only the model selection stage to elicit the finite sample properties of our proposed estimation procedure. We consider a relatively parsimonious set of 20 doses that are equidistant between 0 and 1. Figure 2 plots results for 10 (200), 25 (500), 50 (1,000), and 100 (2,000) observations at each dose (total observations). Note that the true number of observations per dose is half of this amount, as we are using a 2-fold cross-fit estimator.

For a low amount of observations per dose (10), we see that the result in Proposition 3 does not hold; the estimator sometimes chooses a dose value over the true threshold. However, these erroneous choices disappear at 50 observations per dose. We view this as strong evidence that the number of observations per dose required to identify a correct control group is relatively low.

3.5.2 Second Stage Estimation of ATET

Having verified the theoretical properties of our selection of $\hat{d}_c$, we turn to estimation of the ATET. These estimates arise from a cross-fitting procedure in which half of the sample is used to choose $\hat{d}_c$, and the remaining half utilizes this to estimate the ATET. We then flip the halves used for each part of the estimation and average these to arrive at an estimate, which we plot in Figure 3. For low numbers of observations, we see a large variance in estimates; however, once there are 50 observations per dose, our estimates are tightly distributed around the True ATET of 1. Note that, due to the result in Proposition 4, there is a leftward bias as the ATET will be attenuated.

Figure 2: Finite Sample Performance of Threshold Estimation



(a) 10 Observations Per Dose

(b) 25 Observations Per Dose

(c) 50 Observations Per Dose

(d) 100 Observations Per Dose

This figure plots the distribution of threshold estimates across 100 simulations. There are 20 equidistant dose values between 0 and 1, where 0.5 is the true cutoff, and the caption on each figure gives the number of observations per dose. See text for details on data generating process.

