Difference-in-Differences Designs: A Practitioner's Guide

Andrew Baker[*]   Brantly Callaway[†]   Scott Cunningham[‡]

Andrew Goodman-Bacon[§]   Pedro H. C. Sant'Anna[¶]

June 9, 2025

Abstract

Difference-in-differences (DiD) is arguably the most popular quasi-experimental research design. Its canonical form, with two groups and two periods, is well-understood. However, empirical practices can be ad hoc when researchers go beyond that simple case. This article provides an organizing framework for discussing different types of DiD designs and their associated DiD estimators. It discusses covariates, weights, handling multiple periods, and staggered treatments. The organizational framework, however, applies to other extensions of DiD methods as well.

1 Introduction

Dating to the 1840s, difference-in-differences (DiD) is now the most common research design for estimating causal effects in the social sciences.[1] A basic DiD design requires two time periods, one before and one after some treatment begins, and two groups, one that receives a treatment and one that does not. The DiD estimate equals the change in outcomes for the treated group minus the change in outcomes for the untreated group: the difference of two differences. If the average change in the outcomes would have been the same in the two groups had the treatment not occurred, which is referred to as a "parallel trends" assumption, this comparison estimates the average treatment effect among treated units.

---

[*]University of California, Berkeley

[†]University of Georgia

[‡]Baylor University

[§]Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis

[¶]Emory University

[1]Currie, Kleven and Zwiers (2020) find that almost 25% of all NBER empirical working papers and 17% of empirical articles in five leading general-interest economics journals in 2018 mention DiD. The earliest DiD applications we are aware of are from Ignaz Semmelweis from the 1840s (Semmelweis, 1983) and Snow (1855). For a brief overview of the long history of DiD in economics, see Section 2 of Lechner (2011).

In practice, however, researchers apply DiD methods to situations that are more complicated than the classic two-period and two-group (2×2) setup. Most datasets cover multiple periods, and units may enter (or exit) treatment at different times. Treatment might also vary in its amount or intensity. Other variables are often used to make treated and untreated units more comparable. Today's typical DiD study includes at least one of these deviations from the canonical 2×2 setup.

For many years, the common practice in applied research was to estimate complex DiD designs using linear regressions with unit and time fixed effects (two-way fixed effects, henceforth TWFE). Their identifying assumptions and interpretation were informally traced to the fact that, in the 2×2 case, a TWFE estimator gives the same estimate as a DiD estimator calculated directly from sample means, and thus inherits a clear causal interpretation under a specific parallel trends identification assumption. This appeared to justify the use of a single technique for any type of design or specification. Recent research, however, has shown that simple regressions can fail to estimate meaningful causal parameters when DiD designs are complex and treatment effects vary, producing estimates that are not only misleading in their magnitudes but potentially of the wrong *sign*. The significance of these findings is substantial; given the prevalence of DiD analysis in modern applied econometrics work, common empirical practices have almost certainly yielded misleading results in several concrete cases (Baker, Larcker and Wang, 2022).

So, what should applied researchers do instead? This paper proposes a unified framework for discussing and conducting DiD studies that is rooted in the principles of causal inference in the presence of treatment effect heterogeneity. The central conclusion of recent methodological research is that even complex DiD studies can be understood as aggregations of 2×2 comparisons between one set of units for which treatment changes and another set for which it does not. This fact links a wide variety of DiD designs used in practice and guides methodological choices about estimating them. Viewing DiD studies through the lens of 2×2 "building blocks" aids in interpretability by clarifying that they yield causal quantities that aggregate the treatment effects identified by each 2×2 component. It also means that identification comes from the simple parallel trends assumptions required for each 2×2 building block. Practically, this framework suggests first estimating each 2×2 building block and then aggregating them. As long as the effective sample size is large, this approach allows for asymptotically valid inference using standard techniques.

This framework is a "forward-engineering" approach to DiD that embraces treatment effect heterogeneity and constructs estimators that recover well-motivated causal parameters under explicitly stated assumptions. By fixing the goals of the study (the target parameters) and deriving analytical techniques, forward engineering provides clear benefits over "reverse-engineering" approaches that begin with a familiar regression specification and derive the assumptions under which it has *some* causal interpretation. The methods we describe in this paper combine familiar techniques with some newer ones, but expressly avoid the difficulties of interpretation inherent in common regression estimators (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfoeuille, 2020; Sun and Abraham, 2021; Borusyak, Jaravel and Spiess, 2024). Moreover, the interpretation

of common regression estimators changes across specifications, which makes it hard to understand the difference between non-robustness and a shifting target parameter. In contrast, our proposed framework naturally leads to estimation procedures that target the same parameter under different transparent identification assumptions. Thus, two estimates can be distinguished easily by their identifying assumptions. Finally, the principles of the forward-engineering approach provide guidance to good econometric practices even in settings without well-established methodological findings.

This paper is not designed to be a comprehensive literature review; its goal is to provide guidelines for practitioners who want to better understand DiD and its various forms. Because of the tremendous variations in design, data, and specification that practitioners encounter, we opt to focus on three of the most common aspects of modern DiD studies: the use of weights, covariates, and staggered treatment timing. We apply techniques to address these issues to a specific example: the causal effect of recent public health insurance expansions in the US on county-level mortality. In an appendix, we briefly discuss related DiD designs with different treatment variables (ones that turn on and off or take many values), additional comparisons (i.e., triple-difference designs), distributional target parameters, or different data structures (repeated cross-sections or unbalanced panels). Several recent reviews follow the logic laid out here and cover additional DiD-related topics and technical details: Roth, Sant'Anna, Bilinski and Poe (2023); de Chaisemartin and D'Haultfoeuille (2023b); Callaway (2023).

The rest of the paper is structured as follows. Section 2 introduces the Medicaid example. Section 3 discusses the canonical 2×2 DiD setups with and without weights, and Section 4 discusses threats to the identification assumptions, how to assess them, and how to incorporate covariates. Section 5 extends the 2×2 setup to multiple periods with potentially staggered treatment adoption. Section 6 concludes and briefly discusses some extensions that involve more complex DiD designs.

2 Medicaid and mortality: The running example

To make our methodological discussion concrete, we revisit a timely and important causal question: How did the expansion of public health insurance (Medicaid) under the Affordable Care Act (ACA) affect mortality?

Medicaid expansion is a great example of a staggered treatment adoption. The ACA originally mandated that in 2014 all states expand Medicaid eligibility to adults with incomes up to 138% of the federal poverty threshold. In upholding the law's constitutionality in a 2012 decision, however, the Supreme Court made Medicaid expansion optional. As a result, many states expanded Medicaid after 2014, but several have not done so as of 2024.

Columns 1 and 2 of Table 1 illustrate the variation in Medicaid expansion dates.

Table 1: Medicaid Expansion under the Affordable Care Act

Expansion Year	States	Share of States	Share of Counties	Share of Adults (2013)
Pre-2014	DE, MA, NY, VT	0.08	0.03	0.09
2014	AR, AZ, CA, CO, CT, HI, IA, IL, KY, MD, MI, MN, ND, NH, NJ, NM, NV, OH, OR, RI, WA, WV	0.44	0.36	0.45
2015	AK, IN, PA	0.06	0.06	0.06
2016	LA, MT	0.04	0.04	0.02
2019	ME, VA	0.04	0.05	0.03
2020	ID, NE, UT	0.06	0.04	0.02
2021	MO, OK	0.04	0.06	0.03
2023	NC, SD	0.04	0.05	0.03
Non-Expansion	AL, FL, GA, KS, MS, SC, TN, TX, WI, WY	0.20	0.31	0.26

The table shows which states adopted the ACA's Medicaid expansion in each year as well as the share of all states, counties, and adults in each expansion year.

States expanded Medicaid largely because of economic and political considerations (Sommers and Epstein, 2013), which created observable differences between expansion and non-expansion states. For instance, just four out of the 22 states that expanded Medicaid in 2014 are in the southern Census region; conversely, seven out of 10 non-expansion states are in the South. This suggests a potential role for covariates when analyzing Medicaid expansion.

Finally, mortality is measured in jurisdictions like states and counties, which are of very different sizes. Choices about (population) weights determine not only how different estimation approaches average the units within a given expansion group but also how a given estimation technique averages estimated effects across those groups. California, for example, represented 4.5% of the states that expanded Medicaid in 2014, 5.4% of the counties, but 27.7% of the adults ages 20-64; its contribution to "the" average outcome for the 2014 expansion group is very different with weights than without. The final three columns of Table 1 show that, in our data the entire 2014 expansion group contains 44% of the states, 36% of the counties, but 45% of all adults. Weighting will therefore change how important the estimated treatment effects are for the 2014 group.

