Difference-in-Differences with a Continuous Treatment*

Brantly Callaway† Andrew Goodman-Bacon‡ Pedro H. C. Sant'Anna§

First draft on arXiv: July 6, 2021. This draft: December 31, 2025

Abstract

This paper analyzes difference-in-differences designs with a continuous treatment. We show that treatment-on-the-treated-type parameters are identified under a parallel trends assumption analogous to the binary treatment case. However, comparing these parameters across treatments is challenging because parallel trends does not rule out selection bias. We discuss alternative, typically stronger, assumptions that eliminate selection bias. We further show that popular two-way fixed effects estimands admit multiple interpretations, depending on the underlying causal building block, all having important limitations as meaningful summaries of treatment effects. Finally, we introduce alternative estimation procedures that avoid these drawbacks and demonstrate them in an empirical application.

JEL Codes: C14, C21, C23

Keywords: Difference-in-Differences, Continuous Treatment, Multi-Valued Discrete Treatment, Parallel Trends, Two-way Fixed Effects, Multiple Periods, Variation in Treatment Timing, Treatment Effect Heterogeneity

*We thank the participants of many seminars, workshops, and conferences for their comments. We are grateful to Xiaohong Chen for numerous discussions about implementing the data-driven sieve estimator used in this paper, Amy Finkelstein for sharing her data with us, Carol Caetano, Greg Caetano, Stefan Hoderlein, Jo Mullins, Jon Roth, and Abbie Wozniak for their comments, and Honey Batra for valuable research assistance. Code implementing the methods proposed in the paper is available in the R package `contdid`, which is available on CRAN. The views expressed here are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. The Supplementary Appendix is available here.

†University of Georgia. Email: brantly.callaway@uga.edu

‡Federal Reserve Bank of Minneapolis and NBER. Email: andrew@goodman-bacon.com

§Emory University. Email: pedro.santanna@emory.edu

1 Introduction

The canonical difference-in-differences (DiD) research design compares outcomes before and after treatment started (difference one), between treated and untreated groups (difference two). However, in many DiD applications, the treatment does not simply "turn on", it has a "dose" or operates with varying intensity. Pollution dissipates across space, affecting locations near its source more severely than faraway locations. Localities spend different amounts on public goods and services, or set different minimum wages. Students choose how long to stay in school.

Continuous treatments can offer advantages over binary ones.[1] Variation in intensity makes it possible to evaluate treatments that all units receive. A clear "dose-response" relationship between outcomes and treatment intensity can bolster the case for a causal interpretation or test a theoretical prediction. Finally, we may care more about the effect of changes in treatment intensity, such as increased funding, pollution abatement, or expanded eligibility, than about the effect of the existence of a treatment that already exists.

Despite how conceptually useful and practically common continuous DiD designs are, currently available econometric results provide little guidance on applying and interpreting them, except in some specific cases. In this paper, we introduce a set of tools that are suitable for DiD setups with variation in treatment dosage. In particular, we (a) discuss how one can identify a variety of treatment effect parameters by exploiting parallel-trends-type assumptions, (b) demonstrate that a simple linear two-way fixed effects (TWFE) estimand accommodates *multiple* decompositions that are difficult to justify as meaningful summaries of treatment effects, and (c) propose "forward-engineered" estimators that directly target well-defined causal objects, allowing for more transparent interpretation and robust inference in applications with treatment effect heterogeneity.

To foster intuition and simplify exposition, we start by discussing causal parameters in a two-period DiD design in which units move from no treatment to a non-zero dose. We call the difference between a unit's potential outcome under dose d and its untreated potential outcome a *level treatment effect*. We call the difference in a unit's potential outcome with a marginal increase in the dose a *causal response* (Angrist and Imbens, 1995). When treatment is binary, these two notions of treatment effects coincide, but they do not under a continuous treatment. Importantly, level treatment effects and causal responses can have meaningfully different interpretations, and we establish that they generally require different identifying assumptions as well.

Comparisons between treated and untreated units identify average (level) treatment effect parameters under a parallel trends assumption on untreated potential outcomes, similar to binary DiD designs. Comparisons between adjacent dose groups, however, do not identify average causal response parameters under the "standard" parallel trends assumption. They include causal responses but are contaminated by an additional term involving possibly different treatment effects of the same dose for different dose groups—we refer to this additional term as selection bias.[2] We discuss an alterna-

[1] We generally use "continuous" treatments also to mean multi-valued ordered discrete treatments, but make the distinction explicit for certain results.

[2] In applications where units choose their amount of the treatment, it is natural to refer to this term as selection bias. In other applications where the dose measures a unit's amount of exposure to some treatment, a different term,

tive, typically stronger assumption, which we call strong parallel trends, that says that the average evolution of outcomes for the entire treated population under dose d is equal to the path of outcomes that dose group d actually experienced. Thus, strong parallel trends justifies comparing dose groups by restricting treatment effect heterogeneity. Strong parallel trends may not be plausible in many applications. Currently, in empirical work, it is common for papers to write as if they have assumed standard parallel trends and interpret their results as if they have assumed strong parallel trends. Our results clarify what causal questions can be answered under standard parallel trends and what causal questions require stronger assumptions.

The ideas discussed above are in the spirit of what Mogstad and Torgovitsky (2024) call *forward engineering*, where the researcher clearly specifies target parameters and assumptions up front and builds estimators to implement the identification strategy. Our second main contribution is to *reverse engineer* the most common way that practitioners estimate continuous DiD designs, which is to run a TWFE regression that includes time fixed effects (θ_t), unit fixed effects (η_i), and the interaction of a dummy for the post-treatment period ($Post_t$) with a variable that measures unit i's dose or treatment intensity, D_i:

$$Y_{i,t} = \theta_t + \eta_i + \beta^{twfe} D_i \cdot Post_t + v_{i,t}. \tag{1.1}$$

This TWFE specification is clearly motivated by DiD setups with two periods and two treatment groups, though many prominent textbooks suggest using it in more general setups (e.g., Cameron and Trivedi, 2005, Angrist and Pischke, 2008, and Wooldridge, 2010). There are several ways to interpret β^{twfe}, each corresponding to a different type of causal parameter. We decompose it in terms of level effects, scaled level effects, causal responses, and scaled high-versus-low (2×2) effects. Each decomposition is a weighted integral of dose-specific causal parameters, and none provides a clear causal and policy-relevant interpretation of β^{twfe}, at least not when treatment effects are allowed to vary across doses and/or groups.[3]

For instance, we show that β^{twfe} can be expressed as a weighted integral of average level treatment effect parameters but where the weights integrate to zero, indicating that β^{twfe} should not be interpreted as an average (level) treatment effect. Interestingly, however, TWFE puts negative weights on the below-average dose units and positive weights on above-average dose units, and, thus, after re-scaling by a weighted average of the difference between doses for high- and low-dose units, is equivalent to a weighted binary DiD using higher-dose units as the "treated" group and lower-dose units as the "comparison" group, with weights proportional to a unit's absolute distance from the mean dose. Our next decomposition, based on average level treatment effect parameters scaled by their dose, also displays negative weights, though their weights integrate up to one and not zero.

In contrast, a TWFE decomposition in terms of average causal response parameters has weights that integrate up to one and are non-negative, but also includes a selection bias term stemming from

such as "heterogeneity bias", could be more appropriate. For simplicity, throughout the paper, we simply refer to this term as selection bias.

[3]The decompositions that we provide are specific to the particular TWFE regression specification in Equation (1.1), which we focus on due to its ubiquity in empirical work with a continuous treatment. Some of the drawbacks we discuss below, particularly regarding weighting schemes inherited from the TWFE regression, could be addressed by considering a more flexible specification. See, e.g., Wooldridge (2025) for a discussion involving binary treatments.

effect heterogeneity across doses. The strong parallel trends assumption eliminates this selection bias. The weights on causal responses at different doses, however, differ from the distribution of the dose among the treated, which creates a further challenge to interpreting β^{twfe}, even if strong parallel trends holds. The weights are also undesirably sensitive to the size of the untreated group. In our application, if we drop the untreated group, which changes the weights but does not change the underlying average causal responses, our estimate of β^{twfe} shrinks by 78%. Our decomposition of β^{twfe} based on scaled 2×2 average effects as building blocks also highlights limitations of using β^{twfe} as a causal summary parameter.

We demonstrate that these drawbacks are easily avoidable and discuss different DiD estimators that build upon our identification results and recover interpretable causal parameters. When the treatment is discrete, this is as simple as running a linear regression with multiple treatment indicators, which is similar to staggered DiD setups (Callaway and Sant'Anna, 2021). When the treatment is continuous, there are several options, including adopting a parametric, semiparametric, or nonparametric regression model. In particular, we discuss how to adapt the data-driven sieve-based nonparametric regression proposed by Chen, Christensen, and Kankanala (2025) to our context, although we note that other semi/nonparametric procedures are also possible. We also show how to construct causal summary measures of our average treatment effect functions that bypass the TWFE weighting problems by using the dose density as weights. Our results suggest that one can easily summarize average level treatment effects among treated units by comparing the average change in outcomes for all treated units to the average change in outcomes for untreated units. This can be estimated by running a binary DiD with a "treatment dummy" equal to one for any units with positive doses. Summarizing average causal responses using dose density weights involves estimating an average derivative, which is simple to compute using "flexible" linear regressions. We also discuss how to construct event-study results using these summary measures, which can then be used to assess the plausibility of the parallel trends assumptions.

To contrast our proposed estimators with TWFE in practical settings, we revisit Acemoglu and Finkelstein's 2008 study of a 1983 Medicare reform that eliminated labor subsidies for hospitals. The original paper uses a TWFE estimator to compare the change in capital-labor ratios between hospitals whose input prices were more or less affected by the end of the subsidy. It concludes that price regulations favoring capital significantly increase capital use. The distinction between level treatment effect parameters and causal responses is important in this example: a positive level treatment effect shows that the policy as a whole increased the use of capital, while causal responses describe which subsidy levels generated the largest responses. We find that the reform raised capital-labor ratios by about 18 percent (on average), which is 50 percent larger than the comparable TWFE estimate because of the weighting issues highlighted by our decompositions. We also estimate variable average causal response ($ACRT$) parameters that are quite large at low subsidy levels—implying elasticities of substitution greater than 2—yet slightly *negative* for most positive doses. These negative $ACRT$ estimates cast doubt about the plausibility of the strong parallel trends assumption, the simple two-factor model of hospital production, or both. Our results support Acemoglu and Finkelstein's 2008 conclusion that the 1983 Medicare reform led hospitals to favor capital over labor, but suggest

caution in a policy interpretation about which subsidy levels have the largest effects or an economic interpretation in terms of production function parameters.

Related Literature: Our paper contributes to the literature on modern DiD methods; see, e.g., Baker et al. (2025) for an overview. We contribute to this literature by highlighting challenges associated with using TWFE with continuous treatments, discussing the role of different parallel trends assumptions to learn about different causal parameters, and providing easy-to-use estimation procedures that can highlight treatment effect heterogeneity with continuous treatments.

The closest paper to ours is Fricke (2017), which focuses on DiD setups with two time periods and three treatment dosages: H, L, and 0. Fricke (2017) shows that under standard parallel trends, one can identify dose-specific average treatment effect among dose groups in two-period DiD designs. He also considers stronger assumptions that permit causal interpretation of the H vs. L contrast when an untreated group is not available. We generalize his identification results to DiD settings with richer treatment distributions, including continuous cases, multiple time periods, and staggered treatment adoptions. This allows us to (i) discuss a broader set of parameters of interest that are suitable for incremental changes in treatment dose, (ii) discuss event-study and other types of treatment aggregations, (iii) derive decomposition results that question the causal meaning of TWFE estimates under treatment effect heterogeneity, and (iv) offer identification-based estimation templates for researchers to avoid the pitfalls of simple TWFE specifications with continuous/multi-valued treatments.

Our paper is also related to a series of papers on more complicated non-binary DiD setups. de Chaisemartin and D'Haultfoeuille (2018) focuses on fuzzy designs, where a researcher is interested in individual-level effects of a binary treatment that has been aggregated across units into a continuous "treatment rate." In contrast, we study "sharp" designs in which the treatment exposure is itself continuous or multi-valued discrete at the unit-level. The approach proposed in de Chaisemartin and D'Haultfoeuille (2025) can also accommodate continuous treatments, although they focus on aggregated target parameters rather than dose-specific estimators of treatment effect heterogeneity, as we do. They also do not discuss identification, nor estimation, related to average causal response parameters, an important focus of our analysis. de Chaisemartin et al. (2025) builds on de Chaisemartin and D'Haultfoeuille (2025) and considers a DiD setup with continuous treatments with potentially non-staggered (but static) treatments. Our proposal differs in its target parameters and DiD designs. Similarly to de Chaisemartin and D'Haultfoeuille (2025), de Chaisemartin et al. (2025) average effects of discrete rather than marginal changes of treatments. On the other hand, de Chaisemartin et al. (2025) allows for units to already be exposed to the treatment in the first period and considers instrumental variable extensions, which we do not.

Our decompositions also relate to the literature on TWFE bias in heterogeneous treatment effect settings with a binary treatment (e.g., Borusyak, Jaravel, and Spiess, 2024; de Chaisemartin and D'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021) but we extend this logic to continuous treatments and highlight that the same TWFE regression coefficient can have multiple interpretations depending on the "building blocks", and that new "bias" terms may appear, depending on the type of parallel trends assumption being used. A perhaps more practically relevant message

from our decompositions is that, even when all weights are non-negative, TWFE can still provide an unappealing causal summary parameter. Interestingly, if one replaces our DiD setting with a cross-sectional design with a randomly assigned dose, all four of our decomposition results would still hold, highlighting that linear specifications may not be desirable with continuous treatments, even when the dose is fully randomized.

Finally, we note that some of our causal response decomposition builds on Yitzhaki (1996, Proposition 2), which expresses the slope coefficient in a regression of an outcome on a continuous variable as a weighted average of underlying local slopes. Besides differences related to causal interpretations and panel data, we mildly extend those results to accommodate a mass of untreated units.

2 Motivating Continuous DiD from an Empirical Perspective

To fix ideas and provide intuition for our results, we revisit Acemoglu and Finkelstein's 2008 (AF) study of how price regulations affect firms' input choices. When Medicare began in 1965, hospitals received reimbursements from the federal government for a share of their labor and capital expenditures proportional to the fraction of total patient days accounted for by Medicare recipients (m_i). Hospital i thus faced input prices equal to $(1 - s_L m_i)w$ for labor and $(1 - s_K m_i)r$ for capital, where s_L and s_K are the labor and capital subsidy rates and w and r are market wages and rental rates. In 1983, Medicare moved to the Prospective Payment System (PPS), which replaced the labor subsidy with a small payment per episode/diagnosis. This set $s_L = 0$ but left the capital subsidy unchanged. Therefore, the price of labor for a given hospital rose from $(1 - s_L m_i)w$ to w, skewing relative factor prices.

The statutory relationship between a hospital's Medicare volume, m_i, and the change in its price of labor, $s_L m_i w$, motivates AF's use of a continuous DiD design comparing changes in capital/labor ratios before and after 1983 between hospitals with different pre-PPS Medicare inpatient shares. AF's description, estimation, and interpretation of this empirical strategy touch on some of the most common ways of justifying and implementing continuous DiD designs.

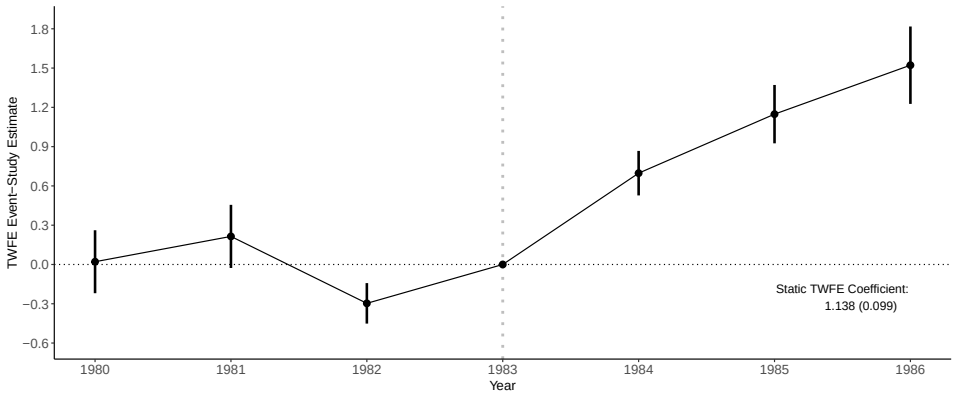
One motivation for this design is practical: variation in a dose (or exposure) permits the evaluation of treatments for which binary DiD is either infeasible or undesirable. In AF's case, about 15 percent of hospitals were "untreated" by the change in Medicare's subsidy policy because they served non-Medicare-eligible populations, like children or psychiatric patients, so one may be concerned about whether these constitute a valid comparison group. AF therefore describe m_i, which is the hospital's Medicare volume in 1983, as an "attractive source of variation" in the price of labor both because it varies substantially—the mean of m_i among treated hospitals is 0.45, and the standard deviation is 0.15—and because hospitals with $m_i > 0$ may be more comparable to each other than treated hospitals are to untreated hospitals.

Another common justification for continuous DiD designs is that a "dose-response" relationship between exposure and outcomes can support a causal interpretation or test a theoretical prediction. Meyer (1995, p. 158), for example, argues that "differences in the intensity of the treatment across different groups allow one to examine if the changes in outcomes differ across treatment levels in the

expected direction." AF lay out a simple theoretical framework in which the move to PPS should (i) raise capital/labor ratios and (ii) do so more strongly for hospitals with higher pre-PPS values of m_i. They view their continuous DiD design as a way to estimate a causal effect of PPS as a whole and test the theoretical predictions of their model.

Finally, researchers often advocate for continuous DiD designs because they can be used to estimate average causal effects of small changes in the dose. In many economic models, price and income elasticities determine optimal policies like tax rates, tax bases, subsidies, and regulations (Hendren, 2016), but these are continuous concepts that can be estimated accurately only with continuous variation. AF's theoretical framework implies that, under some assumptions, DiD estimates provide information about hospitals' elasticity of substitution between capital and labor, although AF do not argue for this kind of "marginal" interpretation.

