Difference-in-Differences with a Continuous Treatment

Brantly Callaway[*] Andrew Goodman-Bacon[†] Pedro H.C. Sant'Anna[‡]

July 27, 2023

WORK-IN-PROGRESS. DO NOT CIRCULATE.

Abstract

This paper analyzes difference-in-differences setups with a continuous treatment. We show that treatment effect on the treated-type parameters can be identified under a generalized parallel trends assumption that is similar to the binary treatment setup. However, interpreting differences in these parameters across different values of the treatment can be particularly challenging due to treatment effect heterogeneity. We discuss alternative, typically stronger, assumptions that alleviate these challenges. We also provide a variety of treatment effect decomposition results, highlighting that parameters associated with popular linear two-way fixed-effect (TWFE) specifications can be hard to interpret, *even* when there are only two time periods. We introduce alternative estimation procedures that do not suffer from these TWFE drawbacks.

JEL Codes: C14, C21, C23

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

[*] 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 between treated and untreated groups (difference one), before and after treatment started (difference two). But 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 locations far away. Localities spend different amounts on public goods and services, or set different minimum wages. Students choose how long to stay in school.

Continuous treatments[1] can offer advantages over binary ones. 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.[2] Finally, we may care more about the effect of changes in treatment intensity (e.g., increased funding, pollution abatement, or expanded eligibility) than about the effect of the existence of a program that already exists.

Despite how conceptually useful and practically common continuous DiD designs are, econometric theory provides little guidance about how researchers should apply and interpret them. For cross-sectional designs, econometric results discuss how to estimate causal effects of small changes in a continuous treatment (Hirano and Imbens, 2004; Florens, Heckman, Meghir, and Vytlacil, 2008), and applied researchers often report this "marginal" interpretation when they use a continuous treatment in a DiD setting (Goodman-Bacon, 2018). But econometric theory research on continuous (and multi-valued) DiD designs is scarce and limited to identification results for individual-level treatment effects from aggregating binary treatment data as in "fuzzy" DiD designs (de Chaisemartin and D'Haultfœuille, 2018), or the causal effects of different multi-level treatments compared to no treatment (see the supplemental appendix of de Chaisemartin and D'Haultfœuille, 2020).[3] Moreover, following the advice in several prominent textbooks (e.g., Cameron and Trivedi, 2005, Angrist and Pischke, 2008, and Wooldridge, 2010), applied researchers almost universally estimate continuous DiD designs using two-way fixed effects (TWFE) regressions, which we now know that are not robust to treatment effect heterogeneity in other complex DiD designs such as staggered timing (Goodman-Bacon, 2021). Therefore, the theoretical gap in our understanding of continuous DiD designs contributes to ambiguity about the best way to implement and interpret such designs in practice. The main goal of this paper is to tackle this problem and provide a new set of well-understood and formally justified tools that are suitable for DiD setups with variations in treatment dosage.

We start our discussion by analyzing DiD designs in which units move from no treatment to a

[1] With some abuse of terminology, we refer to treatments being "continuous" to all cases where treatment can take several different levels. Thus, technically speaking, this includes continuous and multi-valued ordered discrete treatments. Whenever these distinctions are not crucial for the points we are making, we omit them.

[2] In his 1965 presidential address to the Royal Society of Medicine, Sir Austin Bradford Hill, a pioneer in the study of smoking and cancer, included among his criteria for inferring causality from observational data, "a biological gradient, or dose-response curve" and argued that "we should look most carefully for such evidence" (Hill, 1965).

[3] See also D'Haultfoeuille, Hoderlein, and Sasaki (2021) for Changes-in-Changes-type of procedures based on rank-invariance as in Athey and Imbens (2006).

non-zero dose. We first define two types of causal effects. The difference between a unit's potential outcome under dose d and its untreated potential outcome is a *level treatment effect*. The difference in a unit's potential outcome with a marginal increase in the dose is a *causal response* (Angrist and Imbens, 1995). Level treatment effects and causal responses can have meaningfully different interpretations, and we show that they 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, just like in binary DiD designs. Comparisons between adjacent dose groups, however, only identify average causal response parameters under a stronger assumption, which we call "strong parallel trends", that restricts treatment effect heterogeneity so that groups would have responded to the lower dose in the same way. Intuitively, to be a good counterfactual, lower-dose units must reflect how higher-dose units' outcomes would have changed without treatment *and* at the lower level of the treatment.Without the strong parallel trends assumption, comparisons across treatment dosages are "contaminated" with selection bias related to treatment effect heterogeneity.[4] These results come from comparisons between two groups, but also apply to estimators of the entire average level effect or average causal response curves or summary estimates that average across doses.

We use the identification results to evaluate the most common way that practitioners estimate a summary parameter in 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:

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

Under parallel trends, we decompose β^{twfe} into three different weighted sums corresponding to the causal parameter being used as the "building block": level effects, scaled level effects, and causal responses. None of the weighted sum representations provide a clear causal and policy-relevant interpretation of β^{twfe}.

For instance, expressing β^{twfe} as a weighted sum of average level treatment effect parameters shows that it is equivalent to a binary DiD with the treatment group defined as units with above-average doses and a comparison group of units with below-average doses, and with weights proportional to a unit's absolute distance from the mean dose. TWFE, therefore, puts "negative weights" on the treatment effects of lower-dose groups by using them as "controls". When units with below-average treatment dosage have non-zero level treatment effects, it is hard to attach a meaningful causal interpretation to β^{twfe} in terms of average level treatment effects. This is particularly true when the share of untreated units (dosage $d = 0$) is small in the population.[5]

[4] Interpreting comparisons of average treatment effect on the treated at different values of treatment dosage d is related to existing points made on comparing "local" treatment effect parameters to each other, e.g., Oreopoulos (2006), Angrist and Fernandez-Val (2013), and Mogstad, Santos, and Torgovitsky (2018) in the context of local average treatment effects, or Cattaneo, Titiunik, Vazquez-Bare, and Keele (2016) and Cattaneo, Keele, Titiunik, and Vazquez-Bare (2021) in the context of regression discontinuity designs with multiple cutoffs.

[5] We also present a similar decomposition based on average level treatment effect parameters scaled by their dose. It also puts negative weight on effects for below-average dose units, but weights these comparisons slightly differently.

On the other hand, the decomposition in terms of average casual responses parameters has no negative weights, but does include an additional "selection bias" term stemming from heterogeneous treatment effect functions across dose groups.[6] When one imposes the strong parallel trends assumption this "selection bias" term disappears. The weights on causal responses at different doses, however, differ from the distribution of the dose, which creates a further challenge to interpreting β^{twfe} in the presence of heterogeneity, even if strong parallel trends holds.

When TWFE fails to deliver interpretable causal parameters, what is the alternative? We propose nonparametric estimators of the average level treatment effect and average causal response curves based on Chen, Christensen, and Kankanala (2022). These tools are motivated by clearly defined parallel trends assumptions, do not rely on strong functional form assumptions, are easy to implement, are fully data-driven. By leveraging the procedures in Chen, Christensen, and Kankanala (2022), we show that our estimators converge at the fastest possible (i.e., minimax) rate in sup-norm, and our uniform confidence bands are asymptotically narrower (more precise) than those based on undersmoothing, and yet have correct asymptotic coverage and contract at, or within a $\log \log n$ factor of, the minimax rate. We also show how to construct easy-to-interpret summary measures that use the treatment dosage density to average parameters across doses. For average level treatment effects, estimating this summary parameter is as simple as running a binary DiD with a "treatment dummy" equal to one for any units with positive doses.

To show how TWFE performs in practice and to illustrate the benefits of our proposed estimators, we replicate Acemoglu and Finkelstein (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 that favor 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; a positive causal response, under some assumptions, reflects the sign and magnitude of the elasticity of substitution. Decomposing the TWFE estimate in terms of level effects shows that 38 percent of hospitals have negative weights and that they have non-negligible effects. [BLAH]

2 A Running Example: Acemoglu and Finkelstein (2008)

To fix ideas and provide intuition for our theoretical results, we revisit Acemoglu and Finkelstein (2008)'s (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 that was proportional to the share 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

[6] In an appendix, we extend these baseline results to a setup with more than two time periods and where treatment varies in intensity as well as timing, generalizing the results in de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021) to the case with a continuous treatment.

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 he 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.[7] AF's description, estimation, and interpretation of this empirical strategy touch on some of the most common ways that researchers justify and implement continuous DiD designs.

One motivation for this design is practical: variation in a dose permits the evaluation of treatments for which binary DiD is either infeasible or undesirable. In AF's case, about 15 percent of hospitals specifically served non-Medicare-eligible populations such as children so they were "untreated" by the change in subsidy policy. But this meant that they differed from treated hospitals in terms of patient mix and may not necessarily constitute a valid comparison group. AF therefore describe m_i 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.[8]

Another common justification for continuous DiD designs is that a "dose-response" relationship between exposure and outcomes supports a causal interpretation or tests theoretical predictions. Meyer (1995, pg. 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".[9] AF lay out a simple theoretical framework showing that if hospitals have identical homothetic production functions, then 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

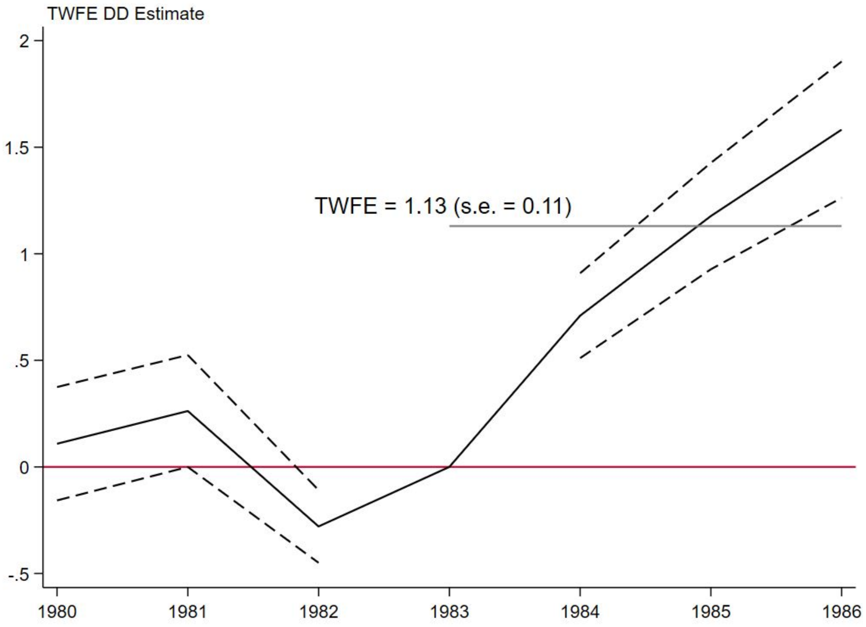
[7] AF use data reported by hospitals each year to the American Hospital Association from 1980 to 1986 (CITE). They proxy for the capital/labor ratio using the depreciation share of total operating expenses, which averages about 4.5 percent in their period.

[8] A good example of an analysis with no untreated units is Card (1992), who exploits geographic differences in the "bite" of a 1991 federal minimum wage increase. In a statutory sense, the federal change affected all workers, so there is no untreated group to use in a binary DiD, and while it should have affected lower-wage workers more directly than higher-wage workers, comparing these groups would require longitudinal data. Instead, Card regresses the change in each state's teen employment rate on the share of teens in that state who earned less than the new minimum wage in the pre-period and are thus "eligible" for a statutory wage increase. Converting the analysis to the state-level DiD creates a continuous treatment variable that can identify effects of a federal policy.

[9] Hill (1965) makes this point in the context of smoking and cancer:

> "The fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers."

He also notes that more deaths among light rather than heavy smokers would weaken the causal claim unless one could "envisage some much more complex relationship to satisfy the cause-and-effect hypothesis."



Notes: The figure plots TWFE event-study coefficients 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 or the year dummies. 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 (CITE).

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

estimate 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 only be estimated accurately with continuous variation. We discuss how AF's theoretical framework implies, under some assumptions, that DiD estimates can be used to learn about hospitals' elasticity of substitution between capital and labor, although AF do not argue for this kind of "marginal" interpretation.

In terms of estimation, AF follow the standard practice for continuous DiD designs: a TWFE regression with hospital and year fixed effects. They follow textbook advice. Wooldridge (2010, pg. 132) observes that a two-period DiD regression estimator "can be easily modified to allow for continuous, or at least nonbinary, 'treatments' ". Angrist and Pischke (2008, pg. 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...the minimum wage is therefore a variable with differing treatment intensity across states and over time".

Figure 1 reproduces AF's DiD event-study coefficients for each calendar year relative to 1983 and the estimate of β^{twfe} from equation (1). The findings clearly show that after 1983 capital/labor ratios rose more strongly for hospitals with higher values of m_i, but there was no differential change in input mix before PPS. 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].

Our impression is that event-study results like those in Figure 1 would usually be interpreted as very strong causal evidence because there are small pre-trend estimates, large differences in outcomes between higher- and lower-dose units after treatment, and tight confidence intervals. What is missing from most continuous (or nonbinary) 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 central issues.

3 Baseline Case: A New Treatment with Two Periods

We illustrate our main points in a setup where a researcher has access to two periods of panel data denoted by t and $t-1$. In the first period, no unit is treated. In the second period, some units receive a treatment "dose" denoted by D_i, and some others remain untreated. We denote the support of D by $\mathcal{D}$. We define potential outcomes for unit i in period $s \in \{t-1, t\}$ by $Y_{is}(d)$. This is the outcome that unit i would experience in period s under dose d. D_i can be (absolutely) continuous or can be multi-valued ordered, but to simplify the exposition we refer to it as "continuous". We assume that all expectations are finite and well-defined.[10]

3.1 Parameters of Interest with a Continuous Treatment

The potential outcomes notation $Y_t(d)$ reflects that treatment can take many values, which also means that each unit can experience many types of causal effects. The *level treatment effect* of dose d in time period t for a given unit equals its potential outcome when $D = d$ minus its untreated potential outcome: $Y_t(d) - Y_t(0)$. This is a straightforward extension of a binary "treatment effect" to a continuous "treatment effect function" or "dose-response function."[11]

But no treatment is not the only relevant counterfactual possible. We define a unit's *causal response* at d as $Y_t'(d)$, the derivative of the potential outcome[12] (when d is continuous) or as the difference in potential outcomes between adjacent doses, $Y_t(d_j) - Y_t(d_{j-1})$ (when d is discrete). 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 or continuous endogenous variables (Angrist and Imbens, 1995, Angrist, Graddy, and Imbens, 2000).

[10] A sufficient condition for this is when all potential outcomes $Y_t(d)$ are bounded.

[11] We include i subscripts for units in expressions that refer to sample quantities but not in theoretical expressions of population quantities.

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

We focus on "building block" parameters that are averages of these two kinds of causal effects. Average level treatment effects extend definitions from the binary case so that they refer to the average effect of being treated with a particular dose compared to not being treated. In particular, we define

$$ATT(d|d') = \mathbb{E}[Y_t(d) - Y_t(0)|D = d'] \quad \text{and} \quad ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)].$$

$ATT(d|d')$ is the average effect of dose d compared to zero dosage, on units that actually experienced dose d'. When $d' = d$, this is the ATT among units that received dose d. $ATE(d)$ is the mean difference between potential outcomes under dose d relative to untreated potential outcomes across all units, not just those that experienced dose d. We henceforth refer to these functions as Average treatment effects instead of Average level treatment effects to simplify the terminology.

Average causal response parameters for absolutely continuous treatments are defined as:

$$ACRT(d|d') = \left.\frac{\partial \mathbb{E}[Y_t(l)|D = d']}{\partial l}\right|_{l=d'} \quad \text{and} \quad ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d},$$

$ACRT(d|d)$ equals the derivative of the average potential outcome for units that received dose d evaluated at d. This is also equivalent to the derivative of $ATT(s|d)$ with respect to s, evaluated at $s = d$. For multi-valued discrete treatments average causal response are defined in the same way with slightly different notation:

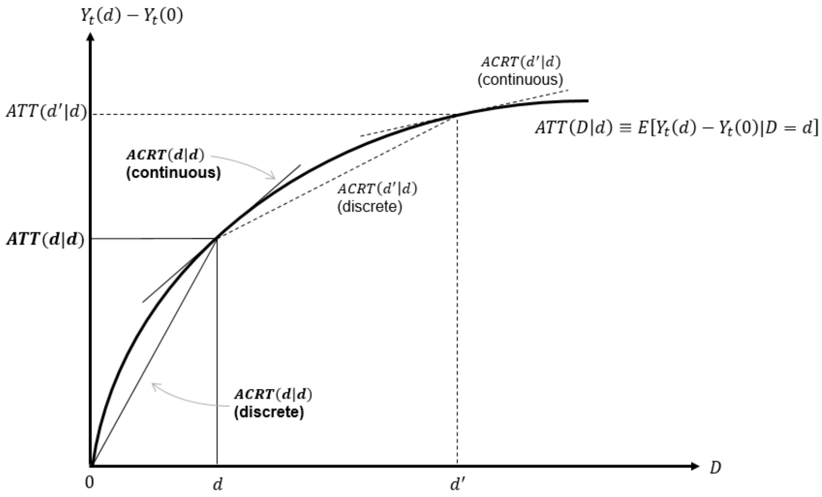
$$ACRT(d_j|d_k) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})|D = d_k] \quad \text{and} \quad ACR(d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})].$$

$ACRT(d_j|d_j)$ equals the difference in mean potential outcomes between dose level d_j and the next lowest dose d_{j-1} (no matter how big the gap between d_j and d_{j-1} is).[13] Note that $ACR(d) \equiv \mathbb{E}[ACRT(d|D)]$ averages causal responses at dose d across the entire population in the same way that $ATE(d)$ averages $ATT(d|d)$ terms. It is therefore not "local" to the units that experienced dose d.

Figure 2 illustrates these parameters graphically. The concave line plots an average treatment effect function against the dose for units actually treated with dose d, $ATT(D|d)$. If we consider dose levels d and d', there are two potential ATT parameters. $ATT(d|d)$, the level of group d's average treatment effect function at d, is an average treatment effect that is "local" to units that experienced dose d. $ATT(d'|d)$ is also "local" to the d group, 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 $ATT(D|d)$ function and the discrete-dose $ACRT$ parameters are the slopes of lines connecting two points on the $ATT(D|d)$ function. As with ATTs, our definitions encompass causal responses dosages other than the one a given group actually receives (i.e., $ACRT(d'|d)$).

A proper interpretation of continuous/multi-valued DiD results hinges on which type of parameter one wants to and can identify and 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 misinter-

[13] Differences in $ATT(d|d)$ between doses that are farther apart than, say, one unit in the discrete case or differ by a finite amount in the continuous case equal averages of the $ACRT$ between the doses in question.


$Y_t(d) - Y_t(0)$
$ACRT(d'|d)$ (continuous)
$ATT(d'|d)$
$ATT(D|d) \equiv E[Y_t(d) - Y_t(0)|D = d]$
$\boldsymbol{ACRT(d|d)}$ **(continuous)**
$ACRT(d'|d)$ (discrete)
$\boldsymbol{ATT(d|d)}$
$\boldsymbol{ACRT(d|d)}$ **(discrete)**
0
d
d'
D

Notes: The figure plots $ATT(D|d)$ (the average effect of experiencing each dose among units that actually experienced dose 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 a tangent line, and for a discrete multi-valued dose when $ACRT$ is a line connecting two discrete points on $ATT(D|d)$.

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

preting a large ATT estimate as an ACR, in this case, would mistakenly conclude that a policy to raise every unit's dose would have large effects. A researcher confusing a small ACR for an ATT would mistakenly conclude that an entire policy was ineffective when it actually just has small effects at the margin.

The above-mentioned causal parameters are functional parameters because they are allowed to vary arbitrarily across treatment dosage groups d', and/or across (counterfactual) dosages d. But researchers will also typically want to aggregate these functionals into an interpretable summary measure or to gain precision. (Regression estimators are one way to do this.) Likely the most natural way to combine many causal parameters across dose groups (or a function defined over doses) is to average using the dose distribution itself. We denote these aggregate parameters by:

$$ATT^* = \mathbb{E}[ATT(D|D)|D > 0] \qquad \text{and} \qquad ATE^* = \mathbb{E}[ATE(D)|D > 0]$$
$$ACRT^* = \mathbb{E}[ACRT(D|D)|D > 0] \qquad \text{and} \qquad ACR^* = \mathbb{E}[ACR(D)|D > 0].$$

Note that $ACRT^*$ and ACR^* are average derivatives, a type of parameter that econometricians have studied in different contexts for a while; see, e.g., Ai and Chen (2007) and Ichimura and Todd (2007) and references therein.

3.2 Identification with a Continuous Treatment in the Baseline Case

The definition of the causal parameters in the previous section narrows down the types of causal questions we attempted to answer in this paper. However, all these causal parameters involve counterfactual quantities, implying that they are not nonparametrically identified without additional assumptions and structure. In this section, we pursue this route and present identification results for the average treatment effects and average causal response-type parameters.[14]

We make the following assumptions:

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

Assumption 2 (Continuous and Multi-valued Treatment)**.** *In period* $t-1$*, no unit is treated, while in period* t*, the treatment dosage is either continuous or multi-valued. More precisely, one of the following is true:*

(a) The support of the treatment D *is given by* $\mathcal{D} = \{0\} \cup \mathcal{D}_+^c$*, where* $\mathcal{D}_+^c = [d_L, d_U]$ *with* $0 < d_L < d_U < \overline{d} < \infty$*, for some* $\overline{d} \in \mathbb{R}$*. In addition,* $\mathbb{P}(D = 0) > 0$*,* $a_f^{-1} < f_D(d) < a_f$ *for some positive constant* $a_f < \infty$ *and all* $d \in \mathcal{D}_+^c$*, and* $\mathbb{E}[\Delta Y_t | D = d]$ *is continuously differentiable on* $\mathcal{D}_+^c$*.*

(b) The support of the treatment D *is given by* $\mathcal{D} = \{0\} \cup \mathcal{D}_+^{mv}$ *where* $\mathcal{D}_+^{mv} = \{d_1, d_2, \ldots, d_J\}$ *where* $0 < d_1 < d_2 < \cdots < d_J < \overline{d} < \infty$*, for some* $\overline{d} \in \mathbb{R}$*. In addition,* $\mathbb{P}(D = d) > 0$ *for all* $d \in \mathcal{D}$*.*

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

$$Y_{it-1} = Y_{it-1}(d) = Y_{it-1}(0) \quad \textit{and} \quad Y_{it} = Y_{it}(D_i).$$

Assumption 1 says that we observe two periods of *iid* panel data. Assumption 2 formalizes that the treatment consists of a mass of units that do not participate in the treatment in both periods, and an otherwise continuous (part a) or multi-valued (part b) treatment. Assumption 2.a allows for the smallest value of the treatment to be strictly larger than zero, which is common in applications. Assumption 3 says that we observe untreated potential outcomes for all units in the first period, as no unit act on future treatment knowledge before treatment starts. 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 treatment effects follows closely from the binary treatment case. In particular, our results rely on an extension of the binary parallel trends assumption:

[14] In this paper, we use identification as synonymous for point identification. That is, we abstract from partial identification results.

Assumption 4 (Parallel Trends)**.** *For all $d \in \mathcal{D}$,*

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

Just like in binary DiD designs, Assumption 4 says that the average path of outcomes that units with any dose d would have experienced without treatment is the same as the path of outcomes that units in the untreated group actually experienced. The following result shows that under Assumption 4, $ATT(d|d)$ is identified; all proofs are in Appendix F. Henceforth, let $\Delta Y_t = Y_t - Y_{t-1}$.

Theorem 3.1. *Under Assumptions 1 to 4, $ATT(d|d)$ is identified for all $d \in \mathcal{D}$, and it is given by*

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

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

The identification results for $ATT(d|d)$ in Theorem 3.1 holds by essentially the same arguments used for binary treatments. Because Assumption 4 ensures that $\mathbb{E}[\Delta Y_t|D = 0]$ is the same as the path of outcomes that treated units would have experienced absent the treatment, $ATT(d|d)$ equals the difference between the change in outcomes for the dose d group and the untreated group. As a direct consequence, by averaging all the $ATT(d|d)$'s over the distribution of non-zero dosages, we have that the ATT^* is identified by simply comparing units with positive treatment dosage with those with zero treatment dosage. That is, even with continuous or multi-valued treatments, one can identify a summary measure of the causal effects among treated units by relying on a binary comparison.

On the other hand, just like in the binary case, Parallel Trends Assumption 4 is *not* strong enough to guarantee the identification of $ATE(d)$.

Proposition 3.1. *Under Assumptions 1 to 4, $ATE(d)$ is are not identified.*

Intuitively, the result holds because $ATE(d)$ is defined as the average of $ATT(d|k)$ across all values of $k > 0$, but Assumption 4 does not allow identification of ATT parameters at doses other than the one each group actually receives. As a consequence, neither $ATE(d)$ nor ATE^* are identified.

We now turn to the identification of average causal response parameters, which differs from identification of ATT parameters because it requires comparisons between dose groups. Estimating $ATT(d|d)$ can be done in two equivalent ways. One approach is to compare estimated $ATT(d|d)$s across values of d; e.g. $ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1})$. This measures how much average treatment effects vary with the dose–a "dose-response" relationship. Alternatively, one could use lower-dose units as a comparison group for higher-dose units: $\mathbb{E}[\Delta Y_t|D = d_j] - \mathbb{E}[\Delta Y_t|D = d_{j-1}]$. Designs without untreated units motivate a continuous/multi-valued DiD strategy this way. Since the change in outcomes for untreated units, $\mathbb{E}[\Delta Y_t|D = 0]$, cancels in the comparison of ATTs, the two approaches are equivalent. We, therefore, refer to both ACR estimators in Theorem 3.2 and throughout the paper.

Our central identification result is that ACR parameters are not identified under Parallel Trends Assumption 4 alone because comparisons between different dose groups are biased when treatment effects vary across groups even when the path of untreated potential outcomes is the same.

Theorem 3.2. *Under Assumptions 1 to 4 ACRT parameters are not identified. Furthermore,*

(a) Under Assumption 2(a), for $d \in \mathcal{D} \setminus \{0\}$,

$$\frac{\partial \mathbb{E}[\Delta Y_t | D = d]}{\partial d} = \frac{\partial ATT(d|d)}{\partial d} = ACRT(d|d) + \underbrace{\frac{\partial ATT(d|l)}{\partial l}\Big|_{l=d}}_{\text{“selection bias”}}.$$

(b) Under Assumption 2(b), for $d_j \in \mathcal{D} \setminus \{0\}$,

$$\begin{aligned}\mathbb{E}[\Delta Y_t | D = d_j] - \mathbb{E}[\Delta Y_t | D = d_{j-1}] &= ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) \\ &= ACRT(d_j|d_j) + \underbrace{ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})}_{\text{“selection bias”}}.\end{aligned}$$

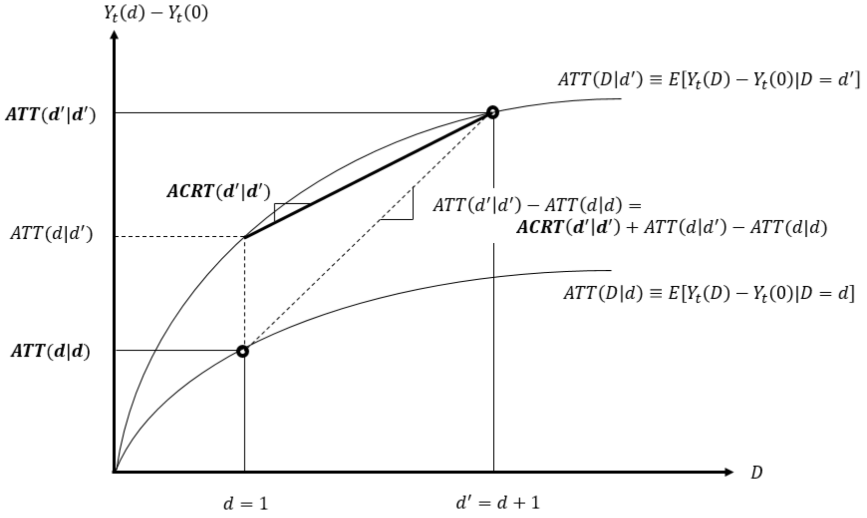
Theorem 3.2 says that under parallel trends, comparisons of outcome paths between higher- and lower-dose groups mix together (i) $ACRT(d|d)$ and (ii) a “selection bias” type of term that comes from differences in average treatment effects across groups. Intuitively, even if untreated outcomes evolve in the same way, observed outcomes may change differently between units with dose d_j and units with dose d_{j-1} both because $d_j > d_{j-1}$, a causal response, and cross-group differences in the effect of the first d_{j-1} dose units, “selection bias”.

Figure 3 illustrates this result for an example with two groups and two doses, $d' = d + 1$. 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.[15] This is captured by the “selection bias” term, which equals the difference in treatment effects at the lower dose: $ATT(d|d') - ATT(d|d)$. “Selection bias” is a version of selection-on-gains. Here, 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 a lower dose. The “selection bias” is not identified because we do not observe $Y_t(d)$ for units that experienced dose d'. Such a result precludes a causal interpretation of ATT differences across doses, at least when one is not willing to further strengthen the Parallel Trends Assumption 4.

3.2.2 Identification under strong parallel trends

The fact that causal responses are not identified under a “traditional” parallel trends assumption suggests that learning about the new kind of parameter that continuous DiD designs introduce requires new assumptions as well. This section introduces a stronger assumption that allows the identification of ACR (and ATE) parameters:

[15] Because we are considering one unit increments, the “bias” can be seen on the y-axis as well. $ATT(d|d') - ATT(d|d)$ is “bias” and $ATT(d'|d') - ATT(d|d')$ is the $ACRT(d'|d')$.


$Y_t(d) - Y_t(0)$
$ATT(D|d') \equiv E[Y_t(D) - Y_t(0)|D = d']$
$\boldsymbol{ATT(d'|d')}$
$\boldsymbol{ACRT(d'|d')}$
$ATT(d'|d') - ATT(d|d) =$
$\boldsymbol{ACRT(d'|d')} + ATT(d|d') - ATT(d|d)$
$ATT(d|d')$
$ATT(D|d) \equiv E[Y_t(D) - Y_t(0)|D = d]$
$\boldsymbol{ATT(d|d)}$
D
$d = 1$
$d' = d + 1$

Notes: The figure shows that comparing adjacent $ATT(d|d)$ estimates 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).

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

Assumption 5 (Strong Parallel Trends)**.** *For all $d \in \mathcal{D}$,*

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

Assumption 5 says that for every dose group, the average change in potential outcomes over time is the same as the population average change in potential outcomes for the same dose change. Alternatively, under no-anticipation as in Assumption 3, one can express Assumption 5 as $\mathbb{E}[Y_t(d) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d]$, suggesting a type of parallel trends that involves changes in potential outcomes from zero to dosage d. These two interpretations highlight that Assumption 5 notably differs from Assumption 4 because it involves potential outcomes under different doses $Y_t(d)$ rather than only untreated potential outcomes, $Y_t(0)$. The Strong Parallel Trends (SPT) Assumption 5 is useful because the left-hand side is not identified, but will turn out to be important in identifying ACR and ATE parameters, while the right-hand side is the observed change in mean outcomes for dose group d.