Several recent papers study the effect of ACA Medicaid expansion on mortality rates for lower-income adults, who are most likely to gain insurance through Medicaid. Miller, Johnson and Wherry (2021) and Wyse and Meyer (2024) use simple DiD methods to provide evidence that Medicaid reduced adult mortality rates for targeted sub-populations. Unfortunately, their analyses require restricted links between income and mortality data, which are important for overcoming the low statistical power in studies using aggregate mortality data (Black, Hollingsworth, Nunes and Simon, 2022). Our goal is to pursue a replicable and shareable example based on a related analysis by Borgschulte and Vogler (2020). They use a sophisticated strategy to select and use covariates in a weighted TWFE regression using restricted access data, and find that Medicaid expansion reduced aggregate county-level mortality rates. We use publicly available data, which

we include in a fully-reproducible replication package, and consider only a handful of intuitive demographic and economic covariates sufficient to illustrate several practical challenges that can arise with DiD. This empirical exercise is meant solely to illustrate how to tackle several common features of DiD designs. The results are pedagogical in spirit and do not represent the best possible estimates of Medicaids effect on adult mortality.

Our outcome variable is the crude adult mortality rate, $Y_{i,t}$, for people ages 20-64 (measured per 100,000) by county (i) from 2009 to 2019 released by the Centers for Disease Control and Prevention (2024).[2] We denote county i's adult population in 2013 by W_i and its socioeconomic covariates in year t (discussed below) by $X_{i,t}$. The information in Table 1 defines the treatment group variable G_i that equals the year in which county i's state expanded Medicaid; $G_i = \infty$ for the non-expansion states. Our final sample contains 2,604 counties in states with complete data on mortality rates from 2009 to 2019 and covariates for 2013 and 2014.

Faced with a setup such as this, researchers need to make a range of tightly related choices. Which treatment groups in Table 1 should be compared with each other and over what time horizons? What must be true for those comparisons to identify causal effects, and how should one empirically evaluate their plausibility? How can other information, such as covariates or pre-period outcomes, be used to improve the credibility of the design? How do these methodological choices affect the causal interpretation of a given analysis? The aim of this review is to demonstrate to practitioners using DiD in realistic scenarios why and how to use state-of-the-art econometric tools to answer these questions.

3   2×2 DiD designs

We begin our discussion by focusing on the canonical 2×2 DiD setup, which has two time periods—one before and one after treatment—and two groups—one that remains untreated in both periods and one that becomes treated in the second period. In our Medicaid example, we focus on comparisons in 2014 and 2013 between the 2014 expansion group (978 counties) and the group that had not expanded by 2019 (1,222 counties). When we consider more complex designs, this kind of comparison will still play a role: it will be one 2 × 2 "building block" among many.

Using these basic ingredients, we can now define a 2×2 DiD *design*, composed of a causal target parameter, a treatment variable, an assumption under which it is identified, and an estimation approach, which will be the classic difference of two differences. This may be familiar territory in the simple case, but it is a crucial framework for building up appropriate techniques in more complex cases.

---

[2]It is common to adjust mortality rates by the county age distribution. Unfortunately, the CDC measurements of age-specific deaths are restricted for counties with fewer than 10 annual deaths. We aim to use publicly available and shareable data for pedagogical purposes; we follow Borgschulte and Vogler (2020) and use the crude mortality rate.

We first define a treatment group dummy D_i that equals one for the treated units (expansion states, $G_i = 2014$) and zero for the untreated units (states that had not expanded by 2019, $G_i > 2019$). The treatment *status* dummy, $D_{i,t} = D_i \times 1\{t \geqslant 2014\}$, then equals one for counties in 2014 expansion states during post-expansion years. To highlight how weights enter different kinds of DiD analyses, we use the following notation for expected values. For generic random variables A and C, for a given set of non-negative weights ω, define $\mathbb{E}_\omega[A|C] = \mathbb{E}[\omega A|C]/\mathbb{E}[\omega|C]$ as the ω-weighted population expectation of A given C. When $\omega = 1$ for all units in the population, we simply write $\mathbb{E}_\omega[A|C] = \mathbb{E}[A|C]$. Henceforth, unless otherwise noted, we assume that we have a balanced panel data random sample of $(Y_{t=1}, \ldots, Y_{t=T}, G, X)$.

3.1   Causal effects and target parameters: The ATT

The first step of any causal analysis is to define the causal quantity of interest, also called the target parameter. We use the potential outcomes framework of Rubin (1974) and Robins (1986) to do so. Let $Y_{i,t}(0,0)$ denote unit i's potential outcome at time t if it remained untreated in both periods. Analogously, let $Y_{i,t}(0,1)$ denote unit i's potential outcome at time t if untreated in the first period but exposed to treatment by the second period. In our example, $Y_{i,t}(0,0)$ is county i's mortality rate in period t in a world in which Medicaid did not expand in its state, and $Y_{i,t}(0,1)$ is its mortality rate in a world in which Medicaid did expand in 2014.[3] To simplify notation, we will write $Y_{i,t}(0) = Y_{i,t}(0,0)$ and $Y_{i,t}(1) = Y_{i,t}(0,1)$, as the potential outcomes are defined by treatment exposure in period two (Medicaid expansion status by 2014). Nonetheless, it will be useful for later discussions that these potential outcomes correspond to treatment sequences.

In practice, we never observe $Y_{i,t}(1)$ and $Y_{i,t}(0)$ for the the same unit. Instead, the data we observe, $Y_{i,t}$, are treated outcomes $Y_{i,t}(1)$ for treated units, and untreated outcomes $Y_{i,t}(0)$ for untreated units, as in the following equation:

$$Y_{i,t} = (1 - D_i)Y_{i,t}(0) + D_i Y_{i,t}(1). \tag{3.1}$$

We additionally assume that county mortality rates were not affected by the Medicaid expansion *before* Medicaid expanded, which is crucial to the validity of the DiD estimator (see, e.g., Abbring and van den Berg, 2003, Malani and Reif, 2015, and Roth et al., 2023). This standard "no anticipation" assumption ensures that we observe untreated potential outcomes before Medicaid expansion takes effect: $Y_{i,2013} = Y_{i,2013}(0)$. It also helps us define effective treatment dates. For instance, if the announcement of Medicaid expansion affected mortality before its actual expansion, "treatment" would begin when the policy was announced rather than implemented. We formally state this assumption for completeness and maintain it throughout the paper.

---

[3]We have implicitly introduced the stable unit treatment value assumption, which holds if the only treatment determining county i's mortality rate is its own. In other words, Medicaid expansion in neighboring counties must not affect deaths in county i. If this fails, then there are effectively many different treatment variables and counterfactuals. The potential outcomes notation and ensuing analysis would then need to account for this.

Assumption NA (No-Anticipation). For all treated units i and all pre-treatment periods t, $Y_{i,t}(1) = Y_{i,t}(0)$.

The potential outcomes define a causal effect for every unit in every time period, $Y_{i,t}(1) - Y_{i,t}(0)$. These describe what Medicaid expansion did to mortality rates in a specific treated county or what it would have done in a specific untreated county. This framework allows for arbitrary heterogeneity in the effects across units and time; that is, it allows the effect of Medicaid expansion to be different in every county and year. But it is hard to learn about this degree of rich heterogeneity without additional strong conditions holding. Instead, DiD analyses typically seek to estimate (weighted) averages of heterogeneous treatment effects. In particular, most DiD designs target the average treatment effect on the treated at time t, or $ATT(t)$:

$$ATT(t) = \mathbb{E}_\omega[Y_{i,t}(1) - Y_{i,t}(0)|D_i = 1]$$
$$= \mathbb{E}_\omega[Y_{i,t}|D_i = 1] - \mathbb{E}_\omega[Y_{i,t}(0)|D_i = 1]. \tag{3.2}$$

Equation (3.2) shows that $ATT(t)$ compares (weighted) average *observed* post-expansion mortality rates among treated counties ($\mathbb{E}_\omega[Y_{i,t}|D_i = 1]$) with the (weighted) average untreated mortality rates for the same treated counties ($\mathbb{E}_\omega[Y_{i,t}(0)|D_i = 1]$). The second quantity is counterfactual because untreated outcomes are never observed for treated counties. Note that by the no-anticipation assumption, $ATT(t) = 0$ for all pre-treatment periods; that is, $ATT(2013) = 0$ in our two-period Medicaid example. This ensures that cross-group outcome comparisons before treatment begins reflect *untreated* potential outcome gaps, which is central to the logic of DiD. Note that we abuse notation and omit the weight index when defining ATT's; we do that to unclutter notation throughout the paper.

Equation (3.2) shows that weighting enters the analysis early on, as part of the definition of the causal parameter. In the Medicaid context, the unweighted $ATT(2014)$ answers the question, "What was the average causal effect of Medicaid expansion on 2014 mortality rates among the 2014 expansion state counties?" The weighted $ATT(2014)$ answers the question, "What was the average causal effect of Medicaid expansion on 2014 mortality rates among adults in counties in states that expanded Medicaid in 2014?" This point interacts with other justifications for weighting, such as improving precision. With heterogeneous treatment effects, adopting a weighting scheme designed to improve precision in the presence of heteroskedasticity in a constant-coefficient regression model will also change the target parameter, potentially by a lot when treatment effects are correlated with the weights (Solon, Haider and Wooldridge, 2015). Comparing weighted and unweighted estimates, therefore, does not show whether weighting matters for estimation or inference; these reflect different target parameters. In our example, the population-weighted parameter is probably more policy-relevant, but we conduct some of our empirical exercises both ways to show how weighting can affect a given DiD result. Which parameter is "of interest" is an argument about theoretical importance, policy relevance, and the use to which it will be put.