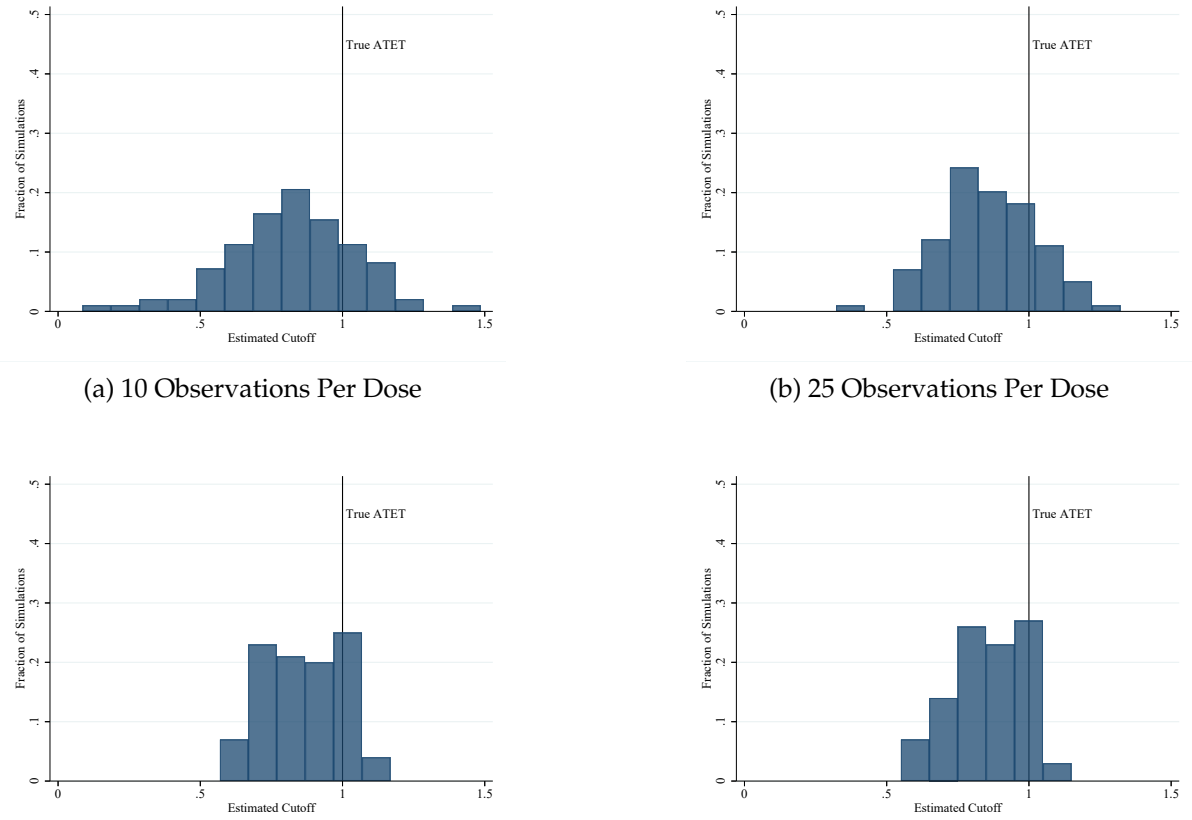
3.5.3 Standard Error Coverage

Table 2 below displays the coverage rate for the Smoothed Bootstrap Estimator of the standard error of the Split-Sample estimator, using 100 simulations and 500 bootstrap replications, across various parameter and sample size values. We find that at very low numbers of observations per dose (10 and 25) there is very poor coverage, likely reflecting the large mass of observations below the true ATET in Figure 3. However, this quickly improves at higher doses, and we recover proper coverage at 100 observations per dose, which is often met in applications.

Table 2: Bootstrap Coverage

M=10	M=25	M=50	M=100
0.040	0.150	0.880	0.980

Figure 3: Finite Sample Performance of ATET Estimation



(a) 10 Observations Per Dose

(b) 25 Observations Per Dose

(c) 50 Observations Per Dose

(d) 100 Observations Per Dose

This figure plots the distribution of ATT estimates across 100 simulations. There are 20 equidistant dose values between 0 and 1, where 0.5 is the true cutoff, and the caption on each figure gives the number of observations per dose. See text for details on data generating process.

4 Extensions

4.1 Continuous Case

We now study the case where Assumption A2.2 holds and there is a continuous dose distribution where no dose is observed more than one time. This clearly precludes the direct use of pairwise comparisons, but Mallik et al. (2011) develops an analogous procedure: instead of taking averages at each dose value, create a grid over the dose space and use standard kernel smoothing estimators to estimate local averages of the outcome. Since there is no first dose to use as a comparison group, we assume that $d_c > F^{-1}(\tau)$ and use $\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i < F^{-1}(\tau)]$ as our reference value.

Let $\hat{\mu}(d)$ denote the Nadayara-Watson estimator at some dose value $D_i = d$ of the change in outcome $\triangle Y$. With this estimator, we test the null hypothesis $H_{0.d} : \mu(d) = \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i < \tau]$

against the alternative $H_{1,d} : \mu(d) > \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i < \tau]$ using the test statistic

$$T(d, \tau_0) = \sqrt{n \cdot h_n} \left[\hat{\mu}(d) - \mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i < \tau]\right]$$

which converges in distribution to a mean-zero normal distribution with variance $\Sigma^2(d, \sigma) = \frac{\sigma^2(d)\overline{K}^2}{f(d)}$ under the null, where $\sigma^2(d)$ is an estimate of the standard error of the regression error term, $f(d)$ is the PDF value at d, and $\overline{K}^2$ is the integrated square of the kernel Li and Racine (2007).

Approximate p-values for this test are given by

$$\tilde{P}(D_i) = 1 - \Phi\left\{\frac{T(d, \hat{\tau}_n)}{\Sigma^2(d, \hat{\sigma})}\right\}$$

In practice, of course, $\mathbb{E}[\triangle Y_{i,\tau}(D_i, 0)|d_i < \tau]$ is not known, so we need to utilize an estimator $\hat{\mathbb{E}}[\triangle Y_{i,\tau}(D_i, 0)|d_i < \tau]$ as well as estimates of the unknown terms in Σ^2. Mallik et al. (2011) shows that the following estimator of the cutoff will converge to the threshold d_c:

$$\tilde{d}_C := \underset{d\in[0,1]}{\arg\max} \sum_{i:D_i \leq d} \left[\tilde{P}(D_i) - \frac{1}{2}\right]^2 + \sum_{i:D_i > d} \left[\tilde{P}(D_i)\right]^2$$

Suppose we use this estimator in a two step procedure. First, estimate the MED cutoff in a first stage using a random subset of the full data. Then, using the second half of the sample, estimate the *ATET* using a standard regression estimator. This comprises a standard two-step M-estimator, and under standard assumptions should allow identification of the *ATET* (Wooldridge, 2010). The difficulty with this approach lies in the speed of convergence of the first stage. Mallik et al. (2013) finds that this estimator converges at the rate $n^{-1/(2k+1)}$, where k is the cusp of the discontinuity of the dose response function at d_c. This is not an issue of having a poor estimator, as this meets the minimax rate for threshold estimation identified in Raimondo (1998). As a result, assuming the dose response function is continuous at d_c implies a convergence rate no faster than $n^{-1/3}$, implying the usual standard errors from a two-step M-estimator are not asymptotically valid.