In practice, Assumption 5 is most easily understood in terms of a *slightly stronger* assumption that imposes that *all* dose groups would have experienced the same path of potential outcomes had they been assigned the same dose. This is a treatment effect homogeneity condition because it implies that $ATT(d|d) = ATT(d|d') = ATE(d)$; see Proposition B.1 in the Appendix. It rules out selection-on-gains into a particular dose level and ensures the observed outcome changes for every dose group reflect what would have happened to all other units had they received that dose. However, we stress that mathematically speaking, Assumption 5 does not *require* ruling out these behaviors, and does not even guarantee that the $ATT(d|d)$'s are identified, as we discuss in Appendix B.[16]

The following theorem shows that the SPT Assumption 5 allows identification of average treatment effect and average causal response parameters at each dose:

Theorem 3.3. *Assume that Assumptions 1 to 3 and 5 hold.*

(a) For $d \in \mathcal{D} \setminus \{0\}$, it follows that

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

(b) When Assumption 2(a) holds (i.e., treatment is continuous), it follows that, for $d \in \mathcal{D} \setminus \{0\}$,

$$ACR(d) = \frac{\partial \mathbb{E}[\Delta Y_t | D = d]}{\partial d} = \frac{\partial ATE(d)}{\partial d},$$

(c) When Assumption 2(b) holds (i.e., treatment is multi-valued), it follows that, for $d_j \in \mathcal{D} \setminus \{0\}$,

$$ACR(d_j) = \mathbb{E}[\Delta Y_t | D = d_j] - \mathbb{E}[\Delta Y_t | D = d_{j-1}] = ATE(d_j) - ATE(d_{j-1}),$$

Theorem 3.3 follows directly from our earlier definition of parameters and of the way that effect heterogeneity biases outcome comparisons across dose groups. For part (a), note that parallel trends identifies $ATT(d|d)$, but strong parallel trends identifies $ATE(d)$. The two parameters differ when there is selection into dose group d on the basis of treatment effects. SPT rules this out and means that comparing average outcome changes of dose group d to the untreated units identifies the $ATE(d)$. For parts (b) and (c), the same implication of SPT ensures that lower dose groups are a valid counterfactual for higher-dose groups. The restriction on potential outcomes that delivers this property, however, also means that ACR estimands apply to all units, not just those treated with dose d.

Strong parallel trends only change the interpretation of the estimand, not its form. Theorem 3.3 makes an explicit connection between assumptions and the types of parameters that different comparisons in a continuous DiD design can identify. One important implication is that conventional pre-tests for differential changes across groups before treatment cannot distinguish between Assumption 4 and Assumption 5. Only untreated potential outcomes are observed before treatment, so these periods cannot test the content of an assumption like SPT that necessarily involves treated potential outcomes.[17]

[16] It turns out that Assumption 5 is not even strictly stronger than Assumption 4 in the way it restricts trends in *untreated* outcomes; see Appendix B for a discussion.

[17] An interesting intermediate assumption between Assumption 4 and Assumption 5 would be to directly assume that the "selection bias" term in Theorem 3.2 (i.e., $\partial ATT(d|l)/\partial l|_{l=d}$) is equal to 0. This would imply that $ACRT(d|d)$ is identified. Another interesting intermediate assumption is that for $d \in \mathcal{D}_s$ where $\mathcal{D}_s \subset \mathcal{D}_+$, $\mathbb{E}[Y_t(d) - Y_{t-1}(0)|D \in \mathcal{D}_s] = \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d]$. This would imply that one could identify parameters such as $\mathbb{E}[Y_t(d) - Y_t(0)|D \in \mathcal{D}_s]$ for $d \in \mathcal{D}_s$ (as well as its derivative). These types of assumptions might be 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 selected similar doses, but the researcher is unwilling to assume that there is no "selection bias" among units that select substantially different doses.

Finally, the identification results in Theorem 3.3 immediately imply that averages of the $ATE(d)$ and $ACR(d)$ building blocks are identified as well. The following lemma states this for the averages that weight by $f_{D>0}(d)$, the density of dosage d, conditional on the dosage being positive. (When the dosage is discrete, we write its probability distribution function among units with positive dosage as $P(D = d_j|D > 0)$.)

Corollary 3.1. *Assume that Assumptions 1 to 3 and 5 hold.*

(a) For $d \in \mathcal{D} \setminus \{0\}$, it follows that

$$ATE^* = \mathbb{E}[\Delta Y_t|D > d] - \mathbb{E}[\Delta Y_t|D = 0]$$

(b) When Assumption 2(a) holds (i.e., treatment is continuous), it follows that, for $d \in \mathcal{D} \setminus \{0\}$,

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

(c) When Assumption 2(b) holds (i.e., treatment is multi-valued), it follows that, for $d_j \in \mathcal{D} \setminus \{0\}$,

$$ACR^* = \sum_{j=1}^{J} \left(\mathbb{E}[\Delta Y_t|D = d_j] - \mathbb{E}[\Delta Y_t|D = d_{j-1}] \right) P(D = d_j|D > 0)$$

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

3.3 What Parameter Does TWFE Estimate in the Baseline Case?

In practice, when empirical researchers approach a continuous DiD design they begin by estimating a single summary parameter using a TWFE regression like Equation (1.1). This section links the TWFE estimator to the identification results for dose-specific average treatment effect or average causal response parameters, describes the assumptions necessary to give TWFE *some* causal interpretation, and discusses what that interpretation is.

In our baseline case, analyzing the TWFE coefficient β^{twfe} from Equation (1.1) is straightforward because it is equivalent to the univariate slope coefficient from a regression of $\Delta Y_{it} = Y_{it} - Y_{it-1}$ on an intercept and D_i. From that point, we present two sets of alternative decompositions of β^{twfe} in terms of weighted sums of different "causal building block" parameters, one for each of the two versions of parallel trends considered.

We start our discussion based on the "usual" Parallel Trends (PT)Assumption 4. One of our decompositions breaks down β^{twfe} into a weighted sum of $ATT(d|d)$'s. Another decomposition we present uses the "per-dosage" scaled $ATT(d|d)$ parameters, $ATT(d|d)/d$, as the building block. A third decomposition uses results from Yitzhaki (1996) to express β^{twfe} as a weighted sum of $ACRT$ type terms.[18] These results stress that the same TWFE coefficient β^{twfe} can have different

[18] Appendix C also decomposes β^{twfe} into a weighted average of DiD comparisons between every pair of dose

interpretations, which crucially depends on the choice of the "building blocks".[19] We repeat the same type of exercise but impose the SPT Assumption 5, focusing on ATE, scaled-ATE, and ACR as the "building blocks" of the decompositions.

The following theorem describes the TWFE estimand with a continuous treatment under Parallel Trends Assumption 4; results for the multi-valued results are in Appendix C.

Theorem 3.4. *Under Assumptions 1, 2(a), 3, and 4, we can decompose the TWFE regression coefficient β^{twfe} in (1.1) in different ways, depending on the choice of the causal estimand that serves as the summand for these decompositions. More explicitly,*

(a) We can decompose β^{twfe} in terms of ATT's as

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

where $w^{lev}(l) = \frac{(l-\mathbb{E}[D])}{\text{Var}(D)} f_D(l)$, $w^{lev}(l) \lesseqgtr 0$ for $l \lesseqgtr \mathbb{E}[D]$, and $\int_{\mathcal{D}} w^{lev}(l)\, dl = 0$.

(b) We can decompose β^{twfe} in terms of scaled-ATT's as

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

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

(c) We can decompose β^{twfe} in terms of ACRT's as

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{acr}(l) \left[ACRT(l|l) + \frac{\partial ATT(l|h)}{\partial h}\bigg|_{h=l} \right] dl + w_0^{acr} \frac{ATT(d_L|d_L)}{d_L}$$

where $w_1^{acr}(l) = \frac{(\mathbb{E}[D|D\geq l]-\mathbb{E}[D])\mathbb{P}(D\geq l)}{\text{Var}(D)}$ and $w_0^{acr} = \frac{(\mathbb{E}[D|D>0]-\mathbb{E}[D])\mathbb{P}(D>0)d_L}{\text{Var}(D)}$. In addition, (i) $w_1^{acr}(l) \geq 0\ \ \forall l \in \mathcal{D}$, $w_0^{acr} > 0$, and (ii) $\int_{d_L}^{d_U} w_1^{acr}(l)\, \mathrm{d}l + w_0^{acr} = 1$

The most important conclusion from Theorem 3.4 is that, with a continuous treatment dose, it is hard to attach an easy-to-understand causal interpretation to the TWFE regression coefficient under PT Assumption 4. The basic reason is that TWFE is "variation hungry"; it exploits all the variation in D, which necessarily means that it compares treated groups to each other. Those comparisons do not identify treatment effect parameters for which the relevant comparison treatment status is *no* treatment, and they do not identify causal response parameters without further restrictions. Parts (a) to (c) from Theorem 3.4 show that these issues arise for interpretations based on treatment effects, scaled treatment effects, and causal responses, albeit in slightly different ways.

Part (a) expresses β^{twfe} as a weighted sum of $ATT(d|d)$ parameters and highlights that the $ATT(d|d)$ for units with treatment dosage below the mean will get negative weights. In other words,

groups, which extends the decomposition from Goodman-Bacon (2021). We do not discuss that result here because the theoretical causal interpretation is less clear than the decompositions based directly on ATT'sand $ACRT$s.

[19] The importance of the choice of building blocks have also been highlighted by Słoczyński (2022) in cross-sectional contexts with binary treatment and unconfoundedness designs; see also Heckman, Urzua, and Vytlacil (2006) for a related discussion based on identification via instrumental variables.

TWFE effectively uses units with doses above $\mathbb{E}[D]$ as the "treatment group" (positive weights) and those with doses below $\mathbb{E}[D]$ as the "comparison group" (negative weights). However, TWFE regressions also weigh and scale these groups differently, making it hard to interpret them. Indeed, from Theorem 3.4(a), some simple algebra, and exploring that $ATT(0|0) = 0$ and $\int_{\mathcal{D}} w^{\text{lev}}(l)\, dl = 0$, it follows that we can re-write β^{twfe} as

$$\begin{aligned}
\beta^{twfe} &= \mathbb{E}\left[w^v(D)\ m_\Delta(D)\middle|\ D \geq \mathbb{E}[D]\right] P(D \geq \mathbb{E}[D]) \\
&\qquad - \mathbb{E}\left[w^v(D)\ m_\Delta(D)\middle|\ D < \mathbb{E}[D]\right] P(D < \mathbb{E}[D]) \qquad (3.1)\\
&= \mathbb{E}\left[w^v(D)\ ATT(D|D)\middle|\ D \geq \mathbb{E}[D]\right] P(D \geq \mathbb{E}[D]) \\
&\qquad - \mathbb{E}\left[w^v(D)\ ATT(D|D)\middle|\ 0 < D < \mathbb{E}[D]\right] P(0 < D < \mathbb{E}[D]), \qquad (3.2)
\end{aligned}$$

with $w^v(D) = \left|\frac{(D-\mathbb{E}[D])}{Var(D)}\right|$, and $m_\Delta(D) = \mathbb{E}[\Delta Y_t|D]$. From (3.1), we see that β^{twfe} compares weights averages of $m_\Delta(s)$'s above and below $\mathbb{E}[D]$, and the weights are proportional to "how far" from $\mathbb{E}[D]$ each group is. These weights are particularly hard to justify when D is not symmetric around its mean, i.e., when $P(D \geq \mathbb{E}[D]) \neq P(D < \mathbb{E}[D])$.

Even when treatment D is symmetric around its mean, (3.2) point out that expressing such comparisons across dosages in terms of $ATT(d|d)$'s can lead to "attenuation" problems or even sign-reversal due to "negative weights".[20] Although negative weighting also appears in the TWFE estimand with a binary staggered treatment (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020), here we show that this drawback can arise even in the simplest two-periods DiD setup with continuous treatment. Just like in the binary staggered case, a sufficiently large untreated group ensures that negative weights are not a first-order concern. Through the lens of Theorem 3.4(a), "large enough" means that the quantity of untreated observations pulls $\mathbb{E}[D]$ below the minimum treated dose, d_L. In that case, the second term in (3.2) drops and $\beta^{twfe} = \mathbb{E}\left[w^v(D)\ ATT(D|D)\right| D \geq \mathbb{E}[D]]\, P(D \geq \mathbb{E}[D])$.

Yet, even when all weights are positive, the way TWFE aggregates $ATT(d|d)$'s is not very intuitive or policy-relevant. A perhaps more natural way to aggregate the $ATT(d|d)$ parameters is to take a simple expectation, i.e., weight them by $f_{D|D>0}(d)$ to get $ATT^* \equiv \mathbb{E}[ATT(D|D)|D > 0]$. Indeed, as indicated by Theorem 3.1, ATT^* is identified, easy to estimate, and does not suffer from these non-standard implicit weighting schemes. In setups where treatment effects may vary with the treatment dose, we recommend favoring ATT^* vis-a-vis β^{twfe}, as the weighting scheme of the latter may be especially misleading.

Part (b) of Theorem 3.4 extends part (a) to apply to scaled ATT parameters, $ATT(d|d)/d$,[21] which gives a "per-treatment-dosage" interpretation. Part (b) shows that when the scaled ATT

[20] Chodorow-Reich, Nenov, and Simsek (2021, pg. 1636) 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."

[21] This is the parameter one would obtain by estimating equation (1) on a sample of units with $D = 0$ or $D = d$.

is the building block of interest, TWFE estimand has negative weights under the same conditions as in part (a). The weights themselves equal $w^{lev}(d)$ weights times the dose, which creates two key differences. First, they integrate to 1 in the treated sample. Second, they weigh the building block parameters for the highest and lowest doses even more heavily than in part (a). In the case of a discrete dose, this result is similar to the one in Theorem S3 of the Supplementary Appendix of de Chaisemartin and D'Haultfœuille (2020). Therefore, using average slopes as the underlying parameter of interest does not eliminate the potential negative weighting issue with the TWFE estimator, and it is still hard to add a convincing justification for such weighting schemes when treatment effects are heterogeneous.

Theorem 3.4(c) shows that when we attempt to use the $ACRT$'s as building blocks of the analysis, the TWFE estimand can be written as combinations of two positively weighted averages: one of $ACRT(d|d)$ parameters, and another of "selection bias" terms due to heterogeneous ATT'sas derived in Theorem 3.2. The sign of this bias depends on how treatment effects vary across groups at a given dose. If units with higher doses always have larger positive treatment effects, for example, then TWFE will be larger than the average of the $ACRT$'s that appear in Theorem 3.4(c). Figure 3 illustrate this case for two groups. The magnitude of the bias depends on how strong the heterogeneity is and its relationship to the weights on each $\frac{\partial ATT(l|h)}{\partial h}\Big|_{h=l}$ term. Heterogeneity comes from the "technology" that generates treatment effects–do $ACRT$'s vary and by how much?–and the allocation mechanism for the dose–how is the ATT function related to the observed dose? This has important econometric implications, but does not come *from* TWFE itself. The weights, however, do inherit their form from ordinary least squares (OLS). Differentiating $w_1^{acr}(d)$ shows that the weights are hump-shaped and centered around $\mathbb{E}[D]$, so selection bias around the average dose affects β^{twfe} the most. Therefore, even when one takes $ACRT$'s as the building block for the parameter of interest, as many applied papers implicitly do, TWFE still does not yield a causal estimand under parallel trends. Part (c) also shows how TWFE handles a discrete jump from 0 to the minimum treated dose, d_L. Causal responses below d_L cannot be estimated, so the scaled treatment effect of d_L, $ATT(d_L|d_L)/d_L$ as in part (b) is averaged into β^{twfe}.

It is worth stressing that our PT Assumption 4 is not strong enough to guarantee the identification of the $ACRT$'s, as already indicated in theorem 3.2. This implies that coming up with a reliable and easy-to-use summary measure of the $ACRT$'s is not as simple as it was for the ATT^*. Putting it simply, we cannot "estimate our way out" when we are interested in causal response type parameters. Following our discussion around Theorem 3.3, though, one may wonder if imposing the SPT Assumption 5 may be enough to give the TWFE coefficient β^{twfe} a causal interpretation. The following theorem describes the TWFE estimands under Assumption 5. It is the analog of Theorem 3.4 but replacing the PT Assumption 4 with the SPT Assumption 5.

Theorem 3.5. *Under Assumptions 1, 2(a), 3, and 5, we can decompose the TWFE regression coefficient β^{twfe} in (1.1) in different ways, depending on the choice of the causal estimand that serves as the summand for these decompositions. More explicitly,* *(a) We can decompose β^{twfe} in terms of ATE's as*

$$\beta^{twfe} = \int_{d_L}^{d_U} w^{lev}(l) ATE(l)\, dl.$$

(b) We can decompose β^{twfe} in terms of scaled-ATE's as

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

(c) We can decompose β^{twfe} in terms of ACR's as

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{acr}(l) ACR(l)\, dl + w_0^{acr} \frac{ATE(d_L)}{d_L}.$$

All weights are defined in Theorem 3.4.

Theorem 3.5 shows that the SPT Assumption 5 eliminates selection bias in the ACR interpretation of β^{twfe} but still does not deliver an interpretation in terms of treatment effects. Assumption 5 restricts treated potential outcomes, but does not change the TWFE estimator. Since negative weights in Theorem 3.4 were a consequence of TWFE, additional assumptions about treatment effects do not change the underlying weighting scheme. Parts (a) and (b) of Theorem 3.5 are thus the same as in Theorem 3.4 except that they aggregate $ATE(d)$'s instead of $ATT(d|d)$'s.

Theorem 3.5(c) is essentially an aggregate version of Theorem 3.3, which shows that Assumption 5 eliminates selection bias in the estimation of each ACR. The particular interpretation of β^{twfe} in terms of ACR's, therefore, hinges on that aggregation, embodied in the weights $w_1^{acr}(d)$. The $w_1^{acr}(l)$ are positive and integrate to 1, so under Assumption 5 β^{twfe} does have a causal interpretation. But β^{twfe} does not estimate a natural target parameter like the $ACR^* \equiv \mathbb{E}[ACR(D)|D > 0]$, because the TWFE weights do not generally equal the dose distribution among treated, $f_{D|D>0}(d)$. As discussed above, the weights are hump-shaped and centered on $\mathbb{E}[D]$, no matter the underlying shape of $f_{D|D>0}(d)$. If D is distributed $U(0,1)$, for example, then relative to ACR^*, β^{twfe} puts more weight on $ACR(d)$ parameters close to the mean and less weight on $ACR(d)$'s closer to 0 or 1.[22] For declining distributions like the exponential, TWFE puts less weight on the most common doses below the mean, and more weight on the rarer doses above the mean.[23] For a bimodal distribution of D with little mass around $\mathbb{E}[D]$, $w_1^{acr}(d)$ will actually put the most weight on ACR's of the least common doses. In general, when $f_{D|D>0}(d)$ is approximately Gaussian, the TWFE estimand is closer to ACR^*. But when the dose distribution is skewed, TWFE weights $ACR(d)$ parameters close to the mean dose more than their population density weights. If the treatment effect function is non-linear or non-monotonic, so that $ACR(d)$ parameters vary widely across doses, then the TWFE estimand may differ meaningfully from ACR^*.

[22] For a uniformly distributed dose we have $w_1^{acr}(d) = 6d(1-d)$. Therefore, the difference in weight on $ACR(d)$ across the two weighting schemes is $f_{D|D>0}(d) - w_1^{acr}(d) = 1 - 6d(1-d)$. This function is concave up and has roots at $1/2 \pm \sqrt{3}/6$, so TWFE puts more over-weights parameters between 0.21 and 0.79.

[23] For $f_D(d) = \lambda e^{-\lambda d}, \quad d \geq 0$, we have $w_1^{acr}(d) = \lambda \ell f_(d)$. The difference between the distribution of D weights and the TWFE weights is $f_{D|D>0}(d) - w_1^{acr}(d) = f_D(d)\lambda(\frac{1}{\lambda} - d)$. This shows that TWFE under-weights ACR's at doses below the mean ($d < \frac{1}{\lambda}$) and over-weights them at doses above the mean ($d > \frac{1}{\lambda}$).

We again stress that providing ex-ante justification for the weights $w_1^{acr}(d)$ is hard, which, in turn, suggests that β^{twfe} may have limited interpretability in terms of ACR's. Instead of letting the estimation method implicitly select how one aggregates these ACR's into a summary measure of the treatment, we recommend that researchers choose these aggregation schemes explicitly. In our view, a natural and econometrically-guided way to aggregate the ACR's into a summary parameter is given by ACR^*, which is identified (as indicated in Corollary 3.1) and can be efficiently estimated, as well.

Remark 1 (No untreated units)**.** *In some applications, all units end up treated; i.e., there are no "never-treated" units. In fact, a lack of untreated units is a frequent justification for using a continuous DiD design. Although $ATT(d|d)$ parameters are not identified in this case, it is still possible to use the results above to characterize β^{twfe} in terms of them. When $\mathbb{P}(D=0)=0$, TWFE necessarily puts negative weights on ATT or scaled ATT parameters of lower-dose groups. The ACR decompositions, on the other hand, are unchanged except that the scaled ATT for dose d_L is not identified (and its weight is zero). In other words, Theorem 3.4 and Theorem 3.5 can still apply in this case, by noting that $\mathbb{P}(D=0)=0$.*

Remark 2 (Restrictions on the shape of $ATT(d|d)$ can aid in identification.)**.** *Suppose that $ATT(d|d)=0$ for $d<d_0$. Then, for those units, $Y(D)=Y(0)$ and, under parallel trends, they can be used to identify ATT's.*[24] *Alternatively, if ATT's were constant and homogeneous (i.e., dosage intensity does not matter; $ATT(d|d)=a$ for all units), then TWFE estimates $\beta^{twfe}=a$ under parallel trends, regardless of the negative weights. Negative weights also do not undermine the causal interpretation based on per-dose effects if ATT is linear and homogeneous in the dose ($ATT(d|d)=b*d$), in which case $\beta^{twfe}=b$. However, we note that these are strong, parametric functional form restrictions.*

Remark 3 (Pre-treatment differences)**.** *TWFE does not estimate pre-treatment differences between treated and untreated units when there are negative weights. To see this, consider a case with three periods, $t-2$, $t-1$, and t, but otherwise the same as the baseline case. Applying equation (3.1) to the change between periods $t-2$ and $t-1$ gives the TWFE estimate of the pre-trend:*

$$\beta_{pre}^{twfe}=\mathbb{E}\left[w^v(D)m_\Delta^{pre}(D)\middle|\, D\geq\mathbb{E}[D]\right]P(D\geq\mathbb{E}[D])-\mathbb{E}\left[w^v(D)m_\Delta^{pre}\middle|\, D<\mathbb{E}[D]\right]P(D<\mathbb{E}[D]),$$

where $m_\Delta^{pre}(D)=\mathbb{E}[Y_{t-2}(0)-Y_{t-1}(0)|D]$. If Assumption 4 holds then $\beta_{pre}^{twfe}=0$, but because the change in outcomes for some dose groups can receive negative weight, this quantity can also equal zero if Assumption 4 is violated. Thus, one should not rely on TWFE to assess the plausibility of PT; see also Sun and Abraham (2021) for related results in staggered designs with a binary treatment.

[24] This kind of assumption has precedent elsewhere. In toxicology, where concerns about dosage drive clinical and regulatory decisions, the threshold below which D has no (detectable) effect is called the No Observable Adverse Effects Level (NOAEL).

4 DiD estimators that can highlight heterogeneity

The results in Theorems 3.4 and 3.5 show that it can be hard to attach a clear causal interpretation to the TWFE regression coefficient β^{twfe} in setups with treatment effect heterogeneity, even when all the weights are positive. Furthermore, in many practical situations, we may be interested in *documenting* the impact of different treatment dosages on a given outcome of interest, and β^{twfe} is clearly not suitable for that. In this section, we discuss alternative data-driven estimation procedures that do not suffer from these drawbacks.

Our first suggested estimation procedure focuses in summarizing treatment effects as weighted averages of $ATT(d|d)$ or $ATE(d)$. In light of Theorems 3.1 and 3.2(a), we can estimate ATT^* or ATE^* by considering the following simple linear regression specification:

$$\Delta Y_i = \beta_0^{bin} + 1\{D_i > 0\}\beta^{bin} + \epsilon_i, \tag{4.1}$$

where $\Delta Y_i = \Delta_{it}$ (as we only have two time periods), $D_i^{>0} = 1\{D_i > 0\}$ is a dummy variable for the treatment dosage being greater than zero, β_0^{bin} and β^{bin} are (unknown) finite-dimensional parameters, and ϵ_i and error term. It is straightforward to show that under Assumptions 1 to 4, $\beta^{bin} = ATT^*$. Thus, one can estimate and make (asymptotically valid) inferences about ATT^* using (4.1), as long as some weak and standard regularity conditions are satisfied.[25] If one imposes the SPT Assumption 5 instead of the PT Assumption 4, then it follows that $\beta^{bin} = ATE^*$. Such an estimation strategy is the simplest possible: it only requires researchers to "binarize" their treatment and rely on familiar regression models. Contrary to β^{twfe}, though, β^{bin} captures a well-defined and easily interpretable causal parameter of interest without relying on additional assumptions.

Although (4.1) is simple and intuitive, it is not flexible enough to highlight treatment effect heterogeneity across dosages, nor suitable to estimate ACR-type parameters. For that, we need to go beyond (4.1). Before we discuss our proposed method that directly builds on Chen, Christensen, and Kankanala (2022), it is worth discussing its general intuition. Somehow, the TWFE specification in (1.1) and the simple "binarize" specification (4.1) are "too restrictive" to accommodate richer forms of heterogeneity. Hence, a natural step forward is to consider more flexible specifications. There are different ways one can achieve that. For instance, one may rely on flexible regression specifications of the type

$$\Delta Y_i = \sum_{k=1}^{K} \psi_{Kk}(D)\beta_{Kk} + \varepsilon_i$$

where $\psi^K(d) = (\psi_{K1}(d), \psi_{K2}(d), \ldots, \psi_{KK}(d))'$ is a K-dimensional vector of flexible (known) transformations of the treatment dosage D, $\beta_K = (\beta_{K1}, \beta_{K2}, \ldots, \beta_{KK})'$ is a vector of finite dimensional (unknown) parameters, and ε_i is an idiosyncratic error term. One could then estimate these unknown β coefficients using OLS, and use the (functional) delta method to form estimators for the

[25] This includes bounded second moments, and $P(D = 0)$ and $P(D > 0)$ being uniformly bounded away from zero. If one wishes to cluster the standard errors at a higher level than i, there should also be sufficiently many treated $(D > 0)$ and untreated $(D = 0)$ clusters to justify the application of a Central Limit Theorem; see Roth, Sant'Anna, Bilinski, and Poe (2023) for a discussion.

different target parameters, $ATT(d|d)$'s, $ATE(d)$'s, $ACR(d)$'s, ATT^*, or ACR^*.

When treatment is multi-valued, as considered in Assumption 2(b), this task is simple. In such cases, $\psi^K(D)$ can be a saturated set of treatment dosage indicators with the no-treatment as the baseline, i.e., with some abuse of terminology,

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

One can then use OLS estimate the β's; denote such estimators by $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_J)'$.[26] Under SPT Assumption 5 (and Assumptions 1 and 3), it is very easy to show that each $\widehat{\beta}_j$ is a consistent (nonparametric) estimator for the $ATE(d_j)$, and $\widehat{\beta}_j - \widehat{\beta}_{j-1}$ is a consistent (nonparametric) estimator for $ACR(d_j)$. It is straightforward to aggregate these $ACR(d)$'s to form a plug-in estimator for the ACR^* using the identification formula in Corollary 3.1(c),[27] i.e.,

$$\widehat{ACR}^* = \sum_{j=1}^{J} \left(\widehat{\beta}_j - \widehat{\beta}_{j-1}\right) \widehat{P}(D = d_j | D > 0), \tag{4.2}$$

where $\widehat{P}(D = d_j|D > 0) = \sum_{i=1}^{n} 1\{D_i = d_j\} / \sum_{i=1}^{n} 1\{D_i > 0\}$. It follows from the delta method, our identification assumptions, and some weak regularity conditions that, as sample size increases, $\sqrt{n}\left(\widehat{ACR}^* - ACR^*\right)$ converges to a normal distribution with mean zero and estimable asymptotic variance, implying that standard inference procedures can be reliably used when treatments are multi-valued. One can follow a similar strategy when using the scaled-$ATE(d)$ as the "building blocks" of the aggregation.

Now, when the treatment dosage is continuous, as considered in Assumption 2(b), things are more complicated, especially when one does not want to impose parametric restrictions on the treatment effect heterogeneity. One needs to be more explicit about the type of "flexible" transformations $\psi^K(d)$: it involves picking the class of transformations (basis functions) and the number of terms K used to implement the method. Poor tuning parameter choices can lead to estimators that converge "too slowly", and confidence bands with inappropriate statistical guarantees. On the other hand, "good" choices of tuning parameters usually require additional knowledge of model structure, such as the smoothness of $ATE(d)$, which, in practice, it is ex-ante unknown. It is thus desirable to have a data-driven estimation method that adapts to these unknown model regularities, yield estimators and confidence bands with solid statistical guarantees and, at the same time, is easy to implement. The important question is: *How to do it?* Fortunately, we can build on Chen, Christensen, and Kankanala (2022), who propose (a) data-adaptive nonparametric estimators for generic structural functions (conditional expectations) and their derivatives that converge at the fastest possible (i.e., minimax) rate in sup-norm, and (b) data-driven uniform confidence bands that have correct asymptotic coverage and contract at the minimax rate. As a direct consequence,

[26] This is equivalent to running one DiD analysis by comparing each dosage d_j with zero dosage; see, e.g., Sun and Shapiro (2022).

[27] When one imposes the PT Assumption 4 instead of the SPT Assumption 5, each $\widehat{\beta}_j$ is a consistent estimator for the $ATT(d_j|d_j)$. However, comparison across $\widehat{\beta}_j$ does not give an $ACRT$-type parameter, as indicated in Theorem 3.2.

our estimators inherited from Chen, Christensen, and Kankanala (2022) these nice statistical guarantees when estimating and making inferences about $ATE(d)$, $ATT(d|d)$, and $ACR(d)$. We also show how to build on the estimators for $ACR(d)$ to construct an (efficient) estimator for the ACR^*.

In what follows, we discuss how we construct our data-adaptive estimator for the $ATE(d)$ and $ACR(d)$ curves under the SPT Assumption 5; if one imposes Assumption 4 instead, they estimate the $ATT(d|d)$ curve instead. First, as discussed above, we must pick a family of basis functions $\psi^K(d)$. We restrict our attention to dyadic (cubic) B-Splines as they are able to achieve minimax sup-norm rates; see, e.g., Belloni, Chernozhukov, Chetverikov, and Kato (2015), Chen and Christensen (2015) and Chen and Christensen (2018).[28]

Before discussing how we pick our data-driven choice $\widehat{K}$ of sieve dimension, we introduce some notation and discuss the form of our proposed estimators. Let $\mathcal{K} = \left\{\left(2^k+3\right)^d : k \in \mathbb{N}_0\right\}$ be the set of possible sieve dimensions for our cubic B-Splines. For a given sieve dimension $K \in \mathcal{K}$, our proposed nonparametric estimator for $ATE(d)$ and $ACR(d)$ are given by

$$\widehat{ATE}_K(d) = \left(\psi^K(d)\right)' \widehat{\beta}_K, \qquad \widehat{ACR}_K(d) = \left(\partial\psi^K(d)\right)' \widehat{\beta}_K, \tag{4.3}$$

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

$$\begin{aligned}\widehat{\beta}_K &= \arg\min_{b_K\in\Theta_K} \mathbb{E}_n\left[\left(\Delta Y - \mathbb{E}_n\left[\Delta Y|D=0\right] - \psi^K(D)'b\right)^2 \Big| D>0\right] \\ &= \mathbb{E}_n\left[1\{D>0\}\psi^K(D)\psi^K(D)'\right]^- \mathbb{E}_n\left[1\{D>0\}\psi^K(D)\left(\Delta Y - \mathbb{E}_n\left[\Delta Y|D=0\right]\right)\right], \end{aligned}\tag{4.4}$$

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

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

Note that $\widehat{\beta}_K$ is simply the OLS estimated coefficient of the regression of the "transformed outcome" $\Delta Y - \mathbb{E}_n\left[\Delta Y|D=0\right]$ onto the K-dimensional B-spline $\psi^K(D)$, in the sub-sample of units that have positive treatment dosage.