Figure 1: Two-Way Fixed Effects Event-Study Estimates of the Effect of Medicare's Reimbursement Reform on Hospital Input Mix



Notes: The figure plots TWFE event-study coefficients and their 95% confidence intervals from regressions with hospital fixed effects, year fixed effects, and the 1983 Medicare inpatient share (m_i) interacted with either a dummy for years after 1983 (static TWFE specification) or the year dummies (event study). The outcome variable is the depreciation share of total operating expenses, a measure of hospitals' capital/labor ratio. The data cover the years 1980-1986 and come from the American Hospital Association's annual survey (American Hospital Association, 1986). The results are not numerically identical to AF's because we drop 860 hospitals (out of 6,741) with missing outcomes for some years.

In terms of estimation, AF use the standard tool for continuous DiD designs: a TWFE regression with hospital and year fixed effects. They follow textbook advice. Wooldridge (2010, p. 132) observes that a two-period DiD regression estimator "can be easily modified to allow for continuous, or at least non-binary, 'treatments'." Angrist and Pischke (2008, p. 234) emphasize "a second advantage of regression DD is that it facilitates the study of policies other than those that can be described by a dummy." They also follow common practice and describe their identifying assumption as an extension of the parallel trends assumption from binary designs: "*Without the introduction of PPS*, hospitals with different m_i's would not have experienced differential changes in their outcomes in the post-PPS period" (emphasis added).

Figure 1 reproduces AF's DiD event-study coefficients for each calendar year, relative to 1983, and the estimate of β^{twfe} from an equation like (1.1). AF interpret these results as indicative that after 1983, capital/labor ratios rose more strongly for hospitals with higher values of m_i, without a substantial differential change in input mix before PPS. Our impression is that event-study results

like those in Figure 1 would usually be interpreted as strong causal evidence because there are (relatively) small pre-trend estimates, large estimates in post-treatment periods, and tight confidence intervals. What is missing from most continuous DiD analyses, however, is a specific statement about *what* causal parameters researchers would like to estimate, the assumptions under which they are identified, and a formal justification for a particular estimator. Our goal is to shed light on these three issues.

3 Baseline Case: A New Treatment with Two Periods

We illustrate our main points in a setup with two periods of panel data, $t = 1$ and $t = 2$. In the first period, no unit is treated. In the second period, some units receive a treatment "dose," denoted by D_i, and others remain untreated. Extensions to multiple periods and staggered setups are discussed in Section 5. We denote the support of D by $\mathcal{D}$. D_i can be (absolutely) continuous or can be multi-valued discrete, but to simplify the exposition, we refer to it as "continuous." We define potential outcomes for unit i in period t as $Y_{i,t}(0, d)$, where potential outcomes are indexed by the treatment sequence (Robins, 1986). As we focus on the setup where all units have $d = 0$ in period $t = 1$, we simplify the potential outcome notation and henceforth write $Y_{i,t}(d)$, where d is the dosage in period $t = 2$. This is the outcome that unit i would experience in period t under (period-two) dose d. In each time period t, the observed outcome for unit i is $Y_{i,t} = Y_{i,t}(D_i)$. We assume that all expectations are finite and well-defined. Henceforth, we omit the unit index i to make the notation less cluttered and define $\Delta Y = Y_{t=2} - Y_{t=1}$.

3.1 Parameters of Interest with a Continuous Treatment

The potential outcomes notation $Y_t(d)$ reflects that treatment can take many values, and so each unit can experience many types of causal effects. The *level treatment effect* of dose d in time period t for a given unit is defined as its potential outcome when $D = d$ minus its untreated potential outcome: $Y_t(d) - Y_t(0)$. Level treatment effects measure the treatment effect at time t from switching treatment dosage from 0 to d. This is a straightforward extension of a binary "treatment effect" to a continuous "treatment effect function" or "dose-response function."

But zero-treatment is not the only relevant counterfactual. We define a unit's *causal response* at d as $Y_t'(d)$, the derivative of the potential outcome with respect to dose d (when the treatment is continuous),[4] or as the difference in potential outcomes between adjacent doses scaled by the difference in the doses, $(Y_t(d_j) - Y_t(d_{j-1}))/(d_j - d_{j-1})$ (when the treatment is discrete). Causal responses measure the treatment effect at time t of a marginal increment of dose d. These two types of treatment effects—the level of $Y_t(d) - Y_t(0)$ or its slope, $Y_t'(d)$—define unit-level causal parameters in continuous designs, and connect to results in the instrumental variables (IV) literature on multi-valued discrete or continuous endogenous variables (Angrist and Imbens, 1995, Angrist, Graddy, and Imbens, 2000).

[4] This is a slight abuse of notation as we do not require $Y_t(d)$ to be differentiable (or even continuous), but rather we mean here the causal effect of a marginal increase in the dose on a unit's outcome: $\lim_{h \to 0^+} (Y_t(d + h) - Y_t(d))/h$.

We focus on "building block" parameters that are averages of these two kinds of causal effects in the post-treatment period, $t = 2$. Average level treatment effects (which we refer to as average treatment effects) extend definitions from the binary case:

$$ATT(d|d') = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)|D = d'] \quad \text{and} \quad ATT(d) = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)|D > 0],$$

where $ATT(d|d')$ is the average effect of dose d compared to zero dosage in the post treatment period $t = 2$ on units that actually experienced dose d'. When $d' = d$, this is the ATT local to units that received dose d. $ATT(d)$ is the average difference between potential outcomes under dose d relative to untreated potential outcomes across all treated units, not just those that experienced dose d, in time period $t = 2$.

Average causal response parameters for absolutely continuous treatments are defined as

$$ACRT(d|d') = \frac{\partial ATT(l|d')}{\partial l}\bigg|_{l=d} = \frac{\partial \mathbb{E}[Y_{t=2}(l)|D = d']}{\partial l}\bigg|_{l=d} \text{ and } ACRT(d) = \frac{\partial ATT(d)}{\partial d} = \frac{\partial \mathbb{E}[Y_{t=2}(d)|D > 0]}{\partial d}.$$

$ACRT(d|d')$ is the average effect of a marginal dose increase from d for dose group d'. It equals the derivative of $ATT(l|d')$ with respect to l, evaluated at $l = d$, which is equivalent to the derivative of the $t = 2$ average potential outcome function with respect to dose d among dose group d'. $ACRT(d)$ is the average causal response of dose d across all treated units. For discrete treatments, average causal responses are defined in a similar way but with slightly different notation:

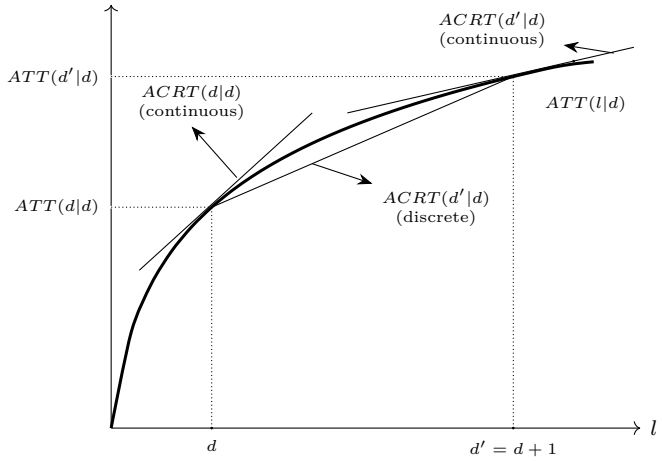
$$ACRT(d_j|d_k) = \frac{\mathbb{E}[Y_{t=2}(d_j) - Y_{t=2}(d_{j-1})|D = d_k]}{d_j - d_{j-1}} \quad \text{and} \quad ACRT(d_j) = \frac{\mathbb{E}[Y_{t=2}(d_j) - Y_{t=2}(d_{j-1})|D > 0]}{d_j - d_{j-1}}.$$

$ACRT(d_j|d_k)$ equals the difference in mean potential outcomes between dose level d_j and the next lowest dose d_{j-1} in period $t = 2$ for dose group d_k, scaled by the difference between the two doses. Similarly, $ACRT(d_j)$ gives the average causal response of dose d_j relative to dose d_{j-1}, but it is for the entire treated group. We note that our definition of $ACRT(d_j|d_k)$ and $ACRT(d_j)$ differs from the definitions in Angrist and Imbens (1995), as it scales the changes in expected potential outcomes by the change in dosage.

Figure 2 illustrates these parameters graphically. The concave line plots an average treatment effect function against the dose for dose group d, $ATT(\cdot|d)$. If we consider dose levels d and d', there are two possible ATT parameters. The first, $ATT(d|d)$, the level of dose group d's average treatment effect function at d, is an average treatment effect that is "local" to units that experienced dose d. The second, $ATT(d'|d)$, is also "local" to dose group d, but refers to the effect they would experience at dose d' even though they did not actually receive that dose. The continuous-dose $ACRT$ parameters are the slopes of tangent lines to the $ATT(\cdot|d)$ function, and the discrete-dose $ACRT$ parameters are the slopes of lines connecting two points on the $ATT(\cdot|d)$ function. As with ATT's, our definitions encompass causal responses to doses other than the one a group actually receives (i.e., $ACRT(d'|d)$).

A proper interpretation of continuous DiD results hinges on which type of parameter one wants, can identify, and can estimate. For instance, even if all $ATT(d|d)$ parameters are large and positive, some $ACRT(d|d)$ parameters could be zero or negative. A researcher misinterpreting a large ATT estimate as an $ACRT$, in this case, would mistakenly conclude that a policy to raise every unit's dose would have large effects. A researcher confusing a small $ACRT$ for an ATT would mistakenly conclude that a policy was ineffective, even though it actually just has small effects at the margin.

Figure 2: Causal Parameters in a Continuous Difference-in-Differences Design



Notes: The figure plots $ATT(\cdot|d)$ (the average effect of experiencing each dose for dose group d). We highlight causal parameters for two doses, d and d'. $ATT(d|d)$ and $ATT(d'|d)$ are average treatment effect on the treated parameters and refer to the height of the curve. $ACRT(d|d)$ and $ACRT(d'|d)$ are average causal response parameters and refer to the slope of the curve. We show them for a continuous dose, when the $ACRT$ is the slope of a tangent line, and for a discrete dose when $ACRT$ is the slope of a line connecting two discrete points on $ATT(d|d)$.

The above-mentioned causal parameters are functional parameters because they are allowed to vary arbitrarily across dose groups d and across (counterfactual) doses d'. This contrasts with β^{twfe} from (1.1), which is a single number. In many applications, it may be desirable to aggregate these functional parameters into lower-dimensional objects that are easier to report and may be more precisely estimated. We focus on aggregations that average the functional parameters discussed above using the distribution of the dose among all treated units,

$$ATT^{\texttt{loc}} = \mathbb{E}[ATT(D|D)|D > 0] \qquad \text{and} \qquad ATT^{\texttt{glob}} = \mathbb{E}[ATT(D)|D > 0]$$
$$ACRT^{\texttt{loc}} = \mathbb{E}[ACRT(D|D)|D > 0] \qquad \text{and} \qquad ACRT^{\texttt{glob}} = \mathbb{E}[ACRT(D)|D > 0].$$

These provide natural ways to summarize the underlying parameters. We use the `loc` superscript to denote that $ATT^{\texttt{loc}}$ and $ACRT^{\texttt{loc}}$ summarize treatment effects that are local effects of particular doses, while we use the superscript `glob` to denote that $ATT^{\texttt{glob}}$ and $ACRT^{\texttt{glob}}$ summarize treatment effects of particular doses globally (i.e., across all treated units). All four of these parameters provide "best" approximations in the sense of minimizing the mean squared distance between the summary parameter and the functional parameters. Also, note that $ACRT^{\texttt{loc}}$ and $ACRT^{\texttt{glob}}$ are average derivative-type parameters, and average derivatives have been widely studied in econometrics.

3.2 Identification with a Continuous Treatment

This section discusses the identification of average treatment effect and average causal response parameters. Toward this end, we make the following assumptions.

Assumption 1 (Random Sampling)**.** *The observed data consist of* $\{Y_{i,t=2}, Y_{i,t=1}, D_i\}_{i=1}^n$*, which is independent and identically distributed.*

Assumption 2 (Treatment)**.** *In period* $t = 1$*, no unit is treated, while in period* $t = 2$*, the treatment dosage* D *has support* $\mathcal{D} = \{0\} \cup \mathcal{D}_+$*, where* $\mathcal{D}_+ \subseteq (0, \infty)$*. In addition,* $\mathbb{P}(D = 0) > 0$*.*

Assumption 3 (No-Anticipation and Observed Outcomes)**.** *For all units, and all* $d \in \mathcal{D}$*,*

$$Y_{i,t=1} = Y_{i,t=1}(d) = Y_{i,t=1}(0) \quad \textit{and} \quad Y_{i,t=2} = Y_{i,t=2}(D_i).$$

Assumption 1 says that we observe two periods of *iid* panel data. Assumption 2 formalizes that a mass of units do not participate in the treatment in either period (we discuss the case with no untreated units in more detail at the end of this section), and the rest receive a positive amount of the treatment that can vary in amount across units. Assumption 3 says that units do not anticipate future treatments, so we observe untreated potential outcomes for all units in the first period. In the second period, we observe the potential outcome corresponding to the actual dose that unit i experienced.

3.2.1 Identification under parallel trends

Identification of average level treatment effects follows closely from the DiD setup with binary treatments. In particular, our results rely on an extension of the binary parallel trends assumption.

Assumption PT (Parallel Trends)**.** *For all* $d \in \mathcal{D}_+$*,*

$$\mathbb{E}[Y_{t=2}(0) - Y_{t=1}(0)|D = d] = \mathbb{E}[Y_{t=2}(0) - Y_{t=1}(0)|D = 0].$$

Assumption PT says that the average evolution of outcomes that units with any dose d would have experienced without treatment is the same as the evolution of outcomes that units in the untreated group actually experienced. Binary DiD designs also rely on assumptions like this. To simplify the exposition below, we often simply refer to Assumption PT as *parallel trends* (PT). The following result shows that under Assumption PT, $ATT(d|d)$ is identified; all proofs are in Appendix A.

Theorem 3.1. *Under Assumptions 1, 2, 3, and PT,* $ATT(d|d)$ *is identified for all* $d \in \mathcal{D}_+$*, and it is given by*

$$ATT(d|d) = \mathbb{E}[\Delta Y|D = d] - \mathbb{E}[\Delta Y|D = 0].$$

Furthermore, $ATT^{loc} = \mathbb{E}[\Delta Y|D > 0] - \mathbb{E}[\Delta Y|D = 0]$*.*

Theorem 3.1 states that $ATT(d|d)$ equals the difference between the change in outcomes for dose group d and the untreated group. It generalizes Fricke (2017)'s result for two doses to richer treatment patterns. As a direct consequence, by averaging all the $ATT(d|d)$'s over the distribution of non-zero dosages, we have that the summary parameter $ATT^{\texttt{loc}}$ is identified by simply comparing units with a positive dose to untreated units. On the other hand, parallel trends, as defined in Assumption PT, is *not* strong enough to guarantee the identification of $ATT(d)$.

The identification of average causal response parameters differs from the identification of ATT parameters because it requires comparisons between dose groups.

Assumption 4 (Continuous or Multi-Valued Discrete Treatment)**.** *The treatment is either continuous or multi-valued discrete. More precisely, one of the following is true:*

(a) $\mathcal{D}_+ = \mathcal{D}_+^c$*, where* $\mathcal{D}_+^c = (d_L, d_U)$ *with* $f_{D|D>0}$ *a Lebesgue density which satisfies* $f_{D|D>0}(d) > 0$ *for all* $d \in \mathcal{D}_+^c$*, and* $\mathbb{E}[\Delta Y|D = d]$ *is differentiable on* $\mathcal{D}_+^c$*.*

(b) $\mathcal{D}_+ = \mathcal{D}_+^{mv}$ *where* $\mathcal{D}_+^{mv} \subset \mathbb{N}_+$*, with* $\mathbb{N}_+ = \{1, 2, 3, \ldots\}$ *denotes the strictly positive natural numbers. Let* d_j *denote the* j^{th} *element of* $\mathcal{D}_+^{mv}$*. In addition,* $\mathbb{P}(D = d) > 0$ *for all* $d \in \mathcal{D}_+^{mv}$*.*

Assumption 4 distinguishes between cases with a continuous 4(a) or discrete 4(b) treatment. Assumption 4(a) allows for the smallest value of the treatment to be arbitrarily close to zero or strictly larger than zero, both of which are common in applications.

Our central identification result is that causal response parameters are not identified under Assumption PT, because comparisons between different dose groups are biased when treatment effects (of the same dose) vary across dose groups, even when the average evolution of untreated potential outcomes is the same.

Theorem 3.2. *Under Assumptions 1, 2, 3, and PT, comparisons of paths of outcomes among different dose groups recover a mix of causal effect parameters and selection bias terms. Specifically,*

(a) For $(h, l) \in \mathcal{D}_+ \times \mathcal{D}_+$*,*

$$\begin{aligned}\mathbb{E}[\Delta Y|D = h] - \mathbb{E}[\Delta Y|D = l] &= ATT(h|h) - ATT(l|l) \\ &= \underbrace{\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D = h]}_{\textit{causal effect}} + \underbrace{\Big(ATT(l|h) - ATT(l|l)\Big)}_{\textit{selection bias}}.\end{aligned}$$

(b) If Assumption 4(a) also holds, then, for $d \in \mathcal{D}_+^c$*,*

$$\frac{\partial \mathbb{E}[\Delta Y|D = d]}{\partial d} = \frac{\partial ATT(d|d)}{\partial d} = ACRT(d|d) + \underbrace{\frac{\partial ATT(d|l)}{\partial l}\Big|_{l=d}}_{\textit{local selection bias}};$$

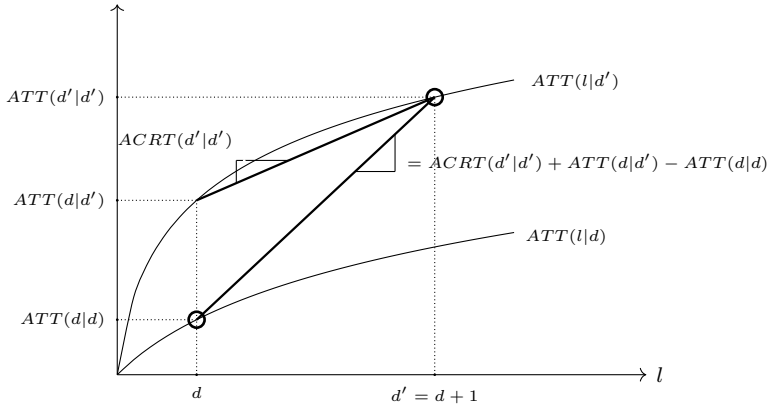
(c) Alternatively, if Assumption 4(b) also holds,

$$\frac{\mathbb{E}[\Delta Y|D = d_j] - \mathbb{E}[\Delta Y|D = d_{j-1}]}{d_j - d_{j-1}} = ACRT(d_j|d_j) + \underbrace{\frac{ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})}{d_j - d_{j-1}}}_{\textit{scaled selection bias}}.$$

Theorem 3.2 says that under parallel trends, comparisons of outcome paths between higher- and lower-dose groups mix together (i) causal responses and (ii) a "selection bias" type of term that comes from differences in average treatment effects of the same dose for different dose groups. Intuitively, even if untreated potential outcomes evolve in the same way, observed paths of outcomes differ between dose groups for two reasons. One is the causal response itself, which comes from differences in doses (h versus l) causing differences in outcomes. The other is a selection bias type of contamination, which comes from differences across dose groups in the average level effect of the particular dose l—parallel trends does not rule out that different dose groups could experience different treatment effects of the same dose.