A different approach, used in Hansen (2000), uses a Bonferroni correction to construct conservative but asymptotically valid confidence intervals. Threshold estimators generally have non-standard asymptotic distributions that cannot be estimated using a standard non-parametric bootstrap (Seijo and Sen, 2011). Mallik et al. (2013) derive asymptotically valid confidence intervals for the approach above, though they contain nuisance parameters that are difficult to estimate. Fur-

ther, while this estimator is consistent under any k, constructing valid standard errors for even the first stage requires the practitioner to make an assumption on what k is, without which the standard errors will not be valid. Even if this moment was known to the researcher, it would likely imply a very slow rate of convergence, leading to right bias unless the sample size was very large, as it is difficult to detect a smooth, small "liftoff" from the starting value.

4.2 Discretizing the Continuous Case

Accordingly, while this route is available to researchers, we recommend a different strategy for dealing with a continuous dose. This is motivated by the core result in Callaway et al. (2024) that the fundamental issues around dose comparisons can be solved with an untreated group.

Instead, we propose discretizing the dose space. Consider an arbitrary partition of the dose space $P_j \in \mathcal{P}$, where we choose a set of exclusive bins P_j that cover $\mathcal{D}$. We associate each of these bins with a pseudo-dose d_j, given by the average dose value within this bin, which is without loss of generality as we only consider indicators for each pseudo-dose. With this structure, we can conduct the same pairwise comparisons as before, but it is unclear what will result. Let P_j^c denote the bin that contains d_c. Given the no crossing property, we know that

$$\mathbb{E}[\triangle Y_{i,t}|D_i \in P_j] = \begin{cases} \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)], & \text{for } P_j \neq P_j^c, d_j < d_c \\ \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)] + \mathbb{E}[\mu_i(D_i)|D_i \in P_j], & \text{for } P_j = P_j^c \\ \mathbb{E}[\triangle Y_{i,\tau}(D_i,0)] + \mathbb{E}[\mu_i(D_i)|D_i \in P_j], & \text{for } P_j \neq P_j^c, d_j > d_c \end{cases}$$

Thus, we retain the same structure as before. However, given the previous results, we know that in the limit this procedure will select a bin strictly below P_j^c, which is the core tradeoff of this procedure. In essence, we are trading the right-bias from a procedure that converges slowly for left-bias from a procedure that eliminates treated observations very quickly.

Putting this all together, we propose the following procedure. First, split the sample into an model selection portion and an estimation portion. Beginning with the model selection sample, discretize the dose space as described above. The results in Sales (2024) require that the minimum number of observations across bins goes to infinity. Thus, we recommend using quantile-spaced bins to ensure that observations are evenly placed across bins. Calonico et al. (2015) provides a data-driven method to choose bins of this type that minimize within-bin integrated mean squared

error. Using these bins, the method described in the previous section can be utilized to estimate a group of untreated bins. Following this, in the estimation sample, we can utilize units estimated as controls as if they are untreated. This allows for the sophisticated method recommended by Callaway et al. (2024) to estimate the dose response through comparisons with untreated units. A similar bootstrap can be utilized to deliver proper standard errors on the second stage estimates.

4.2.1 Estimating BATTs

We can use a similar strategy to estimate any $ATT(d|d)$ of interest, in either the continuous or discrete case. Often times, different parts of the dose distribution are of interest to the researcher. Consider a design which discretizes at the median to compare units above and below this percentile. To understand the impact above, say, the 75th percentile, we might instead discretize at this quartile. This procedure correctly changes the treatment group, by considering all units above a certain researcher-define percentile. However, it is unclear why the comparison group includes units between the 50th and 75th percentile, which were treated in our previous comparison.