Other target parameters are also possible. Designs other than DiD identify different kinds

of average treatment effects, and some DiD methods use quantile regression (Athey and Imbens, 2006; Callaway and Li, 2019) or distribution regression (Fernández-Val, Meier, van Vuuren and Vella, 2024a) approaches to target features of the marginal distributions of $Y_{i,t}(1)$ and $Y_{i,t}(0)$ among treated units. We focus on identification and estimation strategies that target ATT parameters but emphasize that the 2×2 building block framework applies to DiD methods more broadly; see our appendix for more discussions about distributional parameters.

3.2 Identifying assumptions: Parallel trends

A research design is a strategy—a set of assumptions—to identify and estimate specific target parameters. Many different assumptions can identify the missing counterfactual for $ATT(2014)$ in the Medicaid example. For example, mean independence between $Y_{i,2014}(0)$ and D_i implies that the counterfactual equals average 2014 mortality rates in non-expansion counties ($\mathbb{E}_\omega[Y_{i,2014}(0)|D_i = 0]$). Under this assumption, which essentially entails assuming that Medicaid expansion is as good as random, the cross-sectional mortality gap in 2014 between expansion and non-expansion counties is the $ATT(2014)$. Similarly, time invariance of $Y_{i,t}(0)$ among expansion counties (plus the fact that we ruled out anticipatory behavior) implies that the counterfactual equals 2013 mortality rates in expansion counties ($\mathbb{E}_\omega[Y_{i,2013}(0)|D_i = 1]$). Under this assumption, which essentially rules out non-treatment-related changes in the outcome variable, the "time trend" in average mortality in expansion counties is the $ATT(2014)$.

DiD comes from an alternative assumption that identifies the relevant counterfactual even when the average untreated potential outcome differs across treatment groups (which violates mean independence) and changes over time (which violates time invariance). The parallel trends assumption states that in the absence of treatment, the average outcome evolution is the same among treated and comparison groups. For general assumptions and results, we denote time periods by $t = 1, 2$, but continue to be explicit about which years are being used when we reference the Medicaid example.

Assumption PT (2×2 Parallel Trends)**.** The (weighted) average change of $Y_{i,t=2}(0)$ from $Y_{i,t=1}(0)$ is the same between treated and comparison groups; that is,

$$\mathbb{E}_\omega[Y_{i,t=2}(0)|D_i = 1] - \mathbb{E}_\omega[Y_{i,t=1}(0)|D_i = 1] = \mathbb{E}_\omega[Y_{i,t=2}(0)|D_i = 0] - \mathbb{E}_\omega[Y_{i,t=1}(0)|D_i = 0]. \quad (3.3)$$

If the parallel trends assumption holds, then it is easy to construct $\mathbb{E}_\omega[Y_{i,t=2}(0)|D_i = 1]$ from observable quantities—that is, to identify it:

$$\mathbb{E}_\omega[Y_{i,t=2}(0)|D_i = 1] = \mathbb{E}_\omega[Y_{i,t=1}|D_i = 1] + \big(\mathbb{E}_\omega[Y_{i,t=2}|D_i = 0] - \mathbb{E}_\omega[Y_{i,t=1}|D_i = 0]\big). \quad (3.4)$$

In the Medicaid example, assumption PT says that to calculate expansion counties' average 2014 mortality rate in a counterfactual world without Medicaid expansion, start with their average 2013 mortality rate and add the observed change in average mortality rates in non-expansion counties. Substituting (3.4) into the definition of $ATT(2014)$ and replacing potential outcomes with observed

outcomes using (3.1) gives the 2×2 DiD estimand, an expression for the target parameter in terms of four estimable *population* averages:

$$
ATT(2014) = \overbrace{\mathbb{E}_\omega[Y_{i,2014}|D_i = 1]}^{=\mathbb{E}_\omega[Y_{i,2014}(1)|D_i=1]} - \Big(\overbrace{\mathbb{E}_\omega[Y_{i,2013}|D_i = 1] + \big(\mathbb{E}_\omega[Y_{i,2014}|D_i = 0] - \mathbb{E}_\omega[Y_{i,2013}|D_i = 0]\big)}^{=\mathbb{E}_\omega[Y_{i,2014}(0)|D_i=1]} \Big)
$$

$$
= \underbrace{\big(\mathbb{E}_\omega[Y_{i,2014}|D_i = 1] - \mathbb{E}_\omega[Y_{i,2013}|D_i = 1]\big)}_{\text{(weighted) average change for } D_i=1} - \underbrace{\big(\mathbb{E}_\omega[Y_{i,2014}|D_i = 0] - \mathbb{E}_\omega[Y_{i,2013}|D_i = 0]\big)}_{\text{(weighted) average change for } D_i=0}.
$$

$$(3.5)$$

Equation (3.5) highlights what makes DiD so attractive. It is intuitive, it has very mild data requirements (just four means), it answers *ex post* questions like "what did the treatment do?", and its identifying assumption can be stated precisely.

Parallel trends makes DiD distinct from causal designs that are based on statistical independence between treatment and potential outcomes. In designs like randomized trials or instrumental variables, the conditions—mean equalities across groups, for instance—that identify counterfactuals are often a statistical *consequence* of the randomness induced externally (Heckman, 2000). In contrast, parallel trends is just a restriction on untreated potential outcome trends. It does not necessarily come from exogenous variation "outside the model." In fact, because treatment adoption is often chosen by economic actors or policymakers "inside the model," parallel trends need not hold. For this reason, DiD analyses (correctly) devote significant attention to evaluating parallel trends.

We discuss how to generate empirical evidence about the plausibility of parallel trends below, but new theoretical findings about treatment choice behaviors or selection mechanisms also inform the plausibility of parallel trends.[4] These results explicitly connect DiD to behaviors, they provide grounding for stories about why a given DiD analysis is internally valid (or not), and they discipline empirical tests of parallel trends. A full review of this literature is outside the scope of this paper, but some helpful broad themes emerge.[5] For instance, a trade-off exists between the information the agents who choose treatment know and how they act on it, and the time-series properties of untreated potential outcomes. At one extreme, consider someone who knows $Y_{i,t}(0)$ before and after treatment and can opt into or out of treatment based on it. Parallel trends can only hold in this case if, other than common shifts for every unit, $Y_{i,t}(0)$ is constant (Ghanem et al., 2022). In our Medicaid example, neither of these conditions—that state legislatures in 2013 knew their 2014 untreated mortality rates or that untreated mortality rates would all have shifted in parallel—is plausible. The opposite extreme is when treatment timing is random, in which case parallel trends holds without any time-series restrictions. Yet, in this case, DiD is not necessary, and more efficient estimators exist (Roth and Sant'Anna, 2023a). Against the political and economic backdrop of

---

[4]In fact, some of the earliest economic research on DiD methods examined exactly these questions (Ashenfelter and Card, 1985; Heckman and Robb, 1985).

[5]For details on selection models and outcome models that are consistent with parallel trends, see Ghanem, Sant'Anna and Wüthrich (2022) and Marx, Tamer and Tang (2024); Chabé-Ferret (2015) applies these arguments to earnings dynamics.

the early 2010s, it is clear that Medicaid expansion decisions were not random.

Therefore, in realistic scenarios, parallel trends only holds if *some* restrictions on the way untreated outcomes enter the treatment selection mechanism hold as well. As one example, imagine that treatment selection depends on the permanent component of $Y_{i,t}(0)$ (fixed effects) but not on shorter-term fluctuations ("shocks"). For instance, if state legislatures knew and considered their long-run mortality levels only when making their expansion decision, they would be following this kind of selection mechanism. Expansion and non-expansion states might then have large differences in the permanent part of untreated outcomes, which cancel in equation (3.3), and so parallel trends would hold if shocks to $Y_{i,t}(0)$ had a time invariant mean conditional on the fixed effects.[6] State legislatures, however, may have also known whether their 2013 mortality rates were especially high or low when considering expanding Medicaid. If the expansion choice is related to these 2013 mortality shocks as well, parallel trends would hold only if stronger time-series restrictions on $Y_{i,t}(0)$ hold. Ghanem et al. (2022) provide a fuller discussion of the selection/time-series trade-off and theory-driven templates to assess parallel trends, while Marx et al. (2024) discuss economic models that are and are not compatible with parallel trends.

Another implication of the fact that DiD does not rely on statistical independence between $Y_{i,t}(0)$ and treatment status is that there is no guarantee that parallel trends holds across different transformations of $Y_{i,t}(0)$. As stated, it is simply an assumption about averages for a particular $Y_{i,t}(0)$. Roth and Sant'Anna (2023b) show that parallel trends is insensitive to functional form if and only if it holds between groups and across the distribution of $Y_{i,t}(0)$. This can only be true if Medicaid adoption is random, the mortality distribution is constant between 2013 and 2014, or a combination of the two cases. As these conditions are arguably ex-ante implausible, our DiD analysis may depend on our choices to measure $Y_{i,t}$ in rates (deaths per 100,000) as opposed to logs, for example. One approach to this measurement choice is to propose and evaluate a theory that delivers it, though we recognize that this is not always possible. To assess whether cases where parallel trends holding for one functional form come at the cost of ruling other transformations, we recommend that researchers use the Roth and Sant'Anna's (2023b) falsification tests for the null that parallel trends are insensitive to functional form. In our application, we do not reject the null that parallel trends is insensitive to functional form, with p-values above 0.80.