Figure 3 illustrates this result for an example with two dose groups and two doses: d and $d' = d+1$. The slope of the line that connects the points $(d, ATT(d|d))$ and $(d', ATT(d'|d'))$ is steeper than the average causal response of interest, $ACRT(d'|d')$, because it jumps from one ATT function to the other. This is captured by the selection bias term, a version of selection-on-gains that equals the difference in treatment effects at the lower dose: $ATT(d|d') - ATT(d|d)$. It breaks the causal interpretation because observed outcomes for lower-dose units are not a valid counterfactual for what higher-dose units would have experienced at the lower dose. The selection bias is not identified

Figure 3: Non-Identification of Average Causal Response with Treatment Effect Heterogeneity, Two Discrete Doses



Notes: The figure shows that comparing adjacent $ATT(d|d)$'s equals an $ACRT$ parameter (the slope of the higher-dose group's ATT function) and selection bias (the difference between the two groups' ATT functions at the lower dose).

as we do not observe $Y_{t=2}(d)$ for units that experienced dose d'. Such a result precludes a causal interpretation of $ATT(d|d)$ differences across doses under Assumption PT.

3.2.2 Identification under strong parallel trends

This section discusses an alternative, typically stronger assumption that allows for the identification of $ACRT(d)$ and $ATT(d)$ parameters, which we refer to as *strong parallel trends* (SPT).

Assumption SPT (Strong Parallel Trends)**.** *For all $d \in \mathcal{D}$,*

$$\mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D > 0] = \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D = d].$$

Under Assumption 3, the right-hand side of the equation in Assumption SPT is the (observed) average evolution of outcomes for dose group d. Assumption SPT says that the average evolution of outcomes for the entire treated population if all experienced dose d (the left-hand side of the previous equation) is equal to the path of outcomes that dose group d actually experienced. In applications where the treatment is binary, Assumption SPT, like Assumption PT, reduces to the usual parallel trends assumption. Like the case with a binary treatment, it allows for treated units to select into being treated. Among treated units, though, it rules out selection into a particular amount of the treatment. With more complicated treatments, Assumption SPT notably differs from Assumption PT because it involves potential outcomes under different doses, $Y_t(d)$, rather than only untreated potential outcomes, $Y_t(0)$. While Assumption SPT is not strictly stronger than Assumption PT (e.g., notice that it does not require parallel trends in untreated potential outcomes for all dose groups), we refer to it as *strong parallel trends* to indicate that in many applications it would be a stronger, perhaps much stronger, assumption.

An alternative way to think about Assumption SPT is as an assumption that restricts treatment effect heterogeneity.[5] In particular, if one maintains Assumption PT, Assumption SPT is equivalent

[5]There are some instances of versions of strong parallel trends implicitly being discussed in empirical work.

to assuming that $ATT(d|d) = ATT(d)$ for all doses. This condition can also be viewed as a structural assumption in the sense that it effectively allows one to extrapolate treatment effects of dose d among dose group d to treatment effects of dose d for the entire treated population.

In the remainder of this section, we show that Assumption SPT is useful for recovering "global" average causal effect parameters, which are straightforward to compare to each other, and, hence, sidestep the selection bias issues discussed above. Before doing that, it is worth mentioning that we are not proposing Assumption SPT as an assumption that empirical researchers should readily adopt; in fact, in many applications, Assumption SPT may be a strong or implausible assumption. Rather, our aim is to clarify that many natural target parameters in DiD applications with a continuous treatment require stronger assumptions than the parallel trends as defined in Assumption PT.

Theorem 3.3. *Under Assumptions 1, 2, 3, and SPT,*

(a) For $d \in \mathcal{D}_+$, it follows that

$$ATT(d) = \mathbb{E}[\Delta Y|D = d] - \mathbb{E}[\Delta Y|D = 0].$$

(b) For $(h, l) \in \mathcal{D}_+ \times \mathcal{D}_+$,

$$\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D > 0] = ATT(h) - ATT(l) = \mathbb{E}[\Delta Y|D = h] - \mathbb{E}[\Delta Y|D = l]$$

(c) When Assumption 4(a) holds (i.e., treatment is continuous), it follows that, for $d \in \mathcal{D}_+^c$,

$$ACRT(d) = \frac{\partial \mathbb{E}[\Delta Y|D = d]}{\partial d}$$

(d) When Assumption 4(b) holds (i.e., treatment is discrete), it follows that

$$ACRT(d_j) = \frac{\mathbb{E}[\Delta Y|D = d_j] - \mathbb{E}[\Delta Y|D = d_{j-1}]}{d_j - d_{j-1}}$$

For part (a) of Theorem 3.3, recall that $ATT(d|d)$ and $ATT(d)$ differ when there is selection into dose group d on the basis of treatment effects. Strong parallel trends rules out that kind of selection, which means that comparing average outcome changes of dose group d to the untreated group identifies $ATT(d)$. Part (b) says that comparisons of the average change in outcomes over time for different dose groups have a causal interpretation under Assumption SPT. For parts (c) and (d), strong parallel trends ensures that each dose group d serves as a valid counterfactual for the entire treated population under that specific dose d, and, hence, that causal response parameters are identified under Assumption SPT.

Strong parallel trends only changes the interpretation of the estimand, not its form. One important implication is that conventional pre-tests for differential changes across groups before treatment cannot distinguish between Assumption PT and Assumption SPT. That is, because only untreated

Chodorow-Reich, Nenov, and Simsek (2021, p. 1636)'s cross-region study of marginal propensities to consume (MPC) notes the possibility of finding a zero even when the MPC>0 in all areas: "if low wealth areas have high MPCs and high wealth areas have low MPCs, an increase in the stock market could induce the same change in spending in both low and high wealth areas." Similarly, Saez, Slemrod, and Giertz (2012, p. 25) discuss a more restrictive version of strong parallel trends in the context of estimating the elasticity of taxable income for two groups facing different positive tax changes: "if the control group faces a tax change, difference-in-differences estimates will be consistent only if the elasticities are the same for the two groups."

potential outcomes are observed before treatment under Assumption 3, these periods cannot test the additional content of an assumption like SPT that necessarily involves treated potential outcomes.

Finally, the identification results in Theorem 3.3 immediately imply that averages of the $ATT(d)$ and $ACRT(d)$ building blocks are identified as well. The following corollary states these results.

Corollary 3.1. *Under Assumptions 1, 2, 3, and SPT,*

(a) It follows that

$$ATT^{glob} = \mathbb{E}[\Delta Y|D>0] - \mathbb{E}[\Delta Y|D=0].$$

(b) When Assumption 4(a) holds (i.e., treatment is continuous), it follows that

$$ACRT^{glob} = \mathbb{E}\left[\left.\frac{\partial \mathbb{E}[\Delta Y|D=d]}{\partial d}\right|_{d=D} \middle| D>0\right] = \int_{d_L}^{d_U} \left.\frac{\partial \mathbb{E}[\Delta Y|D=d]}{\partial d}\right|_{d=s} f_{D|D>0}(s)ds.$$

(c) When Assumption 4(b) holds (i.e., treatment is multi-valued), it follows that

$$ACRT^{glob} = \sum_{j=1}^{J} \left(\frac{\mathbb{E}[\Delta Y|D=d_j] - \mathbb{E}[\Delta Y|D=d_{j-1}]}{d_j - d_{j-1}}\right) \mathbb{P}(D=d_j|D>0).$$

These results highlight how identification in continuous DiD designs is fundamentally a question about dose-specific building block parameters and the underlying parallel trends assumption, not the aggregation choices that lead to particular summary parameters.

Remark 3.1 (No untreated units)**.** *Researchers often use continuous designs when all units in their sample receive some amount of the treatment, having in mind comparing units that are "more treated" to units that are "less treated". Without untreated units, it is infeasible to compare dose group d to an untreated group, and, hence, it is infeasible to directly recover $ATT(d|d)$ or $ATT(d)$. However, a natural alternative is to compare dose group d to dose group d_L (the lowest possible amount of the treatment). In Appendix SD.1 in the Supplementary Appendix, we show that, under parallel trends, when there are no untreated units,*

$$\mathbb{E}[\Delta Y|D=d] - \mathbb{E}[\Delta Y|D=d_L] = ATT(d|d) - ATT(d_L|d_L).$$

This shows that this comparison is related to underlying causal effect parameters under parallel trends; however, recall from Theorem 3.2 that the expression on the right-hand side mixes together the average causal response of moving from d_L to d with selection bias. Under strong parallel trends, we have instead that

$$\mathbb{E}[\Delta Y|D=d] - \mathbb{E}[\Delta Y|D=d_L] = ATT(d) - ATT(d_L) = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(d_L)|D>0],$$

which does not include selection bias terms. This discussion highlights that (unlike a setting with a binary treatment) continuous variation in the dose can be used to learn about causal effects even if there is no untreated comparison group, but interpreting these results as causal effects of the treatments requires strengthening Assumption PT. See also Fricke (2017) for a related discussion.

3.3 What Parameter Does TWFE Estimate?

Empirical researchers using continuous DiD designs typically estimate a single summary parameter using a linear TWFE regression like Equation (1.1). This section links the TWFE estimand to

our identification results for dose-specific parameters, describes the assumptions necessary to give TWFE *some* causal interpretation, and discusses what that interpretation is. We focus on continuous treatments and defer the discussion of multi-valued discrete treatments to Appendix SD.3 in the Supplementary Appendix.

Our impression is that empirical researchers typically interpret β^{twfe} in three main (and related) ways, implicitly relying on different building blocks. First, β^{twfe} is often directly interpreted as a causal response parameter; that is, how much the outcome causally increases on average when the treatment increases by one unit. This is the causal version of how regression coefficients are often taught to be interpreted in introductory econometrics classes. Second, it is common to pick a representative value for d, to report $d \times \beta^{twfe}$, and interpret this quantity as $ATT(d)$. This is the main interpretation provided in Acemoglu and Finkelstein (2008): "Given that the average hospital has a 38 percent Medicare share prior to PPS, this estimate [i.e., of β^{twfe}, here equal to 1.129] suggests that in its first 3 years, the introduction of PPS was associated with an increase in the depreciation share of about 0.42 ($\approx 1.129 \times 0.38$) for the average hospital." Rearranging this expression shows that under this interpretation $\beta^{twfe} = ATT(d|d)/d$, which relates β^{twfe} to a scaled level effect. Third, it is common to take two different representative values of the dose, d_1 and d_2—a common choice is the 25th percentiles and 75th percentiles of the dose—and interpret β^{twfe} as the average causal response of moving from dose d_1 to dose d_2 scaled by the distance between d_1 and d_2; this is a scaled 2×2 effect. We aim to assess whether such types of interpretations are justified and under which conditions.

Table 1: TWFE Decomposition Weights

Decomposition	$D > 0$ Weights	$D = 0$ Weights
Causal response	$w_1^{\text{acrt}}(l) = \frac{(\mathbb{E}[D\|D \geq l] - \mathbb{E}[D])\mathbb{P}(D \geq l)}{\text{Var}(D)}$	$w_0^{\text{acrt}} = \frac{(\mathbb{E}[D\|D > 0] - \mathbb{E}[D])\mathbb{P}(D > 0)d_L}{\text{Var}(D)}$
Levels	$w_1^{\text{lev}}(l) = \frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	$w_0^{lev} = -\frac{\mathbb{E}[D]\mathbb{P}(D = 0)}{\text{Var}(D)}$
Scaled levels	$w^{\text{s}}(l) = l\frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	
Scaled 2×2	$w_1^{2\times 2}(l, h) = \frac{(h - l)^2 f_D(h) f_D(l)}{\text{Var}(D)}$	$w_0^{2\times 2}(h) = \frac{h^2 f_D(h)\mathbb{P}(D = 0)}{\text{Var}(D)}$

Notes: The table provides the formulas for the weights used in the decompositions of β^{twfe} provided in this section.

The next proposition presents our decompositions of β^{twfe} under parallel trends (Assumption PT) and under strong parallel trends (Assumption SPT). The decompositions differ on the basis of the underlying building block parameters: causal response parameters ($ACRT(d|d)$ and $ACRT(d)$), level treatment effect parameters ($ATT(d|d)$ and $ATT(d)$), scaled level effects ($ATT(d|d)/d$ and $ATT(d)/d$), or scaled 2×2 effects ($\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D = h]/(h - l)$ and $\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D > 0]/(h - l)$). These building blocks are connected with the dose-parameters discussed in Section 3.2 and how empirical researchers interpret β^{twfe}; see Appendix SD.2 in the Supplementary Appendix for additional decompositions based on different building blocks. The weights attached to each of these decompositions are presented in Table 1.

Theorem 3.4. *Under Assumptions 1, 2, 3, 4(a), and PT, β^{twfe} can be decomposed in the following ways:*

(a) Causal Response Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{acrt}(l) \left(ACRT(l|l) + \underbrace{\frac{\partial ATT(l|h)}{\partial h}\Big|_{h=l}}_{\text{selection bias}} \right) dl + w_0^{acrt} \frac{ATT(d_L|d_L)}{d_L}$$

where the weights are always positive and integrate to 1.

(b) Levels Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{lev}(l) ATT(l|l)\, dl,$$

where $w_1^{lev}(l) \lesseqgtr 0$ for $l \lesseqgtr \mathbb{E}[D]$, and $\int_{d_L}^{d_U} w_1^{lev}(l)\, dl + w_0^{lev} = 0$.

(c) Scaled Levels Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w^s(l) \frac{ATT(l|l)}{l}\, dl,$$

where $w^s(l) \lesseqgtr 0$ for $l \lesseqgtr \mathbb{E}[D]$, and $\int_{d_L}^{d_U} w^s(l)\, dl = 1$.

(d) Scaled 2×2 Decomposition

$$\beta^{twfe} = \int_{d_L}^{d_U} \int_{\mathcal{D}, h>l} w_1^{2\times 2}(l,h) \left(\underbrace{\frac{\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h]}{h-l}}_{\text{causal response}} + \underbrace{\frac{ATT(l|h) - ATT(l|l)}{h-l}}_{\text{selection bias}} \right) dh\, dl$$
$$+ \int_{d_L}^{d_U} w_0^{2\times 2}(l) \frac{ATT(l|l)}{l}\, dl,$$

where the weights $w_1^{2\times 2}$ and $w_0^{2\times 2}$ are always positive and integrate to 1.

If one imposes Assumption SPT instead of Assumption PT, then the selection bias terms from Part (a) and Part (d) become zero, and the remainder of the decompositions remain true, except one needs to replace $ACRT(l|l)$ with $ACRT(l)$ in Part (a), $ATT(l|l)$ with $ATT(l)$ in Parts (b), (c) and (d), and $\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h]$ with $\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D>0]$ in Part (d).

Theorem 3.4 shows that the same TWFE estimand yields very different decomposition results, depending on the type of parallel trends used and the particular causal parameter employed as a building block for the analysis. Despite these multiple possible decompositions, one important feature that arises from Theorem 3.4 is that the weighting associated with any of the decompositions of β^{twfe} has some undesirable properties, making β^{twfe} an unappealing causal summary parameter in DiD setups with continuous treatments. Yet, each of these different decompositions highlights distinct concerns, as we discuss below.

Theorem 3.4(a) shows that when causal responses are taken as the building blocks of the analysis, under Assumption PT, β^{twfe} is equal to a convex weighted average of $ACRT(d|d)$ and the same selection bias derived in Theorem 3.2.[6] The sign of this selection bias depends on how treatment

[6]Part (a) is mechanically related to the results in Yitzhaki (1996) on interpreting linear projection coefficients with a continuous regressor when the conditional expectation may be nonlinear. Part (a) also includes a term that shows how TWFE handles a discrete jump from 0 to the minimum treated dose, d_L. Paths of outcomes are not observed for doses below d_L, but the scaled ATT for dose group d_L, $ATT(d_L|d_L)/d_L$, is averaged into β^{twfe}.

effects vary across dose groups at a given dose. If units in higher dose groups would have had larger positive treatment effects at every dose, for example, then β^{twfe} will be larger than the weighted average of the $ACRT(d|d)$'s that appear in Theorem 3.4(a). Figure 3 illustrates this case for two groups. Strong parallel trends eliminates the selection bias term, but does not affect the weights.

Even under strong parallel trends, the particular interpretation of β^{twfe} in terms of $ACRT(d)$'s hinges on the aggregation embodied in the weights $w_1^{acrt}(d)$. Because $w_1^{acrt}(d)$ is positive and integrates to 1, under Assumption SPT β^{twfe} is *weakly causal* (Blandhol, Bonney, Mogstad, and Torgovitsky, 2025). However, it does not estimate a natural target parameter like $ACRT^{\texttt{glob}}$ because the TWFE weights do not generally equal the dose distribution among treated, $f_{D|D>0}(d)$. Interestingly, the weights $w_1^{acrt}(l)$ underlying β^{twfe} depend on the entire distribution of the dose, making it sensitive to the size of the untreated group. This property is rather unappealing. For example, in our application, if we drop the untreated group (dropping the untreated group does not alter the underlying average causal responses), our estimate of β^{twfe} shrinks by 78%. Instead of letting the estimation method implicitly summarize the $ACRT$'s, we recommend that researchers choose these aggregation schemes explicitly. In our view, a natural and econometrically-guided way to aggregate the $ACRT$'s into a summary parameter is given by $ACRT^{\texttt{glob}}$, which is identified (as indicated in Corollary 3.1) and can also be easily estimated.