An extremely simple decomposition can illustrate the issue here. Consider a standard estimator discretized at the 75th percentile, where d_τ denotes the τth percentle of the dose distribution:

$$\hat{\mathbb{E}}[\Delta Y_{it}|d_i > d_{75}] - \hat{\mathbb{E}}[\Delta Y_{it}|d_i \leq d_{75}]$$

This can be rewritten as

$$\hat{\mathbb{E}}[\Delta Y_{it}|d_i > d_{75}] - (2/3)\hat{\mathbb{E}}[\Delta Y_{it}|d_i \leq d_{50}] - (1/3)\hat{\mathbb{E}}[\Delta Y_{it}|d_{75} \leq d_i \leq d_{50}])$$

Trivially, we can write the median estimator as

$$\hat{\mathbb{E}}[\Delta Y_{it}|d_i > d_{50}] - (2/3)\hat{\mathbb{E}}[\Delta Y_{it}|d_i \leq d_{50}] - (1/3)\hat{\mathbb{E}}[\Delta Y_{it}|d_i \leq d_{50}]$$

Differencing these terms, we arrive at

$$(2/3)(\hat{\mathbb{E}}[\Delta Y_{it}|d_i > d_{75}] - \hat{\mathbb{E}}[\Delta Y_{it}|d_i > d_{50}]) - (1/3)(\mathbb{E}[\Delta Y_{it}|d_{75} \leq d_i \leq d_{50}] - \hat{\mathbb{E}}[\Delta Y_{it}|d_i \leq d_{50}])$$

There are certainly reasons we might not trust the comparison in the first term. But comparisons of this form are certainly more muddled with the inclusion of the second term, which arises because

the control group in the first analysis is included as the treatment in the second.

To avoid this, we recommend separating the decision of which units should be considered treated and which should be controls. The treated group is merely a population of interest, and should be specified by the researcher. This could be the ATT at a specific dose or alternatively, the ATT over a range of doses, which we call the *Binned Average Treatment Effect on the Treated* (BATT)

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

By definition, then, the estimators above would be targeting BATT$[d_{50}, \infty]$ and BATT$[d_{75}, \infty]$, respectively, but they would not be consistent without a proper control group. The methods outlined in the previous section can be used to estimate this control group to recover identification.

5 Conclusion

In this paper, we construct a potential outcomes framework for empirical settings where all units are treated. We formalize baseline assumptions needed for inference - the existence of a minimum effective dose, which implies a subset of units experience untreated outcomes. Leveraging this assumption, we propose a non-parametric estimator of the ATET that first estimates the MED in a hold-out sample. We show in simulations that this estimator performs well at a low number of observations and that the bootstrap estimator of the standard error achieves proper coverage.

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.

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, November). 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*.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association 110*(512), 1753–1769.

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

Chernozhukov, V., A. D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

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

de Chaisemartin, C. and X. D'Haultfoeuille (2022). An optimal bandwidth for difference-in-difference estimation with a continuous treatment and an heterogeneous adoption design. hal-03873937ff.

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*.

Dunnett, C. W. and A. C. Tamhane (1992). A step-up multiple test procedure. *Journal of the American Statistical Association 87*(417), 162–170.

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

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

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica 64*(2), 413–430.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics 93*, 345–368.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*(3), 575–603.

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.

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.

Mallik, A., M. Banerjee, and B. Sen (2013). Asymptotics for p-value based threshold estimation in regression settings. *Electronic Journal of Statistics 7*, 2477–2515.

Mallik, A., B. Sen, M. Banerjee, and G. Michailidis (2011). Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika 98*(4), 887–900.

Miao, K., K. Li, and L. Su (2020). Panel threshold models with interactive fixed effects. *Journal of Econometrics 219*, 137–170.

Miao, K., L. Su, and W. Wang (2020). Panel threshold models with latent group structures. *Journal of Econometrics 214*, 451–481.

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

Raimondo, M. (1998). Minimax estimation of sharp change points. *The Annals of Statistics 26*(4), 1379–1397.

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.

Sales, A. C. (2024). Sequential specification tests to choose a model: A change-point approach. *Communications in Statistics - Theory and Methods 53*(12), 4354–4368.

Seijo, E. and B. Sen (2011). Change-point in stochastic design regression and the bootstrap. *Journal of the American Statistical Association 39*(3), 1580–1607.

Seo, M. H. and Y. Shin (2016). Dynamic panels with threshold effect and endogeneity. *Journal of Econometrics 195*, 169–186.

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.

Wang, Q. (2015). Fixed-effect panel threshold model using stata. *The Stata Journal 15*(1), 121–134.

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

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

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

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 $T_i = \mathbb{1}(d_i \geq d_r)$.