The interplay between treatment selection and the properties of the outcome variable characterize the structural basis for a DiD analysis (see DiNardo and Lee, 2011) and engaging with them is essential to any DiD application. While every study will have its own institutions, choices, and outcomes to consider, a rigorous DiD analysis must provide a transparent discussion about the reliability of the underlying identification assumptions. If parallel trends requires implausible behavioral restrictions, one may be better off using an alternative research design.

---

[6]Some researchers may find easier to understand these as "parallel changes" rather than "parallel trends." However, the use of "parallel trends" is now firmly established in the literature, and other influential work has used "changes-in-changes" to refer to an alternative estimator to classical DiD estimation (Athey and Imbens, 2006). To avoid confusion, we use "parallel trends" throughout this paper.

3.3 Estimation and inference: Four means or one regression?

Mapping the DiD estimand in equation (3.5) to the canonical 2×2 DiD estimator follows immediately from replacing population expectations with their sample analogs:

$$\widehat{ATT}(2014) = (\overline{Y}_{\omega,D=1,t=2014} - \overline{Y}_{\omega,D=1,t=2013}) - (\overline{Y}_{\omega,D=0,t=2014} - \overline{Y}_{\omega,D=0,t=2013}), \qquad (3.6)$$

where $\overline{Y}_{\omega,D=g,t=t'} = \dfrac{\sum_{i=1}^{n} \mathbf{1}\{D_i = g, t = t'\}\omega_i Y_{i,t'}}{\sum_{i=1}^{n} \omega_i \mathbf{1}\{D_i = g, t = t'\}}$ is the ω-weighted sample mean of Y for treatment group g in period t'. Equation (3.6) is the classic difference of two differences written in terms of sample means. It is a direct recipe for actually estimating $ATT(t)$ and can be read directly from the following table of average mortality rates in 2013 and 2014 by expansion group.

Table 2: Simple 2×2 DiD

	Unweighted Averages			Weighted Averages		
	Expansion	No Expansion	Gap/DiD	Expansion	No Expansion	Gap/DiD
2013	419.2	474.0	*-54.8*	322.7	376.4	*-53.7*
2014	428.5	483.1	*-54.7*	326.5	382.7	*-56.2*
Trend/DiD	*9.3*	*9.1*	0.1	*3.7*	*6.3*	-2.6

This table reports average county-level mortality rates (deaths among adults aged 20-64 per 100,000 adults) in 2013 (top row) and 2014 (middle row) in states that expanded adult Medicaid eligibility in 2014 (columns 1 and 4) and states that have not expanded by 2019 (columns 2 and 5). The first three columns present unweighted averages and the second three columns present population-weighted averages. Columns 1, 2, 4, and 5 in the third row show time trends in mortality between 2013 and 2014 for each group of states. The first two rows of columns 3 and 6 show the cross-sectional gap in mortality between expansion and non-expansion states in 2013 and 2014. The entries in red text in the bottom row show the simple 2×2 difference-in-differences estimates without weights (column 3) and with them (column 6)

The two across-time changes in equation (3.6) are in the third row of the table. Without weighting, average county-level mortality rates in expansion states rose by 9.3 deaths per 100,000 and 9.1 deaths in non-expansion states, so after rounding, the DiD estimate of $ATT(2014)$ is 0.1 deaths per 100,000. This result implies that the average treatment effect of Medicaid expansion on mortality in 2014 among *counties* that are part of an expansion state was an increase of 0.1 deaths per 100,000. In contrast, the DiD result using population weights suggests that Medicaid expansion caused a reduction of 2.6 deaths per 100,000 for the average *adult* in expansion states.[7]

The same result can be obtained as the (weighted) least squares estimate of $\beta^{2 \times 2}$ in the following linear regression specification (which only uses data from $t = 2013$ and $t = 2014$):

$$Y_{i,t} = \beta_0 + \beta_1 \mathbf{1}\{D_i = 1\} + \beta_2 \mathbf{1}\{t = 2014\} + \beta^{2 \times 2}(\mathbf{1}\{D_i = 1\} \times \mathbf{1}\{t = 2014\}) + \varepsilon_{i,t}, \qquad (3.7)$$

where β's are unknown coefficients and $\varepsilon_{i,t}$ is an idiosyncratic term uncorrelated with D_i. To see why, let's focus on the unweighted case ($\omega_i = 1$) and write each of the four means in $\widehat{ATT}(2)$ in

---

[7]Columns 3 and 6 show cross-group gaps in average mortality in each year. These can also be used to construct the DiD estimate by rearranging equation (3.6): $(\overline{Y}_{D=1,t=2014} - \overline{Y}_{D=0,t=2014}) - (\overline{Y}_{D=1,t=2013} - \overline{Y}_{D=0,t=2013})$.

terms of the estimated coefficients from (3.7):

- Sample average of $Y_{i,t}$ in post-period for treatment group is $\overline{Y}_{D=1,t=2014} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}^{2\times2}$.

- Sample average of $Y_{i,t}$ in pre-period for treatment group is $\overline{Y}_{D=1,t=2013} = \hat{\beta}_0 + \hat{\beta}_1$.

- Sample average of $Y_{i,t}$ in post-period for comparison group is $\overline{Y}_{D=0,t=2014} = \hat{\beta}_0 + \hat{\beta}_2$.

- Sample average of $Y_{i,t}$ in pre-period for comparison group is $\overline{Y}_{D=0,t=2013} = \hat{\beta}_0$.

Substituting these expressions into the definition of $\widehat{ATT}(2)$ yields:

$$\widehat{ATT}(2014) = \left[(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}^{2\times2}) - (\hat{\beta}_0 + \hat{\beta}_1)\right] - \left[(\hat{\beta}_0 + \hat{\beta}_2) - \hat{\beta}_0\right] = \hat{\beta}^{2\times2}.$$

Table 3 demonstrates this equivalence for both unweighted (column 1) and weighted (column 4) regressions. In fact, with balanced panel data, the estimate of $\beta^{2\times2}$ is numerically the same if the regression instead contains fixed effects for each unit (columns 2 and 5) or if one regresses outcome changes on a constant and the treatment group dummy D_i (columns 3 and 6).[8] Provided that one adopts the same notion of uncertainty, standard errors also coincide; such as when one clusters the standard errors at the county level.

Table 3: Regression 2 × 2 DiD

	Unweighted			Weighted		
	Crude Mortality Rate		Δ	Crude Mortality Rate		Δ
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	474.0***		9.1***	376.4***		6.3***
	(4.3)		(2.6)	(7.6)		(1.1)
Medicaid Expansion	-54.8***			-53.7***		
	(6.3)			(11.5)		
Post	9.1***			6.3***		
	(2.6)			(1.1)		
Medicaid Expansion × Post	0.1	0.1	0.1	-2.6*	-2.6*	-2.6*
	(3.7)	(3.7)	(3.7)	(1.5)	(1.5)	(1.5)
County fixed effects	No	Yes	No	No	Yes	No
Year fixed effects	No	Yes	No	No	Yes	No

This table reports the regression 2 × 2 DiD estimate comparing counties that expanded Medicaid in 2014 with counties that did not expand Medicaid by 2019, using only data for the years 2013 and 2014. Columns 1-3 report unweighted regression results, while columns 4-6 weight by county population aged 20-64 in 2013. Columns 1 and 4 report results from regressing the crude mortality rate for adults ages 20-64 on indicators for expansion states (Treat) and post-expansion year (Post); the DiD estimate is the coefficient on the interaction term. Columns 2 and 5 report the corresponding results for the interaction term using county and year fixed effects. Finally, Columns 3 and 6 report the results of the long difference in county mortality rates on a treatment indicator. Standard errors (in parentheses) are clustered at the county level.

The equivalence between calculating a 2 × 2 DiD by hand or with a regression has appealing features. Regressions are simple to run, and they do the averaging and differencing behind the

---

[8]When population weights vary over time, the equivalence between a by-hand DiD estimate and one that comes from a regression with unit fixed effects no longer holds.

scenes. They also allow the use of statistical inference tools from ordinary least squares (OLS), which are themselves the subject of a large econometrics literature that is particularly important when it comes to standard error estimation (Wooldridge, 2003; Bertrand, Duflo and Mullainathan, 2004; Donald and Lang, 2007; Cameron, Gelbach and Miller, 2008; Conley and Taber, 2011; Abadie, Athey, Imbens and Wooldridge, 2020, 2023).

Many inference procedures exist for DiD-type analyses, arising from a combination of choices about the target parameter, details of the data structure and sampling process, and maintained assumptions about the structure of outcomes. In practice, one needs to determine and discuss the forms of uncertainty the standard errors are designed to capture—that is, what is (conceptually) being resampled and what may or may not vary across those resamples. As discussed in Abadie et al. (2020), these details come from the nature of the parameter of interest—whether the focus is on sample-specific average treatment effects or population-level average treatment effects—and the stochastic elements of the model that make the estimator random. Heuristically, this involves a thought experiment (or stochastic process) hypothesized to generate the random components of the model (or the data-generating process). Different inferential frameworks highlight different sources of uncertainty by resampling distinct model components and treating other components as fixed (non-random).