Part (b) expresses β^{twfe} as a weighted integral of $ATT(d|d)$ under parallel trends with weights that integrate to zero rather than one. Therefore, some weights are negative, and, hence, β^{twfe} is not weakly causal when $ATT(d|d)$ is taken as the building block. More significantly, β^{twfe} puts the same amount of negative weight on $ATT(d|d)$'s for doses below $\mathbb{E}[D]$ as it does positive weight on $ATT(d|d)$'s for doses above $\mathbb{E}[D]$. One way to view this result is that TWFE uses above-average dose units as an "effective treated group" and below-average dose units as an "effective comparison group" that potentially includes some treated units. While the cumulative positive weights and negative weights are equal to each other, they do not generally integrate to one within these groups, which means that β^{twfe} does not equal the difference between a weighted average of outcome paths for the effective treated group relative to the effective comparison group. In Appendix SD.2 in the Supplementary Appendix, however, we bridge this gap and derive a corollary of the result in Part (b) that makes the scaling issue related to this interpretation explicit and allows us to express β^{twfe} as the following weighted Wald-estimand:

$$\beta^{twfe} = \frac{\mathbb{E}\Big[w_1^{bin}(D)\Delta Y\Big|D > \mathbb{E}[D]\Big] - \mathbb{E}\Big[w_0^{bin}(D)\Delta Y\Big|D < \mathbb{E}[D]\Big]}{\mathbb{E}\Big[w_1^{bin}(D)D\Big|D > \mathbb{E}[D]\Big] - \mathbb{E}\Big[w_0^{bin}(D)D\Big|D < \mathbb{E}[D]\Big]}. \tag{3.1}$$

The numerator of Equation (3.1) shows that β^{twfe} compares weighted average outcome changes above and below $\mathbb{E}[D]$ with weights proportional to how far a unit's dose is from $\mathbb{E}[D]$.[7] The denominator scales this comparison by the same weighted difference in D. This representation highlights some challenges of using β^{twfe} to summarize the average level-effect of a continuous treatment. First, while

[7]The exact expressions for the weights are $w_1^{bin}(d) = \frac{|d-\mathbb{E}[D]|}{\mathbb{E}\Big[|D-\mathbb{E}[D]|\Big|D>\mathbb{E}[D]\Big]}$ and $w_0^{bin}(d) = \frac{|d-\mathbb{E}[D]|}{\mathbb{E}\Big[|D-\mathbb{E}[D]|\Big|D\leq\mathbb{E}[D]\Big]}$. See Appendix SD.2 in the Supplementary Appendix for more details.

the numerator is (roughly) a weighted level-effect, the denominator shows that β^{twfe} additionally depends on a measure of the average distance between the effective treated and comparison group. Second, the effective comparison group can include treated units. Third, β^{twfe} uses "distance" weights w^{bin}'s to aggregate across dosages. In contrast, $ATT^{\texttt{loc}}$ does not suffer from any of these issues. In applications where the researcher is targeting level-effect parameters, we recommend favoring $ATT^{\texttt{loc}}$ vis-a-vis β^{twfe}.

Parts (c) and (d) of Theorem 3.4 provide interpretations of β^{twfe} taking scaled paths of outcomes as building blocks. For part (c), $ATT(d|d)/d$ (under parallel trends) and $ATT(d)/d$ (under strong parallel trends) are "per-dosage" causal parameters. This part shows that the TWFE estimand includes negative weights under the same conditions as in part (b), though the weights integrate to one. We note that, in the case of a discrete dose, this result in part (c) corresponds to the one in Theorem S3 of the Supplementary Appendix of de Chaisemartin and D'Haultfoeuille (2020). Therefore, using "average slopes" as the underlying parameter of interest eliminates neither TWFE's potential for negative weights nor its non-intuitive weighting scheme. For part (d), when β^{twfe} is interpreted in terms of all possible 2×2 comparisons of changes of outcomes for higher dose groups relative to lower dose groups, the weights are all positive and integrate to 1, but, under parallel trends, these comparisons all mix causal effects of the higher treatment with selection bias terms. Although strong parallel trends removes the selection bias, the weights attached to the causal parameters are still hard to interpret.

Remark 3.2 (Decomposition with no untreated units)**.** *It is straightforward to extend the TWFE decompositions discussed above to settings with no untreated units. For the causal response decomposition (part (a)), the exact same result applies, with the exception that the second term involving w_0^{acrt} is equal to 0. Similarly, for the scaled 2×2 decomposition (part (d)), nothing changes except that the second term involving $w_0^{2\times 2}$ is equal to 0. For the levels decomposition and the scaled levels decomposition (parts (b) and (c)), with no untreated units, $ATT(d|d)$ (or $ATT(d)$) is not identified; instead, along the lines mentioned in Remark 3.1, instead of using the untreated comparison group, we can instead compare to the path of outcomes of the "least treated". Thus, the same decompositions continue to apply except that $ATT(l|l)$ should be replaced by $ATT(l|l) - ATT(d_L|d_L)$. This immediately means that these decompositions (in addition to negative weights) become complicated by issues related to selection bias.*

4 DiD estimators that can highlight or summarize heterogeneity

In this section, we discuss how one can bypass the limitations of the TWFE regression specification in Equation (1.1) by proposing data-driven estimation procedures that target well-defined causal parameters. For simplicity, in this section, we rely on Assumption SPT so we can get all causal parameters under the same identification assumptions. If one is interested in $ATT(d|d)$ or their functionals, one can rely on Assumption PT and use the same estimation procedure for $ATT(d)$ that we discuss below. In this case, though, we stress that one should not interpret derivatives of estimates of $ATT(d|d)$ as estimates of $ACRT(d|d)$.

4.1 Estimating average causal functions among the treated

We start with the estimation of the dose-specific functions, $ATT(d)$ and $ACRT(d)$ under Assumption SPT. First, recall that, from Theorem 3.3, we have that, for a positive dose d,

$$ATT(d) = \mathbb{E}[\Delta Y|D=d] - \mathbb{E}[\Delta Y|D=0],$$

as well as $ACRT(d) = \partial\mathbb{E}[\Delta Y|D=d]/\partial d$ when the treatment is continuous, and $ACRT(d_j) = (\mathbb{E}[\Delta Y|D=d_j] - \mathbb{E}[\Delta Y|D=d_{j-1}]) / (d_j - d_{j-1})$ when the treatment is multi-valued discrete. As $\mathbb{E}[\Delta Y|D=0]$ can be estimated using its sample analog, $\mathbb{E}_n[\Delta Y|D=0] = n_{D=0}^{-1} \sum_{i:D_i=0} \Delta Y_i$, with $n_{D=0} = \sum_{i=1}^n 1\{D_i = 0\}$, the main challenge in estimating all these functions resides in estimating $\mathbb{E}[\Delta Y|D=d]$ among treated units ($d > 0$) and its derivative.

Note that this is a standard regression setup, and, as such, researchers have different options for how to approach it. Examples include adopting a parametric model for $\mathbb{E}[\Delta Y|D=d]$ (e.g., assuming a quadratic model in dose among the treated), or pursuing nonparametric estimators using kernels or sieves/series. We discuss these considerations below.

For simplicity, we start with setups where the treatment is multi-valued discrete, and takes on a relatively few values. In this case, one can estimate $ATT(d_j)$ and $ACRT(d_j)$ for any positive treatment dose d_j in the dose support using a simple saturated regression[8]

$$\Delta Y_i = \beta_0 + \sum_{j=1}^{J} 1\{D_i = d_j\}\beta_j + \varepsilon_i, \tag{4.1}$$

where we use the zero treatment dosage as the omitted category. It will then follow that $\widehat{\beta}_j$ and $\left(\widehat{\beta}_j - \widehat{\beta}_{j-1}\right) / (d_j - d_{j-1})$ are consistent estimators for the $ATT(d_j)$ and $ACRT(d_j)$, respectively, and inference procedures are standard. Note that, in this setup, all that our regression (4.1) is doing is to automate the appropriate comparison of means justified under our identification assumptions.

When the dose (among treated units) is continuous, (4.1) becomes infeasible. One straightforward estimation approach is to impose a parametric functional form restriction on how ΔY varies with D among treated. For instance, one can consider a model in which $\Delta\widetilde{Y}_i = \Delta Y_i - \mathbb{E}_n[\Delta Y|D=0]$ is quadratic in D among treated units, and run the following regression for observations with $D_i > 0$[9]

$$\Delta\widetilde{Y}_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \varepsilon_i. \tag{4.2}$$

When this regression specification is correctly specified, $\widehat{ATT}_{\text{par}}(d) = \widehat{\beta}_0 + \widehat{\beta}_1 d + \widehat{\beta}_2 d^2$ and $\widehat{ACRT}_{\text{par}}(d) = \widehat{\beta}_1 + 2\widehat{\beta}_2 d$ are consistent estimators for $ATT(d)$ and $ACRT(d)$. Pointwise and uniform-in-d inference procedures are standard. Of course, other parametric functional forms can also be adopted.

A limitation of parametric models, such as (4.2), is their reliance on potentially incorrect functional form restrictions. In fact, Theorem 3.4 exactly highlights the consequences of misspecification in the linear case. Provided that the sample size is large, however, researchers can use nonparametric procedures to avoid functional form restrictions. This entails considering a nonparametric regression

[8] One can also use the more flexible TWFE regression specification $Y_{i,t} = \sum_{j=1}^{J} Post_t \cdot \mathbf{1}\{D_i = d_j\}\beta_j + \eta_i + \theta_t + v_{i,t}$, $t = 1, 2$. We also note that we implicitly take $d_0 = 0$.

[9] One could also consider the regression $\Delta Y_i = \alpha + \mathbf{1}\{D > 0\}\left(\beta_0 + \beta_1 D_i + \beta_2 D_i^2\right) + \varepsilon_i$ for all observations.

model of $\Delta\widetilde{Y}_i$ on D_i among treated units,

$$\Delta\widetilde{Y}_i = ATT(D_i) + \varepsilon_i. \tag{4.3}$$

One can estimate (4.3) in any number of ways. In our application, we have adopted the data-driven nonparametric estimators proposed by Chen, Christensen, and Kankanala (2025). An appealing feature of this procedure is that it resembles (4.2) in the sense that, upon computing the optimal sieve-dimension $\widehat{K}$, one runs a linear regression of $\Delta\widetilde{Y}$ on flexible $\widehat{K}$-dimensional transformations of D (cubic B-splines), $\psi^{\widehat{K}}(D)$, in the subsample of units with $D_i > 0$,

$$\Delta\widetilde{Y}_i = \psi^{\widehat{K}}(D)'\beta_{\widehat{K}} + \varepsilon_i, \tag{4.4}$$

and then forming the nonparametric estimators for $ATT(d)$ and $ACRT(d)$ as

$$\widehat{ATT}_{\text{cck}}(d) = \left(\psi^{\widehat{K}}(d)\right)'\widehat{\beta}_{\widehat{K}}, \qquad \widehat{ACRT}_{\text{cck}}(d) = \left(\partial\psi^{\widehat{K}}(d)\right)'\widehat{\beta}_{\widehat{K}}, \tag{4.5}$$

where $\partial\psi^K(s) = (d\psi_{K1}(s)/\,ds, \ldots, d\psi_{KK}(s)/\,ds)'$, and $\widehat{\beta}_{\widehat{K}}$ is the $\widehat{K}$-dimension vector of OLS estimators for $\beta_{\widehat{K}}$.[10] Chen, Christensen, and Kankanala (2025)'s results imply that the nonparametric estimators for $ATT(d)$ and $ACRT(d)$ curves converge at the fastest possible (i.e., minimax) rate in sup-norm, and lead to uniform confidence bands that are asymptotically narrower (more precise) than those based on undersmoothing, and contract at, or within a $\log\log n$ factor of, the minimax rate. See Appendix B for more details on how to compute $\widehat{K}$, as well as how to construct uniform confidence bands based on $\widehat{ATT}_{\text{cck}}(d)$ and $\widehat{ACRT}_{\text{cck}}(d)$. Of course, one can adopt other nonparametric estimation and inference procedures and select tuning parameters using alternative criteria, although these may lead to different statistical guarantees.

4.2 Estimating summary measures of treatment effects

Researchers frequently want to report summary estimates to enhance interpretability and/or statistical precision, or because a lower-dimensional parameter is an input into some model or post-estimation calculation. As we showed in Section 3, however, a linear TWFE regression generally fails to deliver an interpretable summary parameter. In this section, we discuss estimation of $ATT^{\texttt{glob}}$ and $ACRT^{\texttt{glob}}$, which are summary causal effect parameters that have a clear interpretation.

When there are untreated units, part (a) of Corollary 3.1 suggests an extremely simple and familiar estimator of $ATT^{\texttt{glob}}$: the difference between the average change in outcomes among treated units minus the average outcome change for untreated units. This "binarized" DiD estimator can be obtained from the following simple linear regression specification:

$$\Delta Y_i = \beta_0^{bin} + D_i^{>0}\beta^{bin} + \epsilon_i, \tag{4.6}$$

where $D_i^{>0} = 1\{D_i > 0\}$. It is straightforward to show that under the identification assumptions in Corollary 3.1, $\beta^{bin} = ATT^{\texttt{glob}}$. Note that this estimator applies equally to continuous and multi-valued discrete treatments.

[10] As these nonparametric procedures have slower-than-$\sqrt{n}$ rates of convergence, there is no estimation effect from estimating $\mathbb{E}[\Delta Y|D=0]$.

Aggregated average causal response parameters can be constructed easily by weighting the estimated average causal functions across doses using the dose distribution itself. For discrete treatments, it is straightforward to aggregate these $ACRT(d)$'s based on the coefficients from (4.1) to form a plug-in estimator for the $ACRT^{\texttt{glob}}$, using the identification formula in Corollary 3.1(c), i.e.,

$$\widehat{ACRT}^{\texttt{glob}} = \sum_{j=1}^{J} \frac{\widehat{\beta}_j - \widehat{\beta}_{j-1}}{d_j - d_{j-1}} \widehat{\mathbb{P}}(D = d_j | D > 0), \tag{4.7}$$

where $\widehat{\mathbb{P}}(D = d_j|D > 0) = \sum_{i=1}^{n} 1\{D_i = d_j\} / \sum_{i=1}^{n} 1\{D_i > 0\}$. Inference procedures follow from the Delta Method. One can follow a similar strategy when using the scaled $ATT(d)$ as the "building block" of the aggregation. A similar approach applies to estimating $ACRT^{\texttt{glob}}$ with a continuous dose. Our proposed estimator is simple to compute as it is based on the plug-in principle, i.e.,

$$\widehat{ACRT}^{\texttt{glob}} = \mathbb{E}_n \left[\widehat{ACRT}(D) \middle| D > 0 \right] = \frac{1}{n_{D>0}} \sum_{i:D_i>0} \widehat{ACRT}(D_i),$$

with $n_{D>0} = \sum_{i=1}^{n} 1\{D_i > 0\}$ denoting the sample size with a positive dose, and $\widehat{ACRT}(D)$ being a parametric or nonparametric estimator. Under some regularity conditions, one can show that $\widehat{ACRT}^{\texttt{glob}}$ is $\sqrt{n_{D>0}}$ consistent and asymptotically normal; see, e.g., Section 4.1 of Ai and Chen (2007).

We close this section by noticing that it is also possible to consider alternative estimators for $ACRT^{\texttt{glob}}$ using a so-called Neyman-Orthogonal moment representation. More precisely, by exploring the efficient influence function for $ACRT^{\texttt{glob}}$ implied by Theorem 3.1 of Newey and Stoker (1993), it is straightforward to show that

$$ACRT^{\text{glob}} = \mathbb{E}\left[ACRT(D) - (\Delta Y - \mathbb{E}[\Delta Y|D, D > 0]) \frac{f'_{D|D>0}(D)}{f_{D|D>0}(D)} \middle| D > 0 \right]. \tag{4.8}$$

Based on this representation, one can then use flexible nonparametric or machine-learning-based estimators for the nuisance functions and still conduct asymptotically valid inference procedures. This opens the door for leveraging double machine learning procedures to estimate $ACRT^{\text{glob}}$ in DiD contexts. We leave this topic for future research.

5 Extensions

In this section, we briefly summarize several extensions of our main results that are further discussed in the Appendix and Supplementary Appendix.

5.1 Relaxing Strong Parallel Trends

Under traditional DiD assumptions, Assumption PT led to the identification of local $ATT(d|d)$ parameters that are difficult to compare across dosages. On the other hand, the strong parallel trends assumption led to $ATT(d)$ parameters. These can be seen as extreme cases, and it is possible to trade off the strength of assumptions with the type of parameters that can be identified in different ways. The number of these intermediate possibilities is large, however. Here, we sketch what we consider

to be three main ideas to relax strong parallel trends. Appendix SE of the Supplementary Appendix provides substantially more detail.

First, in many cases, researchers may be willing to assume that they know the direction of the selection bias. For example, suppose that a researcher is willing to assume that, for all d and any dose groups $l < h$, $ATT(d|l) \leq ATT(d|h)$, i.e., that higher dose groups would experience larger treatment effects at any value of the dose. In the Supplementary Appendix, we show that this type of assumption leads to bounds on causal effect parameters without requiring strong parallel trends. For example, it implies that, for all d

$$ACRT(d|d) \leq \frac{\partial \mathbb{E}[\Delta Y|D=d]}{\partial d},$$

which provides a bound on $ACRT(d|d)$. See Proposition S7 in the Supplementary Appendix for more details.

A second possibility for relaxing strong parallel trends is to define a sub-region $\mathcal{D}_s \subseteq \mathcal{D}_+$ for which strong parallel trends holds, i.e., to assume that

$$\mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D \in \mathcal{D}_s] = \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D=d] \tag{5.1}$$

holds for all $d \in \mathcal{D}_s$. Under this assumption, we show in Proposition S8 in the Supplementary Appendix that, for $h, l \in \mathcal{D}_s$,

$$\mathbb{E}[\Delta Y|D=h] - \mathbb{E}[\Delta Y|D=l] = \mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D \in \mathcal{D}_s].$$

In other words, comparing the trends in outcomes over time for dose group h to dose group l delivers the average causal effect of dose h relative to dose l among those dose groups in $\mathcal{D}_s$. Under PT, the same comparison would include selection bias terms.