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

□

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 $\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0]$, where $T_i = \mathbb{1}(d_i \geq d_c)$.

$$
\begin{aligned}
\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0] \to_p\ & E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] \\
&= E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] \\
&= E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 1]
\end{aligned}
$$

$$= E[Y_{i,\tau}(D_i, 1)|D_i \geq d_c] - E[Y_{i,\tau}(D_i, 0)|D_i \geq d_c] = \text{ATT}$$

The second equality follows from the assumption of a minimum effective dose, and the third equality follows from the parallel trends assumptions. □

A.4 Proposition 3

This follows directly from Proposition 2 in Sales (2024).

A.5 Proposition 4

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 $T_i = \mathbb{1}(d_i \geq \hat{d}_c)$.

$$\begin{aligned}\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0] &= \hat{E}[\Delta Y_{i\tau}|d_i \geq \hat{d}_c] - \hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c] \\ &= (\hat{E}[\Delta Y_{i\tau}|d_i \geq \hat{d}_c] - E[\Delta Y_{i\tau}|d_i \geq d_c]) - (\hat{E}[\Delta Y_{i\tau}|d_i < d_c] - E[\Delta Y_{i\tau}|d_i < \hat{d}_c]) \\ &+ (E[\Delta Y_{i\tau}|d_i \geq d_c] - E[\Delta Y_{i\tau}|d_i < d_c])\end{aligned}$$

The second term is equal to the ATT, so we focus attention on the first term. Fix $\varepsilon > 0$. Then, by the law of total probability, we can write

$$\mathbb{P}(|\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c] - E[\Delta Y_{i\tau}|d_i < d_c]| > \varepsilon) = \sum_{d_j \in \mathcal{D}} \mathbb{P}(|\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c] - E[\Delta Y_{i\tau}|d_i < d_c]| > \varepsilon|\hat{d}_c = d_j)\mathbb{P}(\hat{d}_c = d_j)$$

If $d_j > d_c$, $\mathbb{P}(\hat{d}_c = d_j)$ goes to 0 in the limit by Proposition 3. If $d_j = d_c$, $\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c]$ converges to $E[\Delta Y_{i\tau}|d_i < \hat{d}_c]$ in probability by results in Proposition 2. If $d_j < d_c$, since $E[\Delta Y_{i\tau}|d_i < d_c] = E[\Delta Y_{i\tau}|d_i < d_j]$ $\forall d_j < d_c$, note that

$$\mathbb{P}(|\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c] - E[\Delta Y_{i\tau}|d_i < d_c]| > \varepsilon|\hat{d}_c = d_j) = \mathbb{P}(|\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c] - E[\Delta Y_{i\tau}|d_i < d_j]| > \varepsilon|\hat{d}_c = d_j)$$

This term converges to 0 in the limit as $\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c]$ converges in probability to $E[\Delta Y_{i\tau}|d_i < d_j]$. Taking this all together, we can take limits of the expression above to see that

$$\lim_{n\to\infty} \mathbb{P}(|\hat{E}[\Delta Y_{i\tau}|d_i < \hat{d}_c] - E[\Delta Y_{i\tau}|d_i < d_c]| > \varepsilon) = 0$$

Turning to the first part, we can utilize a similar decomposition:

$$\hat{E}[\Delta Y_{i\tau}|d_i \geq \hat{d}_c] = \sum_{d_j \in \mathcal{D}} \hat{E}[\Delta Y_{i\tau}|d_i \geq \hat{d}_c, \hat{d}_c = d_j]\mathbb{P}(\hat{d}_c = d_j) = \sum_{d_j \in \mathcal{D}} \hat{E}[\Delta Y_{i\tau}|d_i \geq d_j]\mathbb{P}(\hat{d}_c = d_j)$$

If $d_j > d_c$, $\mathbb{P}(\hat{d}_c = d_j)$ goes to 0 in the limit by Proposition 3. Note that $\hat{E}[\Delta Y_{i\tau}|d_i \geq d_j]$ converges in probability to $E[\Delta Y_{i\tau}|d_i \geq d_j]$, so we have that

$$\lim_{n\to\infty} \hat{E}[\Delta Y_{i\tau}|d_i \geq \hat{d}_c] = \sum_{d_j \leq d_c} E[\Delta Y_{i\tau}|d_i \geq d_j]\mathbb{P}(\hat{d}_c = d_j)$$