Let $K^+ = \min\{k \in \mathcal{K} : k > K\}$ be the smallest sieve dimension in $\mathcal{K}$ exceeding K, and $v_n = \max\left\{1, (0.1\log n)^4\right\}$ (so $v_n = 1$ unless n is bigger than 10 billion). Let $\{\omega_i\}_{i=1}^n$ be iid standard normal draws independent of the data $\{W_i\}_{i=1}^n = \{Y_{it}, Y_{it-1}, D_i\}_{i=1}^n$. In addition, for a given K and K_2, let

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

with

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

[28] We pick cubic B-splines out of convenience, but other choices are also possible and do not affect our theoretical results. We also note that one can use Cohen-Daubechies-Vial wavelet basis and get sup-norm minimax rates, though they are harder to compute than B-splines. On the other hand, not all basis functions lead to minimax rates–for example, polynomials and Fourier series do not lead to minimax rates. See Belloni, Chernozhukov, Chetverikov, and Kato (2015), Chen and Christensen (2015), Chen and Christensen (2018) and Chen, Christensen, and Kankanala (2022) for a discussion.

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, 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 $\sqrt{n}\left(\widehat{ATE}_K(d) - \widehat{ATE}_{K_2}(d)\right)$, and consider the bootstrap process

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

Our data-driven choice $\widehat{K}$ of the sieve dimension K leverages the Lepskii-type selection of Chen, Christensen, and Kankanala (2022) (henceforth, CCK) and can be computed according to the following Algorithm.

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

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

$$\widehat{\mathcal{K}} = \left\{K \in \mathcal{K} : 0.1\left(\log \widehat{K}_{max}\right)^2 \le K \le \widehat{K}_{max}\right\} \tag{4.5}$$

where

$$\widehat{K}_{max} = \min\left\{K \in \mathcal{K} : K\sqrt{\log K}v_n \le 10\sqrt{n} < K^+\sqrt{\log K^+}v_n\right\} \tag{4.6}$$

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*

$$\sup_{(d,K,K_2)\in\mathcal{D}^c_+\times\widehat{\mathcal{K}}\times\widehat{\mathcal{K}}:K_2>K}\left|\mathbb{Z}^*_n(d,K,K_2)\right|. \tag{4.7}$$

Let $\widehat{\gamma}_{1-\widehat{\alpha}}$ *denote the* $(1-\widehat{\alpha})$ *quantile of the sup-t statistic* (4.7) *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*

$$\widehat{K} = \inf\left\{K \in \widehat{\mathcal{K}} \ : \ \sup_{(d,K_2)\in\mathcal{D}^c_+\times\widehat{\mathcal{K}}:K_2>K}\frac{\sqrt{n}\left|\widehat{ATE}_K(d) - \widehat{ATE}_{K_2}(d)\right|}{\widehat{\sigma}_{K,K_2}(d)} \le 1.1\widehat{\gamma}_{1-\widehat{\alpha}}\right\}. \tag{4.8}$$

Our data-driven estimators for the $ATE(d)$ and $ACR(d)$ are therefore given by

$$\widehat{ATE}_{\widehat{K}}(d) = \left(\psi^{\widehat{K}}(d)\right)'\widehat{\beta}_{\widehat{K}}, \qquad \widehat{ACR}_{\widehat{K}}(d) = \left(\partial\psi^{\widehat{K}}(d)\right)'\widehat{\beta}_{\widehat{K}}. \tag{4.9}$$

Before we establish that $\widehat{ATE}_{\widehat{K}}(d)$ and $\widehat{ACR}_{\widehat{K}}(d)$ attain the minimax rate for estimating both $ATE(d)$ and $ACR(d)$, we introduce some define the parameter space for $ATE(\cdot)$. Let $H^p_{\infty,\infty}(M)$ denote the Holder ball of smoothness p and radius M. For given constants $M>0$ and $\underline{p} > \overline{p} > 0.5$, let $\mathcal{H}^p = H^p_{\infty,\infty}(M)$ and $\mathcal{H} = \bigcup_{p\in[\underline{p},\overline{p}]}\mathcal{H}^p$. For each $ATE(\cdot)\in\mathcal{H}$, we let $\mathbb{P}_{ATE}$ denote the distribution of $\{\Delta Y_i, D_i\}_{i=1}^{\infty}$ where each observation is generated by iid draws of of (D,u) from a distribution of (D,u) satisfying Assumptions 1, 2(a), 3, 5, Assumption 6 listed in the Appendix, and setting $\Delta Y - \mathbb{E}[\Delta Y|D=0] = ATE(D) + u$.

Theorem 4.1. *Let Assumptions 1, 2(a), 3, 5, and Assumption 6 listed in the Appendix hold. Then:*

(a) There exist a universal constant $C_1 > 0$ for which

$$\sup_{p\in[\underline{p},\overline{p}]}\ \sup_{ATE(\cdot)\in\mathcal{H}^p} \mathbb{P}_{ATE}\left(\sup_{d\in\mathcal{D}_+^c}\left|\left(\widehat{ATE}_{\widehat{K}} - ATE\right)(d)\right| > C_1\left(\frac{\log n}{n}\right)^{\frac{p}{2p+1}}\right) \to 0.$$

(b) For $\underline{p} > 1$, there exist a universal constant C_1' for which

$$\sup_{p\in[\underline{p},\overline{p}]}\ \sup_{ATE(\cdot)\in\mathcal{H}^p} \mathbb{P}_{ATE}\left(\sup_{d\in\mathcal{D}_+^c}\left|\left(\widehat{ACR}_{\widehat{K}} - ACR\right)(d)\right| > C_1'\left(\frac{\log n}{n}\right)^{\frac{p-1}{2p+1}}\right) \to 0.$$

Importantly, the convergence rates in parts (a) and (b) are the minimax rates for estimating $ATE(d)$ and $ACR(d)$, $d \in \mathcal{D}_+^c$, under sup-norm loss.

Remark 4 (Comparison with CCK)**.** *Our Algorithm 1 slightly differs from Procedure 1 of CCK. For instance, we consider a "transformed outcome" as the regressand of the sieve-based regression, whereas they consider an "observed" outcome as the regressand. We also focus on a specific subpopulation, those with positive treatment. These modifications are important in our DiD contests, as we allow for the causal effect of D on Y to be discontinuous when treatment dosage changes from $D = 0$ to $D = d_l$ (the minimum positive dosage). However, we note that these adaptions of the CCK procedures do not affect the asymptotic properties of the proposed estimators, as $\mathbb{E}_n[\Delta Y|D = 0]$ is $\sqrt{n}$-estimable and can be treated as known when establishing the asymptotic properties of the procedure. Thus, Theorem 4.1 follows from Theorem 4.1 of CCK.*

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

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

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

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

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

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

$$t^* = \sup_{(d,K)\in\mathcal{D}_+^c\times\widehat{\mathcal{K}}_-} |\mathbb{Z}_n^*(d, K, K_2)|, \quad \textit{and} \quad t^{*,acr} = \sup_{(d,K)\in\mathcal{D}_+^c\times\widehat{\mathcal{K}}_-} |\mathbb{Z}_n^{*,acr}(d, K, K_2)|. \tag{4.10}$$

   Let $z_{1-\alpha}$ and $z_{1-\alpha}^{acr}$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 $ATE(d)$ and $ACR(d)$, $d \in \mathcal{D}_+^c$, are respectively given by*

$$C_n(d) = \left[\widehat{ATE}_{\widehat{K}}(d) - \left(z_{1-\alpha}^* + \widehat{A}\,\widehat{\gamma}_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}_{\widehat{K}}(d)}{\sqrt{n}},\ \widehat{ATE}_{\widehat{K}}(d) + \left(z_{1-\alpha}^* + \widehat{A}\,\widehat{\gamma}_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}_{\widehat{K}}(d)}{\sqrt{n}}\right] \tag{4.11}$$

$$C_n^{acr}(d) = \left[\widehat{ACR}_{\widehat{K}}(d) - \left(z^{*,acr}_{1-\alpha} + \widehat{A}\;\widehat{\gamma}_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}^{acr}_{\widehat{K}}(d)}{\sqrt{n}},\; \widehat{ACR}_{\widehat{K}}(d) + \left(z^{*,acr}_{1-\alpha} + \widehat{A}\;\widehat{\gamma}_{1-\widehat{\alpha}}\right)\frac{\widehat{\sigma}^{acr}_{\widehat{K}}(d)}{\sqrt{n}}\right] \quad (4.12)$$

The UCBs described in Algorithm 2 enjoy attractive statistical guarantees such as *honesty* and *adaptivity*. In practice, these mean that these UCBs are guaranteed to have asymptotically corrected coverage over a large (and generic) class of data-generating processes (honesty), and contract at the minimax sup-norm rate (adaptivity). These nice guarantees are established over a generic subclass $\mathcal{G}$ of $\mathcal{H}$, as Low (1997) show that it is impossible to construct UCBs that are honest and adaptive over $\mathcal{H}$. This restriction, though, can be seen as a technical sidestep without major practical consequences, though; See Sections 4.3 and Appendix C.3 of CCK for a more detailed discussion. To save space, we define the class of self-similar functions $\mathcal{G}$ in the Appendix. Let $C_n(d, A)$ and $C_n({}^{acr}d, A)$ denote the UCBs from (4.11) and (4.12) replacing $\widehat{A}$ with a fixed $A > 0$.

The next theorem adapts Theorems 4.2 and 4.4 of CCK to our context.

Theorem 4.2. *Let Assumptions 1, 2(a), 3, 5, and Assumption 6 listed in the Appendix hold. Then:*

(a) There exist a universal constant $C_2 > 0$ and constant A_2^* (independent of α) such that for all $A \geq A_2^*$, we have

$$(i) \qquad \liminf_{n\to\infty} \inf_{ATE(\cdot)\in\mathcal{G}} \mathbb{P}_{ATE}\left(ATE(d) \in C_n(d, A) \quad \forall d \in \mathcal{D}_+^{cont}\right) \geq 1 - \alpha;$$

$$(ii) \qquad \inf_{p\in[\underline{p},\overline{p}]} \inf_{ATE(\cdot)\in\mathcal{G}^p} \mathbb{P}_{ATE}\left(\sup_{d\in\mathcal{D}_+^{cont}} |C_n(d, A)| \leq C_2(1 + A)\left(\frac{\log n}{n}\right)^{\frac{p}{2p+1}}\right) \to 1.$$

(b) For $\underline{p} > 1$, there exist a universal constant $C_2' > 0$ and constant $A_2^{*,\prime}$ (independent of α) such that for all $A \geq A_2^{*,\prime}$, we have

$$(i) \qquad \liminf_{n\to\infty} \inf_{ATE(\cdot)\in\mathcal{G}} \mathbb{P}_{ATE}\left(ACR(d) \in C_n^{acr}(d, A) \quad \forall d \in \mathcal{D}_+^{cont}\right) \geq 1 - \alpha;$$

$$(ii) \qquad \inf_{p\in[\underline{p},\overline{p}]} \inf_{ATE(\cdot)\in\mathcal{G}^p} \mathbb{P}_{ATE}\left(\sup_{d\in\mathcal{D}_+^{cont}} |C_n^{acr}(d, A)| \leq C_2'(1 + A)\left(\frac{\log n}{n}\right)^{\frac{p-1}{2p+1}}\right) \to 1.$$

We end this section discussing how one can build on the data-driven estimators $\widehat{ACR}_{\widehat{K}}(d)$ to construct an efficient estimator for ACR^* that can be used to summarize the ACR's. Our proposed estimator is simple to compute as it is based on the plug-in principle, i.e.,

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

with $n_{D>0} = \sum_{i=1}^n 1\{D_i > 0\}$ denoting the sample size with positive treatment dosage. The next theorem shows that, under some regularity conditions, this estimator is consistent, asymptotically normal, and semiparametrically efficient.[29]

[29] The semiparametric efficiency bound for average derivatives of conditional expectations was derived by Newey (1994).

Theorem 4.3. *Let Assumptions 1, 2(a), 3, 5, and Assumption 6 listed in the Appendix hold. Then,*

$$\sqrt{n}\left(\widehat{ACR}^* - ACR^*\right) \xrightarrow{d} N(0, V_{ACR}),$$

where V_{ACR} is the semiparametric efficiency bound of ACR^* and is given by

$$V_{ACR} = \text{Var}\left[ACR(D) - (\Delta Y - \mathbb{E}[\Delta Y|D=0] - ATE(D))\frac{f'_{D>0}(D)}{f_{D>0}(D)} \middle| D > 0\right]$$

Following Newey (1994) and Ackerberg, Chen, and Hahn (2012), we can form a simple and practical estimator for the V_{ACR} by "pretending" we follow a parametric model for the $ATE(d)$ and $ACR(d)$ functions and use the delta-method.

5 Continuous DiD in Practice: Causal Effects of Medicare PPS

Our results show that clarity about the causal question is crucial because it shapes the choice of an estimator, the identifying assumption, and the interpretation. TWFE does not estimate an interpretable causal parameter, but non-parametric methods can. This section evaluates estimators for treatment effects and causal responses in AF's study of Medicare PPS and discusses the interpretation and assumptions of both.

To begin, consider the profit maximization problem for a hospital with Medicare inpatient share of $M = m$. We follow AF and assume a production function, $F_t(L, K)$, that is homothetic in labor (L), and capital (K).[30] Market wages and rental rates are normalized by the output price, and Medicare subsidies mean that net input prices are $(1 - s_{Lt}M)w$ and $(1 - s_{Kt}M)r$.

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

The solution to this problem generates factor demands $K_t^*((1-s_{Lt}M)w, (1-s_{Kt}M)w)$ and $L_t^*((1-s_{Lt}M)w, (1-s_{Kt}M)w)$, and an associated capital labor ratio that is only a function of the input price ratio.

We define the "dose" or "intensity" of price regulation as the extent to which a hospital's subsidy ratio differs from one: $\frac{(1-s_{Lt}M)}{(1-s_{Kt}M)} - 1 = \frac{(s_{Kt}-s_{Lt})M}{1-s_{Kt}M}$. This reflects the fact that we observe untreated outcomes in all periods for hospitals with no Medicare patients $(M = 0)$ and for all hospitals before PPS when $s_{K,t-1} = s_{L,t-1} = s$.[31] Because PPS set $s_{Lt} = 0$ after 1983, we denote the post-PPS dose as $D \equiv \frac{s_{Kt}M}{1-s_{Kt}M}$. The post-PPS net-of-subsidy input price ratio thus equals $(1 + D)\frac{w}{r}$, and we write potential outcomes for the capital-labor ratio as:

$$Y_{t-1} = Y_{t-1}(0) \equiv \frac{K_t^*((1 - sM)w, (1 - sM)r)}{L_t^*((1 - sM)w, (1 - sM)r)}$$

[30] AF's theoretical analysis includes a productivity shifter/technology choice, an explicit output price in order to model Medicare's prospective payment (output price subsidy) as well as multiple types of labor, and an endogenous choice of m. We set these modelling choices aside because they do not alter out points about how features of the production technology map to econometric assumptions.

[31] We assumed weakly positive doses but $D < 0$ whenever $s_{Kt} < s_{Lt}$. In practice this does not happen under Medicare PPS, so the non-negative dose assumption is satisfied in this example.

$$Y_t = Y_t(D) \equiv Y_i\left(\frac{s_{Kt}M}{1 - s_{Kt}M}\right) = \frac{K_t^*(w, (1 - s_{Kt}M)r)}{L_t^*(w, (1 - s_{Kt}M)r)}$$

Three details of the theoretical set up are worth noting. First, homotheticity allows us to write potential outcomes as a function of one treatment–the subsidy *ratio*–because it implies that scale effects do not alter input ratios. This strong assumption generates sign predictions and structural interpretations of treatment effects and causal responses. Without it, potential outcomes are a function of net labor and capital prices separately and both the theoretical analysis and the definition of causal parameters is more complicated. Second, 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). Third, PPS creates convex relationship between D and M. We define causal parameters in terms of D, but empirical comparisons across values of M will mix together these parameters with the change of variables required to go from M to D.[32]

5.1 Average Treatment Effects of PPS

PPS sought to help hospitals invest in new medical technologies with the aim of improving patient outcomes [CITE THAT OTA REPORT], but costly capital investments were also a key source of medical inflation in the 1980s. The theoretical model predicts that PPS would raised capital-labor ratios for all treated hospitals (the first part of AF's prediction 1). The causal question that AF are primarily interested in–did PPS raise capital-labor ratios?–is thus of first-order importance both theoretically and in terms of policy.

The building block parameter that answers this question is the average treatment effect of PPS on hospitals with $M = m$:

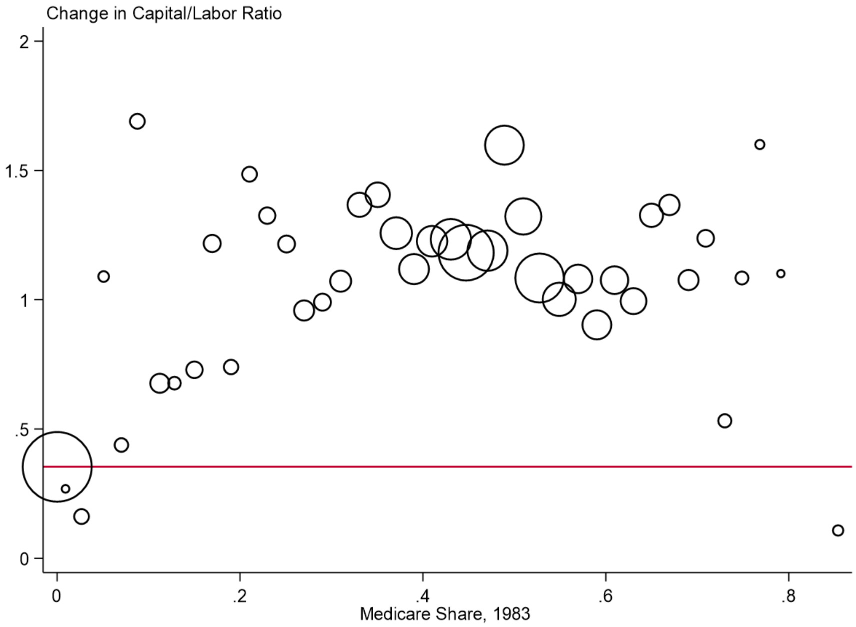
$$ATT(d|d) = \mathbb{E}[Y_t(d) - Y_t(0)|D = d] = E\left[Y_t\left(\frac{s_{Kt}m}{1 - s_{Kt}m}\right) - Y_t(0)\middle| M = m\right]$$

As a rough visualization, Figure 4 presents a binned scatter plot of the change in mean capital labor ratios before and after PPS against the Medicare share of inpatient days m.[33] The red horizontal line equals the mean change in capital-labor ratio for untreated hospitals, each circle is the mean outcome change for a given bin of the Medicare inpatient share, and their size is 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. If Assumption 4 holds, then this suggests that $ATT(d|d)$ is generally positive.

Results from the nonparametric estimator in Figure 9 formalize what the scatter plot in Figure 4 suggests: $ATT(d|d) > 0$ for all observed doses. We do not detect an effect for values of m below 5 percent, but we do reject zero for doses between 0.05 and 0.78. Only one percent of hospitals had

[32] Alternative policies that changed s_L by a different amount or changed the capital subsidy as well, would create a different treatment for the group of hospitals with $M = m$. This highlights the distinction between heterogeneous effects *across* groups of a given policy and heterogeneity *within* a group to alternative policies.

[33] We use bins of 2 percentage points to make the figure readable, but use the reported values in our replication of AF's results and application of our estimators.



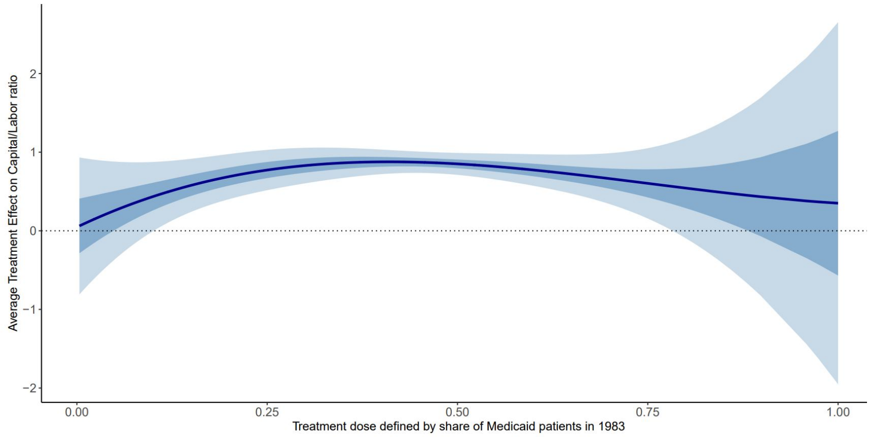
Notes:

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

more than 78 percent of their inpatient days accounted for by Medicare patients, and the $ATT(d|d)$ estimates are smaller and much less precise at these highest doses. We return to this point below when we discuss ACR estimates.

The next choice in a typical continuous DiD analysis is how to aggregate the collection of $ATT(d|d)$ estimates that come from the nonparametric approach. Our view is that the parameter of interest should guide this choice, and the most natural aggregation to answer questions about PPS' treatment effects is based on the actual distribution of the dose (Sun and Shapiro, 2022). The solid line in Figure 10 plots the smoothed density of M (and therefore D) among treated hospitals. Averaging the curve in Figure 9 directly using $f_{D|D>0}(d)$ gives an estimate of $ATT^* = \mathbb{E}[ATT(D|D)|D > 0]$ 0.8. In fact, getting ATT^* is even easier: estimate a 2x2 DiD comparing average outcome changes for treated versus untreated units: $\mathbb{E}[\Delta Y|D > 0] - \mathbb{E}[\Delta Y|D = 0]$. Because both terms can be written as integrals over $\mathcal{D}_+$ weighted by $f_{D|D>0}(d)$, this approach also yields $ATT^* = 0.8$ (s.e. $= 0.06$).

As Theorem 3.4 shows, the TWFE estimate of 1.13 comes from an alternative way to aggregate $ATT(d|d)$ parameters. How does this compare to ATT^*? First, because about 20 percent of hospitals have positive Medicare shares that are still below $\mathbb{E}[M]$, Theorem 3.4 and Theorem 3.5 imply that the TWFE estimate has negative weights. The long-dashed line in Figure 10 shows this directly by plotting $w^{lev}(d)$ for $M > 0$. But since all $ATT(d|d)$'s are positive, negative weighting should *understate* ATT^*, yet β^{twfe} is larger than our preferred estimate of ATT^*. The second



Notes:

Figure 5: Non-Parametric Estimates of $ATT(d|d)$ for Medicare PPS

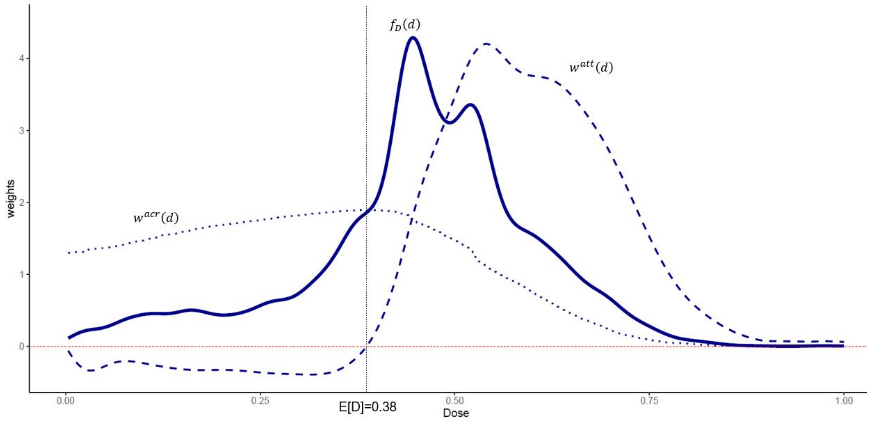
reason that these two estimates differ is that, as discussed above, TWFE scales the (negatively weighted) average of ATT's by a measure of D. Therefore, comparing $\beta^{twfe} = 1.13$ to a similarly scaled parameter weightd by $f_{D|D>0}(d)$, $ATT^*/E[D] = 0.8/0.45 = 1.78$, shows that TWFE is an underestimate.

Identification of treatment effects relies on parallel trends which researchers typically justify by presenting evidence of small *pre*-treatment trends in the outcome (an implication of parallel trends in all periods). Figure 11 plots event-study estimates of ATT_t^* parameters that compare treated to untreated units so that each coefficient equals:

$$\beta_{1983+k}^{twfe} = \mathbb{E}[Y_{1983+k}(0) - Y_{1983}(0)|D > 0] - E[Y_{1983+k}(0) - Y_{1983}(0)|D = 0]$$

An extension of Assumption 4 to all periods implies that these coefficients are zero when $k < 0$, although the converse is not true. The actual pre-treatment estimates are small. There is a statistically significant difference in 1982 of -0.19. This 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 grows in each year following PPS, which matches the fact that PPS' subsidy reforms actually phased in over three years.

Remark 3 above showed that TWFE estimates of pre-trends will tend to understate pre-treatment differences when parallel trends violations all have the same sign and some treated groups receive negative weights. The TWFE estimate in 1982 in Figure 1 is 0.29. In contrast, the estimate from Figure 11 scaled by the mean dose among treated hospitals (0.45) equals 0.42. TWFE thus underestimates the one-year pre-trend by about one third. Although TWFE still rejects the null of zero for the 1982 coefficient, in general it carries the risk of masking evidence of parallel trends violations by putting negative weight on pre-treatment outcome changes of low-dose units.



Notes:

Figure 6: Weighting Schemes: ATT, ACR, and Dose Distribution for AF

5.2 Average Causal Responses to PPS

The average causal response on the treated after PPS equals:

$$ACRT(d_i|d_i) = E\big[Y'_{it}(d_i)\big|D = d_i\big] = E\left[Y'_{it}\left(\frac{s_K m_i}{1 - s_K m_i}\right)\middle| M = m_i\right]$$

Here the mapping between m_i and d_i is important for interpreting a DiD estimand that compares the change in outcomes for two dose groups:

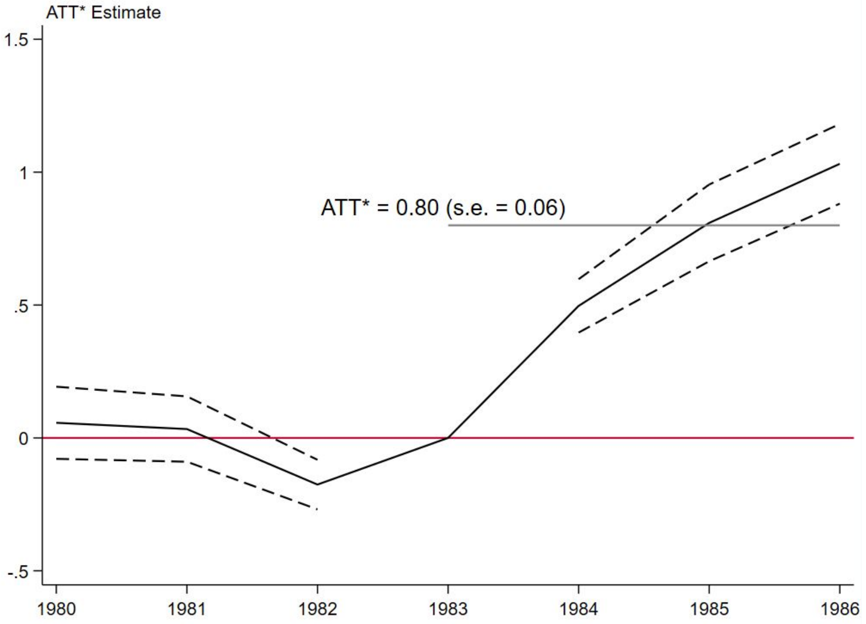
$$\begin{aligned}&\frac{\partial}{\partial m} E\left[Y_t\left(\frac{s_K m}{1 - s_K m}\right) - Y_{t-1}(0)\middle| M = m\right] = \\ &= \frac{s_K}{(1 - s_K m)^2} ACRT(d_i|d_i) + \frac{\partial}{\partial \ell} ATT(d|\ell)\big|_{\ell=m} + \frac{\partial}{\partial \ell}\mathbb{E}[\Delta Y(0)\big|M = \ell]\big|_{\ell=m}\end{aligned}$$

This expression shows that under Assumption 5, which ensures that the second and third terms equal zero, DiD comparisons between hospitals with different Medicare inpatient shares equal $ACR(d_i)$ times the PPS-specific mapping between m_i and the subsidy ratio: $\frac{\partial d}{\partial m} = \frac{s_K}{(1-s_K m)^2}$. Since the dose is convex in m, small differences in m skew subsidy ratios more at high Medicare share than at low ones. A DiD strategy based on m itself therefore yields a parameter that is the product of this mapping and the underlying ACR.

This also highlights the economic content of SPT in the context of the theoretical model.

6 Conclusion

In this paper, we have studied difference-in-differences approaches to identifying and estimating causal effects of a multi-valued or continuous treatment. The paper has a number of results that are potentially useful to empirical researchers, and, to conclude the paper, we briefly summarize these results.

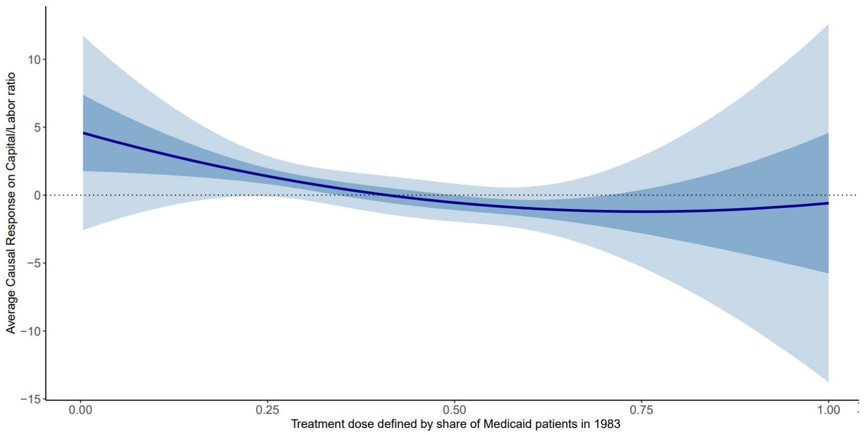


Notes:

Figure 7: Event-Study Estimates of ATT^*

First, while ATT-type parameters can be identified under a standard parallel trends assumption, a fundamental complication in the case with a multi-valued/continuous treatment is that comparisons across different amounts of the treatment are confounded by "selection bias" type terms that make these sorts of comparisons very difficult to interpret. This kind of bias carries over to identifying average causal responses of more dose. These sorts of difficulties can be avoided by invoking alternative parallel trends assumptions, but these assumptions are likely to be substantially stronger than the ones most researchers have in mind when they are using a difference-in-differences identification strategy. In addition, pre-tests commonly used in DiD applications are not able to distinguish between these two types of parallel trends assumptions.