While the assumption in Equation (5.1) is weaker than SPT, the tradeoff is that now only comparisons within the set $\mathcal{D}_s$ have a causal interpretation. In some applications, this assumption could be notably weaker than Assumption SPT—in fact, this assumption should, at least arguably, no longer be called "strong parallel trends" because it is non-trivially non-nested with Assumption PT. For example, suppose that $\mathcal{D}_s$ contains large doses. The assumption in (5.1) says that we can learn about the trend in outcomes for a higher-dose group at a counterfactual lower dose by looking at the trend in outcomes for that lower-dose group, but only for doses in $\mathcal{D}_s$. This could be much more plausible than Assumption PT, which assumes that even very high dose groups would have experienced the same trend in *untreated* potential outcomes as the untreated group, even though these units might be very different from each other. This local version of the SPT assumption is appealing in applications where there is substantial variation in the dose and the researcher is willing to assume that there is no selection bias among units that select similar doses, but the researcher is unwilling to assume that there is no selection bias among units that select substantially different doses.

Finally, in some applications, strong parallel trends may be more plausible after conditioning on some observed covariates X. Under a version of strong parallel trends conditional on covariates, one can show that the conditional average treatment effect, $ATT_x(d) = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)|X=x, D>0]$, is identified. Since this parameter is not local to dose group d, conditional on $X=x$, one can compare $ATT_x(d)$ across different values of the dose without inducing selection bias terms. This

is an intermediate case, however, in that these are more local parameters than $ATT(d)$ because they are local to the particular value of the covariates x. See the discussion in Appendix SE in the Supplementary Appendix for more details.

5.2 Multiple time periods and variation in treatment timing

Although our results so far focus on two-period cases, we can extend them to setups with multiple time periods and variation in treatment timing across units by combining the ideas discussed in Section 3.2 with those in Callaway and Sant'Anna (2021). We consider this setting in detail in Appendix C.

In a setting with staggered treatment adoption (i.e., where once a unit becomes treated with dose d, that unit remains treated with dose d in subsequent periods), knowing the time period that a unit becomes treated with a positive dose (which we denote by G_i and refer to as a unit's *timing group*) and dose D_i (i.e., dose group) fully characterizes a unit's sequence of treatments across all periods. In this context, we need to augment our potential outcomes terminology and write $Y_{i,t}(g,d)$ as the potential outcome of unit i at time t if it were first treated in period g, with dose d; we write $Y_{i,t}(0) = Y_{i,t}(\infty, 0)$ to denote a unit's untreated potential outcome—the potential outcome in time period t if that unit did not participate in the treatment in any available period. With this notation at hand, we can define a multi-period analog of $ATT(d|d)$ as

$$ATT(g,t,d|g,d) = \mathbb{E}[Y_t(g,d) - Y_t(0)|G=g, D=d] \quad \text{and} \quad ACRT(g,t,d|g,d) = \left.\frac{\partial ATT(g,t,l|g,d)}{\partial l}\right|_{l=d}$$

which are the average treatment effect and average causal response in period t of (i) becoming treated in period g and (ii) experiencing dose d among those in timing group g and dose group d.

Under no anticipation and a multiple-period version of the parallel trends assumption, we show in Appendix C that, in post-treatment periods (i.e., periods where $t \geq g$)

$$ATT(g,t,d|g,d) = \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|G=\infty, D=0]. \tag{5.2}$$

The argument is similar to the two-period case discussed earlier. The main difference is that the expression above involves the "long difference" in changes in outcomes over time, i.e., from period $g-1$ to t. The reason for this difference is that $g-1$ is the most recent period for which units in group g were untreated. The expression above uses the never-treated group ($G=\infty$) as the comparison group, but, like the case with a binary treatment, one can use alternative comparison groups such as the not-yet-treated. Under a multiple-period version of the strong parallel trends assumption, one can take the derivative of the right-hand side of Equation (5.2) with respect to d to identify $ACRT(g,t,d|g,d)$.

One complication that arises in the staggered case is that $ATT(g,t,d|g,d)$ and $ACRT(g,t,d|g,d)$ are often relatively high-dimensional objects that can be hard to report (and perhaps hard to estimate precisely). In Appendix C, we discuss two main strategies for aggregating these parameters into lower-dimensional objects. First, we average across timing groups and time periods to target causal effect parameters that are a function of only the dose: $ATT^{dose}(d|d)$, and $ACRT^{dose}(d|d)$—these parameters highlight heterogeneous effects across different doses and are analogous to $ATT(d|d)$ and $ACRT(d|d)$ in the two-period case that we have emphasized above. They can be averaged across the

dose to deliver scalar summary parameters. Second, we consider event-study parameters: $ATT^{es}_{\texttt{loc}}(e)$, and $ACRT^{es}_{\texttt{loc}}(e)$ that average across the dose and highlight how treatment effects and/or causal responses vary with the length of exposure to the treatment—these parameters are the event study analog of $ATT^{\texttt{loc}}$ and $ACRT^{\texttt{loc}}$ in the two period case above. See Callaway, Goodman-Bacon, and Sant'Anna (2024) for alternative, intermediate aggregations. The discussion here focuses on causal effect parameters that are local to a specific dose group and timing group, but, like the two-period case discussed above, it is also possible to recover causal effect parameters across all treated units under strong parallel trends; see Appendix SB in the Supplementary Appendix for more details.

5.3 Interpreting TWFE Regressions with Multiple Periods/Groups

In Appendix SB.2 of the Supplementary Appendix, we also extend our TWFE decomposition results from Theorem 3.4 to cover setups beyond the two-period case, including setups with staggered treatment adoptions with continuous or multi-valued discrete treatments. These results generalize the decompositions in de Chaisemartin and D'Haultfoeuille (2020) and Goodman-Bacon (2021) to the case of a continuous treatment. Those results demonstrate that TWFE regressions with multiple periods and variation in treatment timing (i) continue to suffer from the weighting and selection bias issues that we highlighted in Theorem 3.4, (ii) inherit weighting issues (including possible negative weights) that are prevalent in TWFE regressions with binary, staggered treatment adoption, and (iii) are affected by violations of parallel trends in pre-treatment periods.

5.4 Event-Study and Pre-Treatment Differences

When there are multiple periods of data available, DiD applications typically assess the plausibility of the parallel trends assumption by checking whether or not parallel trends holds in pre-treatment periods. In a setting with a continuous treatment, one can check whether $\mathbb{E}[\Delta Y_t|D=d] = \mathbb{E}[\Delta Y_t|D=0]$ is approximately correct for all pre-treatment time periods t and all d; one can also check $\mathbb{E}[Y_t - Y_{g-1}|D=d] = \mathbb{E}[Y_t - Y_{g-1}|D=0]$, with g being the time of treatment adoption (where we simplify and consider a single treatment date setup). Implementing these tests, however, can be complicated because it involves multiple dose-response nonparametric estimates. A convenient alternative is to report aggregated event study parameters such as $ATT^{es}_{\texttt{loc}}(e)$ or $ACRT^{es}_{\texttt{loc}}(e)$ in pre-treatment periods (i.e., $e<0$). Plotting estimates of $ATT^{es}_{\texttt{loc}}(e)$ and $ACRT^{es}_{\texttt{loc}}(e)$ for pre-treatment periods ($e<0$) can be used to assess the plausibility of parallel trends. We report these for our empirical application in Figures 8 and 10. Having said that, we note that one possible drawback of this test is that it may overlook violations of the parallel trends assumption that these event-study versions of the test would not detect.

An interesting (though subtle) caveat is that in cases where an aggregate level effect such as $ATT^{\texttt{loc}}$ or its event study version $ATT^{es}_{\texttt{loc}}(e)$ is the target parameter of the analysis, it is possible to recover it under "weaker" parallel trends assumptions that allow for violations of parallel trends where the *average* violation of parallel trends across dose groups is equal to zero (rather than the violation of parallel trends being equal to zero for all dose groups)—we refer to the corresponding

averaged version of parallel trends as aggregate parallel trends. If one maintains aggregate parallel trends, then only $ATT^{es}_{\texttt{loc}}(e)$ (and not, e.g., $ACRT^{es}_{\texttt{glob}}(e)$) is relevant for assessing its plausibility using pre-treatment periods. That being said, it is debatable whether or not the violations of parallel trends that can be allowed for under aggregate parallel trends should be counted as evidence against the design.

6 Continuous DiD in Practice: Causal Effects of Medicare PPS

We have so far shown that the causal question of interest shapes identification in a continuous DiD design and argued that it should guide the estimation approach, too. We now apply our preferred average level treatment effect and average causal response estimators to Acemoglu and Finkelstein (2008)'s study of Medicare PPS. To map their setting to our theoretical analysis, we consider the balanced panel data component of Acemoglu and Finkelstein (2008), which comprises 5881 hospitals, and also average all pre-treatment outcomes and post-treatment outcomes over time. Thus, we use $t = 1$ to denote the average of pre-treatment periods (1980-1983), and $t = 2$ to denote the average of post-treatment periods (1984-1986). We also denote treatment dose here by M instead of D, as M is a short-hand notation for the 1983 Medicare inpatient share that determines treatment exposure in the AF application.

To begin, consider the profit maximization problem for a hospital with Medicare inpatient share M. We follow AF and assume a production function, $F_t(L, K)$, that is homothetic in labor (L) and capital (K). Market wages and rental rates are normalized by the output price, and Medicare subsidies mean that net input prices are $(1-s_{L,t}M)w$ and $(1-s_{K,t}M)r$. Firms consider the following profit maximization problem:

$$\max_{L,K} F_t(L, K) - (1 - s_{L,t}M)wL - (1 - s_{K,t}M)rK.$$

The solution to this problem generates factor demands and a capital-labor ratio that is only a function of the input price ratio, $k^*_t\left(\frac{(1-s_{L,t}M)w}{(1-s_{K,t}M)r}\right)$. We write the subsidy ratio, $\frac{(1-s_{L,t}M)}{(1-s_{K,t}M)}$ as $1 + S_t(M) = 1 + \frac{(s_{K,t}-s_{L,t})M}{1-s_{K,t}M}$. This reflects the fact that hospitals with no Medicare patients ($M = 0$), and all hospitals before PPS (when $s_{K,t=1} = s_{L,t=1} = s$) face no relative price distortion. PPS set $s_{L,t} = 0$ in 1983, making $S_{t=2}(M) = \frac{s_{K,t=2}M}{1-s_{K,t=2}M}$.

This structure allows us to define the capital-labor ratio potential outcomes in terms of Medicare inpatient share M:

$$Y_{t=1} = Y_{t=1}(0) = k^*_{t=1}\left(\frac{w}{r}\right) \quad \text{and} \quad Y_{t=2} = Y_{t=2}(M) = k^*_{t=2}\left((1 + S_{t=2}(M))\frac{w}{r}\right).$$

Three details of the theoretical setup are worth noting. First, homotheticity allows us to connect potential outcomes as a function of M to a firm's optimal capital-labor ratio as a function of relative prices (as a function of M). Without this assumption, a hospital's scale affects its input mix, and capital-labor ratios are a function of net labor and capital prices separately, complicating the theoretical interpretation of causal parameters. Second, we define our parameters of interest in terms of causal effects of M on Y. A structural interpretation of those parameters in terms of k^* necessar-

ily involves the non-linear way in which M changes the subsidy ratio, $S_t(M)$ (as well as a kind of exclusion restriction that rules out direct effects of M on outcomes). Third, we use time subscripts to match the fact that PPS changed over time, but this is not a dynamic model. The assumed lack of forward-looking behavior implies the no anticipation assumption (Assumption 3) and allows us to write $Y_{t=1} = Y_{t=1}(0)$. All these details are in line with AF's theoretical model.

6.1 Causal Questions Around Medicare PPS

AF are primarily interested in the question: did PPS raise capital-labor ratios? PPS sought to help hospitals invest in new medical technologies with the aim of improving patient outcomes (Office of Technology Assessment, 1984). But regulators also worried about the "incentive for hospitals to adopt expensive capital equipment that reduces operating costs but raises total costs per case" (Office of Technology Assessment, 1984, p. 14). Thus, Medicare's role in technology investments has important policy implications. Moreover, the theoretical model predicts that PPS would raise capital-labor ratios for all treated hospitals, so the sign of its effects is a test of a simple neoclassical production theory. The building block parameters that answer these questions are the average treatment effect of PPS on hospitals with $M = m$:

$$ATT(m|m) = \mathbb{E}[Y_{t=2}(m) - Y_{t=2}(0)|M = m] = \mathbb{E}\left[k^*_{t=2}\left((1 + S_{t=2}(m))\frac{w}{r}\right) - k^*_{t=2}\left(\frac{w}{r}\right)\middle| M = m\right].$$

Estimating and plotting the entire $ATT(m|m)$ function shows which hospitals responded most to PPS and tests the prediction that *all* treated hospitals increase their capital intensity. Under parallel trends alone, one cannot compare across $ATT(m|m)$'s, as it is not possible to discern whether variation from $ATT(m|m)$ comes directly from subsidy differences or from treatment effect heterogeneity. Averaging this function across treated hospitals yields $ATT^{\texttt{loc}} = \mathbb{E}[ATT(M|M)|M > 0]$, a summary parameter that directly answers the question "did PPS raise capital-labor ratios on average?"

One may also be interested in which subsidy levels have larger causal effects. For example, if technologies are "lumpy", then hospitals may not respond to subsidies too small to cover the minimum investment costs. Improving the design of input subsidies thus requires causal estimates of the responsiveness to different subsidy levels. The causal effects of marginal changes in the subsidy ratio also represent another test of the theoretical model because they are proportional to a hospital's elasticity of substitution, $\sigma_{i,t}(m) = \frac{k^{*\prime}_{i,t}}{k^*_{i,t}} \times (1 + S_t(m)) \times \frac{w}{r}$, which, with two inputs, must be positive. The building block parameters that answer these questions are the average causal responses of PPS:

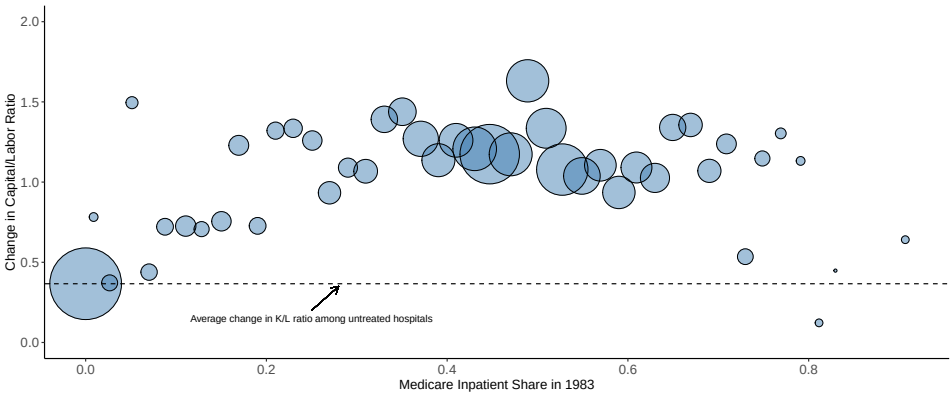
$$\begin{aligned} ACRT(m) = \mathbb{E}\big[Y'_{t=2}(m)\big|M > 0\big] &= \mathbb{E}\left[k^{*\prime}_{t=2}\left((1 + S_{t=2}(m))\frac{w}{r}\right)S'_{t=2}(m)\frac{w}{r}\middle| M > 0\right] \\ &= \mathbb{E}\left[\sigma_{t=2}(m)k^*_{t=2}\left((1 + S_{t=2}(m))\frac{w}{r}\right)\frac{s_k}{1 - s_k m}\middle| M > 0\right] \end{aligned} \tag{6.1}$$

Again, reporting estimates of the entire $ACRT(m)$ function highlights heterogeneity in how hospitals respond to subsidies, and the summary parameter $ACRT^{\texttt{glob}}$ provides a single measure of how much hospitals respond on average to small subsidy differences.

Before turning to our formal estimates, Figure 4 presents a binned scatter plot of the change in mean capital-labor ratios before (1980-1983) and after (1984-1986) PPS against the Medicare share

of inpatient days in 1983, m. Following AF, we measure the capital-labor ratio using the depreciation share of total costs.

Figure 4: Changes in Capital-Labor Ratios before and after 1983 versus the Medicare Inpatient Share



Notes: The figure presents a binned scatter plot of the change in the average depreciation share (capital-labor ratio) between the periods 1980-1983 and 1984-1986 for hospitals in 2-percentage-point bins of the 1983 Medicare inpatient share, M. In the lowest bin, hospitals with $M = 0$ are plotted separately from hospitals with $M \in (0, 0.02]$. We also consider a single bin for all hospitals with $M > 0.84$.

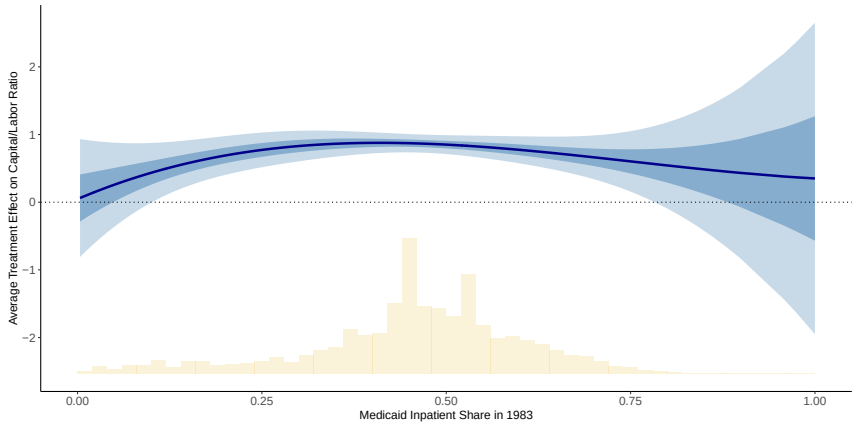
The horizontal line equals the mean change in capital-labor ratio for untreated hospitals (0.37). Each circle is the mean outcome change for a given bin of the Medicare inpatient share, with its size proportional to the number of hospitals in that bin. Almost all groups of treated hospitals had stronger growth in capital intensity than untreated hospitals, consistent with the theoretical prediction. The relationship is nonlinear, however, which indicates heterogeneity in average treatment effects, at least, and perhaps heterogeneity in the sign of average causal responses.

6.2 Average Treatment Effects of PPS

Figure 5 presents our proposed data-adaptive nonparametric estimates of $ATT(m|m)$ based on (4.5). For inference, we cluster at the hospital level. Our data-driven procedure to optimally choose the sieve dimension selected $\widehat{K} = 4$. These estimates formalize what the scatter plot suggests: that $ATT(m|m)$ is positive. We plot pointwise 95% confidence intervals in the dark-shaded region and the wider (honest) 95% uniform confidence bands in the light-shaded region. We do not detect an effect for values of m below 5 percent, but we reject zero for doses between 0.05 and 0.78, which contains 96 percent of treated hospitals. Significant values of $\widehat{ATT}(m|m)$ range from about 0.44 percentage points at $m = 0.1$ to 0.88 percentage points at $m = 0.41$. The average across all doses ($\widehat{ATT}^{\texttt{loc}}$) is 0.80 (s.e. $= 0.05$), or about 18 percent of the 1983 mean outcome (measured by the depreciation share) of 4.5. This evidence suggests that PPS substantially raised capital-labor ratios.