Inferential frameworks on two extremes help cement these concepts. Design-based frameworks treat potential outcomes and covariates as non-random, focus on finite-population parameters (e.g., sample average treatment effects), and consider the allocation of treatment as the only source of the randomness in the model (Imbens and Rubin, 2015).[9] The only thing that is random and thus varies across the hypothetical resamples from this point of view is the treatment allocation. On the other hand, a traditional sampling-based approach to inference presumes that we independently sample units from a superpopulation. In this case, it is customary to focus on population parameters (like $ATT(2)$), treat all variables in the model as random variables, and compute standard errors by clustering at the level in which the (hypothesized) sampling was conducted. In this framework, every variable in the analysis—outcomes, covariates, and treatment—is randomly redrawn across the hypothetical resamples. A drawback of the sampling approach is that sometimes, it is unnatural to think of the data as a random sample from a well-defined population.

A third popular approach to inference—the model-based approach—is more structural and involves taking a stand on the structure of the error component of the model (e.g., imposing a putative model for how shocks affect outcomes and their relationship with treatment and other variables in the model). The uncertainty reflected in this model-based setting entails a thought experiment in which different values of these shocks and the other random variables in the model are drawn from their joint distribution (Abadie et al., 2023). This model-based approach is common in

---

[9]Traditionally, design-based inference procedures are justified when treatment assignment is fully random, which is a much stronger requirement than parallel trends. See Rambachan and Roth (2024) for a discussion on design-based inference for quasi-experimental designs, including a discussion of the Medicaid expansion.

econometrics, and it almost always takes the linear regression specification (or model, in this case) as the starting point of the analysis. Although this is often convenient, it is important to note that imposing model restrictions on the error component of the model necessarily imposes restrictions on treatment effect heterogeneity and on the relationship between potential outcomes; see Section 5 and Appendix A of Roth et al. (2023) for a discussion. Another challenge with the model-based approach is that it is hard to use this framework when adopting estimation strategies other than linear regressions—for example, inverse probability weighting or doubly robust procedures—which we will discuss in Session 4.4.

Ultimately, as the discussion above highlights, each approach has pros and cons, and discussions about the best way to compute standard errors are complex and intrinsically context-specific. A detailed treatment of the topic is outside the scope of this paper. It requires information about the sampling process, the research design and target parameters, what is treated as fixed and random, and the structure of the error components of their models (e.g., presence of spatial or serial correlation), among other factors. We refer interested readers to Abadie et al. (2020, 2023) and Section 5 of Roth et al. (2023) for discussions on these topics, though we also emphasize that further methodological research in this area is warranted. For the remainder of this article, we adopt a sampling perspective for uncertainty and cluster our standard errors at the county level. In our context, this is compatible with treating all variables as random, including treatment groups and potential outcomes. It also allows us to avoid (a) making time-series dependence restrictions on potential (and realized) outcomes—as we are in a short-panel framework with a large number of units and a fixed number of time periods—and (b) taking an explicit stand on the structure of error components of the model, which is particularly appealing as the starting point of our analysis is potential outcomes rather than regression models. It is also worth mentioning that as our treatment in the empirical example is assigned at the state level, clustering at the county level would also be compatible with treating state-specific shocks as fixed (or conditioned on) and assessing if they lead to violations of parallel trends (Roth et al., 2023, Section 5.1). Clustering at the state level would be justified if we were using a design-based perspective (Rambachan and Roth, 2024), though that would require us to treat potential outcomes as fixed (which we do not do in this paper). Our choice of inference procedure is not without controversies, and other inferential approaches may also be rationalized.

We conclude this section by stressing that the appeal of using regressions like (3.7) to estimate ATT in DiD designs comes from the fact that their numerical equivalence to the "by-hand" DiD estimator (3.6), which was explicitly derived from the $ATT(t)$ and the parallel trends assumption. This ensures that the regression specification respects the underlying identifying assumptions and estimates the desired target parameter. Unfortunately, the tight connection between (TWFE) regressions and DiD designs breaks under more complex setups that are ubiquitous in practice. We now turn to some of these issues and how approaching them from the point of view of 2×2 building blocks can guide good econometric practices.

4 Incorporating covariates into 2×2 DiD

So far, we have focused on 2×2 DiD designs that do not leverage any information about covariates, but researchers frequently incorporate them into DiD analyses in one of three ways: checking for balance in variables thought to influence $Y_{i,t}(0)$, controlling for those variables in the main estimates, and estimating treatment effect heterogeneity. For example, in the absence of Medicaid expansion, mortality rates likely would have evolved differently in poorer and richer counties; they certainly did before 2014 (Currie and Schwandt, 2016). Therefore, parallel trends may fail if poverty rates differ between expansion and non-expansion counties. If they do, then one may want to "control for" poverty rates when estimating *ATT* parameters. Finally, because Medicaid expansion reached more people in higher-poverty counties, its average effects on overall mortality may be larger there. This heterogeneity may be of interest in its own right or may be used to assess the plausibility of the overall DiD design.

This section discusses how to use auxiliary information on covariates to evaluate parallel trends, identify *ATT* parameters under potentially weaker assumptions, and study heterogeneity. These approaches stem from viewing a DiD with covariates in terms of 2×2 building blocks that themselves condition on those variables, which creates a clear link to the assumptions, estimators, and interpretation of unconditional 2×2 designs. We also discuss how regression estimators that control for covariates impose extra assumptions and can fail to identify *ATT* parameters.

4.1 Covariate balance: Is unconditional parallel trends plausible?

Assumption PT is fundamentally untestable, as it contains an unobserved counterfactual component. Therefore, "tests" of these assumptions are necessarily indirect and rely on other *observed* variables thought to be related to untreated potential outcome trends. For example, during the 2010s, demographics and economic conditions were strongly correlated with mortality levels and trends. If these relationships would have held in the absence of Medicaid expansion, and if expansion and non-expansion counties differ in those demographic and economic characteristics, then the parallel trends assumption (3.3) may fail to hold. Checking balance in observable determinants of changes in $Y_{i,t}(0)$ is thus a common and sensible way to evaluate parallel trends.

Most DiD analyses check for balance across groups in baseline covariate levels ($\mathbb{E}_\omega[X_{i,t=1}|D_i = 1] - \mathbb{E}_\omega[X_{i,t=1}|D_i = 0]$) or covariate trends before and after treatment ($\mathbb{E}_\omega[\Delta X_{i,t=2}|D_i = 1] - \mathbb{E}_\omega[\Delta X_{i,t=2}|D_i = 0]$, where $\Delta X_{i,t=2} = X_{i,t=2} - X_{i,t=1}$). We consider the following variables, $X_{i,t}$: the percentages of a county's population that are female, white, or Hispanic; the unemployment rate; the poverty rate; and county-level median income (in thousands of dollars).[10] The top panel of Table 4 reports averages of these variables by group in 2013 with and without population

---

[10]We focus on these variables for convenience. Borgschulte and Vogler (2020) use a LASSO procedure that selects more and different covariates to include in their analysis. We replicate their findings when we follow their methodology but diverge here for the sake of brevity.

weights. We also report a measure of imbalance that is comparable across variables: the normalized difference in means between the treatment and comparison groups (Imbens and Rubin, 2015, Chapter 14),

$$\text{Norm. Diff}_\omega = \frac{\overline{X}_{\omega,T} - \overline{X}_{\omega,C}}{\sqrt{(S^2_{\omega,T} + S^2_{\omega,C})/2}},$$

where $\overline{X}_{\omega,T}$ and $\overline{X}_{\omega,C}$ are the sample weighted or unweighted averages for the treatment and comparison groups, respectively, and $S^2_{\omega,T}$ and $S^2_{\omega,C}$ are the sample weighted or unweighted variances of the covariates for the treatment and comparison group. As a general rule of thumb, values of the normalized difference in excess of 0.25 in absolute value indicate a potentially problematic imbalance between the two groups (Imbens and Rubin, 2015, page 277).[11]

Table 4: Covariate Balance Statistics

Variable	Unweighted			Weighted		
	Non-Adopt	Adopt	Norm. Diff.	Non-Adopt	Adopt	Norm. Diff.
2013 Covariate Levels						
% Female	49.43	49.33	-0.03	50.48	50.07	-0.24
% White	81.64	90.48	0.59	77.91	79.54	0.11
% Hispanic	9.64	8.23	-0.10	17.01	18.86	0.11
Unemployment Rate	7.61	8.01	0.16	7.00	8.01	0.50
Poverty Rate	19.28	16.53	-0.42	17.24	15.29	-0.37
Median Income	43.04	47.97	0.43	49.31	57.86	0.68
2014 - 2013 Covariate Differences						
% Female	-0.02	-0.02	0.00	0.02	0.01	-0.09
% White	-0.21	-0.21	0.01	-0.32	-0.33	-0.04
% Hispanic	0.20	0.21	0.04	0.25	0.33	0.29
Unemployment Rate	-1.16	-1.30	-0.21	-1.08	-1.36	-0.55
Poverty Rate	-0.55	-0.28	0.14	-0.41	-0.35	0.05
Median Income	0.98	1.11	0.06	1.10	1.74	0.32

This table reports the covariate balance between adopting and non-adopting states. In the top panel, we report the averages and standardized differences of each variable, measured in 2013, by adoption status. All variables are measured in percentage values, except for median household income, which is measured in thousands of U.S. dollars. In the bottom panel we report the average and standardized differences of the county-level long differences between 2014 and 2013 of each variable. We report both weighted and unweighted measures of the averages to correspond to the different estimation methods of including covariates in a 2 × 2 setting.