Furthermore, two way fixed effects regressions that are commonly used by empirical researchers have a number of drawbacks. In a baseline case with two periods, TWFE regressions deliver a weighted average of causal responses to the treatment. The weights are all positive, but they are driven by the estimation procedure which can result in misleading results in a number of realistic cases. Moreover, in cases where there are multiple time periods, variation in treatment timing and in treatment intensity (which are common in applications), TWFE regressions are additionally sensitive to (i) treatment effect dynamics and (ii) heterogeneous causal responses across timing groups. We propose an identification and estimation strategy that is straightforward to implement and does not suffer from these drawbacks.



Notes:

Figure 8: Non-Parametric Estimates of $ATT(d|d)$ for Medicare PPS

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. ISSN: 00223808.

Ackerberg, Daniel, Xiaohong Chen, and Jinyong Hahn (2012). “A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators”. *Review of Economics and Statistics* 94.2, pp. 481–498.

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.

Angrist, Joshua D (1991). “Grouped-data estimation and testing in simple labor-supply models”. *Journal of Econometrics* 47.2-3, pp. 243–266.

Angrist, Joshua D and Ivan Fernandez-Val (2013). “ExtrapoLATE-ing: External validity and overidentification in the LATE framework”. *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*. Vol. 51. Cambridge University Press, p. 401.

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.

Athey, Susan and Guido Imbens (2006). “Identification and inference in nonlinear difference-in-differences models”. *Econometrica* 74.2, pp. 431–497.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato (2015). “Some new asymptotic theory for least squares series: Pointwise and uniform results”. *Journal of Econometrics* 186.2, pp. 345–366.

Borusyak, Kirill and Xavier Jaravel (2017). “Revisiting event study designs”. Working Paper.

Callaway, Brantly and Pedro H.C. Sant’Anna (2020). “Difference-in-differences with multiple time periods”. *Journal of Econometrics* Forthcoming.

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

Card, David (1992). “Using regional variation in wages to measure the effects of the federal minimum wage”. *Industrial and Labor Relations Review* 46.1, pp. 22–37.

Cattaneo, Matias, Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare (2021). “Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs”. *Journal of the American Statistical Association* 116.536, pp. 1941–1952.

Cattaneo, Matias, Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele (2016). “Interpreting Regression Discontinuity Designs with Multiple Cutoffs”. *Journal of Politics* 78.4, pp. 1229–1248.

Chen, Xiaohong, Timothy Christensen, and Sid Kankanala (2022). “Adaptive Estimation and Uniform Confidence Bands for Nonparametric Structural Functions and Elasticities”. *arXiv:2107.11869*.

Chen, Xiaohong and Timothy M Christensen (2018). “Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression”. *Quantitative Economics* 9.1, pp. 39–84.

— (2015). “Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions”. *Journal of Econometrics* 188.2, pp. 447–465.

Chodorow-Reich, Gabriel, Plamen T. Nenov, and Alp Simsek (May 2021). “Stock Market Wealth and the Real Economy: A Local Labor Market Approach”. *American Economic Review* 111.5, pp. 1613–57.

D’Haultfoeuille, Xavier, Stefan Hoderlein, and Yuya Sasaki (2021). “Nonlinear difference-indifferences in repeated cross sections with continuous treatments”. Working Paper.

de Chaisemartin, Clement and Xavier D’Haultfœuille (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.

Florens, Jean-Pierre, James J Heckman, Costas Meghir, and Edward Vytlacil (2008). “Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects”. *Econometrica* 76.5, pp. 1191–1206.

Goodman-Bacon, Andrew (2018). “Public Insurance and Mortality: Evidence from Medicaid Implementation”. *Journal of Political Economy* 126.1, pp. 216–262.

— (2021). “Difference-in-differences with variation in treatment timing”. *Journal of Econometrics* forthcoming.

Heckman, James J., Sergio Urzua, and Edward Vytlacil (2006). “Understanding Instrumental Variables in Models with Essential Heterogeneity”. *The Review of Economics and Statistics* 88.3, pp. 389–432.

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

Hill, Sir Austin Bradford (1965). “The environment and disease: association or causation?” *Journal of the Royal Society of Medicine* 58.5, pp. 295–300.

Hirano, Keisuke and Guido W Imbens (2004). “The propensity score with continuous treatments”. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* 226164, pp. 73–84.

Ichimura, Hidehiko and Petra E. Todd (2007). “Implementing Nonparametric and Semiparametric Estimators”. *Handbook of Econometrics*. Vol. 6, Part B. Amsterdam: North-Holland: Elsevier. Chap. 74, pp. 5369–5468.

Low, Mark G. (1997). “On nonparametric confidence intervals”. *Annals of Statistics* 25.6, pp. 2547–2554. ISSN: 00905364.

Marcus, Michelle and Pedro HC Sant’Anna (2021). “The role of parallel trends in event study settings: An application to environmental economics”. *Journal of the Association of Environmental and Resource Economists* 8.2, pp. 235–275.

Meyer, Bruce D. (1995). “Natural and Quasi-Experiments in Economics”. *Journal of Business & Economic Statistics* 13.2, pp. 151–161.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky (2018). “Using instrumental variables for inference about policy relevant treatment parameters”. *Econometrica* 86.5, pp. 1589–1619.

Newey, Whitney (1994). “The asymptotic variance of semiparametric estimators”. *Econometrica*, pp. 1349–1382.

Oreopoulos, Philip (2006). “Estimating average and local average treatment effects of education when compulsory schooling laws really matter”. *American Economic Review* 96.1, pp. 152–175.

Roth, Jonathan, Pedro H. C. Sant’Anna, Alyssa Bilinski, and John Poe (2023). “What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature”. *Journal of Econometrics* 235.2, pp. 2218–2244.

Słoczyński, Tymon (2022). “Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights”. *The Review of Economics and Statistics* 104.3, pp. 501–509.

Sun, Liyan and Sarah Abraham (2021). “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects”. *Journal of Econometrics* 225.2.

Sun, Liyang and Jesse M. Shapiro (2022). “A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure”. *Journal of Economic Perspectives* 36.4, pp. 193–204.

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

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

A Additional Assumptions

Let $\Delta Y - \mathbb{E}[\Delta Y|D=0] = h(D) + u$. Under Assumption 4, $h(d) = ATT(d|d)$, whereas under Assumption 5, $h(d) = ATE(d)$. Let $\overline{\sigma}, \underline{\sigma}, \overline{C}, \underline{c}$ be some finite, positive constants, and $\rho \in (0,1)$. Finally, let

$$\sigma_K^2(d) = \psi^K(d)'\mathbb{E}\left[\psi^K(D)\psi^K(D)'\middle| D>0\right]^{-}\mathbb{E}\left[u^2\psi^K(D)\psi^K(D)'\middle| D>0\right]\mathbb{E}\left[\psi^K(D)\psi^K(D)'\middle| D>0\right]^{-}\psi^K(d),$$

and $||\sigma_{d,K}||^2 = \psi^K(d)'\mathbb{E}\left[\psi^K(D)\psi^K(D)'\middle| D>0\right]^{-}\psi^K(d)$, which satisfies $||\sigma_{d,K}|| \asymp \sigma_K(d)$ under Assumption 6(i) below. Let $\left|\left|\sigma_{d,K}^{acr}\right|\right|^2 = (\partial\psi^K(d))'\mathbb{E}\left[\psi^K(D)\psi^K(D)'\middle| D>0\right]^{-}(\partial\psi^K(d))$.

Assumption 6 (Additional regularity conditions)**.**

(i) $\mathbb{P}\left(\mathbb{E}\left[u^4|X,D>0\right] \leq \overline{\sigma}^2\right) = 1$*, and* $\mathbb{P}\left(\mathbb{E}\left[u^2|X,D>0\right] \geq \underline{\sigma}^2\right) = 1$.

(ii) $\underline{c}K \leq \inf_{d\in\mathcal{D}_+^{cont}}||\sigma_{d,K}||^2 \leq \sup_{d\in\mathcal{D}_+^{cont}}||\sigma_{d,K}||^2 \leq \overline{C}K$ *for all* $K \in \mathcal{K}$*;*

(iii) $\limsup_{K\to\infty}\sup_{d\in\mathcal{D}_+^{cont},K_2\in\mathcal{K}:K_2>K}(\sigma_K^2(d)/\sigma_{K_2}^2(d)) < \rho$*;*

(iv) $\underline{c}K^3 \leq \inf_{d\in\mathcal{D}_+^{cont}}\left|\left|\sigma_{d,K}^{acr}\right|\right|^2 \leq \sup_{d\in\mathcal{D}_+^{cont}}\left|\left|\sigma_{d,K}^{acr}\right|\right|^2 \leq \overline{C}K^3$ *for all* $K \in \mathcal{K}$*.*

Assumption 6(iii) is only needed for Theorem 4.2(b), but we keep this assumption here for simplicity.

B Comparing Alternative Parallel Trends Assumptions

It is worth thinking more carefully about the differences between Assumption 4 and Assumption 5. In this section, we show that Assumption 5 is not strictly stronger than Assumption 4 though it is likely to be *stronger in practice* in most applications. Here, we maintain Assumption 3, so $Y_{t-1}(d) = Y_{t-1}(d') = Y_{t-1}(0)$ for any $(d,d') \in \mathcal{D}\times\mathcal{D}$.

To see that Assumption 5 is not strictly stronger, consider the case where there are two doses d_1 and d_2. In this case, Assumption 4 is equivalent to the following conditions

$$\mathbb{E}[\Delta Y_t(0)|D=d_1] = \mathbb{E}[\Delta Y_t(0)|D=d_2] = \mathbb{E}[\Delta Y_t(0)|D=0] \tag{B.1}$$

while Assumption 5 is equivalent to

$$\mathbb{E}[\Delta Y_t(0)] = \mathbb{E}[\Delta Y_t(0)|D=0] \tag{Comp-0}$$

$$\mathbb{E}[Y_t(d_1) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d_1) - Y_{t-1}(0)|D=d_1] \tag{Comp-1}$$

$$\mathbb{E}[Y_t(d_2) - Y_{t-1}(0)] = \mathbb{E}[Y_t(d_2) - Y_{t-1}(0)|D=d_2] \tag{Comp-2}$$

Assumption 4 does not place any restrictions on any potential outcomes besides untreated potential outcomes, and therefore the "extra" conditions in Equations (Comp-1) and (Comp-2) imply that Assumption 5 is not weaker than Assumption 4.

On the other hand, Equation (Comp-0) does not imply Equation (B.1); rather, it implies that

$$\mathbb{E}[\Delta Y_t(0)|D=0] = \mathbb{E}[\Delta Y_t(0)|D=d_1]\frac{\mathbb{P}(D=d_1)}{\mathbb{P}(D=d_1)+\mathbb{P}(D=d_2)} + \mathbb{E}[\Delta Y_t(0)|D=d_2]\frac{\mathbb{P}(D=d_2)}{\mathbb{P}(D=d_1)+\mathbb{P}(D=d_2)}$$

In other words, the trend in untreated potential outcomes does not have to be exactly the same for all doses, but, instead, they have to be the same on average.

In practice, this potentially allows for some units to select their amount of dose on the basis of the path of their untreated potential outcomes (which is not allowed under the standard parallel trends assumption in Assumption 4), but that the amount of selection has to average out across doses to be equal to zero. It seems hard to think of realistic cases where Assumption 5 would be practically weaker than Assumption 4, though.

A related alternative assumption is

Assumption 5-Alt (Alternative Strong Parallel Trends Assumption)**.** *For all $d \in \mathcal{D}$ and $l \in \mathcal{D}$,*

$$\mathbb{E}[Y_t(d) - Y_{t-1}(d)|D = l] = \mathbb{E}[Y_t(d) - Y_{t-1}(d)|D = d]$$

Assumption 5-Alt is a stronger, but related version of the strong parallel trends assumption in Assumption 5. Assumption 5-Alt says that across all potential doses d, the path of potential outcomes $Y_t(d) - Y_{t-1}(d)$ (which, under a no-anticipation assumption, is the path of outcomes that a unit would experience if they experienced dose d in time period t) is, on average, (i) the same across all actual doses experienced, l, and (ii) is equal to the average path of outcomes for units that actually experienced dose d. Further, note that $\mathbb{E}[Y_t(d) - Y_{t-1}(d)|D = l]$ is not identified from the sampling process except in the case where $l = d$ (i.e., the left-hand side of the equation in Assumption 5-Alt is not identified from the sampling process, but the right-hand side is). It is immediately clear that this assumption implies both Assumptions 4 and 5. While it does not place restrictions on the levels of untreated potential outcomes in period $t-1$, it does place (substantial) restrictions on treated potential outcomes and on treatment effect heterogeneity which is demonstrated in the next proposition.

Proposition B.1. *Assumption 5-Alt implies that*

$$ATE(d) = ATT(d|d)$$

Proof. Starting with the definition of $ATE(d)$,

$$\begin{aligned}
ATE(d) &= \mathbb{E}[Y_t(d) - Y_t(0)] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)] \\
&= \mathbb{E}[Y_t(d) - Y_{t-1}(0)|D = d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|D = d] \\
&= \mathbb{E}[Y_t(d) - Y_t(0)|D = d] \\
&= ATT(d|d)
\end{aligned}$$

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t-1}(0)]$, the third equality holds by Assumption 5-Alt (also, notice that this equality does not hold under Assumption 4, nor is $ATE(d)$ generally equal to $ATT(d|d)$ under Assumption 4 alone), the fourth equality holds by canceling the two $\mathbb{E}[Y_{t-1}(0)|D = d]$ terms, and the remaining term in that equality is $ATT(d|d)$ □

Proposition B.1 shows that Assumption 5-Alt implies that the overall average effect of dose d is equal to the average effect of dose d for units who actually experienced dose d. The implication of this result is that Assumption 5-Alt rules out many forms of selection into a particular dose d on the basis of the effect of that amount of dose.

We end this section by noting that although Assumption 5 allows one to identify the $ATE(d)$'s, it does not allow one to identify the $ATT(d|d)$'s parameters. This is because Assumption 5 does not allow identification of $\mathbb{E}[Y_t(0)|D = d]$, only their averages across values of d. This indeed highlights that Assumption 4 and Assumption 5 are non-nested. Of course, as illustrated above in Proposition B.1, Assumption 5-Alt is strong enough to imply these two conditions.

C Alternative Decompositions for TWFE Regression

In this section, we provide three alternative decompositions of the TWFE regression estimator in Equation (1.1) in the baseline case with two periods, where no unit is treated yet in the first period, and where some units remain untreated in the second period.

The first decomposition is one where β^{twfe} equals a weighted average of 2×2 DiD comparisons between pairs of dose groups scaled by the difference in their doses:

Proposition C.1. *Consider β^{twfe} in Equation (1.1) and suppose that Assumption 1 holds.*

(1) If Assumption 2(a) holds, then

$$\beta^{twfe} = \int_{\mathcal{D}_+} \int_{\mathcal{D}, h>l} w_1^{2\times 2, cont}(l,h) \frac{(m_\Delta(h) - m_\Delta(l))}{(h-l)} \, dh \, dl + \int_{\mathcal{D}, h>0} w_0^{2\times 2, cont}(h) \frac{m_\Delta(h) - m_\Delta(0)}{h} \, dh$$

where

$$w_1^{2\times 2, cont}(l,h) = (h-l)^2 (f_D(h) + f_D(l))^2 f_{D|\{h,l\}}(h) f_{D|\{h,l\}}(l) / \mathrm{Var}(D)$$
$$w_0^{2\times 2, cont}(h) = h^2 (f_D(h) + p_0^D)^2 f_{D|\{h,0\}}(h) p^D_{0|\{h,0\}} / \mathrm{Var}(D)$$

and

$$f_{D|\{h,l\}}(h) = f_D(h)/(f_D(h) + f_D(l))$$
$$f_{D|\{h,l\}}(l) = f_D(l)/(f_D(h) + f_D(l))$$
$$f_{D|\{h,0\}}(h) = f_D(h)/(f_D(h) + p_0^D)$$
$$p^D_{0|\{h,0\}} = p_0^D/(f_D(h) + p_0^D)$$

In addition, $w_1^{2\times 2, cont}(l,h) \geq 0$ for all $(l,h) \in \mathcal{D}_+ \times \mathcal{D}_{h>l}$, $w_0^{2\times 2, cont}(h) \geq 0$ for all $h \in \mathcal{D}_+$, and $\int_{\mathcal{D}_+} \int_{\mathcal{D}, h>l} w_1^{2\times 2, cont}(l,h) \, dh \, dl + \int_{\mathcal{D}_+} w_0^{2\times 2, cont}(h) \, dh = 1$.

(2) If Assumption 2(b) holds, then

$$\beta^{twfe} = \sum_{l \in \mathcal{D}} \sum_{h \in \mathcal{D}, h>l} w^{2\times 2, disc}(l,h) \frac{(m_\Delta(h) - m_\Delta(l))}{(h-l)}$$

where

$$w^{2\times 2, disc}(l,h) = (h-l)^2 (p_l^D + p_h^D)^2 p^D_{l|\{g,h\}} (1 - p^D_{l|\{g,h\}}) / \mathrm{Var}(D)$$
$$p^D_{l|\{l,h\}} = \mathbb{P}(D = l | D \in \{l,h\})$$

and $p_h^D = \mathbb{P}(D = h)$, $p_l^D = \mathbb{P}(D = l)$. In addition, $w^{2\times 2, disc}(l,h) \geq 0$ for all $(l,h) \in \mathcal{D}^2$ and $\sum_{l \in \mathcal{D}} \sum_{h \in \mathcal{D}, h>l} w^{2\times 2, disc}(l,h) = 1$.

Proof. From the proof of Theorem 3.4, we have that

$$\begin{aligned} \beta &= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\mathrm{Var}(D)} m_\Delta(D) \right] \\ &= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}} (h - \mathbb{E}[D]) m_\Delta(h) \, dF_D(h) \\ &= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}} \left(h - \int_{\mathcal{D}} l \, dF_D(l) \right) m_\Delta(h) \, dF_D(h) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}} \int_{\mathcal{D}} (h-l) m_{\Delta}(h)\, dF_D(h)\, dF_D(l) \\
&= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}} \int_{\mathcal{D},h>l} (h-l)(m_{\Delta}(h) - m_{\Delta}(l))\, dF_D(h)\, dF_D(l) \\
&= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}} \int_{\mathcal{D},h>l} (h-l)^2 \frac{(m_{\Delta}(h) - m_{\Delta}(l))}{(h-l)}\, dF_D(h)\, dF_D(l) \qquad \text{(C.1)}
\end{aligned}
$$

where the second equality holds by writing the expectation as an integral, the third equality write $\mathbb{E}[D]$ as an integral, the fourth equality rearranges terms, the fifth equality holds because the integrations are symmetric, and the last equality holds by multiplying and dividing by $(h-l)$.

The above arguments hold if treatment is continuous or discrete. Under Assumption 2

$$
\begin{aligned}
\text{Equation (C.1)} = {} & \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}_+} \int_{\mathcal{D},h>l} (h-l)^2 \frac{(m_{\Delta}(h) - m_{\Delta}(l))}{(h-l)} f_D(h) f_D(l)\, dh\, dl \\
& + \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D},h>0} h^2 \frac{m_{\Delta}(h) - m_{\Delta}(0)}{h} f_D(h) p_0^D\, dh
\end{aligned}
$$

which holds by splitting up the first integral in Equation (C.1) by whether $l \in \mathcal{D}_+$ or $l = 0$. Then, the result for part (1) holds by multiplying and dividing the first line by $(f_D(h) + f_D(l))^2$ and by the definition $f_{D|\{h,l\}}$ and multiplying and dividing the second line by $(f_D(h) + p_0^D)^2$ and by the definitions of $f_{D|\{h,0\}}$ and $p^D_{0|\{h,0\}}$.

Under Assumption 2(b),

$$
\begin{aligned}
\text{Equation (C.1)} &= \frac{1}{\mathrm{Var}(D)} \sum_{l \in \mathcal{D}} \sum_{h \in \mathcal{D}, h>l} (h-l)^2 \frac{(m_{\Delta}(h) - m_{\Delta}(l))}{(h-l)} p_h^D p_l^D \\
&= \frac{1}{\mathrm{Var}(D)} \sum_{l \in \mathcal{D}} \sum_{h \in \mathcal{D}, h>l} (h-l)^2 \frac{(m_{\Delta}(h) - m_{\Delta}(l))}{(h-l)} (p_l^D + p_h^D)^2 p^D_{l|\{g,h\}} (1 - p^D_{l|\{g,h\}})
\end{aligned}
$$

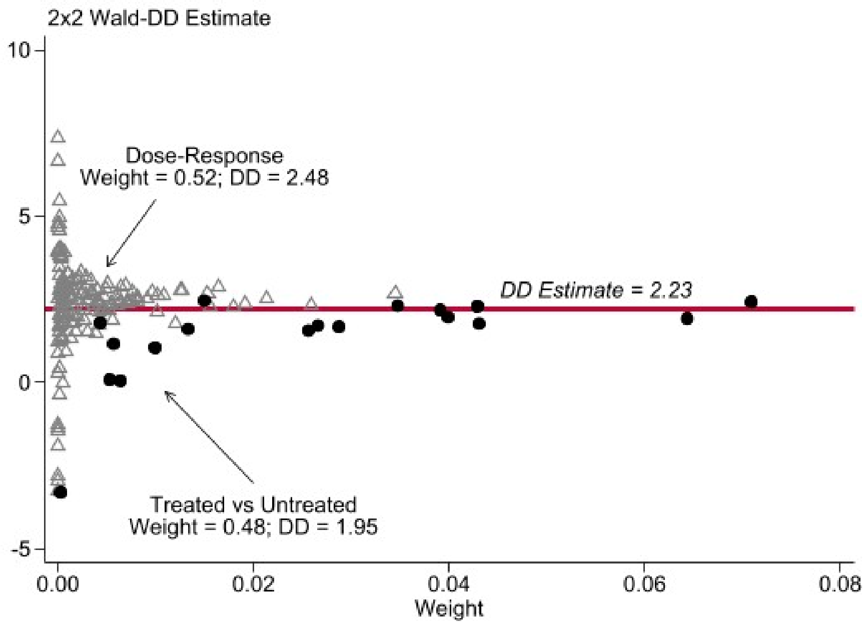
where the first equality holds immediately and the second equality holds by multiplying and dividing by $(p_l^D + p_h^D)^2$ and by the definition of $p^D_{l|\{g,h\}}$.

That the weights are all positive holds immediately by their definitions. That the weights integrate to one holds because

$$
\begin{aligned}
\int_{\mathcal{D}_+} \int_{\mathcal{D},h>l} w_1^{2\times 2,cont}(l,h)\, dh\, dl + \int_{\mathcal{D}_+} w_0^{2\times 2,cont}(h)\, dh &= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}} \int_{\mathcal{D}} \mathbf{1}\{h > l\} (h-l)^2\, dF_D(h)\, dF_D(l) \\
&= 1
\end{aligned}
$$

The same sort of argument holds for the discrete case as well. □

Proposition C.1 is analogous to the decomposition theorem for binary staggered timing designs in Goodman-Bacon (2021) in that it expresses the TWFE coefficient as a variance-weighted average of 2 × 2 DiD comparisons (it is also mechanically very similar to the Wald-IV theorem in Angrist (1991)). Each 2 × 2 term is the change in average outcomes for a group with a higher dose ($m_{\Delta}(h)$) minus the same difference for a group with a lower dose ($m_{\Delta}(l)$), divided by the difference in their doses $(h-l)$. de Chaisemartin and D'Haultfœuille (2018) refer to this as a "Wald-DD" estimator. The weights combine the size of each subsample, $(p_l^D + p_h^D)$, with the variance of the dose in that subsample. The variance depends on the relative size of the two groups, measured by $p^D_{l|\{g,h\}}(1 - p^D_{l|\{g,h\}})$, and the distance between their doses, $(h-l)$. This formula reflects the intuitive



Notes: The figure plots each 2×2 Wald DiD estimate against its weight from Proposition C.1. Closed black circles are comparisons between one dose group and untreated observations: $(\Delta\bar{Y}_h - \Delta\bar{Y}_0)/h$. Open gray triangles are comparisons between two dose groups: $(\Delta\bar{Y}_h - \Delta\bar{Y}_\ell)/(h-\ell)$. The weights are proportional to the share of observations in each subsample $(n_h + n_\ell)^2$ and the variance of the dose in each subsample. The variance of the dose equals the relative size of the two groups $(n_{h\ell}(1-n_{h\ell}))$, and the square of the distance between their doses $(h-\ell)^2$.

Figure 9: **Baseline Case Decomposition: Two-Way Fixed Effects Estimator as a Weighted Average of Wald-DiDs**

way researchers read a scatter plot between $m_\Delta(d)$ and d: each $\left(\frac{m_\Delta(h)-m_\Delta(l)}{k-l}\right)$ is the slope of a line connecting two points and large groups and groups with very different doses (i.e., far apart on the x-axis) have the most influence on the slope.

Using the same simulated data as in Appendix E, Figure 9 represents the decomposition result for β^{twfe} in a different way by plotting each 2×2 Wald-DiD against its weight as in Figure 6 of Goodman-Bacon (2021). Comparisons between each treated group and the untreated group are in black circles and comparisons between two treated groups are in gray triangles. With K non-zero doses and some untreated units there are $(K+1)K/2$ Wald-DiD comparisons in Proposition C.1. With 18 non-zero doses our example has 171 Wald-DiD terms. Because the untreated group is so large (a quarter of the sample), comparisons to the untreated group get about half the weight in this example even though there are just 18 of them, one for each observed dose.

Next, we consider a decomposition that is based on $(m_\Delta(d) - m_\Delta(0))/d$. Under, for example, Assumption 5, this expression is equal to $ATE(d)/d$ which is an alternative way to define an average causal response.

Proposition C.2. *Consider β^{twfe} in Equation (1.1) and suppose that Assumption 1 holds.*

(1) If Assumption 2(a) holds, then

$$\beta^{twfe} = \int_{\mathcal{D}_+} w_1^{alt\text{-}acr,cont}(l) \frac{(m_\Delta(l) - m_\Delta(0))}{l}\, dl$$

where

$$w_1^{\textit{alt-acr,cont}}(l) = \frac{(l - \mathbb{E}[D])l}{\mathrm{Var}(D)} f_D(l)$$

In addition, $\int_{\mathcal{D}} w^{\textit{alt-acr,cont}}(l)\, dl = 1$*, but* $w^{\textit{alt-acr,cont}}(l)$ *can be negative for some values of* $l \in \mathcal{D}$*.*

(2) If Assumption 2(b) holds, then

$$\beta^{twfe} = \sum_{l \in \mathcal{D}_+} w^{\textit{alt-acr,disc}} \frac{(m_\Delta(l) - m_\Delta(0))}{l}$$

where

$$w^{\textit{alt-acr,disc}}(l) = \frac{(l - \mathbb{E}[D])l}{\mathrm{Var}(D)} p_l^D$$

In addition, $\sum_{l \in \mathcal{D}_+} w^{\textit{alt-acr,disc}}(l) = 1$*, but* $w^{\textit{alt-acr,disc}}$ *can be negative for some values of* $l \in \mathcal{D}$*.*

Proof. From the proof of Theorem 3.4, we have that

$$\begin{aligned}
\beta^{twfe} &= \frac{\mathbb{P}(D > 0)}{\mathrm{Var}(D)} \mathbb{E}[(D - \mathbb{E}[D])(m_\Delta(D) - m_\Delta(0))|D > 0] \\
&= \frac{\mathbb{P}(D > 0)}{\mathrm{Var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])(m_\Delta(l) - m_\Delta(0))\, dF_{D|D>0}(l) \\
&= \frac{\mathbb{P}(D > 0)}{\mathrm{Var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])l \frac{(m_\Delta(l) - m_\Delta(0))}{l}\, dF_{D|D>0}(l) \\
&= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])l \frac{(m_\Delta(l) - m_\Delta(0))}{l} f_D(l)\, dl \\
&= \int_{\mathcal{D}_+} w^{\text{alt-acr,cont}}(l) \frac{(m_\Delta(l) - m_\Delta(0))}{l}\, dl
\end{aligned}$$

where the second equality holds by writing the expectation as an integral, the third equality holds by multiplying and dividing by l, the fourth equality holds under Assumption 2, and the last equality holds by the definition of $w^{\text{alt-acr,cont}}$.

For part (2), the first three equalities above continue to hold. The fourth equality replaces the integral with a summation and $f_D(l)$ with p_l^D; then the result holds by the definition of $w^{\text{alt-acr,disc}}$.

In both cases, the weights can be negative because it is possible that $l < \mathbb{E}[D]$ for some values of $l \in \mathcal{D}_+$. That the weights integrate to 1 holds because

$$\begin{aligned}
\int_{\mathcal{D}_+} w^{\text{alt-acr,cont}}(l)\, dl &= \left(\int_{\mathcal{D}_+} (l - \mathbb{E}[D])l\, dF_D(l) + (0 - \mathbb{E}[D])0P(D = 0) \right) \Big/ \mathrm{Var}(D) \\
&= \left(\int_{\mathcal{D}} (l - \mathbb{E}[D])l\, dF_D(l) \right) \Big/ \mathrm{Var}(D) \\
&= 1
\end{aligned}$$

An analogous argument applies for $w^{\text{alt-acr,disc}}$. □

Finally, we provide a decomposition in terms of levels of paths of outcomes: $m_\Delta(d) - m_\Delta(0)$.

Proposition C.3. *Consider* β^{twfe} *in Equation (1.1) and suppose that Assumption 1 holds.*

(1) If Assumption 2(a) holds, then

$$\beta^{twfe} = \int_{\mathcal{D}} w^{levels,cont}(l)(m_\Delta(l) - m_\Delta(0))\, dl$$

where

$$w^{levels,cont}(l) = \frac{(l - \mathbb{E}[D])}{\mathrm{Var}(D)} f_D(l)$$

In addition, $\int_{\mathcal{D}} w^{levels,cont}(l)\, dl = 0$*, and* $w^{levels,cont}(l)$ *can be negative for some values of* $l \in \mathcal{D}$*.*

(2) If Assumption 2(b) holds, then

$$\beta^{twfe} = \sum_{l \in \mathcal{D}_+} w^{levels,disc}(m_\Delta(l) - m_\Delta(0))$$

where

$$w^{levels,disc}(l) = \frac{(l - \mathbb{E}[D])}{\mathrm{Var}(D)} p_l^D$$

In addition, $\sum_{l \in \mathcal{D}_+} w^{levels,disc}(l) = 0$*, but* $w^{levels,disc}$ *can be negative for some values of* $l \in \mathcal{D}$*.*

Proof. From the proof of Theorem 3.4, we have that

$$\begin{aligned}
\beta^{twfe} &= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \mathbb{E}[(D - \mathbb{E}[D])(m_\Delta(D) - m_\Delta(0))|D>0] \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])(m_\Delta(l) - m_\Delta(0))\, dF_{D|D>0}(l) \\
&= \frac{1}{\mathrm{Var}(D)} \int_{\mathcal{D}_+} (l - \mathbb{E}[D])(m_\Delta(l) - m_\Delta(0)) f_D(l)\, dl \\
&= \int_{\mathcal{D}_+} w^{levels,cont}(l)(m_\Delta(l) - m_\Delta(0))\, dl
\end{aligned}$$

where the second equality holds by writing the expectation as an integral, the third equality holds under Assumption 2, and the last equality holds by the definition of $w^{levels,cont}$.