For comparison, we report in Figure 6 parametric estimates for $ATT(m|m)$ that use the quadratic regression specification in Equation 4.2. Different from Figure 5, the interpretability of $\widehat{ATT}_{\text{par}}(m|m)$ in Figure 6 depends on the quadratic specification being correctly specified. When we know that is the case, it is clear from Figure 6 that this results in substantially more precise estimates, as they now fully leverage the functional form. Importantly, these gains in precision are more substantial

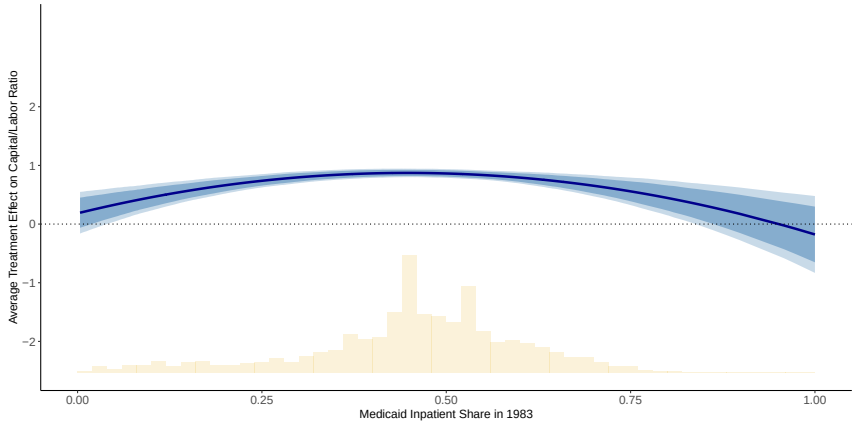
Figure 5: Nonparametric Estimates of $ATT(m|m)$ for Medicare PPS



Notes: The figure plots nonparametric estimate of $ATT(m|m)$ that adapts the Chen, Christensen, and Kankanala (2025) data-driven estimator to our context, as discussed in Section 4.1 and Appendix B. The dark-shaded region is the 95-percent point-wise confidence interval, and the lighter-shaded region is the 95-percent honest and sup-norm rate-adaptive uniform confidence band. We display the histogram of the treatment dose among the treated in yellow.

in the regions where data for particular treatment doses are scarce, e.g., for treatment doses above 0.75. Overall, we have 4987 observations with a positive treatment dose. Among these, only 57 have a treatment dose above 0.75, 20 above 0.80, and 3 above 0.90. The rationale for this is very simple: parametric models are good at extrapolating, whereas nonparametric procedures are more cautious about it. The reliability of the extrapolation, once again, crucially depends on the parametric model for $ATT(m|m)$ being correctly specified.

Figure 6: Parametric Estimates of $ATT(m|m)$ for Medicare PPS using quadratic specification



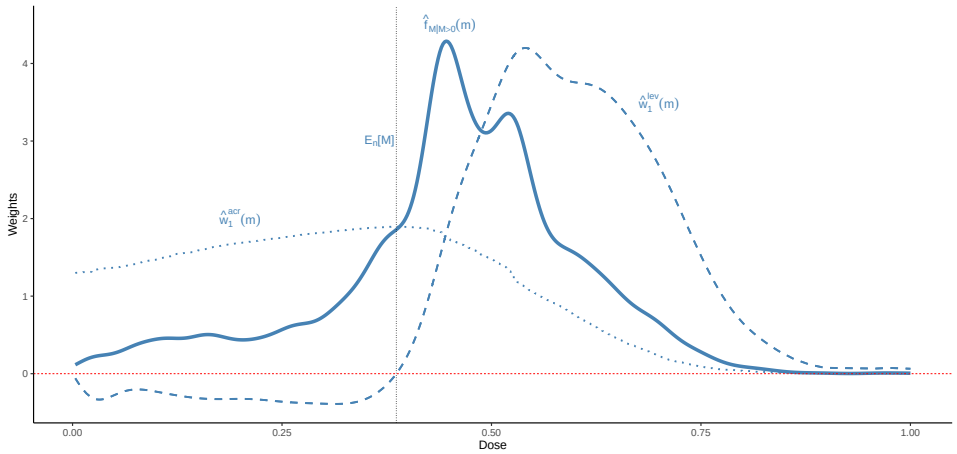
Notes: The figure plots parametric estimate of $ATT(m|m)$ that use the quadratic regression specification in Equation 4.2. The dark-shaded region is the 95-percent point-wise confidence interval, and the lighter-shaded region is the 95-percent uniform-in-treatment-dose confidence band. We display the histogram of the treatment dose among the treated in yellow. We use the same y-scale as in Figure 5.

Although gains in precision are desirable, we caution against using nonparametric results to pick a parametric specification. This, to some extent, resembles a pre-testing problem, and inference based on the parametric model could be misleading. In fact, the appeal of the uniform confidence bands from Chen, Christensen, and Kankanala (2025) that we report in light-shaded blue in Figure 5

is that they account for this type of pre-testing issue and are honest, i.e., they are guaranteed to have asymptotically correct coverage over a large (and generic) class of data-generating processes. The uniform confidence bands in Figure 6 are uniform only in treatment dosage, highlighting that it reflects a narrower type of uncertainty than those in Figure 5. Henceforth, as we find it challenging to ex ante motivate a parametric functional form for $ATT(m|m)$ using arguments grounded in economic theory, we focus our attention on our nonparametric estimators.

In Section 3.3, we argued that β^{twfe} should not be relied upon to summarize level effects. However, the TWFE coefficient is 1.14—roughly similar to our estimate of $ATT^{\texttt{loc}}$. What accounts for their similarity? One explanation comes from Equation (3.1). The numerator compares weighted averages of the paths of outcomes for the "effective" treated group (those with above-average doses) to the "effective" comparison group (those with below-average doses). However, in our example, slightly more than half of the weight on paths of outcomes in the effective comparison group falls on hospitals with a positive dose. This biases β^{twfe} downward relative to $ATT^{\texttt{loc}}$—our estimate of the numerator in Equation (3.1) is 0.60. In contrast, the "weighted distance" between the effective treated and comparison groups in the denominator of Equation (3.1) is estimated to be 0.53, and dividing by 0.53 results in β^{twfe} being upward biased relative to $ATT^{\texttt{loc}}$. That these two biases work in opposite directions and have similar magnitudes in our particular application happens to result in $\widehat{\beta}^{twfe}$ being fairly close to $\widehat{ATT}^{\texttt{loc}}$. Interestingly, though, if we instead were to code a hospital's dose on a scale of 0 to 100, our estimate of β^{twfe} shrinks to $0.0114 = 1.14/100$ while our estimate of $ATT^{\texttt{loc}}$ remains unchanged.

Figure 7: Weighting Schemes for TWFE and Dose Distribution Among Treated

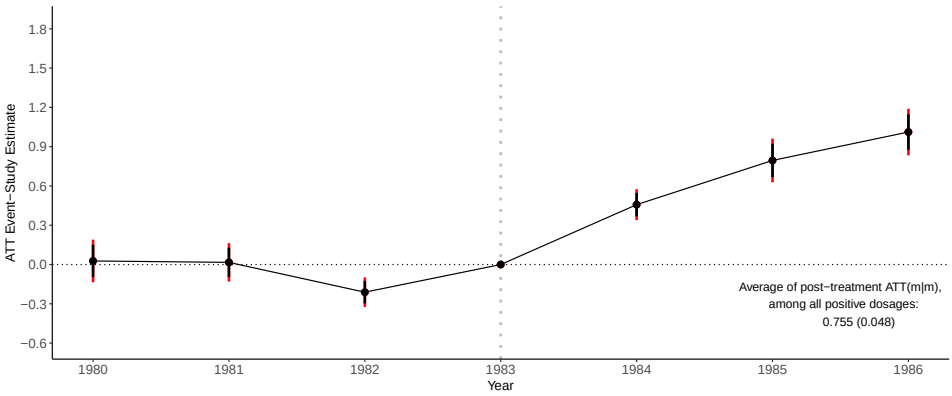


Notes: The dashed lines are the weights that TWFE puts on $ATT(m|m)$ and $ACRT(m)$ parameters, as in Theorem 3.4. The solid line is a smoothed estimate of the density of the Medicare inpatient share, M.

Figure 5 abstracts from dynamics since it is based on average outcomes in the pre- and post-treatment periods. As an alternative, Figure 8 plots estimates of event-study summary parameters, $ATT^{es}_{\texttt{loc}}(e) = \mathbb{E}[Y_{t=e} - Y_{t=1983}|D > 0] - \mathbb{E}[Y_{t=e} - Y_{t=1983}|D = 0]$, using 1983 as the baseline year. The patterns are similar to the TWFE event-study in Figure 1, but their magnitudes reflect proper

averages of year-specific $ATT(m|m)$ parameters.[11]

Figure 8: Event-Study Estimates of ATT



Notes: The figure plots the event-study estimates of $ATT^{es}_{\texttt{loc}}(e)$, with their 95% pointwise confidence intervals reported in black, and the 95% uniform confidence bands reported in red.

6.3 Average Causal Responses to PPS

Figure 9 plots our proposed data-adaptive nonparametric estimate of the slope of the function estimated in Figure 5. Under Assumption SPT, the function in Figure 5 is the $ATT(m)$ and its slope in Figure 9 equals the $ACRT(m)$. The hump shape in Figure 5 is reflected in an $ACRT(m)$ function that starts positive, and declines through most of its support. We estimate negative $ACRT(m)$ parameters for doses above $m = 0.41$, a range that includes 71 percent of treated hospitals. The 95% uniform confidence band covers zero everywhere, although we are able to detect positive $ACRT(m)$ values for doses below the mean as well as negative $ACRT(m)$ values for doses between about 0.5 and 0.7 using pointwise confidence intervals.

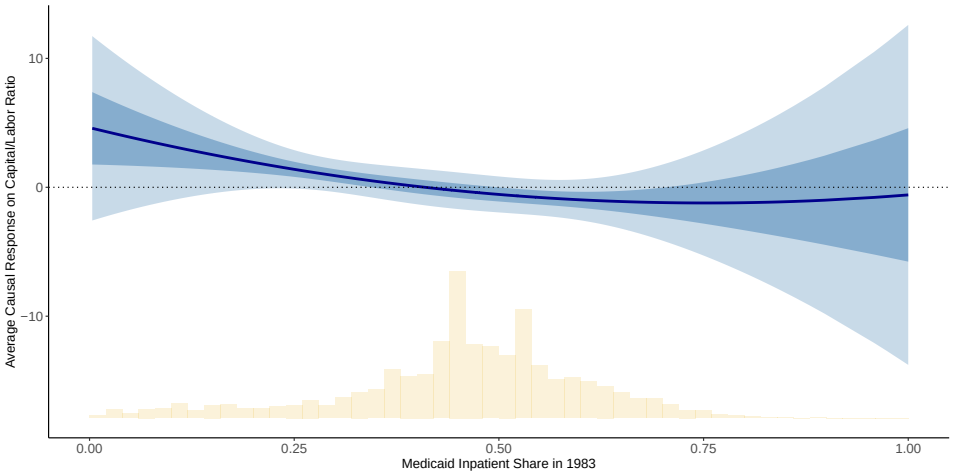
PPS' average causal response parameter weighted by the actual dose distribution of treated hospitals is $\widehat{ACRT}^{\texttt{glob}} = -0.08$ (s.e. = 0.19) and is not significantly different from zero.[12] This differs substantially from the TWFE coefficient, $\widehat{\beta}^{twfe} = 1.14$. From Theorem 3.4(a), the difference between these estimates is fully driven by differences in the weighting scheme. Our estimate of $ACRT^{\texttt{glob}}$ comes from mapping the estimates of $ACRT(m)$ in Figure 9 to the dose distribution weights, $\widehat{f}_{M|M>0}(m)$, in Figure 7; our estimate of β^{twfe} comes from mapping the estimates of $ACRT(m)$ to the TWFE causal response weights, $\widehat{w}^{acrt}_1(m)$, in Figure 7. As discussed in Theorem 3.4(a), the TWFE causal response weights are positive for all values of the dose and integrate to one, providing a reason to hope that estimates of $ACRT^{\texttt{glob}}$ and β^{twfe} would be similar. However, the TWFE weighting scheme turns out to be much different from the dose distribution weighting

[11] The negative pre-PPS coefficient may reflect the fact that PPS was passed in April 1983 and partially took effect in that calendar year, and also that hospitals report labor and capital costs for different fiscal years. Therefore, some 1983 outcomes may include post-treatment months. The results also show that the $ATT^{es}_{\texttt{loc}}(e)$ grows each year following PPS, which matches the fact that PPS' subsidy reforms actually phased in over three years. We also note, however, that these can represent other types of violations of parallel trends.

[12] We treat the sieve dimension used to compute $\widehat{ACRT}^{\texttt{glob}}$ as a non-random sequence, which is in line with the theoretical justification in Ai and Chen (2007). A formal theoretical treatment that accounts for the stochastic nature of our Lepski-type selection is interesting but left for future research.

scheme. Combining these differences with the high degree of heterogeneity in $ACRT(m)$ across m is what leads to the sharp differences in the estimates. Another reason to emphasize the large difference between these estimates is that the literature has often viewed negative weights as a dividing line between an "unreasonable" or "reasonable" weighting scheme (see, e.g., Angrist (1998), de Chaisemartin and D'Haultfoeuille (2020), and Blandhol, Bonney, Mogstad, and Torgovitsky (2025) for related discussions of this point in different contexts). The results here suggest that, at least in our context, articulating a well-defined causal effect parameter and targeting that parameter directly is likely to be more important than checking that weights are all positive and integrate to one.

Figure 9: Nonparametric Estimates of $ACRT(m)$ for Medicare PPS



Notes: The figure plots nonparametric estimate of $ACRT(m)$ that adapts the Chen, Christensen, and Kankanala (2025) data-driven estimator to our context, as discussed in Section 4.1 and Appendix B. The dark-shaded region is the 95-percent point-wise confidence interval, and the lighter-shaded region is the 95-percent honest and sup-norm rate-adaptive uniform confidence band. We display the histogram of the treatment dose among the treated in yellow.

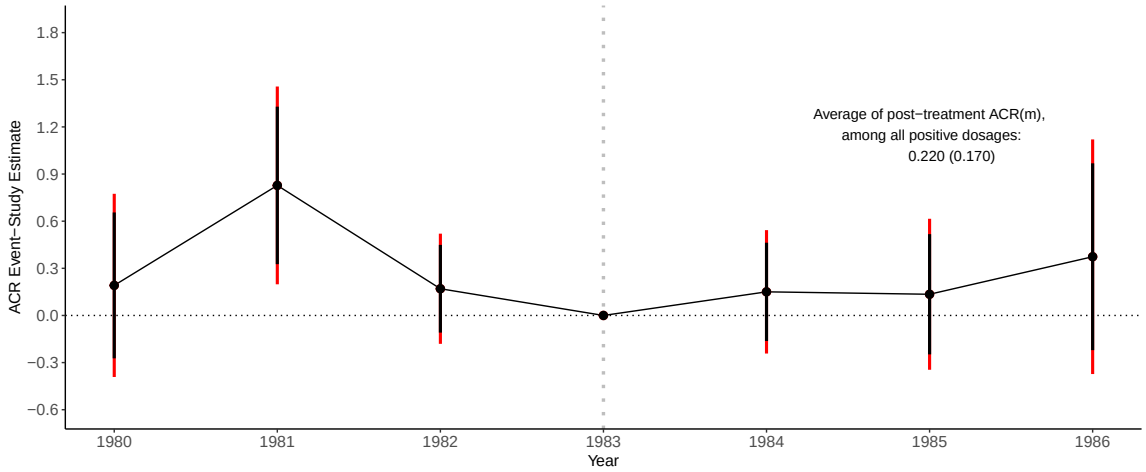
Under Assumption SPT, one policy implication of these estimates is that Medicare could have achieved similar, if not greater, capital investments while providing lower capital subsidies. Figure 9 shows that marginal increments in the subsidy ratio increase capital intensity only for those with low subsidy levels. Hospitals that received large capital subsidies under PPS responded with smaller increases in capital intensity than hospitals with slightly smaller subsidies, a fact easily seen in the binned scatter plot in Figure 4. The strong parallel trends assumption means that these estimated responses are "externally valid" for all treated hospitals, which means that only low subsidies matter for hospitals' input choices. Because higher subsidy ratios do not create further investments in capital, capping capital subsidies may not affect input choices very much.

An important economic implication, however, is that negative $ACRT(m)$ estimates contradict AF's two-factor economic model. $ACRT(m)$ is proportional to the average derivative of the optimal capital-labor ratio for hospitals with Medicare share equal to m, and Equation (6.1) shows specifically how it relates to the elasticity of substitution, $\sigma_{i,t}(m)$. To approximate $\mathbb{E}[\sigma_{i,t=2}(m)|M > 0]$, we separate out the two terms in (6.1) and construct $\frac{ACRT(m)}{\mathbb{E}[Y_{i,t=2}|M=m]} \frac{1-s_k m}{s_k}$ assuming that $s_k = 0.75$. With only two inputs, a rise in the relative price of one must lead to a reduction in its relative use: the elasticity of substitution must be positive. The point estimates of $\mathbb{E}[\sigma_{i,t=2}(m)|M > 0]$ do not fit that

prediction, although our uniform confidence bands do not reject an average elasticity of substitution of zero. Alternative models, such as a three-factor production function (which AF consider in their working paper) or non-homothetic production, could potentially rationalize this finding.

Finally, both the policy and structural interpretations of Figure 9 depend on the strong parallel trends assumption. Without SPT, the slope of $ATT(m|m)$ may be negative for higher-Medicare-share hospitals simply because their treatment effect functions are systematically lower. Medicare might not have been able to achieve similar capital increases with lower subsidy rates if high-subsidy hospitals just responded differently to low subsidy levels than low-subsidy hospitals did. A negative slope also does not necessarily reject a two-factor production model; just a constant-coefficient model with homogeneous firms, as considered by AF.