If $d_j = d_c$, this limit is equivalent to $E[\Delta Y_{i\tau}|d_i \geq d_c]$. If $d_j < d_c$, we can write

$$\begin{aligned} E[\Delta Y_{i\tau}|d_i \geq d_j] &= E[\Delta Y_{i\tau}|d_i \geq d_c]\mathbb{P}(d_i \geq d_c) + E[\Delta Y_{i\tau}|d_j \leq d_i \leq d_c]\mathbb{P}(d_j \leq d_i \leq d_c) \\ &= E[\Delta Y_{i\tau}|d_i \geq d_c]\mathbb{P}(d_i \geq d_c) + E[\Delta Y_{i\tau}|d_i = d_c]\mathbb{P}(d_j \leq d_i \leq d_c) \end{aligned}$$

Where the second line follows from the fact that d_c is the MED. Then,

$$\begin{aligned} \lim_{n\to\infty} \hat{E}[\Delta Y_{i\tau}|d_i \geq \hat{d}_c] - E[\Delta Y_{i\tau}|d_i \geq d_c] &= \sum_{d_j \leq d_c} E[\Delta Y_{i\tau}|d_i \geq d_j]\mathbb{P}(\hat{d}_c = d_j) - E[\Delta Y_{i\tau}|d_i \geq d_c] \\ &= \sum_{d_j \leq d_c} (E[\Delta Y_{i\tau}|d_i \geq d_j] - E[\Delta Y_{i\tau}|d_i \geq d_c])\mathbb{P}(\hat{d}_c = d_j) \\ &= \sum_{d_j < d_c} (E[\Delta Y_{i\tau}|d_i = d_c] - E[\Delta Y_{i\tau}|d_i \geq d_c])\mathbb{P}(d_j \leq d_i \leq d_c)\mathbb{P}(\hat{d}_c = d_j) \\ &= -\text{ATT} \times \sum_{d_j < d_c} \mathbb{P}(d_j \leq d_i \leq d_c)\mathbb{P}(\hat{d}_c = d_j) \\ &:= -\lambda\text{ATT} \end{aligned}$$

Where $\lambda = \sum_{d_j < d_c} \mathbb{P}(d_j \leq d_i \leq d_c)\mathbb{P}(\hat{d}_c = d_j) \in (0, 1)$.

Putting this all together,

$$\lim_{n\to\infty} \hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0] = \text{ATT}(1 - \lambda)$$

□

A.6 Linear Dose Example

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

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

B Appendix

B.1 Misspecification Analysis

The results in this section are based on a misspecification analysis where we explore alternative assumptions that might underly the researcher design to use some threshold d_r to create a treatment and control group. Throughout this section, we assume that a minimum effective dose d_c exists but is unknown to and ignored by the researcher. Broadly, the hope is that, even without knowledge of d_c, throwing away dose variation can nevertheless recover an attenuated version of the ATET. However, we show that this is generally not true, and only holds under restrictive assumptions on the heterogeneity of the dose response function across individuals. To this end, we introduce a formal assumption of homogeneity

A7 **Dose Response Function Homogeneity:** Dose response functions are the same across units and given by $\mu_i(D_i) = \mu(D_i) \forall i$.

In the simplest case, the dose response function is identical for all units, and the dose response function is constant, so that $\mu(D_i) = \beta$. In this case, even without knowledge of d_c, a binned design will recover an attenuated version of the ATT

Proposition 5. *Suppose assumptions A1-A4 hold, and that the dose response function is homogeneous (A7 holds) and constant, so that $\mu_i(D_i) = \beta$. Then the binned estimator under an arbitrary cutoff d_r gives*

$$\widehat{b}_1^{BIN} \to^p ATT \times min \left\{ \frac{1 - F(d_c)}{1 - F(d_r)}, \frac{F(d_c)}{F(d_r)} \right\}$$

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 $T_i = \mathbb{1}(d_i \geq d_r)$.

$$\begin{aligned}
&\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0] \\
&= E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] \\
&= (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0]) \\
&- (E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 0]) \\
&= (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 1]) \\
&- (E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0] - E[Y_{i,\tau}(D_i, 0) - Y_{i,\tau-1}(D_i, 0)|T_i = 0]) \\
&= E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau}(D_i, 0)|T_i = 1] - E[Y_{i,\tau}(D_i, 1) - Y_{i,\tau-1}(D_i, 0)|T_i = 0]
\end{aligned}$$

$$= E[\mu_i(D_i)\mathbb{1}(D_i \geq d_c)|T_i = 1] - E[\mu_i(D_i)\mathbb{1}(D_i \geq d_c)|T_i = 0]$$
$$= \beta(P(D_i \geq d_c|T_i = 1) - P(D_i \geq d_c|T_i = 0))$$

There are two possibilities.