For part (2), the first two equalities above continue to hold. For the third equality, replace the integral with a summation and $f_D(l)$ with p_l^D; then the result holds by the definition of $w^{levels,disc}$.

In both cases, the weights can be negative since l can be less than $\mathbb{E}[D]$. That the weights integrate to 0 holds because

$$\begin{aligned}
\int_{\mathcal{D}_+} w^{levels,cont}(l)\, dl &= \left(\int_{\mathcal{D}_+} (l - \mathbb{E}[D])\, dF_D(l) + (0 - \mathbb{E}[D]) 0 \mathbb{P}(D=0) \right) \Big/ \mathrm{Var}(D) \\
&= \left(\int_{\mathcal{D}} (l - \mathbb{E}[D])\, dF_D(l) \right) \Big/ \mathrm{Var}(D) \\
&= (\mathbb{E}[D] - \mathbb{E}[D]) / \mathrm{Var}(D) \\
&= 0
\end{aligned}$$

An analogous argument applies for $w^{levels,disc}$. □

Proposition C.3 suggests that it would be inappropriate to interpret β^{twfe} as approximating the level effect of the dose.

D Additional Details for Multiple Periods and Variation in Treatment Timing

In this section, we consider alternative identifying assumptions for treatment effect parameters of interest in the case with multiple periods, variation in treatment timing, and where the dose can vary across units.

As in the baseline case with two time periods, identifying ATT parameters involves untreated groups that serve as a valid counterfactual for treated groups. We first define a parallel trend assumption similar to Assumption 4 whose parts correspond to different comparison groups and time periods where one may believe that parallel trends in untreated potential outcomes holds.

Assumption 4-MP (Parallel Trends with Multiple Periods and Variation in Treatment Timing)**.**

(a) For all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(b) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(c) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|G = k]$ for all groups $k \in \mathcal{G}$ such that $t < k$ (i.e., pre-treatment periods for group k).

Assumption 4-MP(a) is the strongest assumption about paths of untreated potential outcomes. It says that paths of untreated potential outcomes are the same for all groups and for all doses across all time periods. Assumption 4-MP(b) says that the path of outcomes for group g in post treatment time periods is the same as the path of untreated potential outcomes among never-treated units. Parallel pre-trends need not hold under part (b). Assumption 4-MP(c) says that the path of outcomes for group g in post treatment time periods is the same as the path of outcomes among all groups that are not treated yet in that period — this includes both the untreated group as well as groups that will eventually be treated but that are not treated yet. Based on the results in earlier sections, note that each parallel trends assumption in Assumption 4-MP is directed towards identifying $ATT(g, t, d|g, d)$ rather than $ATE(g, t, d)$.

Next, we provide an analogous set of assumptions that target identifying $ATE(g, t, d)$.

Assumption 5-MP-Extended (Strong Parallel Trends with Multiple Periods and Variation in Treatment Timing)**.**

(a) For all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$, and $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(b) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$

(c) For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g, d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|G = k]$ for all groups $k \in \mathcal{G}$ such that $t < k$ (i.e., pre-treatment periods for group k).

Parts (a), (b), and (c) of the assumption correspond to the same parts in Assumption 4-MP and differ based on which group is used as the comparison group in terms of untreated potential outcomes. Part (a) additionally corresponds to Assumption 5-MP in the main text. Finally, the reason that there are two parts to these assumptions rather than just one as in Assumption 4-MP is that, in the setup of this section, conditional on being in group g with $t \geq g$, there are no untreated units in the group; thus, the second part of the assumption handles untreated potential outcome slightly differently than treated potential outcomes.

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

(1a) If, in addition, either Assumption 4-MP(a) or (c) holds, then

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

(1b) If, in addition, Assumption 4-MP(b) holds, then

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

(2a) If, in addition, either Assumption 5-MP-Extended(a) or (c) holds, then

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

(2b) If, in addition, Assumption 5-MP-Extended(b) holds, then

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

The proof of Theorem D.1 is provided in Appendix G. Part (1a) of Theorem D.1 says that $ATT(g,t,d|g,d)$ — the average effect of participating in the treatment in time period t among units who became treated in period g and experienced dose d — is identified under a parallel trends assumption and that it is equal to the average path of outcomes experienced by units in group g under dose d adjusted by the average path of outcomes experienced among units that are not-yet-treated by period t. The results in the other parts are similar as well. For part (1b), the weaker parallel trends assumption in Assumption 4-MP(b) implies that the never-treated group should be used as the comparison group (this is a smaller comparison group relative to the not-yet-treated group). Parts (2a) and (2b) show that under Assumption 5-MP-Extended the same estimands identify $ATE(g,t,d)$.

Remark 5. *The parallel trends assumptions in Assumption 4-MP are not the only possible ones. Interestingly, with a multi-valued/continuous treatment, there are some possible (and reasonable) comparison groups that are available that are not available with a binary treatment. For example, one could assume that*

For all $g \in \mathcal{G}$, $t = g, \ldots, \mathcal{T}$, $d \in \mathcal{D}$, $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|G = k, D = d]$ for all groups $k \in \mathcal{G}$ such that $t < k$ (i.e., pre-treatment periods for group k).

This sort of assumption amounts to using as a comparison group the set of units that are not yet treated but will eventually experience the same dose. It is straightforward to adapt the approach described in Theorem D.1 to this sort of case and propose related estimators that can deliver consistent estimates of $ATT(g,t,d|g,d)$.

Remark 6. *If a researcher is interested in targeting a particular $ATT(g,t,d|g,d)$ or $ATE(g,t,d)$, it is generally possible to weaken Assumption 4-MP or 5-MP-Extended. For example, one could make parallel trends directly about long differences, $(Y_t - Y_{g-1})$, rather than all short differences (this sort of assumption is generally weaker), or, in part (c) of each assumption, use more aggregated comparison groups instead of imposing parallel trends for all possible comparison groups (which is also weaker), or alternatively only make parallel trends assumptions for the particular dose being considered.*

E Multiple Periods and Variation in Treatment Timing and Dose

DiD applications often use more than two time periods in which case treatments, whether binary or not, can turn on at different times for different units. This section extends the results from the previous sections to allow for multiple time periods ($t = 1, ..., \mathcal{T}$) with variation in the time when units become treated ($G = g \in \mathcal{G}$). By convention, we set $G = \mathcal{T} + 1$ for units that remain untreated across all time periods, and we exclude units that are treated in the first period so that $\mathcal{G} \subseteq \{2, \ldots, \mathcal{T} + 1\}$.[34] Treated units receive dose $D = d \in \mathcal{D}$. We also focus on the case that treatment is an absorbing state (or where units do not "forget" their treatment experience).

In this section, we somewhat modify the potential outcomes notation from the previous section to allow for variation in treatment timing. For each unit, we define potential outcomes $Y_{it}(g, d)$ indexed by both treatment timing and dose. Note that treated potential outcomes at time t depend on when a unit first becomes treated—i.e., $Y_{it}(g, d)$ may not equal $Y_{it}(g', d)$ for $g \neq g'$—which allows for general treatment effect dynamics. $Y_{it}(\mathcal{T} + 1, 0)$ is the outcome that unit i would experience if they did not participate in the treatment in any period. For simplicity, we define $Y_{it}(0) = Y_{it}(\mathcal{T} + 1, 0)$ and refer to this as a unit's untreated potential outcome.[35] We also define the variable $W_{it} = D_i \mathbf{1}\{t \geq G_i\}$ which is the amount of dose that unit i experiences in time period t; $W_{it} = 0$ 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* $\{Y_{i1}, \ldots, Y_{i\mathcal{T}}, D_i, G_i\}_{i=1}^n$ *which is independent and identically distributed.*

Assumption 2-MP (Support)**.** *(a) The support of* D, $\mathcal{D} = \{0\} \cup \mathcal{D}_+$. *In addition,* $\mathbb{P}(D = 0) > 0$ *and* $dF_{D|G}(d|g) > 0$ *for all* $(g, d) \in (\mathcal{G} \setminus \{\mathcal{T} + 1\}) \times \mathcal{D}_+$.

(b) $\mathcal{D}_+ = [d_L, d_U]$ *with* $0 < d_L < d_U < \infty$.

(c) For all $g \in (\mathcal{G} \setminus \{\mathcal{T} + 1\})$ *and* $t = 2, \ldots, \mathcal{T}$, $\mathbb{E}[\Delta Y_t | G = g, D = d]$ *is continuously differentiable in* d *on* $\mathcal{D}_+$.

Assumption 3-MP (No Anticipation / Staggered Adoption)**.** *(a) For all* $g \in \mathcal{G}$ *and* $t = 1, \ldots, \mathcal{T}$ *with* $t < g$ *(i.e., in pre-treatment periods),* $Y_{it}(g, d) = Y_{it}(0)$.

(b) $W_{i1} = 0$ *almost surely and for* $t = 2, \ldots, \mathcal{T}$, $W_{it-1} = d$ *implies that* $W_{it} = d$.

Assumption 1-MP says that we have access to $\mathcal{T}$ periods of panel data and observe each unit's dose and treatment timing. Assumption 2-MP extends our definitions of the support of D to the case with multiple periods and variation in treatment timing. As in earlier sections, many of our identification results only require part (a) (which allows for very general treatment regimes) while some of our results are specialized to the continuous case as in parts (b) and (c).[36] Assumption 2-MP also imposes a kind of common support of the dose across timing groups, though it allows for the distribution of the dose to vary across timing groups in otherwise unrestricted ways; that

[34] We could alternatively use $G = \infty$ for units that remain untreated across all time periods.

[35] The analysis in this section could be extended to allow for units to be "treated" at time g but with $d = 0$. For example, units may live in a jurisdiction that implements a program at time g for which they are not eligible. Similarly, we could allow for units to have dose d but remain untreated $g = \mathcal{T} + 1$. This would make sense if a program's dose was based on a known formula so that it was possible to observe d even for units not actually selected for treatment.

[36] For the results in this section that are specialized to the case where the treatment is continuous, it is straightforward to adjust them to allow for a multi-valued discrete treatment along the same lines as in the previous section.

said, it appears to be straightforward to relax this part of the assumption at the cost of additional notation.

Assumption 3-MP(a) rules out that units anticipate experiencing the treatment in ways that affect their outcomes before they actually participate in the treatment. It would be relatively straightforward to extend our arguments in this section to allow for anticipation along the lines of Callaway and Sant'Anna (2020) (in the case of a binary treatment). Assumption 3-MP(b) implies that we consider the case with staggered adoption which means that once units become treated with dose d they remain treated with dose d in all subsequent periods. This allows us to fully categorize a unit by the timing of their treatment adoption and the amount of dose that they experience.

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

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

In other words, we observe a unit's untreated potential outcomes in time periods before they participate in the treatment, and we observe treated potential outcomes in post-treatment time periods that can depend on the timing of the treatment and the amount of the dose.

E.1 Parameters of Interest 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. The average treatment effect parameters of dose d, for group g, in time period t are:

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

$ATT(g,t,d|g,d)$ is the average effect of dose d, for timing group g, in time period t, among units in group g that experienced dose d. $ATE(g,t,d)$ is the average effect of dose d among all units in timing group g (not all units in the population though), in time period t. $ATT(g,t,d|g,d)$ and $ATE(g,t,d)$ are similar to the group-time average treatment effects discussed in Callaway and Sant'Anna (2020) except they are also specific to a dose, and allow for the effect of dose to vary arbitrarily across timing groups and time periods.

ACR 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 the ACR parameters are:

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

$$ACR(g,t,d) = \frac{\partial \mathbb{E}\left[Y_t(g,d)|G = g\right]}{\partial d}.$$

For discrete treatments the ACR parameters are:

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

The two parameters—$ACRT(g,t,d|g,d)$ and $ACR(g,t,d)$—correspond to $ATT(g,t,d|g,d)$ and $ATE(g,t,d)$ in that they are either local to a specific timing group and dose or refer to the entire population.

In many applications, $ACR(g,t,d)$ is relatively high-dimensional and challenging to report. There are a number of possible aggregations that reduce dimensionality and result in parameters that are easier to interpret. We focus on aggregations into an overall causal response across doses, timing groups, and treated periods, as well as into an event study; see Callaway and Sant'Anna (2020) for additional possible aggregations along these lines. Also, note that the aggregations considered below are identified if $ACR(g,t,d)$'s are identified.

To start with, we define an overall causal response of experiencing dose d, for timing group g, across all post-treatment time periods

$$ACR^{group}(g,d) = \frac{1}{\mathcal{T} - g + 1} \sum_{t=g}^{\mathcal{T}} ACR(g,t,d).$$

These can be further aggregated by averaging across timing groups,

$$ACR^{overall}(d) = \sum_{g \in \mathcal{G}} ACR^{group}(g,d) P(G = g | G \leq \mathcal{T}, D = d)$$

$ACR^{overall}(d)$ is the average causal response of dose d across all timing groups that participate in the treatment in any period. It averages $ACR(g,t,d)$ across all observed doses, groups, and treated periods (in other words, all doses at each event-time and then across all event-times). This is a natural analogue of $ACR(d)$ in the two period case.

Another further aggregation is to average across the distribution of the dose (of all timing groups that participate in the treatment)

$$ACR^{*,mp} = \mathbb{E}\left[ACR^{overall}(D) | G \leq \mathcal{T}\right].$$

$ACR^{*,mp}$ is the overall average causal response (averaged across doses and and over all timing groups that participate in the treatment in any time period). $ACR^{*,mp}$ is a single number; it is the analogue of ACR^* from the two period case and is arguably a natural target parameter for a TWFE regression.

Next, we consider an event study type of aggregation.

$$ACR^{es}(e,d) = \sum_{g \in \mathcal{G}} \mathbf{1}\{g + e \leq \mathcal{T}\} ACR(g, g+e, d) \mathbb{P}(G = g | G + e \leq \mathcal{T}, D = d).$$

$ACR^{es}(e,d)$ is the average causal response of dose d among units that have been exposed to the treatment for exactly e periods. This can be further aggregated across the distribution of the dose

$$ACR(e) = \mathbb{E}[ACR^{es}(e,D) | G \leq \mathcal{T}],$$

which is the average partial effect (averaged across all doses) among units that have been exposed to the treatment for exactly e periods. Importantly, this keeps the distribution of the dose constant across different lengths of exposure to the treatment; the distribution of the dose is set to be equal to the distribution of the dose among the group of units that ever participate in the treatment. For values of $e \geq 0$, $ACR(e)$ can be interpreted as dynamic effects; but it is also interesting to consider cases where $e < 0$ which can be interpreted as a pre-test of the parallel trends assumption.

Remark 7. *The aggregations above are related to ACR(g,t,d), but similar arguments would apply to other parameters discussed in the paper including $ATT(g,t,d|g,d)$, $ATE(g,t,d)$, and $ACRT(g,t,d|g,d)$.*

E.2 Identification with a Continuous Treatment and Staggered Timing

With multiple time periods and variation in treatment timing, there are several possible versions of parallel trends and strong parallel trends assumptions that one could make because there are many ways to compare groups with different changes in their dose over time.

We focus on a version of strong parallel trends in this section and we provide a number of alternative parallel trends assumptions (and corresponding identification results) in Appendix D.

Assumption 5-MP (Strong Parallel Trends with Multiple Periods and Variation in Treatment Timing)**.** *For all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$, and $d \in \mathcal{D}$, $\mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G = g, D = d] = \mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G = g]$ and $\mathbb{E}[\Delta Y_t(0)|G = g, D = d] = \mathbb{E}[\Delta Y_t(0)|D = 0]$.*

Assumption 5-MP is an extension of Assumption 5 to the case with multiple time periods. In particular, it restricts paths of treated potential outcomes (not just paths of untreated potential outcomes) so that all dose groups treated at time g would have had the same path of potential outcomes at every dose.

Theorem E.1. *Under Assumptions 1-MP, 2-MP(a), 3-MP, and 5-MP, and for all $g \in \mathcal{G}$, $t = 2, \ldots, \mathcal{T}$ such that $t \geq g$, and for all $d \in \mathcal{D}_+$.*

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

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

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

The proof of Theorem E.1 is provided in Appendix F. The result is broadly similar to the one in the case with two periods. It says that $ATE(g,t,d)$ 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. Relative to the case with two time periods, the main difference is that the "pre-period" is $g-1$. The reason to use the base period $g-1$ is that this 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.

Remark 8. *Theorem E.1 identifies $ATE(g,t,d)$ and $ACR(g,t,d)$ under a version of strong parallel trends. In Appendix D, we discuss identifying $ATT(g,t,d|g,d)$ and $ACRT(g,t,d|g,d)$ under a version of parallel trends that only involves untreated potential outcomes; in this case, like in the two period case, $ATT(g,t,d|g,d)$ is identified, comparisons of $ATT(g,t,d|g,d)$ across different values of d do not deliver a causal effect of moving from one dose to another (as they additionally include "selection bias" terms), and derivative of paths of outcomes over time do not recover $ACRT(g,t,d|g,d)$ due to the same "selection bias" terms.*

Remark 9. *It is natural to estimate $ATE(g,t,d)$ by simply replacing the population averages in Theorem E.1 by their sample counterpart. This approach is very simple and intuitive, but in some cases, it may be possible to develop more efficient estimators using GMM. See the discussion in Marcus and Sant'Anna (2021) in the context of a binary treatment. When treatment d is continuous, some smoothing is required. However, one can use off-the-shelf standard nonparametric estimations procedures based on kernels or sieves to estimate these target causal parameters.*

E.3 TWFE estimators with multiple time periods and variation in treatment timing

In applications with multiple periods and variation in treatment timing, empirical researchers almost always estimate a TWFE regression

$$Y_{it} = \theta_t + \eta_i + \beta^{twfe} W_{it} + v_{it}. \tag{E.1}$$

Equation (E.1) is exactly the same as the TWFE regression in the baseline case with two periods in Equation (1.1) only with the notation slightly adjusted to match this section. The results in this section generalize the results in several recent papers on TWFE estimates including Goodman-Bacon (2021) and de Chaisemartin and D'Haultfœuille (2020) to our DiD setup with variation in treatment intensity.

To start with, write population versions of TWFE adjusted variables by

$$\ddot{W}_{it} = (W_{it} - \bar{W}_i) - \left(\mathbb{E}[W_t] - \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[W_t]\right), \quad \text{where} \quad \bar{W}_i = \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} W_{it}.$$

The population version of the TWFE estimator is

$$\beta^{twfe} = \frac{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[Y_{it}\ddot{W}_{it}]}{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2]}. \tag{E.2}$$

As in the previous section, we present both a "mechanical" decomposition of the TWFE estimator and a "causal" decomposition of the estimand that relates assumptions to interpretation.

In order to define these decompositions, we introduce a bit of new notation. First, define the fraction of periods that units in group g spends treated as

$$\bar{G}_g = \frac{\mathcal{T} - (g-1)}{\mathcal{T}}.$$

For the untreated group $g = \mathcal{T} + 1$ so that $\bar{G}_{\mathcal{T}+1} = 0$.

Next, we define time periods over which averages are taken. For averaging variables across time periods, we use the following notation, for $t_1 \leq t_2$,

$$\bar{Y}_i^{(t_1,t_2)} = \frac{1}{t_2 - t_1 + 1}\sum_{t=t_1}^{t_2} Y_{it}.$$

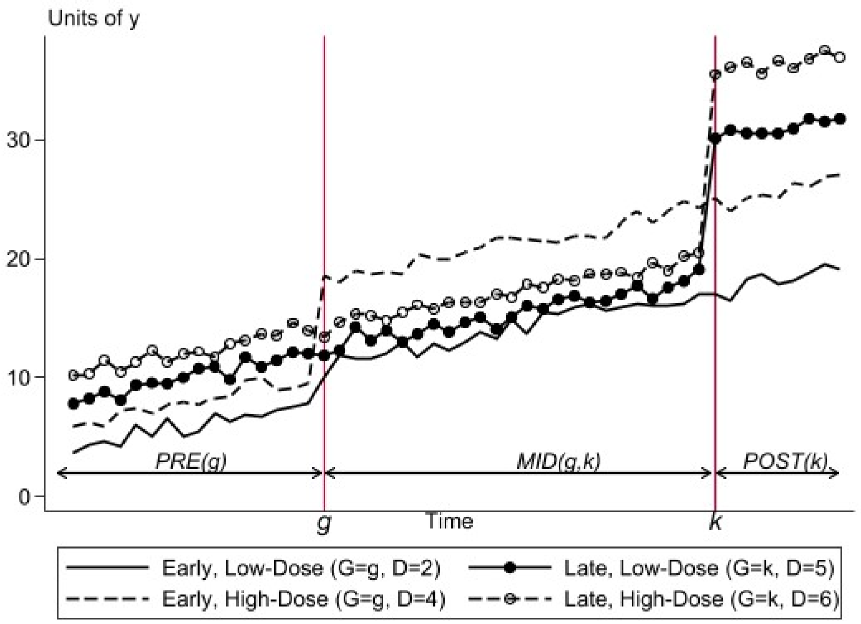
It is also convenient to define some particular averages across time periods. For two time periods g and k, with $k > g$, (below, g and k will often index groups defined by treatment timing), we define

$$\bar{Y}_i^{PRE(g)} = \bar{Y}_i^{(1,g-1)}, \quad \bar{Y}_i^{MID(g,k)} = \bar{Y}_i^{(g,k-1)}, \quad \bar{Y}_i^{POST(k)} = \bar{Y}_i^{(k,\mathcal{T})}.$$

$\bar{Y}_i^{PRE(g)}$ is the average outcome for unit i in periods 1 to $g-1$, $\bar{Y}_i^{MID(g,k)}$ is the average outcome for unit i in periods g to $k-1$, and $\bar{Y}_i^{POST(g,k)}$ is the average outcome for unit i in periods k to $\mathcal{T}$. Below, when g and k index groups, $\bar{Y}_i^{PRE(g)}$ is the average outcome for unit i in periods before units in either group are treated, $\bar{Y}_i^{MID(g,k)}$ is the average outcome for unit i in periods after group g has become treated but before group k has been treated, and $\bar{Y}_i^{POST(k)}$ is the average outcome for unit i after both groups have become treated.

To fix ideas about how the staggered-timing/continuous treatment case works, consider a setup with two timing groups, g and k with $k > g$. Some units in the "early -treated" group have $d = 2$ and others have $d = 4$. Some units in the late treated group have $d = 5$ and others have $d = 6$. Thus, the four groups are early-treated/high-dose, early-treated/low-dose, late-treated/high-dose, and late-treated/low-dose. Figure 10 plots constructed outcomes for these groups with a treatment effect that is a one-time shift equal to $d^{1.5}$.

Following Goodman-Bacon (2021), we motivate the decomposition of the TWFE estimand by



Notes: The figure plots simulated data for four groups: early-treated/high-dose, early-treated/low-dose, late-treated/high-dose, and late-treated/low-dose.

Figure 10: A Simple Set-Up with Staggered Timing and Variation in the Dose

considering the four types of simple DiD estimands that can be formed using only one source of variation. The first comparison is a within-group comparison of paths of outcomes among units that experienced different amounts of the treatment.

$$\delta^{WITHIN}(g) = \frac{\text{Cov}(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G=g)}{\text{Var}(D|G=g)}. \tag{E.3}$$

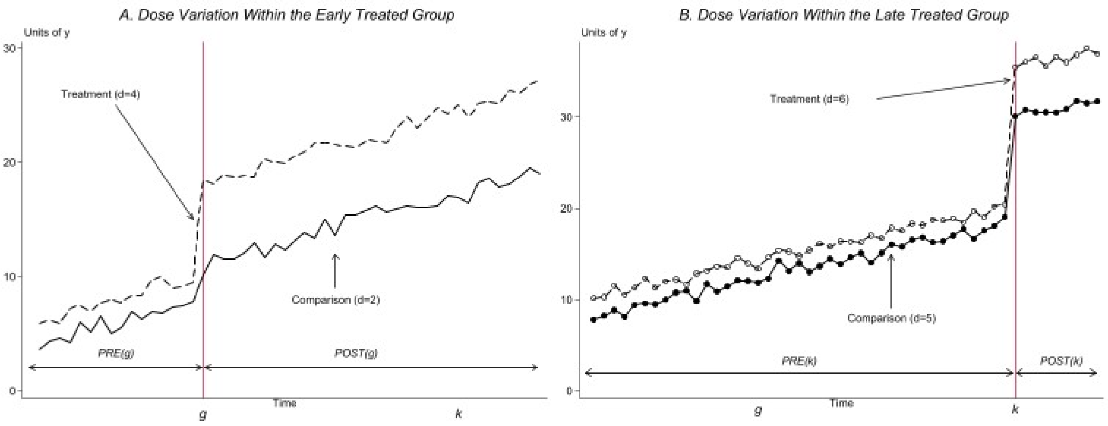
This term is essentially the same as the expression for the TWFE estimand in the baseline two-period case. It equals the OLS (population) coefficient from regressing the change in average outcomes before and after g for units treated at time g on their dose, d. Figure 11 uses the four-group example to show how $\delta^{WITHIN}(g)$ and $\delta^{WITHIN}(k)$ use higher-dose units as the "treatment group" and lower-dose units as the "comparison group".

The second comparison is based on treatment timing. It compares paths of outcomes between a particular timing group g and a "later-treated" group k (i.e., $k > g$) in the periods after group g is treated but before group k becomes treated relative to their common pre-treatment periods.[37]

$$\delta^{MID,PRE}(g,k) = \frac{\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=k\right]}{\mathbb{E}[D|G=g]}. \tag{E.4}$$

Panel A of Figure 12 plots the outcomes used in this comparison with timing-group averages in black and the specific dose groups from Figure 10 in light gray. Under a parallel trends assumption, we show below that this term corresponds to a reasonable treatment effect parameter because the

[37] Each of the following expressions also includes a term in the denominator. Below, this term is useful for interpreting differences across groups as partial effects of more treatment, but, for now, we largely ignore the expressions in the denominator.



Notes: The figure shows the within-timing group comparison between higher- and lower-dose units defined by $\delta^{WITHIN}(g)$ and $\delta^{WITHIN}(k)$.

Figure 11: Within-Timing-Group Comparisons Across Doses

path of outcomes for group k (which is still in its pre-treatment period here) is what the path of outcomes would have been for group g if it had not been treated. Also note that this term encompasses comparisons of group g to the "never-treated" group.

The third comparison is between paths of outcomes for the "later-treated" group k in its post-treatment period relative to a pre-treatment period adjusted by the same path of outcomes for the "early -treated" group g.

$$\delta^{POST,MID}(g,k) = \frac{\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G=k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G=g\right]}{\mathbb{E}[D|G=k]}. \tag{E.5}$$

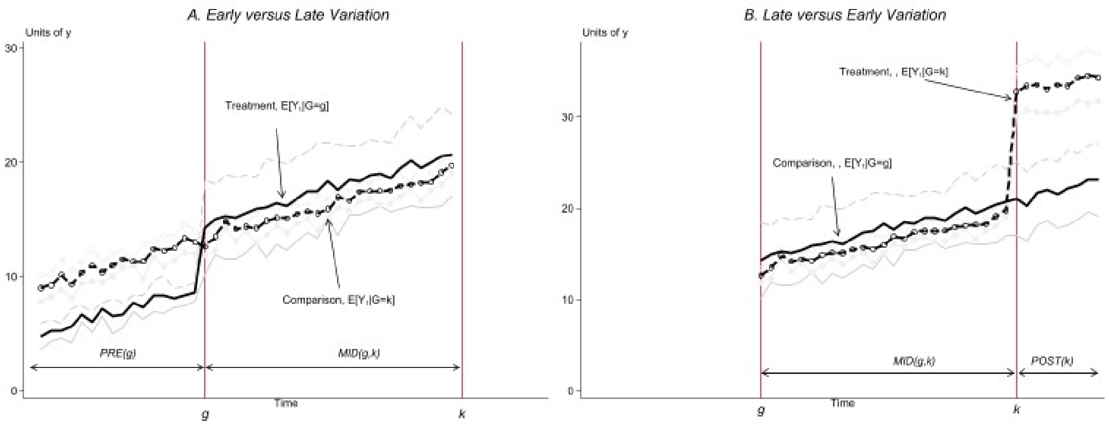
These terms use the already-treated group g as the comparison group for group k. Panel B of Figure 12 plots the outcomes used in this term. Mechanically, the TWFE regression exploits this comparison because group g's treatment status/amount is not changing over these time periods. However, these are post-treatment periods for group g and parallel trends assumptions do not place restrictions on paths of post-treatment outcomes, which are subtracted in Equation (E.5). Below we discuss assumptions about treatment effect heterogeneity over time that are necessary to deal with this issue.[38]

The final comparison that shows up in the TWFE estimator is between paths of outcomes between "early" and "late" treated groups in their common post-treatment periods relative to their common pre-treatment periods. In other words, this comparison only uses periods in which treatment status differs and focusing only on the "endpoints" where the two timing groups are either both untreated or both treated with potentially different average doses.

$$\delta^{POST,PRE}(g,k) = \frac{\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=k\right]}{\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k]}. \tag{E.6}$$

Figure 13 shows the outcomes that determine the comparisons that show up in this term. The reason that this term shows up in β^{twfe} is that differences in the paths of outcomes between groups

[38] This sort of comparison also shows up in the case with a binary, staggered treatment. See, e.g., Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021).



Notes: The figure shows the between-timing-group comparisons that average the outcomes in groups g and k across dose levels and compare the early group to the later group (panel C) or the later group to the early group (panel D).