Figure 10: Event-Study Estimates of $ACRT$



Notes: The figure plots the event-study estimates of $ACRT^{es}_{\texttt{glob}}(e)$, with their 95% pointwise confidence intervals reported in black, and the 95% uniform confidence bands reported in red

Another indirect way to assess the plausibility of SPT that justifies a causal interpretation of $ACRT^{\texttt{glob}}$ is to compute $ACRT^{es}_{\texttt{glob}}(e)$, the event-study version of $ACRT^{\texttt{glob}}$. These parameters can be estimated using the same procedure discussed in Section 4, and we plot these in Figure 10. The no-anticipation assumption means that prior to treatment, when all observed outcomes are untreated potential outcomes, both Assumptions PT and SPT have the same implication: that the average relationship between outcome changes for adjacent dose groups should be zero. Our estimates of these pre-trends reject this in 1981, which is a pre-treatment period. Figure 10 corroborates our conclusions about the implausibility of SPT based on implausibly high implied elasticities of substitution.

In summary, our empirical results align with AF's conclusion that the 1983 Medicare reform led hospitals to favor capital over labor. We find evidence against parallel trends in pre-treatment periods, though the magnitudes of these violations are small relative to estimated effects in post-treatment periods. Finally, our negative estimates of $ACRT(m)$ at high values of m cut against the theoretical predictions of the model discussed above; this provides a piece of evidence that casts doubt on the plausibility of strong parallel trends in this application, indicating that one should be cautious when interpreting $ACRT$ parameters.

References

Acemoglu, Daron and Amy Finkelstein (2008). "Input and technology choices in regulated industries: Evidence from the health care sector". *Journal of Political Economy* 116.5, pp. 837–880.

Ai, Chunrong and Xiaohong Chen (2007). "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables". *Journal of Econometrics* 141, pp. 5–43.

American Hospital Association (1986). *AHA Annual Survey Database*. Tech. rep. Health Forum, LLC.

Angrist, Joshua D (1998). "Estimating the labor market impact of voluntary military service using Social Security data on military applicants". *Econometrica* 66.2, pp. 249–288.

Angrist, Joshua D, Kathryn Graddy, and Guido W Imbens (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish". *The Review of Economic Studies* 67.3, pp. 499–527.

Angrist, Joshua D and Guido W Imbens (1995). "Two-stage least squares estimation of average causal effects in models with variable treatment intensity". *Journal of the American Statistical Association* 90.430, pp. 431–442.

Angrist, Joshua D and Jorn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H.C. Sant'Anna (2025). "Difference-in-differences designs: A practitioner's guide". *Journal of Economic Literature* Forthcoming.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2025). "When is TSLS actually LATE?" *Review of Economic Studies* Forthcoming.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2024). "Revisiting event-study designs: Robust and efficient estimation". *Review of Economic Studies* 91.6, pp. 3253–3285.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna (2024). "Event studies with a continuous treatment". *AEA Papers and Proceedings* 114, pp. 601–605.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with multiple time periods". *Journal of Econometrics* 225.2, pp. 200–230.

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chen, Xiaohong, Timothy Christensen, and Sid Kankanala (2025). "Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities". *Review of Economic Studies* 92.1, pp. 162–196.

Chodorow-Reich, Gabriel, Plamen T. Nenov, and Alp Simsek (2021). "Stock market wealth and the real economy: A local labor market approach". *American Economic Review* 111.5, pp. 1613–57.

de Chaisemartin, Clement and Xavier D'Haultfoeuille (2018). "Fuzzy differences-in-differences". *The Review of Economic Studies* 85.2, pp. 999–1028.

— (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". *American Economic Review* 110.9, pp. 2964–2996.

de Chaisemartin, Clément and Xavier D'Haultfoeuille (2025). "Difference-in-differences estimators of intertemporal treatment effects". *Review of Economics and Statistics*, pp. 1–45.

de Chaisemartin, Clément, Xavier D'Haultfoeuille, Félix Pasquier, Doulo Sow, and Gonzalo Vazquez-Bare (2025). "Difference-in-differences for continuous treatments and instruments with stayers". Working Paper.

Fricke, Hans (2017). "Identification based on difference-in-differences approaches with multiple treatments". *Oxford Bulletin of Economics and Statistics* 79.3, pp. 426–433.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* 225.2, pp. 254–277.

Hendren, Nathaniel (2016). "The policy elasticity". *Tax Policy and the Economy* 30.1, pp. 51–89.

Meyer, Bruce D. (1995). "Natural and quasi-experiments in economics". *Journal of Business & Economic Statistics* 13.2, pp. 151–161.

Mogstad, Magne and Alexander Torgovitsky (2024). "Instrumental variables with unobserved heterogeneity in treatment effects". *Handbook of Labor Economics*. Vol. 5. Elsevier, pp. 1–114.

Newey, Whitney K. and Thomas M. Stoker (1993). "Efficiency of weighted average derivative estimators and index models". *Econometrica* 61.5, pp. 1199–1223.

Office of Technology Assessment (1984). "Medical Technology and Costs of the Medicare Program". OTA-H-227.

Robins, James (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Mathematical Modelling* 7.9, pp. 1393–1512.

Saez, Emmanuel, Joel Slemrod, and Seth H. Giertz (2012). "The elasticity of taxable income with respect to marginal tax rates: A critical review". *Journal of Economic Literature* 50.1, pp. 3–50.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* 225.2, pp. 175–199.

Wooldridge, Jeffrey M (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

— (2025). "Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators". *Empirical Economics* Forthcoming.

Yitzhaki, Shlomo (1996). "On using linear regressions in welfare economics". *Journal of Business & Economic Statistics* 14.4, pp. 478–486.

A Proofs of Main Results

A.1 Proofs of Results in Section 3.2

This section contains the proofs of the results in Section 3.2 on identifying causal effect parameters such as $ATT(d|d)$ and $ATT(d)$ under parallel trends assumptions and with a continuous treatment or multi-valued discrete treatment.

Proof of Theorem 3.1

Proof. To show the result, notice that

$$\begin{aligned} ATT(d|d) &= \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)|D=d] \\ &= \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D=d] - \mathbb{E}[Y_{t=2}(0) - Y_{t=1}(0)|D=d] \\ &= \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D=d] - \mathbb{E}[Y_{t=2}(0) - Y_{t=1}(0)|D=0] \\ &= \mathbb{E}[\Delta Y|D=d] - \mathbb{E}[\Delta Y|D=0] \end{aligned} \tag{A.1}$$

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t=1}(0)|D=d]$, the third equality holds by Assumption PT, and the last equality holds because $Y_{t=2}(d)$ and $Y_{t=1}(0)$ are observed potential outcomes when $D=d$ and $Y_{t=2}(0)$ and $Y_{t=1}(0)$ are observed potential outcomes when $D=0$. That $ATT^{\texttt{loc}}$ is identified holds immediately given its definition and that $ATT(d|d)$ is identified. To derive the particular expression for $ATT^{\texttt{loc}}$, notice that

$$ATT^{\texttt{loc}} = \mathbb{E}\Big[ATT(D|D)\Big|D>0\Big]$$

$$= \mathbb{E}\Big[\Big(\mathbb{E}[\Delta Y|D] - \mathbb{E}[\Delta Y|D=0]\Big)\Big|D>0\Big]$$
$$= \mathbb{E}[\Delta Y|D>0] - \mathbb{E}[\Delta Y|D=0]$$

where the first equality is the definition of $ATT^{\texttt{loc}}$, the second equality holds from Equation (A.1), the first part of the third equality holds by an implication of the law of iterated expectations, and the second part of the third equality holds because $\mathbb{E}[\Delta Y|D=0]$ is non-random. □

Proof of Theorem 3.2

Proof. To prove part (a), notice that

$$\begin{aligned}\mathbb{E}[\Delta Y|D=h] - \mathbb{E}[\Delta Y|D=l] &= \Big(\mathbb{E}[\Delta Y|D=h] - \mathbb{E}[\Delta Y|D=0]\Big) - \Big(\mathbb{E}[\Delta Y|D=l] - \mathbb{E}[\Delta Y|D=0]\Big) \\ &= ATT(h|h) - ATT(l|l) \end{aligned} \tag{A.2}$$

where the first equality holds by adding and subtracting $\mathbb{E}[\Delta Y|D=0]$, and the second equality holds by Theorem 3.1. Next,

$$\begin{aligned} ATT(h|h) - ATT(l|l) &= \mathbb{E}[Y_{t=2}(h) - Y_{t=2}(0)|D=h] - \mathbb{E}[Y_{t=2}(l) - Y_{t=2}(0)|D=l] \\ &= \mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h] \\ &\quad + \mathbb{E}[Y_{t=2}(l) - Y_{t=2}(0)|D=h] - \mathbb{E}[Y_{t=2}(l) - Y_{t=2}(0)|D=l] \\ &= \mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h] + \Big(ATT(l|h) - ATT(l|l)\Big) \end{aligned} \tag{A.3}$$

where the first equality holds by the definition of $ATT(d|d)$, the second equality holds by adding and subtracting $\mathbb{E}[Y_{t=2}(l)|D=h]$, and the third equality holds by the definition of $ATT(l|h)$ and $ATT(l|l)$. Notice that $\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h]$ is a causal response of going from dose l to dose h for dose group h. An alternative expression for this term is

$$\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h] = ATT(h|h) - ATT(l|h) \tag{A.4}$$

Next, we prove part (b). Using a similar argument as above, notice that, for $d \in \mathcal{D}_+^c$ and $(d+h) \in \mathcal{D}_+^c$,

$$\begin{aligned}\frac{\mathbb{E}[\Delta Y|D=d] - \mathbb{E}[\Delta Y|D=d+h]}{h} &= \frac{ATT(d|d) - ATT(d+h|d+h)}{h} \\ &= \frac{ATT(d|d) - ATT(d+h|d)}{h} + \frac{ATT(d+h|d) - ATT(d+h|d+h)}{h}\end{aligned}$$

where the first equality holds using the same argument as for Equation (A.2), and the second equality holds by using the arguments in Equations (A.3) and (A.4). The result holds by taking the limit as $h \to 0$ and the definition of $ACRT(d|d)$.

Finally, the second result in part (c) involving a discrete treatment holds by taking $h = d_j$ and $l = d_{j-1}$ in Equations (A.2) and (A.3) and by the definition of $ACRT(d_j|d_j)$. □

Proof of Theorem 3.3

Proof. For part (a), notice that

$$ATT(d) = \mathbb{E}[Y_{t=2}(d) - Y_{t=2}(0)|D>0]$$

$$
\begin{aligned}
&= \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D > 0] - \mathbb{E}[Y_{t=2}(0) - Y_{t=1}(0)|D > 0] \\
&= \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D = d] - \mathbb{E}[Y_{t=2}(0) - Y_{t=1}(0)|D = 0] \\
&= \mathbb{E}[\Delta Y|D = d] - \mathbb{E}[\Delta Y|D = 0]
\end{aligned}
$$

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t=1}(0)|D > 0]$, the third equality holds by Assumption SPT, and the fourth equality holds because $Y_{t=2}(d)$ and $Y_{t=1}(0)$ are observed outcomes when $D = d$.

Next, we prove the first part of part (b). First, notice that

$$
\begin{aligned}
ATT(h) - ATT(l) &= \mathbb{E}[Y_{t=2}(h) - Y_{t=2}(0)|D > 0] - \mathbb{E}[Y_{t=2}(l) - Y_{t=2}(0)|D > 0] \\
&= \mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D > 0]
\end{aligned}
$$

where the first equality holds by the definition of $ATT(d)$, and the second equality holds by cancelling the terms involving $Y_{t=2}(0)$. For the second part, notice that, from part (a), we have that

$$
\begin{aligned}
ATT(h) - ATT(l) &= \Big(\mathbb{E}[\Delta Y|D = h] - \mathbb{E}[\Delta Y|D = 0]\Big) - \Big(\mathbb{E}[\Delta Y|D = l] - \mathbb{E}[\Delta Y|D = 0]\Big) \\
&= \mathbb{E}[\Delta Y|D = h] - \mathbb{E}[\Delta Y|D = l]
\end{aligned}
$$

Now, for part (c), notice that for $d \in \mathcal{D}_+^c$ and $(d + h) \in \mathcal{D}_+^c$,

$$
\frac{ATT(d) - ATT(d + h)}{h} = \frac{\mathbb{E}[\Delta Y|D = d] - \mathbb{E}[\Delta Y|D = d + h]}{h}
$$

which follows from part (b). The result holds by taking the limit as $h \to 0$ and from the definition of $ACRT(d)$. Finally, the result in part (d) involving a discrete treatment holds from part (b) by taking $h = d_j$ and $l = d_{j-1}$ and by the definition of $ACRT(d_j)$. □

Proof of Corollary 3.1

Proof. The result holds immediately by averaging the results in Theorem 3.3 over the distribution of the dose among dose groups that experienced any positive amount of the treatment. □

B Adapting CCK to DiD Contexts

In this Appendix, we provide more details on how to adapt the Chen, Christensen, and Kankanala (2025) (henceforth, CCK) data-driven nonparametric estimation and inference procedures in our DiD context. As discussed in Section 4.1, the CCK estimator for $ATT(d)$ and $ACRT(d)$ are given by

$$
\widehat{ATT}_{\text{cck}}(d) = \left(\psi^{\widehat{K}}(d)\right)' \widehat{\beta}_{\widehat{K}}, \qquad \widehat{ACRT}_{\text{cck}}(d) = \left(\partial\psi^{\widehat{K}}(d)\right)' \widehat{\beta}_{\widehat{K}},
$$

where $\psi^K(d)$ is a K-dimensional vector of cubic B-splines basis functions, $\partial\psi^K(s) = (d\psi_{K1}(s)/\,ds, \ldots, d\psi_{KK}(s)/\,ds)'$, $\widehat{\beta}_{\widehat{K}}$ is the $\widehat{K}$-dimension vector of OLS estimators for $\beta_{\widehat{K}}$, and $\widehat{K}$ is the CCK data-driven estimator for the optimal sieve dimension. Henceforth, let $n_{D>0} = \sum_{i=1}^n 1\{D_i > 0\}$ be the sample size with positive treatment dose.

To discuss the optimal choice of the sieve dimension K derived in CCK, we need to add more

notation. Let $\mathcal{K} = \left\{\left(2^k + 3\right) : k \in \mathbb{N}_+ \cup 0\right\}$ be the set of possible sieve dimensions for the cubic B-splines. Let $K^+ = \min\{k \in \mathcal{K} : k > K\}$ be the smallest sieve dimension in $\mathcal{K}$ exceeding K, and $v_n = \max\{1, (0.1 \log n)^4\}$. Let $\{\omega_i\}_{i=1}^n$ be iid standard normal draws independent of the data $\{W_i\}_{i=1}^n = \{Y_{i,t=2}, Y_{i,t=1}, D_i\}_{i=1}^n$. In addition, let

$$\widehat{\varphi}_K(W_i, d) = \left(\psi^K(d)\right)' \widehat{\phi}_K(W_i),$$

with

$$\widehat{\phi}_K(W_i) = \mathbb{E}_n\left[1\{D > 0\} \cdot \psi^K(D)\psi^K(D)'\right]^- 1\{D_i > 0\}\psi^K(D_i)\widehat{u}_{i,K},$$

and $\widehat{u}_{i,K} = \Delta Y_i - \mathbb{E}_n[\Delta Y | D = 0] - \left(\psi^K(D_i)\right)' \widehat{\beta}_K$. Finally, for a given K and K_2, let

$$\widehat{\sigma}^2_{K,K_2}(d) = \frac{1}{n}\sum_{i=1}^n \left(\widehat{\varphi}_K(W_i, d) - \widehat{\varphi}_{K_2}(W_i, d)\right)^2$$

be an estimator of the (asymptotic) variance of the contrast $\sqrt{n}\left(\widehat{ATT}_K(d) - \widehat{ATT}_{K_2}(d)\right)$, and consider the bootstrap process

$$\mathbb{Z}^*_n(d, K, K_2) = \frac{1}{\widehat{\sigma}_{K,K_2}(d)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\widehat{\varphi}_K(W_i, d) - \widehat{\varphi}_{K_2}(W_i, d)\right) \cdot \omega_i\right).$$

For a given sieve dimension $K \in \mathcal{K}$, let

$$\widehat{ATT}_K(d) = \left(\psi^K(d)\right)' \widehat{\beta}_K, \qquad \widehat{ACRT}_K(d) = \left(\partial\psi^K(d)\right)' \widehat{\beta}_K, \tag{B.1}$$

where $\partial\psi^K(s) = (\, d\psi_{K1}(s)/\, ds, \ldots, d\psi_{KK}(s)/\, ds)'$,

$$\begin{aligned}\widehat{\beta}_K &= \arg\min_{b_K \in \Theta_K} \mathbb{E}_n\left[\left(\Delta Y - \mathbb{E}_n\left[\Delta Y | D = 0\right] - \psi^K(D)' b_K\right)^2 \middle| D > 0\right] \\ &= \mathbb{E}_n\left[1\{D > 0\}\psi^K(D)\psi^K(D)'\right]^- \mathbb{E}_n\left[1\{D > 0\}\psi^K(D)\left(\Delta Y - \mathbb{E}_n\left[\Delta Y | D = 0\right]\right)\right],\end{aligned} \tag{B.2}$$

and A^- denote the Moore-Penrose inverse of a generic matrix A, and for a generic variable B,

$$\mathbb{E}_n[B | D > 0] = \frac{\sum_{i=1}^n 1\{D_i > 0\}B_i}{\sum_{i=1}^n 1\{D_i > 0\}}.$$

The next algorithm adapts Procedure 1 of CCK to our DiD context and provides the Lepski-type data-driven selection $\widehat{K}$ of the sieve dimension K.