Case 1: $d_r > d_c$. In this case, $P(D_i \geq d_c|T_i = 1) = 1$ and $P(D_i \geq d_c|T_i = 0)) = [F(d_r) - F(d_c)]/F(d_r)$. So, this term simplifies to $\beta F(d_c)/F(d_r)$.

Case 2: $d_c > d_r$. In this case, $P(D_i \geq d_c|T_i = 1) = [1-F(d_c)]/[1-F(d_r)]$ and $P(D_i \geq d_c|T_i = 0)) = 0$. So, this term simplifies to $\beta[1 - F(d_c)]/[1 - F(d_r)]$.

To conclude the proof, note that $d_r > d_c$ implies that $F(d_c)/F(d_r) < 1 < [1 - F(d_c)]/[1 - F(d_r)]$ and that $d_c > d_r$ implies that $[1 - F(d_c)]/[1 - F(d_r)] < 1 < F(d_c)/F(d_r)$. □

We find that the researcher guess leads to an attenuated estimate of the ATET (β), akin to the result of classical measurement error. If the researcher guess is too low ($d_r < d_c$), the intuition is very straightforward. Some of the "treatment" units are actually untreated - by definition, they receive a treatment effect of 0 - and as a result, the estimated treatment effect is a mixture of β and 0. If the researcher guess is too high ($d_c > d_r$), the result is more convoluted. Some of the "control" units are actually treated, and this contamination works to attenuate the estimated treatment effect. In addition, "treatment" units are a higher dose subset (above d_r) of the true treated group (above d_c); since $\mu_i(D_i)$ does not depend on the dose (or individual), this conditioning does not impact the average difference in outcomes in the "treated" group, and we obtain the symmetric result in Proposition 5.

We can also replicate this result if dose assignment is randomized:

A8 **Random Dose Assignment:** Dose assignment is independent of the observed dose response so that $D_i \perp \mu_i(D_i)$

Proposition 6. *Suppose that assumptions A1-A4 hold, and that the dose is assigned randomly (A8 holds). Then, the binned estimator under an arbitrary cutoff point d_r will give*

$$\widehat{b}_1^{BIN} \to^p ATT \times min\left\{\frac{1 - F(d_c)}{1 - F(d_r)}, \frac{F(d_c)}{F(d_r)}\right\}$$

Proof. Under identical arguments to that in Proposition 5, we know that

$$\begin{aligned}
&\hat{E}[\Delta Y_{i\tau}|T_i = 1] - \hat{E}[\Delta Y_{i\tau}|T_i = 0] \\
&\to_P E[\mu_i(D_i)\mathbb{1}(D_i \geq d_c)|T_i = 1] - E[\mu_i(D_i)\mathbb{1}(D_i \geq d_c)|T_i = 0] \\
&= E[\mu_i(D_i)|T_i = 1]E[\mathbb{1}(D_i \geq d_c)|T_i = 1] - E[\mu_i(D_i)|T_i = 0]E[\mathbb{1}(D_i \geq d_c)|T_i = 0] \\
&= E[\mu_i(D_i)](E[\mathbb{1}(D_i \geq d_c)|T_i = 1] - E[\mathbb{1}(D_i \geq d_c)|T_i = 0])
\end{aligned}$$

The result follows by the same arguments that conclude Proposition 5. □

B.1.1 Example: (ATT)enuation Bias at Worst?

The attenuation result breaks down if we begin to relax any part of the assumption of a uniform binary dose response function, in which case the average difference in outcomes above d_r is not identical to that for units above d_c. To see this, consider an example with a binary dose response function for all units, where the level of the response might vary across units; that is, $\mu_i(D_i) = \beta_i$. For simplicity, suppose the dose distribution is $U[0, 1]$, the treatment cutoff $d_c = 1/2$ for all units, and dose response functions are given by

$$\mu_i(D_i) = \begin{cases} 0 & \text{if } D_i < 1/2 \\ 6 \times \mathbb{1}(D_i \geq 1/2) & \text{if } D_i \in [1/2, 3/4) \\ 24 \times \mathbb{1}(D_i \geq 1/2) & \text{if } D_i \in [3/4, 1) \end{cases} \tag{1}$$