Figure 12: Between-Timing-Group Comparisons

that have different distributions of the treatment are informative about β^{twfe}. For example, if more dose tends to increase outcomes and group g's dose is higher on average than group k's, then outcomes may increase more among group g than group k resulting in $\delta^{POST,PRE}(g,k)$ not being equal to 0.[39]

Next, we show how β^{twfe} weights these simple DiD terms together and discuss its theoretical interpretation under a parallel trends assumptions. To characterize the weights, first, define

$$p_{g|\{g,k\}} = P(G = g|G \in \{g,k\}),$$

which is the probability of being in group g conditional on being in either group g or k. We also define the following weights, which measure the variance of the treatment variable used to estimate each of the simple DiD terms in equations Equations (E.3) to (E.6).

$$w^{g,within}(g) = \text{Var}(D|G=g)(1-\bar{G}_g)\bar{G}_g p_g \Bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2],$$

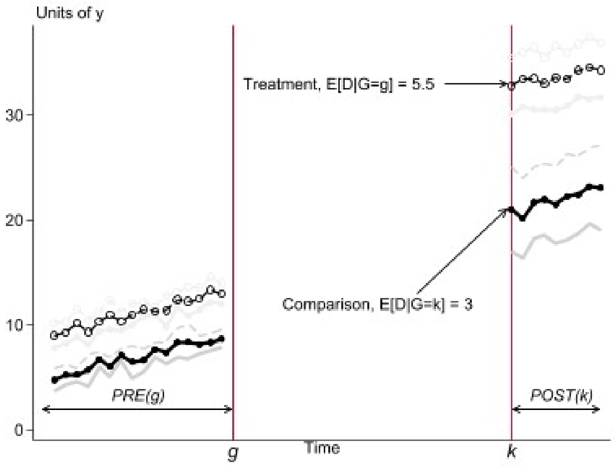
$$w^{g,post}(g,k) = \mathbb{E}[D|G=g]^2(1-\bar{G}_g)(\bar{G}_g - \bar{G}_k)(p_g+p_k)^2 p_{g|\{g,k\}}(1-p_{g|\{g,k\}}) \Bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2],$$

$$w^{k,post}(g,k) = \mathbb{E}[D|G=k]^2\bar{G}_k(\bar{G}_g - \bar{G}_k)(p_g+p_k)^2 p_{g|\{g,k\}}(1-p_{g|\{g,k\}}) \Bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2],$$

$$w^{long}(g,k) = (\mathbb{E}[D|G=g]-\mathbb{E}[D|G=k])^2\bar{G}_k(1-\bar{G}_g)(p_g+p_k)^2 p_{g|\{g,k\}}(1-p_{g|\{g,k\}}) \Bigg/ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2].$$

These weights are similar to the ones in Goodman-Bacon (2021) in the sense that they combine the size of the sample and the variance of treatment used to calculate each simple DiD term. In $w^{g,within}(g)$, for example, $\text{Var}(D|G=g)$ measures how much the dose varies across units with

[39] To be more precise, this term involves comparisons between groups g and k for the group with a higher dose on average to the group with a smaller dose on average. When $\mathbb{E}[D|G=g] > \mathbb{E}[D|G=k]$, this corresponds to the expression in Equation (E.6). When $\mathbb{E}[D|G=g] < \mathbb{E}[D|G=k]$, one can multiply both the numerator and denominator by -1 so that we effectively make a positive-weight comparison for the group that experienced more dose relative to the group that experienced less dose.



Notes: The figure shows the comparisons between timing groups in the $POST(k)$ window when both are treated with potentially different average doses and the $PRE(g)$ window when neither group is treated.

Figure 13: Long Comparisons Between Timing Groups

$G = g$, $(1-\bar{G}_g)\bar{G}_g$ measures the variance that comes from timing which falls when g is closer to 0 or $\mathcal{T}$, and p_g measures the share of units with $G = g$ (i.e.,. subsample size). Since they only compare outcomes between timing-groups, $w^{g,post}(g,k)$ and $w^{k,post}(g,k)$ do not contain a within-timing-group variance of D, but they do include $\mathbb{E}[D|G = k]^2$ which reflects the fact that timing groups with higher average doses get more weight. The rest of the timing weights have the same interpretation as in Goodman-Bacon (2021). Finally, $w^{long}(g,k)$ includes the square of the difference in mean doses between groups g and k—$(\mathbb{E}[D|G = g] - \mathbb{E}[D|G = k])^2$—which shows that the "endpoint" comparisons only influence β^{twfe} to the extent that timing groups have different average doses. Two timing groups with the same average dose do not contribute a $\delta^{POST,PRE}(g,k)$ term because there is no differential change in their doses between the $PRE(g)$ window (when both groups are untreated) and the $POST(k)$ window (when both groups have $\mathbb{E}[D|G = g] = E[D|G = k]$).

Our next result combines the simple DiD terms and their variance weights to provide a mechanical decomposition of β^{twfe} in DiD setups with variation in treatment timing and variation in treatment intensity.

Proposition E.1. *Under Assumptions 1-MP, 2-MP(a), and 3-MP, β^{twfe} in Equation (E.1) can be written as*

$$\beta^{twfe} = \sum_{g\in\mathcal{G}} w^{g,within}(g)\delta^{WITHIN}(g)$$
$$+\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{w^{g,post}(g,k)\delta^{MID,PRE}(g,k) + w^{k,post}(g,k)\delta^{POST,MID}(g,k) + w^{long}(g,k)\delta^{POST,PRE}(g,k)\right\}.$$

In addition, (i) $w^{g,within}(g) \geq 0$, $w^{g,post}(g,k) \geq 0$, $w^{k,post}(g,k)$, and $w^{long}(g,k) \geq 0$ for all $g \in \mathcal{G}$ and $k \in \mathcal{G}$ with $k > g$, and (ii) $\sum_{g\in\mathcal{G}} w^{g,within}(g)+\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\left\{w^{g,post(g,k)}(g,k)+w^{k,post}(g,k)+ w^{long}(g,k)\right\} = 1$.

Proposition E.1 generalizes the decomposition theorem for binary staggered timing designs in Goodman-Bacon (2021) to our setup with variation in treatment intensity.[40] Notice that it does

[40] In particular, in the special case of a staggered, binary treatment, $w^{g,within}(g)\delta^{WITHIN}(g) = 0$ (since there is no

not require Assumption 2-MP(b) or (c), and is therefore compatible with a binary, multi-valued, continuous, or mixed treatment. It says that β^{twfe} can be written as a weighted average of the four comparisons in Equations (E.3) to (E.6). These weights are all positive and sum to one.

Proposition E.1 is a new, explicit description of what kinds of comparisons TWFE uses to compute β^{twfe}, but it does not on its own provide guidance on how to interpret TWFE estimates. Our baseline results, for example, show that simple estimators like $\delta^{WITHIN}(g)$ equal averages of $ACRT$ parameters plus "selection bias" that arises from heterogeneous treatment effect functions. Similarly, the terms that compare outcomes across timing groups necessarily average over the dose-specific treatment effects of units within that timing group. We analyze the theoretical interpretation of each of these simple DiD estimand under different assumptions and then discuss what this implies about the (arguably implicit) identifying assumptions and estimand for TWFE.

To begin we define additional weights that apply to the underlying causal parameters in the DiD terms in Equations (E.3) through (E.6):

$$w_1^{within}(g,l) = \frac{\Big(\mathbb{E}[D|G=g, D \geq l] - \mathbb{E}[D|G=g]\Big)}{\text{Var}(D|G=g)}\mathbb{P}(D \geq l|G=g),$$

$$w_1(g,l) = \frac{\mathbb{P}(D \geq l|G=g)}{\mathbb{E}[D|G=g]}, \qquad w_0(g) = \frac{d_L}{\mathbb{E}[D|G=g]},$$

$$w_1^{across}(g,k,l) = \frac{(\mathbb{P}(D \geq l|G=g) - \mathbb{P}(D \geq l|G=k))}{(\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])}, \qquad \tilde{w}_0^{across}(g,k) = \frac{d_L}{(\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])},$$
$$\tilde{w}_1^{across}(g,k,l) = \frac{\mathbb{P}(D \geq l|G=k)}{(\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k])}.$$

In addition, define the following differences in paths of outcomes over time

$$\pi^{POST(\tilde{k}),PRE(\tilde{g})}(g) = \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|G=g\right] - \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|D=0\right],$$
$$\pi^{MID(\tilde{g},\tilde{k}),PRE(\tilde{g})}(g) = \mathbb{E}\left[\big(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|G=g\right] - \mathbb{E}\left[\big(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|D=0\right],$$
$$\pi^{POST(\tilde{k}),MID(\tilde{g},\tilde{k})}(g) = \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})}\big)|G=g\right] - \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})}\big)|D=0\right],$$

and, similarly,

$$\pi_D^{POST(\tilde{k}),PRE(\tilde{g})}(g,d) = \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|G=g, D=d\right] - \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|D=0\right],$$
$$\pi_D^{MID(\tilde{g},\tilde{k}),PRE(\tilde{g})}(g,d) = \mathbb{E}\left[\big(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|G=g, D=d\right] - \mathbb{E}\left[\big(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\big)|D=0\right],$$
$$\pi_D^{POST(\tilde{k}),MID(\tilde{g},\tilde{k})}(g,d) = \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})}\big)|G=g, D=d\right] - \mathbb{E}\left[\big(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})}\big)|D=0\right],$$

which are the same paths of outcomes but conditional on having dose d.

The following result is our main result on interpreting TWFE estimates with continuous treatment.

Theorem E.2. *Under Assumptions 1-MP, 2-MP, and 3-MP,*

within group variation in the dose in this case), and $w^{long}(g,k)\delta^{POST,PRE}(g,k) = 0$ (because the distribution of the dose is the same across all groups). Then, Proposition E.1 collapses to Theorem 1 in Goodman-Bacon (2021) because the terms $w^{g,post(g,k)}\delta^{MID,PRE}(g,k)$ and $w^{k,post}(g,k)\delta^{POST,MID}(g,k)$ correspond exactly to between-timing-group comparisons.

(1) The four comparisons in Equations (E.3) to (E.6) can be written as

$$\delta^{WITHIN}(g) = \int_{\mathcal{D}_+} w_1^{within}(g,l)\frac{\partial \pi_D^{POST(g),PRE(g)}(g,l)}{\partial l}\, dl,$$

$$\begin{aligned}\delta^{MID,PRE}(g,k) &= \int_{\mathcal{D}_+} w_1(g,l)\frac{\partial \pi_D^{MID(g,k),PRE(g)}(g,l)}{\partial l}\, dl + w_0(g)\frac{\pi_D^{MID(g,k),PRE(g)}(g,d_L)}{d_L} \\ &\quad - w_0(g)\frac{\pi^{MID(g,k),PRE(g)}(k)}{d_L},\end{aligned}$$

$$\begin{aligned}\delta^{POST,MID}(g,k) &= \int_{\mathcal{D}_+} w_1(k,l)\frac{\partial \pi_D^{POST(k),MID(g,k)}(k,l)}{\partial l}\, dl + w_0(k)\frac{\pi^{POST(k),MID(g,k)}(k,d_L)}{d_L} \\ &\quad - w_0(k)\left(\frac{\pi^{POST(k),PRE(g)}(g) - \pi^{MID(g,k),PRE(g)}(g)}{d_L}\right),\end{aligned}$$

$$\begin{aligned}\delta^{POST,PRE}(g,k) &= \int_{\mathcal{D}_+} w_1^{across}(g,k,l)\frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l}\, dl \\ &\quad - \left\{\int_{\mathcal{D}_+} \tilde{w}_1^{across}(g,k,l)\left(\frac{\partial \pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l}\right) dl \right. \\ &\qquad \left. + \tilde{w}_0^{across}(g,k)\left(\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L}\right)\right\}.\end{aligned}$$

(2) If, in addition, Assumption 5-MP holds, then

$$\delta^{WITHIN}(g) = \int_{\mathcal{D}_+} w_1^{within}(g,l)\overline{ACR}^{POST(g)}(g,l)dl,$$

$$\delta^{MID,PRE}(g,k) = \int_{\mathcal{D}_+} w_1(g,l)\overline{ACR}^{MID(g,k)}(g,l)\, dl + w_0(g)\frac{\overline{ATE}^{MID(g,k)}(g,d_L)}{d_L},$$

$$\begin{aligned}\delta^{POST,MID}(g,k) &= \int_{\mathcal{D}_+} w_1(k,l)\overline{ACR}^{POST(k)}(k,l)\, dl + w_0(k)\frac{\overline{ATE}^{POST(k)}(k,d_L)}{d_L} \\ &\quad - w_0(k)\left(\frac{\pi^{POST(k),PRE(g)}(g) - \pi^{MID(g,k),PRE(g)}(g)}{d_L}\right),\end{aligned}$$

$$\begin{aligned}\delta^{POST,PRE}(g,k) &= \int_{\mathcal{D}_+} w_1^{across}(g,k,l)\overline{ACR}^{POST(k)}(g,l)\, dl \\ &\quad - \left\{\int_{\mathcal{D}_+} \tilde{w}_1^{across}(g,k,l)\left(\frac{\partial \pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l}\right) dl \right. \\ &\qquad \left. + \tilde{w}_0^{across}(g,k)\left(\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L}\right)\right\}.\end{aligned}$$

In addition, (i) $w_1^{within}(g,d) \geq 0$, $w_1(g,d) \geq 0$, and $w_0(g) \geq 0$, for all $g \in \mathcal{G}$ and $d \in \mathcal{D}_+$ and (ii) $\int_{\mathcal{D}_+} w_1^{within}(g,l)\, dl = 1$, $\int_{\mathcal{D}_+} w_1(g,l)\, dl + w_0(g) = 1$, and $\int_{\mathcal{D}_+} w_1^{across}(g,k,l)\, dl = 1$.

Part (1) of Theorem E.2 links the four sets of comparisons in the TWFE estimator in Proposition E.1 to derivatives of conditional expectations (this is analogous to Theorem 3.4 in the baseline case above) along with some additional (nuisance) paths of outcomes.

Part (2) of Theorem E.2 imposes Assumption 5-MP. Under Assumption 5-MP, $\delta^{WITHIN}(g)$ and $\delta^{MID,PRE}(g,k)$ both deliver weighted averages of ACR-type parameters. However, $\delta^{POST,MID}(g,k)$ and $\delta^{POST,PRE}(g,k)$ still involve non-negligible nuisance terms. Under Assumption 5-MP, the additional term in $\delta^{POST,MID}(g,k)$ involves the difference between treatment effects for group g in group k's post-treatment periods relative to treatment effects for group g in the periods after group g is treated but before group k is treated — that is, treatment effect dynamics. Parallel trends assumptions do not imply that this term is equal to 0. And, in the special case where the treatment is binary, this term corresponds to the "problematic" term related to treatment effect dynamics in Goodman-Bacon (2021).

The additional nuisance term in $\delta^{POST,PRE}(g,k)$ involves differences in partial effects of more treatment across groups in their common post-treatment periods. Parallel trends does not restrict these partial effects to be equal to each other. This term does not show up in the case with a binary treatment because, by construction, the distribution of the dose is the same across groups. It is helpful to further consider where this expression comes from. For simplicity, temporarily suppose that the partial effect of more dose is positive and constant across groups, time, and dose. In this case, if group g has more dose on average than group k, then its outcomes should increase more from group g and k's common pre-treatment period to their common post-treatment period. This is the comparison that shows up in $\delta^{POST,PRE}(g,k)$. However, when partial effects are not the same across groups and times (which is not implied by any parallel trends assumption), then, for example, it could be the case that the partial effect of dose is positive for all groups and time periods but greater for group k relative to group g. If these differences are large enough, it could lead to the cross-group, long-difference comparisons in $\delta^{POST,PRE}(g,k)$ having the opposite sign.

Next, we engage on how one could potentially "rescue" TWFE procedures such that β^{twfe} would always recover a weighted average of reasonable treatment effect parameters. To do so, one must further restrict different types of treatment effect heterogeneity.

Assumption 7. *(a) [No Treatment Effect Dynamics] For all $g \in \mathcal{G} \setminus (\mathcal{T}+1)$ and $t \geq g$ (i.e, post-treatment periods for group g), $ACR(g,t,d)$ and $ATE(g,t,d_L)$ do not vary with t.*

(b) [Homogeneous Causal Responses across Groups] For all $g \in \mathcal{G} \setminus (\mathcal{T}+1)$ with $t \geq g$ and $k \in \mathcal{G} \setminus (\mathcal{T}+1)$ with $t \geq k$, $ACR(g,t,d) = ACR(k,t,d)$ and $ATE(g,t,d_L) = ATE(k,t,d_L)$.

(c) [Homogeneous Causal Responses across Dose] For all $g \in \mathcal{G}\setminus(\mathcal{T}+1)$ with $t \geq g$, $ACR(g,t,d)$ does not vary across d, and, in addition, $ATE(g,t,d_L)/d_L = ACR(g,t,d)$.

Assumption 7 introduces three additional conditions limiting treatment effect heterogeneity. Assumption 7(a) imposes that, within a timing-group, the causal response to the treatment does not vary across time which rules out treatment effect dynamics. Assumption 7(b) imposes that, for a fixed time period, causal responses to the treatment are constant across timing-groups. Assumption 7(c) imposes that, within timing-group and time period, the causal response to more dose is constant across different values of the dose.

Proposition E.2. *Under Assumptions 1-MP, 2-MP, 3-MP, and 5-MP,*

(a) If, in addition, Assumption 7(a) holds, then

$$\delta^{POST,MID}(g,k) = \int_{\mathcal{D}_+} w_1(k,l)\overline{ACR}^{POST(k)}(k,l)\, dl + w_0(k)\frac{\overline{ATE}^{POST(k)}(k,d_L)}{d_L}.$$

(b) If, in addition, Assumption 7(b) holds, then

$$\delta^{POST,PRE}(g,k) = \int_{\mathcal{D}_+} w_1^{across}(g,k,l)\overline{ACR}^{POST(k)}(g,l)\,dl.$$

(c) If, in addition, Assumption 7(a), (b) and (c) hold, then

$$\beta^{twfe} = ACR^{*,mp}.$$

Proposition E.2 provides additional conditions under which the nuisance terms in $\delta^{POST,MID}(g,k)$ and $\delta^{POST,PRE}(g,k)$ are equal to 0. For $\delta^{POST,MID}(g,k)$, these nuisance terms can be eliminated by ruling out treatment effect dynamics; that is, by assuming that, within a particular group, the causal response to more dose does not vary across time. Ruling out these sort of treatment effect dynamics is analogous to the kinds of conditions that are required to interpret TWFE estimates with a binary treatment. In order for the nuisance terms in $\delta^{POST,PRE}(g,k)$ to be equal to 0, we impose homogeneous causal responses across groups — that the causal response to more dose is the same across groups conditional on having the same amount of dose and being in the same time period. Neither of these assumptions are implied by any of the parallel trends assumptions that we have considered, and they are both potentially very strong. Therefore, under both Assumption 7(a) and (b), β^{twfe} is equal to a weighted average of average causal response parameters, but these weights continue to be driven by the TWFE estimation strategy and, like in the baseline two period case, can continue to deliver poor estimates of the overall average causal response to the treatment. Imposing Assumption 7(a), (b), and (c) implies that $ACR(g,t,d)$ does not vary by timing group, time period, or the amount of dose, and part (c) of Proposition E.2 says that β^{twfe} is equal to the overall average causal response under these additional, strong conditions.

Remark 10. *The results in Part (2) of Theorem E.2 and in Proposition E.2 relied on the multi-period version of strong parallel trends in Assumption 5-MP. In Theorem E.2-Extended in Appendix F, we additionally show that, under a multi-period version of standard parallel trends (this is analogous to Assumption 4 in the two period case and details are provided in Assumption 4-MP(a) in Appendix D), similar results hold except that* $\overline{ACR}^{\cdot}(\cdot,d)$ *should be replaced by* $\overline{ACRT}^{\cdot}(\cdot,d|\cdot,d) + \frac{\partial \overline{ATT}^{\cdot}(\cdot,d|\cdot,l)}{\partial l}\Big|_{l=d}$ *where the second term is a "selection bias" term, and* $\overline{ATE}^{\cdot}(\cdot,d_L)$ *should be replaced by* $\overline{ATT}^{\cdot}(\cdot,d_L|\cdot,d_L)$*. This implies that, under a standard version of parallel trends, all four comparisons in Equations* (E.3) *to* (E.6) *include "selection bias" terms.*

E.4 Discussion

The results in this section suggest three important weaknesses of TWFE estimands in a difference-in-differences framework with multiple time periods, and variation in treatment intensity and timing of adoptions. First, like the TWFE estimands considered above in the case with two time periods, TWFE estimands have weights that are driven by the estimation method. As above, these weights may have undesirable properties in setups where treatment effect heterogeneity is the rule rather than the exception.

Second, in addition to reasonable treatment effect parameters, TWFE estimands also include undesirable components due to treatment effect dynamics and heterogeneous causal responses across groups and time periods. That these show up in the TWFE estimand is potentially problematic and can possibly lead to very poor performance of the TWFE estimator. Ruling out these problems requires substantially stronger conditions in addition to any kind of parallel trends assumption.

Finally, even when these extra conditions hold (i.e., the best case scenario for TWFE), if a researcher invokes a standard parallel trends assumption, the TWFE estimand delivers weighted averages of derivatives of ATT-type parameters which are themselves hard to interpret because, like in the two period case, they include both actual causal responses and “selection bias” terms.

Of these three weaknesses, the first two can be completely avoided by using the estimands presented in Theorem E.1. These estimands rely only on parallel trends assumptions; in particular, they are available without imposing any conditions on treatment effect dynamics or how causal responses vary across groups. The third weakness, though, is a more fundamental challenge of difference-in-differences approaches with variation in treatment intensity as comparing treatment effect parameters across different values of the dose appears to fundamentally require imposing stronger assumptions that rule out some forms of selection into different amounts of the treatment. Although undesirable, we are not aware of any other practical solution to this empirically relevant DiD problem. Thus, we urge practitioners to transparently discuss their assumptions, potentially exploiting context-specific knowledge to justify the plausibility of a stronger parallel trends assumption in the given application.

F Proofs

F.1 Proofs of Results in Section 3.2

This section contains the proofs of the results in Section 3.2 on identifying $ATT(d|d)$ and $ATE(d)$ under parallel trends assumptions and with a multi-valued/continuous treatment.

Proof of Theorem 3.1

Proof. To show the result, notice that

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

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t-1}(0)|D = d]$, the third equality holds by Assumption 4, and the last equality holds because $Y_t(d)$ and $Y_{t-1}(0)$ are observed potential outcomes when $D = d$ and $Y_t(0)$ and $Y_{t-1}(0)$ are observed potential outcomes when $D = 0$. □

Proof of Proposition 3.1

Proof. To show the result, notice that

$$\begin{aligned}
ATE(d) &= \mathbb{E}[Y_t(d) - Y_t(0)] \\
&= \mathbb{E}\Big[\mathbb{E}[Y_t(d) - Y_t(0)|D]\Big] \\
&= \int_{\mathcal{D}} ATT(d|l)\, dF_D(l)
\end{aligned}$$

where the second equality holds by the law of iterated expectations, and the third equality holds by the definition of $ATT(d|l)$. Then, the result holds because $ATT(d|l)$ is only identified under Assumption 4 when $d = l$. □

Proof of Theorem 3.2

Proof. For Equation (a-Cont), notice that, for $d \in \mathcal{D}_+$ and $(d+h) \in \mathcal{D}_+$,

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

where the first equality holds by Theorem 3.1 and the second equality holds by **??**. The result holds by taking the limit as $h \to 0$ and the definition of $ACRT(d|d)$.

For Equation (a-MV) and for $d_j \in \mathcal{D}_+$,

$$\begin{aligned}\mathbb{E}[\Delta Y_t|D=d_j] - \mathbb{E}[\Delta Y_t|D=d_{j-1}] &= \Big(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)\Big) + \Big(ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})\Big) \\ &= ACRT(d_j|d_j) + \Big(ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})\Big)\end{aligned}$$

where the first equality holds by Theorem 3.1 and **??** and the second equality holds by the definition of $ACRT(d_j|d_j)$.

Similarly, the result in Equation (b-Cont) holds by noting that for $d \in \mathcal{D}_+$ and $(d+h) \in \mathcal{D}_+$,

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

which follows from **??** and then by following the same arguments as for Equation (a-Cont).

For Equation (b-MV) and for $d_j \in \mathcal{D}_+$,

$$\mathbb{E}[\Delta Y_t|D=d_j] - \mathbb{E}[\Delta Y_t|D=d_{j-1}] = \Big(ATE(d_j) - ATT(d_{j-1})\Big)$$

which holds by **??**. □

Proof of Theorem 3.3

Proof. Notice that

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

where the second equality holds by adding and subtracting $\mathbb{E}[Y_{t-1}(0)]$, the third equality holds by Assumption 5, and the fourth equality holds because $Y_t(d)$ and $Y_{t-1}(0)$ are observed outcomes when $D=d$. □

F.2 Proofs of Results from Section 3.3

This section contains the proofs of the results in Section 3.3 on interpreting TWFE regressions with a multi-valued/continuous treatment.

Proof of Theorem 3.4

To conserve on notation, we define

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

and write $\Delta Y_i = \Delta Y_{it}$, since we have only two time periods.

Proof. First, notice that Equation (1.1) is equivalent to

$$\Delta Y_i = (\theta_t - \theta_{t-1}) + \beta^{twfe} D_i + \Delta v_{it} \tag{F.1}$$

which holds by taking first differences and because all units are untreated in the first period. Therefore, it immediately follows that

$$\begin{aligned}
\beta^{twfe} &= \frac{\mathbb{E}[\Delta Y (D - \mathbb{E}[D])]}{\mathrm{Var}(D)} \\
&= \mathbb{E}\left[\frac{(D-\mathbb{E}[D])}{\mathrm{Var}(D)}(m_\Delta(D) - m_\Delta(0))\right] \\
&= \mathbb{E}\left[\frac{(D-\mathbb{E}[D])}{\mathrm{Var}(D)}(m_\Delta(D) - m_\Delta(0)) | D > 0\right]\mathbb{P}(D > 0) \\
&= \mathbb{E}\left[\frac{(D-\mathbb{E}[D])}{\mathrm{Var}(D)}(m_\Delta(D) - m_\Delta(d_L)) | D > 0\right]\mathbb{P}(D > 0) + \mathbb{E}\left[\frac{(D-\mathbb{E}[D])}{\mathrm{Var}(D)}(m_\Delta(d_L) - m_\Delta(0)) | D > 0\right]\mathbb{P}(D > \\
&= A_1 + A_2
\end{aligned}$$

where the first equality holds because Equation (F.1) is a simple linear regression of ΔY on an intercept and D, the second equality holds because $\mathbb{E}[(D - \mathbb{E}[D]) m_\Delta(0)] = 0$, the third equality holds because $\mathbb{E}[m_\Delta(D) - m_\Delta(0) | D = 0] = 0$, and the fourth equality holds by adding and subtracting $m_\Delta(d_L)$.

We consider A_1 and A_2 separately next. First, for A_1,

$$\begin{aligned}
A_1 &= \mathbb{E}\left[\frac{(D-\mathbb{E}[D])}{\mathrm{Var}(D)}(m_\Delta(D) - m_\Delta(d_L)) | D > 0\right]\mathbb{P}(D > 0) \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{d_L}^{d_U} (k - \mathbb{E}[D])(m_\Delta(k) - m_\Delta(d_L))\, dF_{D|D>0}(k) \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{d_L}^{d_U} (k - \mathbb{E}[D]) \int_{d_L}^{k} m'_\Delta(l)\, dl\, dF_{D|D>0}(k) \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{d_L}^{d_U} (k - \mathbb{E}[D]) \int_{d_L}^{d_U} \mathbf{1}\{l \leq k\} m'_\Delta(l)\, dl\, dF_{D|D>0}(k) \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{d_L}^{d_U} m'_\Delta(l) \int_{d_L}^{d_U} (k - \mathbb{E}[D]) \mathbf{1}\{l \leq k\}\, dF_{D|D>0}(k)\, dl \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{d_L}^{d_U} m'_\Delta(l) \mathbb{E}[(D - \mathbb{E}[D]) \mathbf{1}\{l \leq D\} | D > 0]\, dl \\
&= \frac{\mathbb{P}(D>0)}{\mathrm{Var}(D)} \int_{d_L}^{d_U} m'_\Delta(l) \mathbb{E}[(D - \mathbb{E}[D]) | D \geq l] \mathbb{P}(D \geq l | D > 0)\, dl \\
&= \int_{d_L}^{d_U} m'_\Delta(l) \frac{(\mathbb{E}[D|D \geq l] - \mathbb{E}[D])\mathbb{P}(D \geq l)}{\mathrm{Var}(D)}\, dl
\end{aligned} \tag{F.2}$$

where the first equality is the definition of A_1, the second equality holds by rearranging terms and writing the expectation as an integral, the third equality holds by the fundamental theorem of calculus, the fourth equality rewrites the inner integral so that it is over d_U to d_L, the fifth equality holds by changing the order of integration and rearranging terms, the sixth equality holds by rewriting the inner integral as an expectation, the seventh equality holds by the law of iterated expectations (and since $D \geq l \implies D > 0$), and the last equality holds by combining terms.

Next, for A_2, it immediately holds that

$$\begin{aligned} A_2 &= \mathbb{E}\left[\frac{(D-\mathbb{E}[D])}{\text{Var}(D)}(m_\Delta(d_L)-m_\Delta(0))|D>0\right]\mathbb{P}(D>0) \\ &= \frac{(\mathbb{E}[D|D>0]-\mathbb{E}[D])\mathbb{P}(D>0)d_L}{\text{Var}(D)}\frac{(m_\Delta(d_L)-m_\Delta(0))}{d_L} \end{aligned} \tag{F.3}$$

where the first equality is the definition of A_2, and the second equality holds by multiplying and dividing by d_L.

Then, the first result in Theorem 3.4 holds by combining Equations (F.2) and (F.3). That the weights are all positive holds immediately since $(\mathbb{E}[D|D\geq l]-\mathbb{E}[D])>0$ for all $l\geq d_L$, $\mathbb{P}(D\geq l)>0$ for all $l\geq d_L$, $(\mathbb{E}[D|D>0]-\mathbb{E}[D])>0$, $\mathbb{P}(D>0)>0$, and $\text{Var}(D)>0$.

Next, we next show that $\int_{d_L}^{d_U} w_1(l)\,dl+w_0=1$. First, notice that

$$\begin{aligned} \int_{d_L}^{d_U} w_1(l)\,dl+w_0 = \frac{1}{\text{Var}(D)}\Bigg\{ & \int_{d_L}^{d_U}\mathbb{E}[D|D\geq l]\mathbb{P}(D\geq l)\,dl \\ & -\mathbb{E}[D]\int_{d_L}^{d_U}\mathbb{P}(D\geq l)\,dl \\ & +\mathbb{E}[D|D>0]\mathbb{P}(D>0)d_L \\ & -\mathbb{E}[D]\mathbb{P}(D>0)d_L\Bigg\} \\ = \frac{1}{\text{Var}(D)}\Big\{ & B_1-B_2+B_3-B_4\Big\} \end{aligned}$$

and we consider B_1, B_2, B_3, and B_4 in turn.

For B_1, first notice that for all $l\in\mathcal{D}_+$,

$$\begin{aligned} \mathbb{E}[D|D\geq l]P(D\geq l) &= \mathbb{E}[D\mathbf{1}\{D\geq l\}|D\geq l]\mathbb{P}(D\geq l) \\ &= \mathbb{E}[D\mathbf{1}\{D\geq l\}] \end{aligned} \tag{F.4}$$

which holds by the law of iterated expectations and implies that

$$\begin{aligned} B_1 &= \int_{d_L}^{d_U}\mathbb{E}[D|D\geq l]\mathbb{P}(D\geq l)\,dl \\ &= \int_{d_L}^{d_U}\int_{\mathcal{D}} d\mathbf{1}\{d\geq l\}\,dF_D(d)\,dl \\ &= \int_{\mathcal{D}} d\left(\int_{d_L}^{d_U}\mathbf{1}\{l\leq d\}\,dl\right)dF_D(d) \\ &= \int_{\mathcal{D}} d(d-d_L)\,dF_D(d) \\ &= \mathbb{E}[D^2]-\mathbb{E}[D]d_L \end{aligned} \tag{F.5}$$

where the first line is the definition of B_1, the second equality holds by Equation (F.4), the third equality holds by changing the order of integration, the fourth equality holds by carrying out the inner integration, and the last equality holds by rewriting the integral as an expectation.