We find meaningful imbalance in several baseline measures. Expansion counties in 2013 were

---

[11]As discussed in Austin (2009, Section 3.2), the normalized difference was initially proposed in the psychological literature and is sometimes referred to as Cohens effect size index. In this context, normalized differences of 0.2, 0.5, and 0.8 are sometimes used to represent small, medium, and large imbalances; however, normalized differences as small as 0.1 are sometimes considered worrisome, depending on how important the covariate in question is. Ultimately, there is no universally accepted threshold for what value indicates important imbalances. See Ho, Imai, King and Stuart (2007) and Austin (2009) for additional discussions.

whiter and had higher unemployment rates despite lower poverty and higher median income than non-expansion counties. Because DiD uses *changes* in outcomes, researchers sometimes argue that the effect of pre-treatment variables is differenced out. This logic does not hold, though, if baseline covariates are related to untreated potential outcome trends themselves. The imbalance in the top panel of Table 4 will lead to violations of parallel trends to the extent that counties with different racial composition or income distributions would have had different mortality changes even without Medicaid expansion.

Nevertheless, checks of balance in covariate changes can be informative about parallel trends as well. The bottom panel of Table 4 reports average changes by group between 2013 and 2014 as well as normalized differences. Many of the imbalances evident in baseline levels change, or even flip signs, when measured in changes. Unemployment, for example, was higher in expansion states in 2013 but fell faster. To the extent that these changes are important determinants of $\Delta Y_{i,t}(0)$, then these results could suggest that Assumption PT is violated.

Why do we say "could"? A major challenge in interpreting cross-group gaps in $\Delta X_{i,t}$ involves deciding which variables are truly covariates and which are mechanisms/outcomes. If an element of $X_{i,t}$ cannot be affected by the treatment, it is a (strictly exogenous) covariate, and differential changes in exogenous covariates may indicate a PT violation. Since the treatment cannot have caused $X_{i,t}$ to change (by assumption), something else that differs across groups and over time must have. Since little research suggests an effect of Medicaid expansion on unemployment, this may be a good assumption. On the other hand, if Medicaid expansion can change the demographic and economic composition of its counties, then differential changes in these variables may actually be a consequence of the expansion itself.[12] If so, then differential post-treatment changes in them would not necessarily indicate a parallel trends violation; they could partially reflect a causal effect. As with the plausibility of Assumption PT itself, whether something is a covariate or a mechanism is not a data question per se. It requires context-specific knowledge about how the treatment works.

4.2 DiD with covariates: Identification under conditional parallel trends

Having detected covariate imbalance that casts doubt on Assumption PT, how should we proceed to estimate $ATT(2)$? Because the imbalance documented in Table 4 suggested that unconditional parallel trends may not hold, our goal is to develop a DiD identification strategy based on an assumption that accounts for this imbalance. Working from a conditional parallel trends assumption shows how to construct $ATT(2)$ from 2×2 comparisons that are each conditioned on specific covariate values, thus addressing the imbalance problem.

Let X_i be a vector of observed determinants of changes in $Y_{i,t}(0)$. Here, we purposefully

---

[12] In fact, comparing mean covariate changes in expansion and non-expansion is the same as using $X_{i,t}$ as the outcome in a 2×2 DiD estimator.

omit the time subscript on X_i because the covariates in this section can be time-invariant, such as fixed variations or baseline values $(X_{i,t=1})$, or time-varying in the sense of including values from the second period, $X_{i,t=2}$. The empirical content of a "new" identification assumption that incorporates X_i, henceforth conditional parallel trends (CPT) assumption, is formalized as follows.

Assumption CPT (2×2 Conditional Parallel Trends). The (weighted) average change of $Y_{i,t=2}(0)$ from $Y_{i,t=1}(0)$ is the same between treated and comparison units that share the same covariate values,

$$\mathbb{E}_\omega[Y_{i,t=2}(0) - Y_{i,t=1}(0)|X_i, D_i = 1] = \mathbb{E}_\omega[Y_{i,t=2}(0) - Y_{i,t=1}(0)|X_i, D_i = 0]. \tag{4.1}$$

Assumption CPT has the same structure as Assumption PT but states that PT holds within each covariate-specific stratum rather than across the whole population. With respect to the baseline covariates in Table 4, this amounts to assuming parallel trends in $Y_{i,t}(0)$ between expansion and non-expansion counties that in 2013 had the same female, white, and Hispanic shares, as well as the same unemployment and poverty rates and median income. This does not restrict *how* $Y_{i,t}(0)$ changes in different covariate strata, nor is not necessary to estimate these trends. Assumption CPT only requires that covariate-specific trends are common between expansion and non-expansion counties.

For both expectations in Assumption CPT to be well-defined for all values of X_i, there must be both untreated and treated units in the population at each covariate value. If, for some covariate values, there are only treated units, for example, then the right-hand side of (4.1) is undefined. A formal statement of this assumption, called "common support" or "strong overlap," is as follows.[13]

Assumption SO (Strong overlap). The conditional (weighted) probability of belonging to the treatment group, given observed covariates X_i, which are determinants of untreated potential outcome growth, is uniformly bounded away from zero and one. That is, for some $\epsilon > 0$, $\epsilon < P_\omega[D_i = 1|X_i] < 1 - \epsilon$.

Assumptions CPT and SO allow us to identify the $ATT(2)$:

$$
\begin{aligned}
ATT(2) &= \mathbb{E}_\omega[Y_{i,t=2}(1)|D_i = 1] - \mathbb{E}_\omega[Y_{i,t=2}(0)|D_i = 1] \\
&= \mathbb{E}_\omega[Y_{i,t=2}|D_i = 1] - \mathbb{E}_\omega\left[\mathbb{E}_\omega[Y_{i,t=2}(0)|X_i, D_i = 1]\Big|D_i = 1\right] \\
&= \mathbb{E}_\omega[Y_{i,t=2}|D_i = 1] - \mathbb{E}_\omega\left[\mathbb{E}_\omega[Y_{i,t=1}(0)|X_i, D_i = 1] + \mathbb{E}_\omega[\Delta Y_{i,t=2}(0)|X_i, D_i = 0]\Big|D_i = 1\right] \\
&= \mathbb{E}_\omega[\Delta Y_{i,t=2}|D_i = 1] - \mathbb{E}_\omega\left[\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]\Big|D_i = 1\right]. \tag{4.2}
\end{aligned}
$$

The first line restates the definition of $ATT(2)$, and the second line uses the law of iterated expectations to write the counterfactual mean for the whole treatment group as an average of

---

[13]$ATT(2)$ is still identified under conditional parallel trends if some values of X_i have only untreated observations. We require $P_\omega[D_i = 1|X_i]$ to be bounded away from zero and one to avoid irregular inference procedures; see Khan and Tamer (2010) for additional details.

counterfactual means conditional on X_i. These quantities are exactly the ones that appear in the conditional parallel trends assumption (and the overlap condition). Therefore, as in section 3, the third line uses Assumption CPT to rewrite the conditional counterfactuals in terms of observable population quantities under no anticipation (Assumption NA). Equation (4.2) then uses the law of iterated expectations again to group terms and express $ATT(2)$ in terms of observed variables $(Y_{i,t=2}, Y_{i,t=1}, G_i, X_i)$; that is, it establishes that $ATT(2)$ is nonparametrically identified under our assumptions. This expression has a clear intuition: the $ATT(2)$ is equal to the path of outcomes experienced by the treated group (the term on the left) minus the average path of outcomes in the comparison group for each value of the covariates, averaged over the treated group's distribution of covariates (the term on the right).

4.3 DiD estimation with covariates: TWFE

Moving from the population identification result in equation (4.2) to sample analogs is a challenge unless the covariates are discrete and the conditional expectations themselves are easily calculable. This difficulty does not arise in an unconditonal DiD. With continuous covariates, or many discrete ones, it may not be feasible to construct $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$. Conditional DiD estimation, therefore, uses additional econometric techniques to bridge this gap. We begin, however, by discussing how regression DiD estimators that include covariates relate to the assumptions used for identification in (4.2).

Because the TWFE specification in (3.7) recovers the $ATT(2)$ in 2×2 DiD setups without covariates, it is natural to extend this logic to regressions with covariates. Indeed, this is by far the most popular approach adopted by practitioners, arguably because it is both easy and familiar. A typical regression specification is

$$Y_{i,t} = \theta_t + \eta_i + \beta_{treat}D_{i,t} + X'_{i,t}\beta_{covs} + e_{i,t}. \tag{4.3}$$

where the unit and time fixed effects, treatment status, and covariates have already been defined, $e_{i,t}$ is an error term, and β_{treat} is interpreted as the parameter of interest. A related specification explicitly controls for baseline covariates by replacing $X_{i,t}$ with interactions of the pre-treatment covariates $X_{i,t=1}$ and a post-treatment dummy,

$$Y_{i,t} = \theta_t + \eta_i + \beta_{treat,2}D_{i,t} + (\mathbf{1}\{t = 2\}X_{i,t=1})\beta_{covs,2} + e_{i,t}, \tag{4.4}$$

In Table 5, we report the OLS and weighted least squares estimates of the unconditional 2×2 DiD estimate, β_{covs} from (4.4), $\beta_{covs,2}$ from (4.3), and their cluster-robust standard errors, using the covariates from Table 4.