Algorithm 1 (Computation of data-driven choice of sieve-dimension K based on CCK.)**.**

1. *Compute the data-driven index set of sieve dimensions*

$$\widehat{\mathcal{K}} = \left\{K \in \mathcal{K} : 0.1\left(\log \widehat{K}_{max}\right)^2 \leq K \leq \widehat{K}_{max}\right\} \tag{B.3}$$

where

$$\widehat{K}_{max} = \min\left\{K \in \mathcal{K} : K\sqrt{\log K}v_n \leq 10\sqrt{n} < K^+\sqrt{\log K^+}v_n\right\} \tag{B.4}$$

2. *Let* $\widehat{\alpha} = \min\left\{0.5, \sqrt{\log \widehat{K}_{max}/\widehat{K}_{max}}\right\}$*. For each independent draw of* $\{\omega_i\}_{i=1}^n$*, compute*

$$\sup_{(d,K,K_2) \in \mathcal{D}^c_+ \times \widehat{\mathcal{K}} \times \widehat{\mathcal{K}} : K_2 > K} \left|\mathbb{Z}^*_n(d, K, K_2)\right|. \tag{B.5}$$

Let $\gamma^*_{1-\widehat{\alpha}}$ *denote the* $(1 - \widehat{\alpha})$ *quantile of the sup-t statistic* (B.5) *across a large number of independent draws of* $\{\omega_i\}_{i=1}^n$*, say, 1,000.*

3. The data-driven choice of the sieve dimension is

$$\widehat{K} = \inf\left\{ K \in \widehat{\mathcal{K}} \ : \sup_{(d,K_2)\in \mathcal{D}_+^c \times \widehat{\mathcal{K}}: K_2 > K} \frac{\sqrt{n}\left|\widehat{ATT}_K(d) - \widehat{ATT}_{K_2}(d)\right|}{\widehat{\sigma}_{K,K_2}(d)} \leq 1.1\gamma^*_{1-\widehat{\alpha}} \right\}. \tag{B.6}$$

Next, we show how one can form data-driven uniform confidence bands (UCBs) for both $ATT(d)$ and $ACRT(d)$ by adapting Procedure 2 of CCK to our DiD context. Toward this end, let $\widehat{A} = \log\log \widehat{K}$ and set $\widehat{\mathcal{K}}_- = \{K \in \widehat{\mathcal{K}} : K < \widehat{K}\}$. Define the bootstrap processes

$$\mathbb{Z}^*_n(d,K) = \frac{1}{\widehat{\sigma}_K(d)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\varphi}_K(W_i,d)\cdot \omega_i, \quad \text{and} \quad \mathbb{Z}^{*,acr}_n(d,K) = \frac{1}{\widehat{\sigma}^{acrt}_K(d)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\varphi}^{acrt}_K(W_i,d)\cdot \omega_i.$$

where $\widehat{\varphi}^{acrt}_K(W_i,d) = \left(\partial\psi^K(d)\right)' \widehat{\phi}_K(W_i)$,

$$\widehat{\sigma}^2_K(d) = \frac{1}{n}\sum_{i=1}^n \widehat{\varphi}_K(W_i,d)^2, \qquad \text{and} \qquad \widehat{\sigma}^{acrt,2}_K(d) = \frac{1}{n}\sum_{i=1}^n \widehat{\varphi}^{acrt}_K(W_i,d)^2.$$

Algorithm 2 (Computation of UCBs for $ATT(\cdot)$ and $ACRT(d)$ based on CCK.)**.**

4. For each independent draw of $\{\omega_i\}_{i=1}^n$, compute

$$t^* = \sup_{(d,K)\in \mathcal{D}_+^c \times \widehat{\mathcal{K}}_-} \left|\mathbb{Z}^*_n(d,K)\right|, \quad \textit{and} \quad t^{*,acr} = \sup_{(d,K)\in \mathcal{D}_+^c \times \widehat{\mathcal{K}}_-} \left|\mathbb{Z}^{*,acr}_n(d,K)\right|. \tag{B.7}$$

Let $z^*_{1-\alpha}$ and $z^{*,acr}_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the sup-t statistic t^* and $t^{*,acr}$, respectively, across a large number of independent draws of $\{\omega_i\}_{i=1}^n$, say, 1,000.

5. The data-driven $100(1-\alpha)\%$ UCB for $ATT(d)$ and $ACRT(d)$, $d \in \mathcal{D}_+^c$, are respectively given by

$$C_n(d) = \left[\widehat{ATT}_{\widehat{K}}(d) - \left(z^*_{1-\alpha} + \widehat{A}\ \gamma^*_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}_{\widehat{K}}(d)}{\sqrt{n}},\ \widehat{ATT}_{\widehat{K}}(d) + \left(z^*_{1-\alpha} + \widehat{A}\ \gamma^*_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}_{\widehat{K}}(d)}{\sqrt{n}}\right] \tag{B.8}$$

$$C^{acrt}_n(d) = \left[\widehat{ACRT}_{\widehat{K}}(d) - \left(z^{*,acr}_{1-\alpha} + \widehat{A}\ \gamma^*_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}^{acrt}_{\widehat{K}}(d)}{\sqrt{n}},\ \widehat{ACRT}_{\widehat{K}}(d) + \left(z^{*,acr}_{1-\alpha} + \widehat{A}\ \gamma^*_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}^{acrt}_{\widehat{K}}(d)}{\sqrt{n}}\right] \tag{B.9}$$

C Multiple Periods and Variation in Treatment Timing and Dose

DiD applications often use more than two time periods, wherein treatments, whether binary or not, can turn on at different times for different units. This section extends the results from the main text to allow for multiple time periods ($t = 1, ..., T$) with variation in the time when units become treated. We refer to the time period when a unit becomes treated as a unit's *timing group*, which we denote by G_i, which takes values in the set $\mathcal{G}$. By convention, we set $G = \infty$ for units that remain untreated across all time periods, and we exclude units that are treated in the first period so that $\mathcal{G} \subseteq \{2, \ldots, T, \infty\}$; we also set $\bar{\mathcal{G}} = \mathcal{G} \setminus \{\infty\}$ to be the set of all timing groups that ever participate in the treatment. Treated units receive dose $D = d \in \mathcal{D}_+$. As in the two-period case, the dose actually experienced, D, also defines a unit's *dose group*.

We extend the potential outcomes notation from the previous section to allow for variation in treatment timing. Therefore, potential outcomes $Y_{i,t}(g,d)$ denote the outcome for unit i at time period t when such a unit is first treated in period g with dosage d. Note that treated potential outcomes at time t depend on when a unit first becomes treated—i.e., $Y_{i,t}(g,d)$ may not equal $Y_{i,t}(g',d)$ for

$g \neq g'$—which allows for general treatment effect dynamics. $Y_{i,t}(\infty, 0)$ is the outcome that unit i would experience if it did not participate in the treatment in any period. We write $Y_{i,t}(0) = Y_{i,t}(\infty, 0)$ and refer to this as a unit's untreated potential outcome. We also define the variable $W_{i,t} = D_i \mathbf{1}\{t \geq G_i\}$. which is the amount of dose that unit i experiences in time period t; $W_{i,t} = 0$ for all units that are not yet treated by time period t.

Throughout this section, we make the following assumptions.

Assumption 1-MP (Random Sampling)**.** *The observed data consists of* $\{Y_{i1}, \ldots, Y_{iT}, D_i, G_i\}_{i=1}^n$ *which is independent and identically distributed.*

Assumption 2-MP (Support)**.**

(a) The support of D, $\mathcal{D} = \{0\} \cup \mathcal{D}_+$ *where* $\mathcal{D}_+ \subseteq (0, \infty)$. *In addition,* $\mathbb{P}(D = 0) > 0$ *and* $dF_{D|G}(d|g) > 0$ *for all* $(g, d) \in \bar{\mathcal{G}} \times \mathcal{D}_+$.

(b) $\mathcal{D}_+^c = (d_L, d_U) \subset \mathcal{D}_+$. *In addition, for all* $g \in \bar{\mathcal{G}}$ *and* $t = 2, \ldots, T$, $\mathbb{E}[\Delta Y_t | G = g, D = d]$ *is continuously differentiable in* d *on* $\mathcal{D}_+^c$.

Assumption 3-MP (No Anticipation / Staggered Adoption)**.**

(a) For all $g \in \mathcal{G}$ *and* $t = 1, \ldots, T$ *with* $t < g$ *(i.e., in pre-treatment periods),* $Y_{i,t}(g, d) = Y_{i,t}(0)$.

(b) $W_{i,1} = 0$ *almost surely, and, for* $t = 2, \ldots, T$ *and* $d \in \mathcal{D}_+$, $W_{i,t-1} = d$ *implies that* $W_{i,t} = d$.

We next introduce versions of Assumption PT and SPT that are suitable for the setting with multiple periods and variation in treatment timing. [13]

Assumption PT-MP (Parallel Trends with Multiple Periods and Variation in Treatment Timing)**.** *For all* $g \in \bar{\mathcal{G}}$, $t = 2, \ldots, T$, $d \in \mathcal{D}_+$, $\mathbb{E}[\Delta Y_t(0) | G = g, D = d] = \mathbb{E}[\Delta Y_t(0) | G = \infty, D = 0]$.

Assumption SPT-MP (Strong Parallel Trends with Multiple Periods and Variation in Treatment Timing)**.** *For all* $g \in \bar{\mathcal{G}}$, $t = 2, \ldots, T$, *and* $l, d \in \mathcal{D}_+$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(g, d) | G = g, D = l] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(g, d) | G = g, D = d]$ *and* $\mathbb{E}[\Delta Y_t(0) | G = g, D = d] = \mathbb{E}[\Delta Y_t(0) | G = \infty, D = 0]$.

For each unit, we observe their outcome in period t, $Y_{i,t}$, which is given by

$$Y_{i,t} = Y_{i,t}(0) \mathbf{1}\{t < G_i\} + Y_{i,t}(G_i, D_i) \mathbf{1}\{t \geq G_i\}.$$

C.1 Identification with a Staggered Continuous Treatment

The causal parameters of interest are the same as in our baseline case, except that they are separately defined for each timing group and in each post-treatment time period:

$$ATT(g, t, d | g, d) = \mathbb{E}[Y_t(g, d) - Y_t(0) | G = g, D = d], \quad \text{and} \quad ATT(g, t, d) = \mathbb{E}[Y_t(g, d) - Y_t(0) | G = g, D > 0].$$

[13] Besides differences related to multiple periods and variation in treatment timing, the version of strong parallel trends made here is slightly different from Assumption SPT in the main text. Part of the difference comes from there being no untreated units in group $g \in \bar{\mathcal{G}}$, which is why there is a separate part of the assumption for untreated potential outcomes. The other difference is that both parts of the assumption hold for all dose groups rather than on average (i.e., we condition on dose group l in the first part and on dose group d in the second part for untreated potential outcomes). The version here is stronger and is made for clarity and because we target $ACRT(g, t, d | g, d)$ rather than $ACRT(g, t, d)$ in this part of the paper.

Causal response parameters are similarly defined as the effect of a marginal change in the dose on the outcomes of timing group g in period t. For continuous treatments, these are defined as

$$ACRT(g,t,d|g,d) = \frac{\partial ATT(g,t,l|g,d)}{\partial l}\bigg|_{l=d} = \frac{\partial \mathbb{E}\left[Y_t(g,l)|G=g,D=d\right]}{\partial l}\bigg|_{l=d},$$
$$ACRT(g,t,d) = \frac{\partial ATT(g,t,d)}{\partial d} = \frac{\partial \mathbb{E}\left[Y_t(g,d)|G=g\right]}{\partial d}.$$

For discrete treatments, these are defined as

$$ACRT(g,t,d_j|g,d_j) = \mathbb{E}[Y_t(g,d_j) - Y_t(g,d_{j-1})|D=d_j, G=g]\Big/(d_j - d_{j-1}),$$
$$ACRT(g,t,d_j) = \mathbb{E}[Y_t(g,d_j) - Y_t(g,d_{j-1})|D>0, G=g]\Big/(d_j - d_{j-1}),$$

For brevity, henceforth we focus on the "local" causal effect parameters $ATT(g,t,d|g,d)$ and $ACRT(g,t,d|g,d)$, which are analogous to the local causal effect parameters $ATT(d|d)$ and $ACRT(d|d)$ in the two-period case that we emphasized in the main text.

Theorem C.1. *Under Assumptions 1-MP, 2-MP(a), 3-MP, and PT-MP, and for all $g \in \bar{\mathcal{G}}$, $t = 2,\ldots,T$ such that $t \geq g$, and for all $d \in \mathcal{D}_+$,*

$$ATT(g,t,d|g,d) = \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0].$$

If, in addition, Assumptions 2-MP(b) and SPT-MP hold, then, for all $d \in \mathcal{D}_+^c$,

$$ACRT(g,t,d|g,d) = \frac{\partial \mathbb{E}[Y_t - Y_{g-1}|G=g,D=d]}{\partial d}.$$

The proof of Theorem C.1 is provided in Appendix SC in the Supplementary Appendix. The result is broadly similar to the one in the case with two periods. The first part says that, under Assumption PT-MP, $ATT(g,t,d|g,d)$ can be recovered by a DiD comparison between the path of outcomes from period $g-1$ to period t for units in group g treated with dose d and the path of outcomes among units that have not participated in the treatment yet (the setup in this section also rationalizes using the never-treated group, $G=\infty$, as the comparison group as was mentioned in Section 5). Relative to the case with two time periods, the main difference is that the "base period" is $g-1$. The reason for using the base period $g-1$ is that it is the most recent time period when the researcher observes untreated potential outcomes for units in group g. Thus, the result is very much like the case with two time periods: take the most recent untreated potential outcomes for units in a particular group, impute the path of outcomes that they would have experienced in the absence of participating in the treatment from the group of not-yet-treated units (these steps yield mean untreated potential outcomes that units in group g would have experienced in time period t) and compare this to the outcomes that are actually observed for units in group g that experienced dose d. The second part says that, under Assumption SPT-MP, $ACRT(g,t,d|g,d)$ can be recovered by taking the derivative of the average path of outcomes from period $g-1$ to period t among timing group g that experienced dose d. Similarly to the arguments in the main text, if Assumption PT-MP held rather than SPT-MP, then the same derivative term would additionally include selection bias terms.

Given the results in Theorem C.1, it follows that causal summary parameters that are aggregations of these dose-and-timing-group-specific parameters are also identified. Of course, one can consider many different types of aggregation, as discussed in Callaway and Sant'Anna (2021) and Callaway, Goodman-Bacon, and Sant'Anna (2024), for example. Here, given that the treatment is continuous, we discuss some aggregations that can remain dose-specific and help highlight heterogeneity in treatment dosages. We provide these estimands as being explicitly about them if useful for estimation and inference, and provide more transparency when comparing across procedures (Baker et al., 2025).

We start discussing natural aggregated parameters, including the average treatment effect of dose d across post-treatment periods for dose group d,

$$ATT^{dose}(d|d) = \mathbb{E}\Big[\overline{TE}(d)\Big|D = d, G \leq T\Big] = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=2}^{T} \omega^{dose}(g,t,d)ATT(g,t,d|g,d),$$

where $\omega^{dose}(g,t,d) = \frac{\mathbf{1}\{t\geq g\}}{T-g+1}\mathbb{P}(G=g|D=d,G\leq T)$, and

$$\overline{TE}_i(d) = \frac{1}{T-G_i+1}\sum_{t=G_i}^{T}\Big(Y_{i,t}(G_i,d) - Y_{i,t}(0)\Big).$$

We can likewise define a causal response parameter

$$ACRT^{dose}(d|d) = \left.\frac{\partial ATT^{dose}(l|d)}{\partial l}\right|_{l=d} = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=2}^{T}\omega^{dose}(g,t,d)ACRT(g,t,d|g,d),$$

and even further aggregate these parameters into scalar summary parameters:

$$ATT^{\texttt{loc}} = \mathbb{E}\Big[ATT^{dose}(D|D)\Big|G\leq T\Big] \quad \text{and} \quad ACRT^{\texttt{loc}} = \mathbb{E}\Big[ACRT^{dose}(D|D)\Big|G\leq T\Big].$$

Empirical researchers are also often interested in analyzing how average treatment effects vary with elapsed treatment timing and consider event-study-type parameters. In our context, with continuous and staggered treatments, one can consider the following dose-specific event study parameters,

$$\widetilde{ATT}^{dose,es}(d|d,e) = \mathbb{E}\Big[TE(d|e)\Big|D=d, G+e\in[2,T], G\leq T\Big] = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=2}^{T}\omega^{dose,es}(g,t,d|e)ATT(g,t,d|g,d),$$

$$\widetilde{ACRT}^{dose,es}(d|d,e) = \left.\frac{\partial\widetilde{ATT}^{dose,es}(l|d,e)}{\partial l}\right|_{l=d} = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=2}^{T}\omega^{dose,es}(g,t,d|e)ACRT(g,t,d|g,d)$$

where $\widetilde{ATT}^{dose,es}(d|d,e)$ and $\widetilde{ACRT}^{dose,es}(d|d,e)$ are the average treatment effect of dose d and average causal response of dose d among those in dose group d for those that have been exposed to the treatment for e periods, $TE_i(d|e) = Y_{i,G_i+e}(G_i,d) - Y_{i,G_i+e}(0)$, $\omega^{dose,es}(g,t,d|e) = \mathbf{1}\{g+e\in[2,T]\}\mathbf{1}\{g+e=t\}\pi_g(e,d)$ and $\pi_g(e,d) = \mathbb{P}(G=g|D=d,G+e\in[2,T],G\leq T)$,

When one also wants to aggregate over treatment dosages to get an easier-to-estimate causal parameter of interest,

$$ATT^{es}_{\texttt{loc}}(e) = \mathbb{E}\Big[\widetilde{ATT}^{dose,es}(D|D,e)\Big|G+e\in[2,T],G\leq T\Big]$$
$$ACRT^{es}_{\texttt{loc}}(e) = \mathbb{E}\left[\widetilde{ACRT}^{dose,es}(D|D,e)\Big|G+e\in[2,T],G\leq T\right]$$

which provide event study versions of average treatment effects and average causal responses across different lengths of exposure to the treatment. For values of $e \geq 0$, $ATT^{es}_{\texttt{loc}}(e)$ and $ACRT^{es}_{\texttt{loc}}(e)$ are related to treatment effect dynamics. It is also interesting to consider cases where $e < 0$, which can be interpreted as a pre-test of the parallel trends assumption. See also Callaway, Goodman-Bacon, and Sant'Anna (2024) for a discussion.

Remark C.1. *We do not provide formal estimation results for the setting with multiple periods and variation in treatment timing. However, we note that, if one bases estimation on the sample analog of the results in Theorem C.1, then the results in the main text for the case with two periods apply directly to the disaggregated parameters* $ATT(g,t,d|g,d)$ *and* $ACRT(g,t,d|g,d)$*. Then, one can estimate any of the aggregated parameters discussed above as the appropriate weighted average of* $ATT(g,t,d|g,d)$ *or* $ACRT(g,t,d|g,d)$*. Interestingly, given the results in Corollary 3.1 and Callaway and Sant'Anna (2021), when* $ATT^{es}_{loc}(e)$ *is the target parameter, one can binarize the treatment (i.e., classify units as being treated if they experience any positive amount of the treatment) and simply rely on the event-study procedures proposed by Callaway and Sant'Anna (2021).*