Next, for term B_2,

$$B_2=\mathbb{E}[D]\int_{d_L}^{d_U}\mathbb{P}(D\geq l)\,dl$$

$$
\begin{aligned}
&= \mathbb{E}[D]\mathbb{P}(D>0)\int_{d_L}^{d_U} \mathbb{P}(D \geq l | D>0)\, dl \\
&= \mathbb{E}[D]\mathbb{P}(D>0)\int_{d_L}^{d_U} \int_{d_L}^{d_U} \mathbf{1}\{d \geq l\}\, dF_{D|D>0}(d)\, dl \\
&= \mathbb{E}[D]\mathbb{P}(D>0)\int_{d_L}^{d_U} \left(\int_{d_L}^{d_U} \mathbf{1}\{l \leq d\}\, dl \right) dF_{D|D>0}(d) \\
&= \mathbb{E}[D]\mathbb{P}(D>0)\int_{d_L}^{d_U} (d - d_L)\, dF_{D|D>0}(d) \\
&= \mathbb{E}[D]\mathbb{P}(D>0)\Big(\mathbb{E}[D|D>0] - d_L\Big) \\
&= \mathbb{E}[D]^2 - \mathbb{E}[D]\mathbb{P}(D>0)d_L
\end{aligned} \tag{F.6}
$$

where the first equality is the definition of B_2, the second equality holds by the law of iterated expectations, the third equality holds by writing $\mathbb{P}(D \geq l | D > 0)$ as an integral, the fourth equality changes the order of integration, the fifth equality carries out the inside integration, the sixth equality rewrites the integral as an expectation, the last equality holds by combining terms and by the law of iterated expectations.

Next,

$$
\begin{aligned}
B_3 &= \mathbb{E}[D|D>0]\mathbb{P}(D>0)d_L \\
&= \mathbb{E}[D]d_L
\end{aligned} \tag{F.7}
$$

which holds by the law of iterated expectations. And finally, recall that

$$
B_4 = \mathbb{E}[D]\mathbb{P}(D>0)d_L \tag{F.8}
$$

Thus, from Equations (F.5) to (F.8), it follows that

$$
B_1 - B_2 + B_3 + B_4 = \mathbb{E}[D^2] - \mathbb{E}[D]^2 = \text{Var}(D)
$$

which implies the result.

For part (2), the proof is similar as for part (1), but we provide the details here for completeness. Notice that,

$$
\begin{aligned}
\beta^{twfe} &= \mathbb{E}\left[\frac{(D - \mathbb{E}[D])}{\text{Var}(D)} (m_\Delta(D) - m_\Delta(0)) \right] \\
&= \frac{1}{\text{Var}(D)} \sum_{d \in \mathcal{D}} (d - \mathbb{E}[D])(m_\Delta(d) - m_\Delta(0)) p_d^D \\
&= \frac{1}{\text{Var}(D)} \sum_{d \in \mathcal{D}} (d - \mathbb{E}[D]) p_d^D \sum_{d_j \in \mathcal{D}_+} \mathbf{1}\{d_j \leq d\}(m_\Delta(d_j) - m_\Delta(d_{j-1})) \\
&= \frac{1}{\text{Var}(D)} \sum_{d_j \in \mathcal{D}_+} (m_\Delta(d_j) - m_\Delta(d_{j-1})) \sum_{d \in \mathcal{D}} (d - \mathbb{E}[D]) \mathbf{1}\{d \geq d_j\} p_d^D \\
&= \sum_{d_j \in \mathcal{D}_+} (m_\Delta(d_j) - m_\Delta(d_{j-1})) \frac{(\mathbb{E}[D|D \geq d_j] - \mathbb{E}[D])\mathbb{P}(D \geq d_j)}{\text{Var}(D)} \\
&= \sum_{d_j \in \mathcal{D}_+} w_1(d_j)(d_j - d_{j-1}) \frac{(m_\Delta(d_j) - m_\Delta(d_{j-1}))}{(d_j - d_{j-1})}
\end{aligned}
$$

where the second equality holds by writing the expectation as a summation, the third equality holds by adding and subtracting $m_\Delta(d_j)$ for all d_j's between 0 and d, the fourth equality holds by changing the order of the summations, the fifth equality writes the second summation as an expectation, and the last equality holds by the definition of the weights and by multiplying and dividing by $(d_j - d_{j-1})$. That $w_1(d_j)(d_j - d_{j-1}) > 0$ holds immediately since $w_1(d_j) \geq 0$ for all $d_j \in \mathcal{D}_+$ and $d_j > d_{j-1}$. Further,

$$\begin{aligned}\sum_{d_j \in \mathcal{D}_+} w_1(d_j)(d_j - d_{j-1}) &= \left(\sum_{d_j \in \mathcal{D}_+} \mathbb{E}[\mathbf{1}\{D \geq d_j\}D](d_j - d_{j-1}) - \mathbb{E}[D] \sum_{d_j \in \mathcal{D}_+} \mathbb{P}(D \geq d_j)(d_j - d_{j-1}) \right) / \mathrm{Var}(D) \\ &= (A - B)/\mathrm{Var}(D)\end{aligned}$$

We consider each of these terms in turn:

$$\begin{aligned} A &= \sum_{d_j \in \mathcal{D}_+} \sum_{d_k \in \mathcal{D}} \mathbf{1}\{d_k \geq d_j\} d_k p^D_{d_k} (d_j - d_{j-1}) \\ &= \sum_{d_k \in \mathcal{D}} p^D_{d_k} d_k \sum_{d_j \in \mathcal{D}_+, d_j \leq d_k} (d_j - d_{j-1}) \\ &= \sum_{d_k \in \mathcal{D}} p^D_{d_k} d_k (d_k - 0) \\ &= \mathbb{E}[D^2] \end{aligned}$$

where the first equality holds by writing the expectation for Term A as a summation, the second equality holds by re-ordering the summations, the third equality holds by canceling all the duplicate d_j terms across summations (and because $d_0 = 0$), and the last equality holds by the definition of $\mathbb{E}[D^2]$.

Next,

$$\begin{aligned} B &= \mathbb{E}[D] \sum_{d_j \in \mathcal{D}_+} \sum_{d_k \in \mathcal{D}} \mathbf{1}\{d_k \geq d_j\} p^D_{d_k} (d_j - d_{j-1}) \\ &= \mathbb{E}[D] \sum_{d_k \in \mathcal{D}} p^D_{d_k} \sum_{d_j \in \mathcal{D}_+, d_j \leq d_k} (d_j - d_{j-1}) \\ &= \mathbb{E}[D] \sum_{d_k \in \mathcal{D}} d_k p^D_{d_k} \\ &= \mathbb{E}[D]^2 \end{aligned}$$

where the first equality holds by writing the expectation for Term B as a summation, the second equality holds by re-ordering the summations, the third equality holds by canceling all the duplicate d_j terms across summations (and because $d_0 = 0$), and the last equality holds by the definition of $\mathbb{E}[D]$.

This implies that $A - B = \mathrm{Var}(D)$ which implies that the weights sum to 1. □

Proof of ??

Proof. The result holds immediately by plugging in the result in Theorem 3.2 into the result in Theorem 3.4 as well as noting that $\mathbb{E}[\Delta Y_t | D = d_L] - \mathbb{E}[\Delta Y_t | D = 0] = ATT(d_L|d_L)$ (under Assumption 4) and that $\mathbb{E}[\Delta Y_t | D = d_L] - \mathbb{E}[\Delta Y_t | D = 0] = ATE(d_L)$ (under Assumption 5). □

G Proofs of Results from Section E

This section contains the proofs of results from Appendix E on DiD with a multi-valued/continuous treatment and with multiple periods and variation in treatment timing.

Proof of Theorems D.1 and E.1

This section proves Theorem D.1; note that Theorem E.1, in the main text, corresponds to Part (2a) of Theorem D.1 (under Assumption 5-MP-Extended(a)).

For part (1a), we show the result under Assumption 4-MP(c) which is strictly weaker than Assumption 4-MP(a). First, notice that,

$$\begin{aligned}
ATT(g,t,d|g,d) &= \mathbb{E}[Y_t(d) - Y_t(0)|G=g, D=d] \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G=g, D=d] - \mathbb{E}[Y_t(0) - Y_{g-1}(0)|G=g, D=d] \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G=g, D=d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g, D=d] \qquad \text{(G.1)} \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G=g, D=d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|W_t=0] \\
&= \mathbb{E}[Y_t(d) - Y_{g-1}(0)|G=g, D=d] - \mathbb{E}[Y_t(0) - Y_{g-1}(0)|W_t=0] \\
&= \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|W_t=0]
\end{aligned}$$

where the first equality is the definition of $ATT(g,t,d|g,d)$, the second equality holds by adding and subtracting $\mathbb{E}[Y_{g-1}(0)|G=g, D=d]$, the third equality holds by adding and subtracting $\mathbb{E}[Y_s(0)|G=g, D=d]$ for $s=g,\ldots,(t-1)$, the fourth equality holds under Assumption 4-MP(c), the fifth equality holds by canceling all the terms involving $\mathbb{E}[Y_s(0)|W_t=0]$ for $s=g,\ldots,(t-1)$ (i.e., from the reverse of the argument for the third equality), and the last equality holds from writing the potential outcomes in terms of their observed counterparts.

For part (1b), in Equation (G.1), $\sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g, D=d] = \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|D=0]$ under Assumption 4-MP(b). Then, the result holds by otherwise following the same arguments as in part (1a).

For part (2a), we show the result under Assumption 5-MP-Extended(c) which is strictly weaker than Assumption 5-MP-Extended(a). First, notice that

$$\begin{aligned}
ATE(g,t,d) &= \mathbb{E}[Y_t(g,d) - Y_t(0)|G=g] \\
&= \mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G=g] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|G=g] \\
&= \mathbb{E}[Y_t(g,d) - Y_{t-1}(0)|G=g, D=d] - \mathbb{E}[Y_t(0) - Y_{t-1}(0)|G=g] \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \mathbb{E}[Y_{t-1}(0) - Y_{g-1}(0)|G=g, D=d] \\
&\quad - \Big(\mathbb{E}[Y_t(0) - Y_{g-1}(0)|G=g] - \mathbb{E}[Y_{t-1}(0) - Y_{g-1}(0)|G=g]\Big) \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g] \qquad \text{(G.2)} \\
&\quad - \sum_{s=g}^{t-1} \Big(\mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g, D=d] - \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g]\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|W_t = 0] \\
&= \mathbb{E}[Y_t(g,d) - Y_{g-1}(0)|G=g, D=d] - \mathbb{E}[Y_t(0) - Y_{g-1}(0)|W_t = 0] \\
&= \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0]
\end{aligned}
$$

where the first equality holds by the definition of $ATE(g,t,d)$, the second equality adds and subtracts $\mathbb{E}[Y_{t-1}(0)|G=g]$, the third equality holds by Assumption 5-MP-Extended(c), the fourth equality adds and subtracts both $\mathbb{E}[Y_{g-1}(0)|G=g, D=d]$ and $\mathbb{E}[Y_{g-1}(0)|G=g]$, the fifth equality holds by writing "long differences" as summations over "short differences" and by rearranging terms, the sixth equality holds by Assumption 5-MP-Extended(c) and by canceling terms, the seventh equality holds by rewriting the sum of short differences as a long difference, and the last equality holds by writing potential outcomes in terms of their corresponding observed outcomes and is the result.

The expression for $ACR(g,t,d)$ comes from taking the partial derivative of $ATE(g,t,d) = \mathbb{E}[Y_t - Y_{g-1}|G=g, D=d] - \mathbb{E}[Y_t - Y_{g-1}|W_t = 0]$ with respect to d and by noting that $\mathbb{E}[Y_t - Y_{g-1}|W_t = 0]$ does not depend on d.

Finally, for part (2b), in Equation (G.2), $\sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|G=g] = \sum_{s=g}^{t} \mathbb{E}[Y_s(0) - Y_{s-1}(0)|D=0]$ under Assumption 5-MP-Extended(b). The result then follows using the same subsequent arguments as in part (2a).

G.1 Proofs of Proposition E.1, Theorem E.2, and Proposition E.2

This section contains the proofs for interpreting TWFE regressions in the case with a continuous treatment, multiple periods, and variation in treatment timing as in Appendix E.

Before proving the main results in this section, we introduce some additional notation.

$$
v(g,t) = \mathbf{1}\{t \geq g\} - \bar{G}_g \tag{G.3}
$$

where the term $\mathbf{1}\{t \geq g\}$ is equal to one in post-treatment time periods for units in group g and recalling that we defined $\bar{G}_g = \frac{\mathcal{T}-g+1}{\mathcal{T}}$ which is the fraction of periods that units in group g are exposed to the treatment (and notice that this latter term does not depend on the particular time period t). Further, notice that $v(g,t)$ is positive in post-treatment time periods and negative in pre-treatment time periods for units in a particular group. Finally, also note that, for the "never-treated" group, $g = \mathcal{T}+1$ (which we set by convention and is helpful to unify the notation in this section) so that both terms in the expression for v are equal to 0 for the "never-treated" group.

Furthermore, recall that, for $1 \leq t_1 \leq t_2 \leq \mathcal{T}$, we defined

$$
\bar{Y}_i^{(t_1,t_2)} = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} Y_{it}
$$

where below (and following the notation used throughout the paper), we sometimes leave the subscript i implicit.

We next state and prove some additional results that are helpful for proving the main results. The first lemma re-writes (overall) expected dose experienced in period t adjusted by the overall expected dose (across periods and units) in a form that is useful in proving later results.

Lemma G.1. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$
\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbb{E}[W_s] = \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d v(g,t) \, dF_{D|G}(d|g) p_g
$$

Proof. First, notice that

$$\begin{aligned}\mathbb{E}[W_t] &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} \mathbb{E}[W_t | G = g, D = d] \, dF_{D|G}(d|g) p_g \\ &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\mathbf{1}\{t \geq g\} \, dF_{D|G}(d|g) p_g \qquad \text{(G.4)}\end{aligned}$$

where the first equality holds by the law of iterated expectations and the second equality holds because, after conditioning on group and dose, W_t is fully determined.

Thus,

$$\begin{aligned}\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbb{E}[W_s] &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\mathbf{1}\{t \geq g\} \, dF_{D|G}(d|g) p_g - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d\mathbf{1}\{s \geq g\} \, dF_{D|G}(d|g) p_g \\ &= \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left(\mathbf{1}\{t \geq g\} - \mathbf{1}\{s \geq g\} \right) \, dF_{D|G}(d|g) p_g \\ &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left\{ \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbf{1}\{t \geq g\} - \mathbf{1}\{s \geq g\} \right\} \, dF_{D|G}(d|g) p_g \\ &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left\{ \mathbf{1}\{t \geq g\} - \frac{\mathcal{T} - g + 1}{\mathcal{T}} \right\} \, dF_{D|G}(d|g) p_g \\ &= \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d v(g,t) \, dF_{D|G}(d|g) p_g\end{aligned}$$

where the first equality applies Equation (G.4) to both terms, the second equality combines terms by averaging the first term across time periods, the third equality re-orders the summations/integrals, the fourth equality holds because $\mathbf{1}\{t \geq g\}$ does not depend on s and by counting the fraction of periods where $s \geq g$, and the last equality holds by the definition of $v(g,t)$. □

The next lemma provides an intermediate result for the expression for the numerator of β^{twfe} in Equation (E.1).

Lemma G.2. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \mathbb{E}[Y_{it} \ddot{W}_{it}] = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \sum_{g \in \mathcal{G}} \int_{\mathcal{D}} d \left(\mathbb{E}[Y_t | G = g, D = d] - \mathbb{E}[Y_t] \right) v(g,t) \, dF_{D|G}(d|g) p_g \right\}$$

Proof. Starting with the numerator for β^{twfe} in Equation (E.1)

$$\begin{aligned}&\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \mathbb{E}[Y_{it} \ddot{W}_{it}] \\ &= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \mathbb{E}[Y_{it} W_{it}] - \mathbb{E}[Y_{it} \bar{W}_i] - \mathbb{E}[Y_t] \left(\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbb{E}[W_s] \right) \right\} \\ &= \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left\{ \mathbb{E}[Y_t D \mathbf{1}\{t \geq G\}] - \mathbb{E}\left[Y_t D \frac{\mathcal{T} - G + 1}{\mathcal{T}} \right] - \mathbb{E}[Y_t] \left(\mathbb{E}[W_t] - \frac{1}{\mathcal{T}} \sum_{s=1}^{\mathcal{T}} \mathbb{E}[W_s] \right) \right\}\end{aligned}$$

$$= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\Bigg\{ \sum_{g\in\mathcal{G}}\int_{\mathcal{D}} \left(\mathbb{E}[Y_t d\mathbf{1}\{t \geq g\}|G=g, D=d] - \mathbb{E}\left[Y_t \frac{\mathcal{T}-g+1}{\mathcal{T}} d \Big| G=g, D=d\right]\right) dF_{D|G}(d|g)p_g$$

$$- \mathbb{E}[Y_t]\left(\mathbb{E}[W_t] - \frac{1}{\mathcal{T}}\sum_{s=1}^{\mathcal{T}}\mathbb{E}[W_s]\right)\Bigg\}$$

$$= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{ \sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g, D=d]v(g,t)\Big)\, dF_{D|G}(d|g)p_g - \mathbb{E}[Y_t]\left(\mathbb{E}[W_t] - \frac{1}{\mathcal{T}}\sum_{s=1}^{\mathcal{T}}\mathbb{E}[W_s]\right)\right\}$$

$$= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{ \sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g, D=d]v(g,t)\Big)\, dF_{D|G}(d|g)p_g - \mathbb{E}[Y_t]\left(\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} dv(g,t)\, dF_{D|G}(d|g)p_g\right)\right\}$$

$$= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{ \sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\left(\mathbb{E}[Y_t|G=g, D=d] - \mathbb{E}[Y_t]\right) v(g,t)\, dF_{D|G}(d|g)p_g\right\}$$

where the first equality holds by the definition of $\ddot{W}_{it}$, the second equality holds by plugging in for W_{it} and $\bar{W}_i$, the third equality holds by the law of iterated expectations, the fourth equality holds by the definition of $v(g,t)$, the fifth equality holds by Lemma G.1, and the sixth equality just combines terms.

□

Next, based on the result in Lemma G.2, we can write the numerator in the expression for β^{twfe} as

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[Y_{it}\ddot{W}_{it}]$$

$$= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{ \sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\left(\mathbb{E}[Y_t|G=g, D=d] - \mathbb{E}[Y_t]\right) v(g,t)\, dF_{D|G}(d|g)p_g\right\}$$

$$= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g, D=d] - \mathbb{E}[Y_t|G=g]\Big)v(g,t)\, dF_{D|G}(d|g)p_g \tag{G.5}$$

$$+ \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g] - \mathbb{E}[Y_t]\Big)v(g,t)\, dF_{D|G}(d|g)p_g \tag{G.6}$$

where the first equality holds from Lemma G.2 and the second equality holds by adding and subtracting $\mathbb{E}[Y_t|G=g]$.

The expression in Equation (G.5) involves comparisons between units in the same group but that have different doses. The expression in Equation (G.6) involves comparisons across different groups. We consider each of these terms in more detail below.

Lemma G.3. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g, D=d] - \mathbb{E}[Y_t|G=g]\Big)v(g,t)\, dF_{D|G}(d|g)p_g$$

$$= \sum_{g\in\mathcal{G}}\left\{(1-\bar{G}_g)\bar{G}_g \text{Cov}\left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G=g\right)\right\}p_g$$

Proof.

$$
\begin{aligned}
&\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g,D=d]-\mathbb{E}[Y_t|G=g]\Big)v(g,t)\,dF_{D|G}(d|g)p_g \\
&\quad= \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\mathbb{E}[Y_t(D-\mathbb{E}[D|G=g])|G=g]v(g,t)p_g\right\} \\
&\quad= \sum_{g\in\mathcal{G}}\left\{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[Y_t(D-\mathbb{E}[D|G=g])|G=g]v(g,t)\right\}p_g \\
&\quad= \sum_{g\in\mathcal{G}}\Bigg\{-\frac{1}{\mathcal{T}}\frac{(T-g+1)}{T}\sum_{t=1}^{g-1}\mathbb{E}[Y_t(D-\mathbb{E}[D|G=g])|G=g] \\
&\qquad\qquad +\frac{1}{\mathcal{T}}\frac{(g-1)}{\mathcal{T}}\sum_{t=g}^{\mathcal{T}}\mathbb{E}[Y_t(D-\mathbb{E}[D|G=g])|G=g]\Bigg\}p_g \\
&\quad= \sum_{g\in\mathcal{G}}\Bigg\{\frac{g-1}{\mathcal{T}}\frac{(T-g+1)}{T}\Bigg(\frac{1}{\mathcal{T}-g+1}\sum_{t=g}^{\mathcal{T}}\mathbb{E}[Y_t(D-\mathbb{E}[D|G=g])|G=g] \\
&\qquad\qquad -\frac{1}{g-1}\sum_{t=1}^{g-1}\mathbb{E}[Y_t(D-\mathbb{E}[D|G=g])|G=g]\Bigg)\Bigg\}p_g \\
&\quad= \sum_{g\in\mathcal{G}}\left\{\frac{g-1}{\mathcal{T}}\frac{(T-g+1)}{T}\left(\mathbb{E}\big[\big(\bar{Y}^{POST(g)}-\bar{Y}^{PRE(g)}\big)(D-\mathbb{E}[D|G=g])|G=g\big]\right)\right\}p_g \\
&\quad= \sum_{g\in\mathcal{G}}\left\{(1-\bar{G}_g)\bar{G}_g\left(\mathbb{E}\big[\big(\bar{Y}^{POST(g)}-\bar{Y}^{PRE(g)}\big)(D-\mathbb{E}[D|G=g])|G=g\big]\right)\right\}p_g \\
&\quad= \sum_{g\in\mathcal{G}}\left\{(1-\bar{G}_g)\bar{G}_g\mathrm{Cov}\left(\bar{Y}^{POST(g)}-\bar{Y}^{PRE(g)},D|G=g\right)\right\}p_g
\end{aligned}
$$

where the first equality holds by the law of iterated expectations (and combining terms involving d and Y_t), the second equality changes the order of the summations, the third equality holds by splitting the summation involving t in time period g and plugs in for $v(g,t)$ (which is constant within group g and across time periods from $1,\dots,g-1$ and from $g,\dots,\mathcal{T}$), the fourth equality multiplies and divides by terms so that the inside expressions can be written as averages, the fifth equality holds by changing the order of the expectation and averaging over time periods, the sixth equality holds by the definition of $\bar{G}_g$, and the last equality holds by the definition of covariance. □

Lemma G.3 shows that part of the TWFE estimator comes from a weighted average of post- vs. pre-treatment outcomes within group but who experienced different doses. In particular, notice that, for units in group g, $\bar{Y}_i^{POST(g)}$ is their average post-treatment outcome while $\bar{Y}_i^{PRE(g)}$ is their average pre-treatment outcome.

Next, we consider the expression from Equation (G.6) above which arises from differences in outcomes across groups. We handle this term over several following results.

Lemma G.4. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Big(\mathbb{E}[D|G=g]v(g,t)-\mathbb{E}[D|G=k]v(k,t)\Big)\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)p_kp_g\right\}$$

Proof. Notice that

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)\,dF_{D|G}(d|g)p_g\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\mathbb{E}[D|G=g]\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)p_g\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\mathbb{E}[D|G=g]\Big(\mathbb{E}[Y_t|G=g]-\sum_{k\in\mathcal{G}}\mathbb{E}[Y_t|G=k]p_k\Big)v(g,t)p_g\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G}}\mathbb{E}[D|G=g]v(g,t)\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)p_kp_g\right\}$$

$$=\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\left\{\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Big(\mathbb{E}[D|G=g]v(g,t)-\mathbb{E}[D|G=k]v(k,t)\Big)\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)p_kp_g\right\}$$

where the first equality holds by integrating over $\mathcal{D}$, the second equality holds by the law of iterated expectations, the third equality holds by combining terms, and the last equality holds because all combinations of g and k occur twice. □

Lemma G.4 is helpful because it shows that the cross-group part of the TWFE estimator can be written as comparisons for each group relative to later-treated groups.

Next, we provide an important intermediate result. Before stating this result, we define the following weights

$$\begin{aligned}
\tilde{w}^{g,within}(g) &= \text{Var}(D|G=g)(1-\bar{G}_g)\bar{G}_g p_g \tilde{w}^{g,post}(g,k) \qquad &= \mathbb{E}[D|G=g]^2(1-\bar{G}_g)(\bar{G}_g-\bar{G}_k)p_kp_g \\
\tilde{w}^{k,post}(g,k) &= \mathbb{E}[D|G=k]^2\bar{G}_k(\bar{G}_g-\bar{G}_k)p_kp_g \\
\tilde{w}^{long}(g,k) &= (\mathbb{E}[D|G=g]-\mathbb{E}[D|G=k])^2\bar{G}_k(1-\bar{G}_g)p_kp_g
\end{aligned}$$

which correspond to $w^{g,post}$, $w^{k,post}$, and $w^{long}(g,k)$ in the main text except they do not divide by $\mathcal{T}^{-1}\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\ddot{W}_{it}^2]$. In addition, notice that

$$\begin{aligned}
&\mathbb{E}[D|G=g]v(g,t)-\mathbb{E}[D|G=k]v(k,t) \\
&\quad=\begin{cases}-\mathbb{E}[D|G=g]\bar{G}_g+\mathbb{E}[D|G=k]\bar{G}_k & \text{for } t<g<k \\ \mathbb{E}[D|G=g](1-\bar{G}_g)+\mathbb{E}[D|G=k]\bar{G}_k & \text{for } g\leq t<k \\ \mathbb{E}[D|G=g](1-\bar{G}_g)-\mathbb{E}[D|G=k](1-\bar{G}_k) & \text{for } g<k\leq t\end{cases}
\end{aligned} \tag{G.7}$$

which holds by the definition of v and is useful for the proof of the following lemma.

Lemma G.5. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$
\begin{aligned}
&\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\Bigg\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)\,dF_{D|G}(d|g)p_g\Bigg\}\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{\tilde{w}^{g,post}(g,k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)}\big)|G=g\Big]-\mathbb{E}\Big[\big(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)}\big)|G=k\Big]\Bigg)\\
&\qquad\qquad+\tilde{w}^{k,post}(g,k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)}\big)|G=k\Big]-\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)}\big)|G=g\Big]\Bigg)\\
&\qquad\qquad+\tilde{w}^{long}(g,k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{PRE(g)}\big)|G=g\Big]-\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{PRE(g)}\big)|G=k\Big]\Bigg\}
\end{aligned}
$$

Proof. The result holds as follows

$$
\begin{aligned}
&\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\Bigg\{\sum_{g\in\mathcal{G}}\int_{\mathcal{D}} d\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t]\Big)v(g,t)\,dF_{D|G}(d|g)p_g\Bigg\}\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\Big(\mathbb{E}[D|G=g]v(g,t)-\mathbb{E}[D|G=k]v(k,t)\Big)\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\Bigg\}p_kp_g\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{\frac{1}{\mathcal{T}}\left(-\mathbb{E}[D|G=g]\bar{G}_g+\mathbb{E}[D|G=k]\bar{G}_k\right)\sum_{t=1}^{g-1}\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\\
&\qquad+\frac{1}{\mathcal{T}}\left(\mathbb{E}[D|G=g](1-\bar{G}_g)+\mathbb{E}[D|G=k]\bar{G}_k\right)\sum_{t=g}^{k-1}\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\\
&\qquad+\frac{1}{\mathcal{T}}\left(\mathbb{E}[D|G=g](1-\bar{G}_g)-\mathbb{E}[D|G=k](1-\bar{G}_k)\right)\sum_{t=k}^{\mathcal{T}}\Big(\mathbb{E}[Y_t|G=g]-\mathbb{E}[Y_t|G=k]\Big)\Bigg\}p_kp_g\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{(1-\bar{G}_g)\left(-\mathbb{E}[D|G=g]\bar{G}_g+\mathbb{E}[D|G=k]\bar{G}_k\right)\Big(\mathbb{E}[\bar{Y}^{PRE(g)}|G=g]-\mathbb{E}[\bar{Y}^{PRE(g)}|G=k]\Big)\\
&\qquad+(\bar{G}_g-\bar{G}_k)\left(\mathbb{E}[D|G=g](1-\bar{G}_g)+\mathbb{E}[D|G=k]\bar{G}_k\right)\Big(\mathbb{E}[\bar{Y}^{MID(g,k)}|G=g]-\mathbb{E}[\bar{Y}^{MID(g,k)}|G=k]\Big)\\
&\qquad+\bar{G}_k\left(\mathbb{E}[D|G=g](1-\bar{G}_g)-\mathbb{E}[D|G=k](1-\bar{G}_k)\right)\Big(\mathbb{E}[\bar{Y}^{POST(k)}|G=g]-\mathbb{E}[\bar{Y}^{POST(k)}|G=k]\Big)\Bigg\}p_kp_g\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{(1-\bar{G}_g)\left(-\mathbb{E}[D|G=g](\bar{G}_g-\bar{G}_k)+(\mathbb{E}[D|G=k]-\mathbb{E}[D|G=g])\bar{G}_k\right)\Big(\mathbb{E}[\bar{Y}^{PRE(g)}|G=g]-\mathbb{E}[\bar{Y}^{PRE(g)}|G=k]\Big)\\
&\qquad+(\bar{G}_g-\bar{G}_k)\left(\mathbb{E}[D|G=g](1-\bar{G}_g)+\mathbb{E}[D|G=k]\bar{G}_k\right)\Big(\mathbb{E}[\bar{Y}^{MID(g,k)}|G=g]-\mathbb{E}[\bar{Y}^{MID(g,k)}|G=k]\Big)\\
&\qquad+\bar{G}_k\left((\mathbb{E}[D|G=g]-\mathbb{E}[D|G=k])(1-\bar{G}_g)-\mathbb{E}[D|G=k](\bar{G}_g-\bar{G}_k)\right)\Big(\mathbb{E}[\bar{Y}^{POST(k)}|G=g]-\mathbb{E}[\bar{Y}^{POST(k)}|G=k]\Big)\Bigg\}p_kp_g\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{\mathbb{E}[D|G=g](1-\bar{G}_g)(\bar{G}_g-\bar{G}_k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)}\big)|G=g\Big]-\mathbb{E}\Big[\big(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)}\big)|G=k\Big]\Bigg)\\
&\qquad+\mathbb{E}[D|G=k]\bar{G}_k(\bar{G}_g-\bar{G}_k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)}\big)|G=k\Big]-\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)}\big)|G=g\Big]\Bigg)\\
&\qquad+(\mathbb{E}[D|G=g]-\mathbb{E}[D|G=k])\bar{G}_k(1-\bar{G}_g)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{PRE(g)}\big)|G=g\Big]-\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{PRE(g)}\big)|G=k\Big]\Bigg)\Bigg\}p_kp_g\\
&\quad=\sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{\tilde{w}^{g,post}(g,k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)}\big)|G=g\Big]-\mathbb{E}\Big[\big(\bar{Y}^{MID(g,k)}-\bar{Y}^{PRE(g)}\big)|G=k\Big]\Bigg)\\
&\qquad+\tilde{w}^{k,post}(g,k)\Bigg(\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)}\big)|G=k\Big]-\mathbb{E}\Big[\big(\bar{Y}^{POST(k)}-\bar{Y}^{MID(g,k)}\big)|G=g\Big]\Bigg)
\end{aligned}
$$

$$+ \tilde{w}^{long}(g,k)\bigg(\mathbb{E}\Big[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=g\Big] - \mathbb{E}\Big[(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)})|G=k\Big]\bigg)\bigg\}$$

where the first equality uses the result in Lemma G.4, the second equality changes the order of the summations (splitting them at g and k where the value of $v(g,t)$ and $v(k,t)$ change) and uses Equation (G.7), the third equality holds by averaging over time periods (which involves multiplying and dividing by $g-1$ in the first line, multiplying and dividing by $k-g$ in the second line, and multiplying and dividing by $\mathcal{T}-k+1$ in the last line), the fourth equality rearranges the expressions for the weights, the fifth equality holds by rearranging terms with common weights, and the last equality holds by the definitions of $\tilde{w}^{g,post}$, $\tilde{w}^{k,post}$, and $\tilde{w}^{long}$ and by noticing that

$$p_k p_g = (p_g + p_k)^2 p_{g|\{g,k\}}(1 - p_{g|\{g,k\}})$$

which holds by multiplying and dividing both p_k and p_g by $(p_g + p_k)$ and by the definition of $p_{g|\{g,k\}}$. □

The result in Lemma G.5 is very closely related to the result on interpreting TWFE regressions with a binary treatment and multiple time periods and variation in treatment timing in Goodman-Bacon (2021).[41] In particular, it says that, even with a continuous/multi-valued treatment, the TWFE regression estimator involves comparisons between (i) the path of outcomes for units that become treated relative to the path of outcomes for units that are not treated yet, (ii) the path of outcomes for units that become treated relative to the path of outcomes for units that have already been treated, and (iii) comparisons of the paths of outcomes across groups from their common pre-treatment periods to their common post-treatment periods. Intuitively, the first set of comparisons are very much in the spirit of DiD, but the second and third sets of comparisons are not (except under additional specialized conditions). We formalize this intuition in the proof of Theorem E.2 below.