Table 5: Regression 2 × 2 DiD with Covariates

| | Unweighted | | | Weighted | | |
| | No Covs | $X_{i,t=2013}$ | $X_{i,t}$ | No Covs | $X_{i,t=2013}$ | $X_{i,t}$ |
	(1)	(2)	(3)	(4)	(5)	(6)
Medicaid Expansion	0.12	-2.35	-0.49	-2.56*	-2.56	-1.37
	(3.75)	(4.29)	(3.83)	(1.49)	(1.78)	(1.62)

This table reports the regression 2 × 2 DiD estimate comparing counties that expanded Medicaid in 2014 with counties that did not, adjusting for the inclusion of covariates (percent female, percent white, percent hispanic, the unemployment rate, the poverty rate, and median household income). Columns 1-3 report unweighted regression results, while columns 4-6 weight by county population aged 20-64 in 2013. Columns 1 and 4 report results for expansion states without covariates, columns 2 and 5 adjust for the baseline levels of the covariates in 2013, and columns 3 and 6 control for the time-varying covariate values in 2014 and 2013. Standard errors (in parentheses) are clustered at the county level.

Although only one covariate-adjusted estimate in Table 5 is (marginally) statistically significant, the point estimates differ noticeably. In the unweighted case, adjusting for the 2013 levels of the covariates decreases the estimated effect of Medicaid expansion on short-run mortality rates from a point estimate of roughly 0.12 to -2.35. However, if we include their time-varying values instead, we estimate an effect of -0.49, a large difference. We find a similar result when using weighted regressions; while the coefficient remains constant (-2.56) when using 2013 values of the covariates, it attenuates to -1.37 if we use (4.3).

The jump from the conditional DiD identification result in (4.2) to the TWFE *estimators* in (4.3) and (4.4) skips a crucial question about β_{treat} or $\beta_{treat,2}$: do they equal the target parameter $ATT(2)$ under the conditional parallel trends assumption? It turns out that the close relationship between regression DiD, $ATT(2)$, and parallel trends in a design without covariates does not hold with covariates. The issues come from exactly what kinds of covariates are effectively being "controlled for" in these specifications and how the regression estimator combines outcome trends for covariate sub-groups.

Note that in our two-period setup, (4.3) and (4.4) are respectively equivalent to (with some abuse of notation),

$$\Delta Y_{i,t=2} = \alpha + \beta_{treat} D_i + \Delta X'_{i,t=2} \beta_{covs} + \Delta e_{i,t=2},$$

$$\Delta Y_{i,t=2} = \alpha + \beta_{treat,2} D_i + X'_{i,t=1} \beta_{covs,2} + \Delta e_{i,t=2}.$$

The first thing that is clear from these representations is that because time-invariant variables drop out of equation (4.3), a TWFE specification can account for differential trends related to baseline covariate levels only if they enter as interactions with the post-treatment dummy as in equation (4.4). The exact regression specification, therefore, determines the implied conditional parallel trends assumption. Controlling for annual poverty rates really means controlling for poverty

changes, and areas that are poor are not the same as areas that are becoming poor.

Another limitation evident in (4.3) relates to "bad controls." Whenever $X_{i,t=2}$ is affected by the treatment, then conditioning on it (in any way) can bias estimates of the $ATT(2)$. If Medicaid expansion lowered poverty rates, for example, then including 2014 poverty rates or the 2013-2014 change in poverty rates as a covariate is problematic. This echoes our discussion about testing balance in $\Delta X_{i,t}$ in the sense that time-varying covariates must be unaffected by treatment in order to interpret imbalance in their trends as a source of bias, and to be able to control for them to address that bias. See Caetano, Callaway, Payne and Rodrigues (2022) for a discussion.

Suppose we have decided on which variables to include in a conditional parallel trends assumption and whether to measure them in levels or changes. If Assumptions CPT and SO hold with respect to this set of covariates, does β_{treat} recover the $ATT(2)$? In the DiD context, Caetano and Callaway (2024) tackle exactly this question. They show that β_{treat} equals a weighted average of conditional average treatment effects, defined as $ATT_{x_k}(2) \equiv \mathbb{E}_\omega\big[Y_{i,t=2}(1) - Y_{i,t=2}(0)|D_i = 1, X_i = x_k\big]$, with weights that may not be convex, plus three bias terms reflecting misspecification either in the set of control variables or the fact that they are included linearly. These conclusions relate to recent findings about the properties of regression estimators in the presence of heterogeneity in other contexts, including instrumental variables (Mogstad and Torgovitsky, 2024), cross-sectional designs (Angrist, 1998; Aronow and Samii, 2015; Sloczynski, 2022; Goldsmith-Pinkham, Hull and Kolesár, 2024), and panel data (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfoeuille, 2020; Sun and Abraham, 2021; Poirier and Sloczynski, 2024).

The weighting results suggest that even when the covariates are correctly selected, measured, and added with the correct functional form, β_{treat} could be negative even when $ATT_{X_i}(2)$ is positive for all values of covariates. Short of this extreme sign-reversal case, β_{treat}'s weighting scheme generally does not yield the $ATT(2)$ target parameter and instead puts too much weight on $ATT_{X_i}(2)$ for X_i's that are relatively uncommon among the treated group relative to the untreated group and puts too little weight on $ATT_{X_i}(2)$ for X_i's that are relatively common among the treated group relative to the untreated group (Sloczynski, 2022; Caetano and Callaway, 2024).

Taken together, these results imply that β_{treat} identifies $ATT(2)$ under the *additional* assumption that treatment effects across covariate strata are constant. To see why, write the conditional ATT in period two given $\Delta X_{i,t=2}$ as

$$ATT_{\Delta X_{i,t=2}}(2) = \mathbb{E}_\omega[Y_{i,t=2}(1) - Y_{i,t=2}(0)|D_i = 1, \Delta X_{i,t=2}],$$

and note that, under Assumptions CPT and SO, it is identified by

$$ATT_{\Delta X_{i,t=2}}(2) = \mathbb{E}_\omega[\Delta Y_{i,t=2}|D_i = 1, \Delta X_{i,t=2}] - \mathbb{E}_\omega[\Delta Y_{i,t=2}|D_i = 0, \Delta X_{i,t=2}].$$

If we take (4.3) to be a correctly specified regression, then

$$ATT_{\Delta X_{i,t=2}}(2) = \big(\beta_{treat} + \Delta X'_{i,t=2}\beta_{covs}\big) - \big(\Delta X'_{i,t=2}\beta_{covs}\big) = \beta_{treat}.$$

In other words, (4.3) implicitly rules out that treatment effects can vary across covariate-strata,

which makes the weighting issues identified by Caetano and Callaway (2024) irrelevant to the interpretation of β_{treat}. Research on the Medicaid expansion using data on mortality rates by income, however, shows clear evidence of heterogeneous effects (Miller et al., 2021; Wyse and Meyer, 2024).

One way to avoid these limitations would be to make (4.4) (or (4.3)) more flexible by including interactions of the covariates with treatment group, time, and treatment-group-by-time. An alternative possibility is to adopt a "forward-engineering" perspective (Mogstad and Torgovitsky, 2024) and derive an estimator for $ATT(2)$ that directly leverages Assumptions NA, CPT and SO. In some situations, the forward-engineering approach will also use regressions. Still, the target parameter and the identifying assumptions will guide the specification we should use (and not the other way around).

4.4 DiD estimators with covariates that target the ATT(2)

Fortunately, TWFE is not the only way to bring covariates into DiD estimation. We now discuss alternative ways to use covariates in a DiD analysis that start from the identification result in (4.2):

$$ATT(2) = \mathbb{E}_\omega[\Delta Y_{i,t=2}|D_i = 1] - \mathbb{E}_\omega\Big[\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]\Big|D_i = 1\Big].$$

This equation provides an intuitive recipe for estimating $ATT(2)$. The first term in $ATT(2)$ is the same as in an unconditional 2×2 design and can be replaced by its sample analog as in equation (3.6): $(\overline{Y}_{\omega,D=1,t=2} - \overline{Y}_{\omega,D=1,t=1})$.