In this example, all units above $d_c = 1/2$ are treated, and the ATET is 15, which will be estimated properly at the correct cutoff $d_c = 1/2$. Now, we can consider alternate researcher cut-off choices. Table 3 shows the estimated treatment effect on the treated, $\widehat{\text{ATET}}$, as a function of the researcher choice of threshold. For a guess that is too low (e.g. $1/4$), we will always recovered an attenuated treatment effect, as the researcher treatment group is mixing all treated units with some fraction of the control units. However, for a guess that is too high (e.g. $3/4$), we are biased upwards, as higher dose units have a larger (flat) dose response. This bias occurs because, unlike in Proposition 5, the average change in outcomes above $3/4$ is higher than the average change in outcomes above $1/2$. This is because the dose response β_i is positively correlated with the dose D_i.

If we have a heterogeneous dose response function and we are not willing to assume that dose

Table 3: Binned Estimator Under Selection

d_r	$[0, 1/4)$	$[1/4, 1/2)$	$[1/2, 3/4)$	[3/4,1)	$\widehat{\text{ATET}}$
1/4	Control	Treatment	Treatment	Treatment	10
1/2	Control	Control	Treatment	Treatment	15
3/4	Control	Control	Control	Treatment	22

assignment is random, the above example illustrates that even with a very simple (binary) heterogeneous dose response function our estimator will have an unknown bias. Unfortunately, this issue cannot be dealt with even with the assumption of a homogenous dose response function. To see this, consider the following functional form

$$\mu(D_i) = \begin{cases} 0 & \text{if } D_i < 1/2 \\ 6 & \text{if } D_i \in [1/2, 3/4) \\ 24 & \text{if } D_i \in [3/4, 1) \end{cases} \tag{2}$$

This gives rise to a set of observed outcomes identical to that in (1), an as a result, the $ATET$ estimates in Table 3 will remain the same. While there may be a set of dose response functions that only admit attenuation bias, it is troubling that an example that only takes on two values can lead to bias of an unknown sign.

B.2 Full Dose Regression

70% of papers estimate a full dose regression, which is equivalent to the bivariate regression

$$\Delta Y_{i,t} = \alpha + \beta \cdot D_i + \Delta \epsilon_{i,t} \tag{3}$$

As Callaway et al. (2024) point out, interpreting $\hat{\beta}$ in this model is difficult and varies across practitioners. Of particular interest for our results is equation (3.1) which states

$$\hat{\beta} \to^p \frac{\mathbb{E}[w_1^{bin}(D_i)\Delta Y_{it} | D_i > \mathbb{E}[D_i]] - \mathbb{E}[w_0^{bin}(D_i)\Delta Y_{it} | D_i < \mathbb{E}[D_i]]}{\mathbb{E}[w_1^{bin}(D_i) D_i | D_i > \mathbb{E}[D_i]] - \mathbb{E}[w_0^{bin}(D_i) D_i | D_i < \mathbb{E}[D_i]]}$$

where the weights are given by

$$w_1^{bin}(d) = \frac{|d - \mathbb{E}[D_i]|}{\mathbb{E}\left[|D_i - \mathbb{E}[D_i]|\middle| D_i > \mathbb{E}[D_i]\right]}$$
$$w_0^{bin}(d) = \frac{|d - \mathbb{E}[D_i]|}{\mathbb{E}\left[|D_i - \mathbb{E}[D_i]|\middle| D_i \leq \mathbb{E}[D_i]\right]}$$

The full dose estimator recovers a Wald-like parameter, taking the difference between a weighted average of the outcome above and below the mean, and scaling by the average distance in doses between these two groups. The key issue for interpretation is that $\hat{\beta}$ is not invariant to scaling of the dose variable, through its impact on the denominator. In the appendix, we show it is always possible to apply a transformation to the dose distribution so the denominator equals one

$$\hat{\beta}^{scaled} \to_p \mathbb{E}[w_1^{bin}(D_i)\Delta Y_{it}|D_i > \mathbb{E}[D_i]] - \mathbb{E}[w_0^{bin}(D_i)\Delta Y_{it}|D_i < \mathbb{E}[D_i]]$$

The full dose regression in (3) is thus similar to a binned design where the cutoff is chosen to be the mean of the dose distribution. Scaling does not change "treatment" and "control" groups, nor does it affect the weights. See appendix for details. Up to a linear transformation of the dose, these estimators only vary in their choice of weights.