Lemma G.6. *Under Assumptions 1-MP, 2-MP(a), and 3-MP,*

$$\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2] = \sum_{g\in\mathcal{G}} \tilde{w}^{g,within}(g) + \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Big\{\tilde{w}^{g,post}(g,k) + \tilde{w}^{k,post}(g,k) + \tilde{w}^{long}(g,k)\Big\}$$

Proof. To start with, notice that $\mathbb{E}[\ddot{W}_{it}^2] = \mathbb{E}[W_{it}\ddot{W}_{it}]$. Then, we can apply the arguments of Lemmas G.2 to G.5 but with W_{it} replacing Y_{it}. This implies that

$$\begin{aligned}
&\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2] \\
&\quad = \sum_{g\in\mathcal{G}} \tilde{w}^{g,within}(g)\frac{\mathrm{Cov}(\bar{W}^{POST(g)} - \bar{W}^{PRE(g)}, D|G=g)}{\mathrm{Var}(D|G=g)} \\
&\quad + \sum_{g\in\mathcal{G}}\sum_{k\in\mathcal{G},k>g}\Bigg\{\tilde{w}^{g,post}(g,k)\frac{\mathbb{E}\left[(\bar{W}^{MID(g,k)} - \bar{W}^{PRE(g)})|G=g\right] - \mathbb{E}\left[(\bar{W}^{MID(g,k)} - \bar{W}^{PRE(g)})|G=k\right]}{\mathbb{E}[D|G=g]} \\
&\qquad + \tilde{w}^{k,post}(g,k)\frac{\mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{MID(g,k)})|G=k\right] - \mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{MID(g,k)})|G=g\right]}{\mathbb{E}[D|G=k]} \\
&\qquad + \tilde{w}^{long}(g,k)\frac{\mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{PRE(g)})|G=g\right] - \mathbb{E}\left[(\bar{W}^{POST(k)} - \bar{W}^{PRE(g)})|G=k\right]}{\mathbb{E}[D|G=g] - \mathbb{E}[D|G=k]}\Bigg\}
\end{aligned}$$

[41] One difference worth noting is that the weights are slightly different due to the terms involving $\mathbb{E}[D|G=g]$ and $\mathbb{E}[D|G=k]$. With a binary treatment, these expectations are equal to each other by construction, but with a continuous treatment these terms are no longer generally equal to each other. This also implies that the third term does not show up in the case with a binary treatment.

$$= \sum_{g\in\mathcal{G}} \tilde{w}^{g,within}(g) + \sum_{g\in\mathcal{G}} \sum_{k\in\mathcal{G},k>g} \left\{ \tilde{w}^{g,post}(g,k) + \tilde{w}^{k,post}(g,k) + \tilde{w}^{long}(g,k) \right\}$$

where the last equality holds by noting that $\bar{W} = D$ in post-treatment periods and $\bar{W} = 0$ in pre-treatment periods, and then by canceling terms. □

Proof of Proposition E.1

Proof. Proposition E.1 immediately holds by combining the results in Lemma G.2, from Equations (G.5) and (G.6), and by Lemmas G.3 to G.5 (which all concern the numerator in the expression for β^{twfe} in Equation (E.2)), and then dividing by $(1/\mathcal{T})\sum_{t=1}^{\mathcal{T}} \mathbb{E}[\ddot{W}_{it}^2]$ (which corresponds to the denominator in the expression for β^{twfe} in Equation (E.2)). That the weights are all positive holds immediately by their definitions. That they sum to one holds by the definitions of the weights and by Lemma G.6. □

Next, we move to proving Theorem E.2. To do this we provide expressions for each of the comparisons that show up in Proposition E.1 in terms of derivatives of paths of outcomes. These results invoke Assumption 2-MP(b) and (c) and, therefore, use that the treatment is actually continuous, but they do not invoke any parallel trends assumptions. That said, it would be straightforward to adapt these results to the case with a discrete multi-valued treatment along the lines of the baseline two period case considered above.

It is also useful to note that

$$\frac{\partial \pi_D^{POST(\tilde{k}),PRE(\tilde{g})}(g,d)}{\partial d} = \frac{\partial \mathbb{E}\left[\left(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\right)|G=g, D=d\right]}{\partial d}$$

$$\frac{\partial \pi_D^{MID(\tilde{g},\tilde{k}),PRE(\tilde{g})}(g,d)}{\partial d} = \frac{\mathbb{E}\left[\left(\bar{Y}^{MID(\tilde{g},\tilde{k})} - \bar{Y}^{PRE(\tilde{g})}\right)|G=g, D=d\right]}{\partial d}$$

$$\frac{\partial \pi_D^{POST(\tilde{k}),MID(\tilde{g},\tilde{k})}(g,d)}{\partial d} = \frac{\partial \mathbb{E}\left[\left(\bar{Y}^{POST(\tilde{k})} - \bar{Y}^{MID(\tilde{g},\tilde{k})}\right)|G=g, D=d\right]}{\partial d}$$

which holds because the second parts of each π_D term do not vary with the dose.

Next, we consider a result for the main term in $\delta^{WITHIN}(g)$ in Equation (E.3).

Lemma G.7. *Under Assumptions 1-MP, 2-MP, and 3-MP,*

$$\text{Cov}\left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G=g\right)$$

$$= \int_{\mathcal{D}_+} \left(\mathbb{E}[D|G=g, D\geq l] - \mathbb{E}[D|G=g]\right)\mathbb{P}(D\geq l|G=g)\frac{\partial \mathbb{E}[\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l}\, dl$$

Proof. First, notice that

$$\text{Cov}\left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}, D|G=g\right) = \mathbb{E}\left[\left(\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}\right)(D - \mathbb{E}[D|G=g])|G=g\right]$$

Then, the proof follows essentially the same arguments as in **??** with $\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}$ replacing ΔY and the other arguments relating to the distribution of the dose holding conditional on being in group g. The second term, involving d_L, in **??** does not show up here as, by construction, there are no untreated units in group g. □

Lemma G.7 says that part of $\delta^{WITHIN}(g)$ in the TWFE regression estimator comes from a weighted average of $\frac{\partial \mathbb{E}[\bar{Y}^{POST(g)} - \bar{Y}^{PRE(g)}|G=g, D=d]}{\partial d}$.

Next, we consider the main term in the expression for $\delta^{MID,PRE}(g,k)$ in Equation (E.4). This term is quite similar to the baseline two-period case considered in **??** because units in group k have not been treated yet.

Lemma G.8. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $k > g$,*

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G=g) \frac{\partial \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l}\, dl \\
&\qquad + d_L \frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g, D=d_L] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L} \\
&\qquad - d_L \frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=k] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L}
\end{aligned}$$

Proof. To start with, notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] \\
&\quad = \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D=0\right] \\
&\qquad - \left(\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D=0\right]\right) \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G=g) \frac{\partial \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l}\, dl \\
&\qquad + d_L \frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g, D=d_L] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L} \\
&\qquad - d_L \frac{\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=k] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L}
\end{aligned}$$

where the first equality holds by adding and subtracting $\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D=0\right]$. For the second equality, notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D=0\right] \\
&\quad = \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g, D=d_L\right] \\
&\qquad + \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g, D=d_L\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D=0\right]
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g, D=d_L\right] \\
&\quad = \int_{\mathcal{D}_+} \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g, D=d\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g, D=d_L\right] dF_{D|G}(d|g) \\
&\quad = \int_{\mathcal{D}_+} \int_{\mathcal{D}_+} \mathbf{1}\{l \leq d\} \frac{\partial \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G=g, D=l\right]}{\partial l}\, dl\, dF_{D|G}(d|g) \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G=g) \frac{\partial \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l}\, dl
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second equality holds by the fundamental theorem of calculus, and the last equality holds by changing the order of integration and simplifying.

Combining the above expressions implies the result. □

Next, we consider the main term for $\delta^{POST,MID}(g,k)$ in Equation (E.5) which comes from comparing paths of outcomes for newly treated groups relative to already-treated groups.

Lemma G.9. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $k > g$,*

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = g\right] \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G = k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|G = k, D = l]}{\partial l}\, dl \\
&\qquad + d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|G = k, D = d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|D = 0]}{d_L} \\
&\quad - \Big\{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = g] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0] \\
&\qquad - \left(\mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|G = g] - \mathbb{E}[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}|D = 0]\right)\Big\}
\end{aligned}$$

Proof. Notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = g\right] \\
&\quad = \left(\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|D = 0\right]\right) \\
&\qquad - \left(\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|D = 0\right]\right) \\
&\quad = \left(\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|G = k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|D = 0\right]\right) \qquad \text{(G.8)} \\
&\qquad - \Bigg\{\left(\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]\right) \\
&\qquad\quad - \left(\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]\right)\Bigg\} \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G = k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}|G = k, D = l]}{\partial l}\, dl \qquad \text{(G.9)} \\
&\qquad + d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{MID}(g,k)|G = g, D = d_L] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|D = 0\right]}{d_L} \\
&\quad - \Bigg\{\left(\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]\right) \\
&\qquad - \left(\mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|G = g\right] - \mathbb{E}\left[\left(\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)}\right)|D = 0\right]\right)\Bigg\}
\end{aligned}$$

where the first equality holds by adding and subtracting $\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{MID(g,k)}\right)|D = 0\right]$, the second equality holds by adding and subtracting both $\mathbb{E}\left[\bar{Y}^{PRE(g)}|G = g\right]$ and $\mathbb{E}\left[\bar{Y}^{PRE(g)}|D = 0\right]$, and the last equality holds by applying the same sort of arguments as in the proof of Lemma G.8. □

The expression in Lemma G.9 appears complicated and is worth explaining in some more detail. Consider Equation (G.8) in the proof of Lemma G.9. There are three parts of this expression. The first part compares the path of outcomes in post-treatment periods relative to some pre-treatment periods for units in group k to the path of outcomes for units that never participate in the treatment. This sort of comparison is very much in the spirit of DiD and will correspond to a reasonable treatment effect parameter under appropriate parallel trends assumptions. Similarly, under suitable parallel trends assumptions, the terms in the second and third lines will correspond to treatment effects for group g between periods k and $\mathcal{T}$ (the second line) and treatment effects for group g between periods g and $k-1$ (the third line). Therefore, the difference between these terms can be thought of as some form of treatment effect dynamics. That means, in general, for this overall term to correspond to a treatment effect parameter for group k, there needs to be no treatment effect dynamics for group g. Ruling out treatment effect dynamics is not implied by any sort of parallel trends assumption and therefore involves an additional (and potentially very strong) assumption.

Finally, we consider the main term for $\delta^{POST,PRE}(g,k)$ in Equation (E.6).

Lemma G.10. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for $k > g$,*

$$
\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] \\
&\quad = \int_{\mathcal{D}_+} \left(\mathbb{P}(D \geq l|G=g) - \mathbb{P}(D \geq l|G=k)\right) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l}\, dl \\
&\quad - \left\{ \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G=k) \left(\frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=k, D=l]}{\partial l} - \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l} \right) dl \right. \\
&\qquad + d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=k, D=d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L} \\
&\qquad \left. - d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L} \right\}
\end{aligned}
$$

Proof. First, by adding and subtracting terms

$$
\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] \\
&\quad = \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|D=0\right] \\
&\quad - \left(\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|D=0\right]\right)
\end{aligned}
$$

Then, using similar arguments as in Lemma G.8 above, one can show that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=g\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|D=0\right] \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G=g) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l}\, dl \\
&\qquad + d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D=0]}{d_L}
\end{aligned}
$$

and that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|G=k\right] - \mathbb{E}\left[\left(\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}\right)|D=0\right] \\
&\quad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G=k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=k, D=l]}{\partial l}\, dl
\end{aligned}
$$

$$+ d_L \frac{\mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G = k, D = d_L] - \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|D = 0]}{d_L}$$

Then, the result holds by adding and subtracting $\int_{\mathcal{D}_+} \mathbb{P}(D \geq l|G = k) \frac{\partial \mathbb{E}[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)}|G=g, D=l]}{\partial l} \, dl$ and combining terms. □

Proof of Part (1) of Theorem E.2

Proof. Starting from the result in Proposition E.1, the expression for $\delta^{WITHIN}(g)$ comes from its definition, the result in Lemma G.7, and the definition of the weights $w_1^{within}(g, l)$. The expression for $\delta^{MID,PRE}(g, k)$ comes from its definition, the result in Lemma G.8, and the definitions of $w_1(g, l)$ and $w_0(g)$. The expression for $\delta^{POST,MID}(g, k)$ comes from combining its definition with the result in Lemma G.9, and the definitions of $w_1(k, l)$ and $w_0(k)$. Finally, the expression for $\delta^{POST,PRE}(g, k)$ comes from its definition, the result in Lemma G.10, and the definitions of $w_1^{across}(g, k, l)$, $\tilde{w}_1^{across}(g, k, l)$, and $\tilde{w}_0^{across}(g, k)$.

That $w_1^{within}(g, d) \geq 0$, $w_1(g, 0) \geq 0$, $w_0(g) \geq 0$ for all $g \in \mathcal{G}$ and $d \in \mathcal{D}_+$ all hold immediately from the definitions of the weights. That $\int_{\mathcal{D}_+} w_1^{within}(g, l) \, dl = 1$, $\int_{\mathcal{D}_+} w_1(g, l) \, dl + w_0(g) = 1$, and $\int_{\mathcal{D}_+} w_1^{across}(g, k, l) \, dl = 1$ hold from the same sorts of arguments used to show that the weights integrate to 1 in the proof of Theorem 3.4. □

Notice that none of the previous results have invoked any sort of parallel trends assumption. Next, we push forward the previous results once a researcher invokes parallel trends assumptions; in the main text, we considered the case where the researcher invoked Assumption 5-MP, but here we consider both that assumption and Assumption 4-MP(a). To further understand this, for $1 \leq t_1 < t_2 \leq \mathcal{T}$ define

$$\bar{Y}_i^{(t_1,t_2)}(g, d) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} Y_{it}(g, t, d)$$

which averages potential outcomes from time periods t_1 to t_2 for unit i if they were in group g and experienced dose d. Note that $\bar{Y}_i^{(t_1,t_2)} = \bar{Y}_i^{(t_1,t_2)}(G_i, D_i)$. Next, for $t_1 \leq t_2$, define

$$\overline{ATT}^{(t_1,t_2)}(g, d|g, d) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} ATT(g, t, d|g, d)$$

which is the average treatment effect experienced by units in group g who experienced dose d averaged across periods from t_1 to t_2. Likewise, define

$$\overline{ATE}^{(t_1,t_2)}(g, d) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} ATE(g, t, d)$$

which is the average treatment effect of dose d among all units in group g averaged across periods from t_1 to t_2. An alternative expression for $\overline{ATT}^{(t_1,t_2)}(g, d|g, d)$ is given by

$$\overline{ATT}^{(t_1,t_2)}(g, d|g, d) = \mathbb{E}\left[\bar{Y}^{(t_1,t_2)}(g, d) - \bar{Y}^{(t_1,t_2)}(0)|G = g, D = d\right]$$

which holds by the definition of $ATT(g, t, d|g, d)$ and changing the order of the expectation and the average over time periods; here, $\mathbb{E}[\bar{Y}^{(t_1,t_2)}(0)|G = g, D = d]$ is the average outcome that units in group g that experienced dose d would have experienced if they had not participated in the treatment between time periods t_1 and t_2. Similarly, for $\overline{ATE}^{(t_1,t_2)}(g,d)$,

$$\overline{ATE}^{(t_1,t_2)}(g,d) = \mathbb{E}\left[\bar{Y}^{(t_1,t_2)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g\right]$$

In addition, define

$$\overline{ACRT}^{(t_1,t_2)}(g,d|g,d) = \frac{\partial \overline{ATT}(g,l|g,d)}{\partial l}\Big|_{l=d} \quad \text{and} \quad \overline{ACR}^{(t_1,t_2)}(g,d) = \frac{\partial \overline{ATE}(g,d)}{\partial d}$$

which are the average causal response to a marginal increase in the dose among units in group g conditional on having dose experienced dose d (for $\overline{ACRT}(g,d|g,d)$) and the average causal response to a marginal increase in the dose among all units in group g.

The next result connects derivatives of conditional expectations to $ACRT$ and ACR parameters under parallel trends assumptions.

Lemma G.11. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for* $1 \leq t_1 \leq t_2 < g \leq t_3 \leq t_4 \leq \mathcal{T}$ *(i.e., t_1 and t_2 are pre-treatment periods for group g, and t_3 and t_4 are post-treatment periods for group g), and for $d \in \mathcal{D}_+$,*

(1) If, in addition, Assumption 4-MP(a) holds, then

$$\frac{\partial \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g,D=d\right]}{\partial d} = \overline{ACRT}^{(t_3,t_4)}(g,d|g,d) + \frac{\partial \overline{ATT}^{(t_3,t_4)}(g,d|g,l)}{\partial l}\Big|_{l=d}$$

(2) If, in addition, Assumption 5-MP holds, then

$$\frac{\partial \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g,D=d\right]}{\partial d} = \overline{ACR}^{(t_3,t_4)}(g,d)$$

Proof. For part (1), notice that, for $1 \leq t_1 \leq t_2 < g \leq t_3 \leq t_4 \leq \mathcal{T}$ (i.e., for group g, t_1 and t_2 are pre-treatment time periods while t_3 and t_4 are post treatment time periods), we can write

$$\begin{aligned}
\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g,D=d\right] &= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right] \\
&= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_3,t_4)}(0)|G=g,D=d\right] \\
&\qquad - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right] \\
&= \overline{ATT}^{(t_3,t_4)}(g,d|g,d) \\
&\qquad - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right]
\end{aligned}$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by adding and subtracting $\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0)|G=g,D=d\right]$, and the last equality holds by the definition of $\overline{ATT}^{(t_3,t_4)}(g,d|g,d)$.

This equation looks very similar to DiD-type equations in simpler cases such as when there are two periods and two groups. The left hand side is immediately identified. The right hand side involves a causal effect parameter of interest and an unobserved path of untreated potential outcomes that would typically be handled using a parallel trends assumption.

In particular, under Assumption 4-MP(a) (though notice that Assumption 4-MP(b) and (c) are not generally strong enough here),

$$\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g,D=d\right] = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|D=0\right]$$

which, importantly, does not vary across d or g.

This suggests that, under Assumption 4-MP(a),

$$\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d\right] = \overline{ATT}^{(t_3,t_4)}(g,d|g,d) - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|D=0\right]$$

Taking derivatives of both sides of the previous equation with respect to d implies the result.

For part (2), notice that,

$$\begin{aligned}
\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d\right] &= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g, D=d\right] \\
&= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_1,t_2)}(0)|G=g\right] \\
&= \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d) - \bar{Y}^{(t_3,t_4)}(0)|G=g\right] \\
&\qquad + \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=g\right] \\
&= \overline{ATE}^{(t_3,t_4)}(g,d) + \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|D=0\right]
\end{aligned}$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by Assumption 5-MP, the third equality holds by adding and subtracting $\mathbb{E}[\bar{Y}^{(t_3,t_4)}(0)|G=g]$, and the last equality holds by the definition of $\overline{ATE}^{(t_3,t_4)}(g,d)$ and by Assumption 5-MP. Taking derivatives of both sides implies the result for part (2). □

The result in Lemma G.11 says that, under Assumption 4-MP(a), the derivative of the path of outcomes (averaged over some post-treatment periods) relative to some pre-treatment periods corresponds to $ACRT(g,t,d|g,d)$ plus the derivative of a selection bias-type term with respect to d across some post-treatment time periods for units in group g. Similarly, under Assumption 5-MP, the derivative of the averaged path of outcomes over time in some post-treatment periods relative to the same average path of outcomes in some pre-treatment periods corresponds to an average of $ACR(g,d)$ with respect to d across the same post-treatment time periods.

The intuition for this sort of result is very similar to that of Theorem 3.2 in the baseline case with two time periods.

Lemma G.12. *Under Assumptions 1-MP, 2-MP, and 3-MP, and for* $1 \leq t_1 \leq t_2 < g \leq t_3 \leq t_4 < k$ *(i.e., t_1 and t_2 are pre-treatment periods for both groups g and k, group g is treated before group k, and t_3 and t_4 are post-treatment periods for group g but pre-treatment periods for group k),*

(1) If, in addition, Assumption 4-MP(a) holds, then

$$d_L \frac{\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=k\right]}{d_L} = d_L \frac{\overline{ATT}^{(t_3,t_4)}(g,d_L|g,d_L)}{d_L}$$

(2) If, in addition, Assumption 5-MP holds, then

$$d_L \frac{\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=k\right]}{d_L} = d_L \frac{\overline{ATE}^{(t_3,t_4)}(g,d_L)}{d_L}$$

Proof. For part (1), notice that

$$\begin{aligned}
&\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=g, D=d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)}|G=k\right] \\
&\quad = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_1,t_2)}(0)|G=g, D=d_L\right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0)|G=k\right] \\
&\quad = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g,d_L) - \bar{Y}^{(t_3,t_4)}(0)|G=g, D=d_L\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad + \left\{ \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0) | G = g, D = d_L \right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0) | G = k \right] \right\} \\
&= \overline{ATT}^{(t_3,t_4)}(g, d_L)
\end{aligned}
$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by adding and subtracting $\mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0)|G = g, D = d_L\right]$, and the last equality holds by the definition of $\overline{ATT}^{(t_3,t_4)}(g, d_L)$ and because the difference between the two terms involving paths of untreated potential outcomes on the second line of the previous equality is equal to 0 under Assumption 4-MP(a). Then, the result holds by multiplying and dividing by d_L.

For part (2),

$$
\begin{aligned}
&\mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)} | G = g, D = d_L \right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)} - \bar{Y}^{(t_1,t_2)} | G = k \right] \\
&\quad = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g, d_L) - \bar{Y}^{(t_1,t_2)}(0) | G = g, D = d_L \right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0) | G = k \right] \\
&\quad = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g, d_L) - \bar{Y}^{(t_1,t_2)}(0) | G = g \right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0) | G = k \right] \\
&\quad = \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(g, d_L) - \bar{Y}^{(t_3,t_4)}(0) | G = g \right] \\
&\qquad + \left\{ \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0) | G = g \right] - \mathbb{E}\left[\bar{Y}^{(t_3,t_4)}(0) - \bar{Y}^{(t_1,t_2)}(0) | G = k \right] \right\} \\
&\quad = \overline{ATE}^{(t_3,t_4)}(g, d_L)
\end{aligned}
$$

where the first equality holds by writing observed outcomes in terms of their corresponding potential outcomes, the second equality holds by Assumption 5-MP, the third equality holds by adding and subtracting $\mathbb{E}[\bar{Y}^{(t_3,t_4)}(0)|G = g]$, and the last equality holds by Assumption 5-MP. The result holds by multiplying and dividing by d_L. □

Proof of Part (2) of Theorem E.2

Proof. The result holds immediately by using the results of Lemmas G.11 and G.12 in each of the expressions for $\delta^{WITHIN}(g)$, $\delta^{MID,PRE}(g, k)$, $\delta^{POST,MID}(g, k)$, and $\delta^{POST,PRE}(g, k)$ in part (1) of Theorem E.2. □

Proof of Proposition E.2

Proof. For part (a), using similar arguments as in Lemma G.8 and then under Assumption 5-MP, it follows that

$$
\begin{aligned}
&\mathbb{E}\left[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)} | G = g \right] - \mathbb{E}\left[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g)} | D = 0 \right] \\
&\qquad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l | G = g) \overline{ACR}^{POST(k)}(g, l) \, dl + d_L \frac{\overline{ATE}^{POST(k)}(g, d_L)}{d_L}
\end{aligned}
$$

and that

$$
\begin{aligned}
&\mathbb{E}\left[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)} | G = g \right] - \mathbb{E}\left[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g)} | D = 0 \right] \\
&\qquad = \int_{\mathcal{D}_+} \mathbb{P}(D \geq l | G = g) \overline{ACR}^{MID(g,k)}(g, l) \, dl + d_L \frac{\overline{ATE}^{MID(g,k)}(g, d_L)}{d_L}
\end{aligned}
$$

Under Assumption 7(a), $ACR(g,t,d)$ and $ATE(g,t,d_L)$ do not vary over time which implies that, for all $g \in \mathcal{G}$ and $k \in \mathcal{G}$ with $k > g$, $\overline{ACR}^{POST(k)}(g,l) = \overline{ACR}^{POST(k)}(g,l)$ for all $l \in \mathcal{D}_+$ and $\overline{ATE}^{POST(k)}(g,d_L) = \overline{ATE}^{MID(g,k)}(g,d_L)$. This implies that $\mathbb{E}\left[\bar{Y}^{POST(k)} - \bar{Y}^{PRE(g))}|G=g\right] = \mathbb{E}\left[\bar{Y}^{MID(g,k)} - \bar{Y}^{PRE(g))}|G=g\right]$ which implies the result for part (a).

For part (b), notice that, under Assumption 5-MP,

$$\begin{aligned}
\frac{\partial \pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l} &= \overline{ACR}^{POST(k)}(k,l) - \overline{ACR}^{POST(k)}(g,l) \\
&= 0
\end{aligned}$$

for $l \in \mathcal{D}_+$ and where the second equality holds by Assumption 7(b) (which implies that, for a particular time period, $ACR(g,t,d)$ does not vary across groups).

The same sort of arguments imply that

$$\begin{aligned}
\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L} &= \frac{\overline{ATE}^{POST(k)}(k,d_L) - \overline{ATE}^{POST(k)}(g,d_L)}{d_L} \\
&= 0
\end{aligned}$$

Finally, for part (c), under Assumption 7(a), (b), and (c), $ACR(g,t,d)$ does not vary across groups, time periods, or dose; since this does not vary, we denote it by ACR for the remainder of the proof. Moreover, from Theorem E.2, we have that $\int_{\mathcal{D}_+} w_1^{within}(g,l)\, dl = 1$, $\int_{\mathcal{D}_+} w_1(g,l)\, dl + w_0(g) = 1$, and that $\int_{\mathcal{D}_+} w_1^{across}(g,k,l) = 1$. From the first two parts of the current result, we also have that the nuisance paths of outcomes in $\delta^{POST,MID}(g,k)$ and $\delta^{POST,PRE}(g,k)$ are both equal to 0 under Assumption 7(a) and (b). This implies that, under the conditions for part (c), $\delta^{WITHIN}(g) = \delta^{MID,PRE}(g,k) = \delta^{POST,MID}(g,k) = \delta^{POST,PRE}(g,k) = ACR$. Finally, from Proposition E.1, we have that β^{twfe} is a weighted average of $\delta^{MID,PRE}(g,k)$, $\delta^{POST,MID}(g,k)$, $\delta^{POST,MID}(g,k)$, and $\delta^{POST,PRE}(g,k)$. That these are all equal to each other implies that $\beta^{twfe} = ACR = ACR^{*,mp}$. □

Next, we provide a version of Theorem E.2 extended to the case where Assumption 4-MP(a) (which is the multi-period version of standard parallel trends that only involves untreated potential outcomes) holds.

Theorem E.2-Extended. *Under Assumptions 1-MP, 2-MP, 3-MP, and 4-MP(a),*

$$\delta^{WITHIN}(g) = \int_{\mathcal{D}_+} w_1^{within}(g,l) \left(\overline{ACRT}^{POST(g)}(g,l|g,l) + \frac{\partial \overline{ATT}^{POST(g)}(g,l|g,h)}{\partial h}\Big|_{h=l} \right) dl$$

$$\begin{aligned}
\delta^{MID,PRE}(g,k) = \int_{\mathcal{D}_+} w_1(g,l) &\left(\overline{ACRT}^{MID(g,k)}(g,l|g,l) + \frac{\partial \overline{ATT}^{MID(g,k)}(g,l|g,h)}{\partial h}\Big|_{h=l} \right) dl \\
&+ w_0(g) \frac{\overline{ATT}^{MID(g,k)}(g,d_L|g,d_L)}{d_L}
\end{aligned}$$

$$\begin{aligned}
\delta^{POST,MID}(g,k) = \int_{\mathcal{D}_+} w_1(k,l) &\left(\overline{ACRT}^{POST(k)}(k,l|k,l) + \frac{\partial \overline{ATT}^{POST(k)}(k,l|k,h)}{\partial h}\Big|_{h=l} \right) dl \\
&+ w_0(k) \frac{\overline{ATT}^{POST(k)}(k,d_L|k,d_L)}{d_L} - w_0(k) \left(\frac{\pi^{POST(k),PRE(g)}(g) - \pi^{MID(g,k),PRE(g)}(g)}{d_L} \right)
\end{aligned}$$

$$
\begin{aligned}
\delta^{POST,PRE}(g,k) = & \int_{\mathcal{D}_+} w_1^{across}(g,k,l) \left(\overline{ACRT}^{POST(k)}(g,l|g,l) + \frac{\partial \overline{ATT}(g,l|g,h)}{\partial h}\Big|_{h=l} \right) dl \\
& - \Bigg\{ \int_{\mathcal{D}_+} \tilde{w}_1^{across}(g,k,l) \left(\frac{\partial \pi_D^{POST(k),PRE(g)}(k,l)}{\partial l} - \frac{\partial \pi_D^{POST(k),PRE(g)}(g,l)}{\partial l} \right) dl \\
& \qquad + \tilde{w}_0^{across}(g,k) \left(\frac{\pi_D^{POST(k),PRE(g)}(k,d_L) - \pi_D^{POST(k),PRE(g)}(g,d_L)}{d_L} \right) \Bigg\}
\end{aligned}
$$

where the weights are the same as in Theorem E.2 and satisfy the same properties.

Proof. The result holds immediately by plugging in the results of part (1) of Lemmas G.11 and G.12 for $\delta^{WITHIN}(g)$, $\delta^{MID,PRE}(g,k)$, $\delta^{POST,MID}(g,k)$, and $\delta^{POST,PRE}(g,k)$ in part (1) of Theorem E.2. □