One way to obtain the second term is to first estimate the inner conditional expectation, $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$. This object is just an (unknown) function that relates average outcome trends for untreated units to their covariates. The most common way to proceed, especially in cases where X_i contains many variables or continuous ones, is to specify a working model, $\mu_{\omega,\Delta,D=0}(X_i)$, with parameters that are simple to estimate. A natural and empirically friendly choice is a linear model, $\mu_{\omega,\Delta,D=0}(X_i) = X_i'\beta_{D=0}$, whose parameters come from a (weighted) regression of $\Delta Y_{i,t=2}$ on X_i in the sample of untreated units. The results of this regression describe *untreated* outcome trends as a function of X_i. The fitted model then generates predicted values, $\widehat{\mu}_{\omega,\Delta,D=0}(X_i) = X_i'\widehat{\beta}_{D=0}$, for all units in the sample, including treated units. With these fitted values we can estimate $\mathbb{E}_\omega\Big[\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]\Big|D_i = 1\Big]$ using the plug-in principle; that is, by replacing $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$ with its fitted value $\widehat{\mu}_{\omega,\Delta,D=0}(X_i)$, and replacing population (weighted) expectations with their sample analogs: $\frac{\sum_{i=1}^n D_i\ \omega_i\ \widehat{\mu}_{\omega,\Delta,D=0}(X_i)}{\sum_{i=1}^n D_i\ \omega_i}$.

Putting the pieces together gives the following estimator for $ATT(2)$:

$$\widehat{ATT}_{ra}(2) = \frac{\sum_{i=1}^n D_i\ \omega_i\ \big(\Delta Y_{i,t=2} - \widehat{\mu}_{\omega,\Delta,D=0}(X_i)\big)}{\sum_{i=1}^n D_i\ \omega_i}. \tag{4.5}$$

This strategy is often referred to as the regression-adjustment (RA) or outcome regression approach

to DiD; see, for example, Heckman, Ichimura and Todd (1997a).

We apply this strategy in our application using only the baseline covariates from Table 4, and report the parameters $\widehat{\beta}_{D=0}$ of our working model in columns (1) and (3) of Table 6. In practice, mortality changes in non-expansion states are only weakly related to our baseline covariates. Multiplying the weighted coefficients in Table 6 times the weighted treatment group 2013 means in Table 4 gives a predicted change in untreated mortality rates for the average treated county of 7.2 deaths per 100,000, the second term in equation (4.5), compared to the observed (weighted) change among untreated counties of 6.3 deaths. The observed weighted trend in mortality for expansion counties from Table 2 is 3.7 deaths. Together, these results imply that this approach, based only on assumptions CPT, SO, NA, the choice of baseline covariates, and a linear model for $\mathbb{E}_{\omega}[\Delta Y_{i,t=2}|X_i, D_i = 0]$, yields an estimated $ATT(2014)$ of -3.5. This matches the formal RA DiD estimate we report in column 4 of Table 7 (labeled as "Regression"). Column 1, however, gives an unweighted estimate from the same procedure of -1.62.

Table 6: Outcome Regression and Propensity Score Models

	Unweighted		Weighted	
	Regression	Propensity Score	Regression	Propensity Score
	(1) OLS	(2) Logit	(3) OLS	(4) Logit
Constant	-20.91	-10.00***	-4.62	-8.17***
	(69.85)	(1.35)	(44.98)	(0.01)
% Female	0.04	-0.04**	-0.09	-0.19***
	(0.82)	(0.02)	(0.68)	(0.00)
% White	0.15	0.06***	0.20*	0.04***
	(0.22)	(0.00)	(0.11)	(0.00)
% Hispanic	-0.08	-0.02***	-0.08	-0.02***
	(0.20)	(0.00)	(0.08)	(0.00)
Unemployment Rate	1.14	0.32***	0.88	0.68***
	(1.56)	(0.03)	(0.99)	(0.00)
Poverty Rate	0.21	0.03*	-0.13	0.11***
	(0.98)	(0.02)	(0.53)	(0.00)
Median Income	0.09	0.08***	-0.05	0.15***
	(0.52)	(0.01)	(0.24)	(0.00)

This table reports the outcome regression propensity score models that enter into the estimator from Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021). The first two columns report the results for unweighted regressions and the second two report results from weighted regression models. The regression model predicts changes in the outcome variable (mortality rates) on the 2013 covariate values for just the counties that do not expand Medicaid in 2014. The propensity score model uses data for 2013 and estimates a logit model of an expansion indicator variable on the 2013 covariate levels. Standard errors (in parentheses) are clustered at the county level.

To compute standard errors, we need to account for the fact that (4.5) is a two-step estimation procedure and take into account the uncertainty associated with estimating the working model

$\mu_{\omega,\Delta,D=0}(X_i)$. This is standard, though, and most statistical software automates this process; see, for example, Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021).

Table 7: DiD Estimates with Covariates

	Unweighted			Weighted		
	Regression	IPW	Doubly Robust	Regression	IPW	Doubly Robust
Medicaid Expansion	-1.62	-0.86	-1.23	-3.46	-3.84	-3.76
	(4.73)	(4.76)	(4.61)	(2.29)	(3.22)	(3.59)

This table reports the 2 × 2 DiD estimate comparing counties that expand Medicaid in 2014 to counties that do not expand Medicaid, adjusting for the inclusion of 2013 covariate values using the methodologies discussed in Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021). The first column reports results using regression adjustment, the second column uses inverse probability weighting based on a propensity score model using the included covariates, and the third column uses the doubly robust combination of the two approaches. Standard errors (in parentheses) are clustered at the county level.

A linear working model for $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$ is a familiar choice, but not the only one. More flexible, or even fully nonparametric, working models are possible, and the procedure is the same: estimate the model on untreated units, get its fitted values for the covariate values of the treated units, and then estimate the $ATT(2)$ using (4.5). An important consideration when choosing the working models is sample size. Large samples permit more flexible estimators that do not sacrifice too much precision. In smaller samples, a parametric linear model may be more appealing. Ultimately, the reliability of DiD RA estimators for $ATT(2)$ depends on how well $\hat{\mu}_{\omega,\Delta,D=0}(X_i)$ approximates $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$. If the working model is misspecified—for example by omitting relevant nonlinear terms—the resulting DiD RA estimator will be biased.

Table 6 showed that our covariates did not strongly predict untreated mortality trends and thus do little to change our potential biased unconditional DiD estimates. However, $ATT(2)$ can be estimated conditional on covariates in a different way without needing to specify which variables determine outcome trends. Instead, one can improve the comparability of the comparison group directly by selecting a model for the conditional probability of being treated and applying an inverse probability weighted (IPW) DiD procedure (Abadie, 2005). The logic of IPW builds on the balance checks we conducted in Table 4: if imbalance in covariates is the source of parallel trends violations, then adjusting the comparison group to be balanced on covariates can address that bias. The adjustment takes the form of re-weighting the observed changes in adult mortality rates for non-expansion counties to ensure that the expansion and non-expansion counties are similar on covariates, thus addressing the compositional source of bias.

To implement the IPW DiD procedure and construct these "balancing weights," we need to model $p_\omega(X_i) = P_\omega(D_i = 1|X_i)$, the (weighted) conditional probability of belonging to the treatment group, known as the "propensity score." IPW weights are a function of $p_\omega(X_i)$, and for estimating $ATT(2)$, they take a form that forces the underlying weighted distribution of covariates

for comparison units to match the distribution for treatment units (Rosenbaum and Rubin, 1983). Intuitively, these weights are formed so that if we find some units that were likely to be observed in the treatment groups (based on their covariate values) but ended up in the comparison group, we give these untreated observations "extra" weight.

Formally, we can show that, under Assumptions CPT and SO, when panel data are available,

$$ATT(2) = \mathbb{E}\left[\left(w_{\omega,D=1}(D_i) - w_{\omega,D=0}(D_i, X_i)\right)\Delta Y_{i,t=2}\right], \tag{4.6}$$

where

$$w_{\omega,D=1}(D) = D\omega \Big/ \mathbb{E}[\omega D], \text{ and } w_{\omega,D=0}(D, X) = \frac{\omega(1-D)p_\omega(X)}{1-p_\omega(X)} \Big/ \mathbb{E}\left[\frac{\omega(1-D)p_\omega(X)}{1-p_\omega(X)}\right]; \tag{4.7}$$

see, for example, Abadie (2005) and Sant'Anna and Zhao (2020). The structure of the weights in equation (4.7) and the way they enter equation (4.6) highlight several intuitive features of how the IPW estimator works. First, $w_{\omega,D=1}(D)$ is only non-zero for treated units, and $w_{\omega,D=0}(D, X)$ is only non-zero for untreated units, which means the estimator subtracts a particular mean of outcome trends for untreated units from a particular mean of outcome trends for treated units. Second, the $w_{\omega,D=1}(D)$ weights do not involve X_i and simply lead simply to a (ω-weighted) mean for the treatment group. Third, the $w_{\omega,D=0}(D, X)$ weights are functions of X_i that, as (4.7) shows, give increasingly more weight to untreated units with high propensity scores. This specific weighting function is what allows IPW weights to build a comparison group whose covariate distribution matches the treatment group. Fourth, equation (4.7) clarifies that two types of weights both enter an IPW analysis. We already discussed how the ω weights shape the parameter of interest, parallel trends assumptions, and estimation. These simply multiply the IPW weights which act to address imbalance (and their product is rescaled to integrate to one within group). Finally, the appeal of (4.6) is that if we have a better sense of how units sort into treatment than of the factors that shape outcome trends, we may be more comfortable modeling $p_\omega(X_i)$ than $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$.

To leverage the characterization of the $ATT(2)$ in (4.6) for estimation and inference purposes, we need a working model for the true propensity score $p_\omega(X_i)$. When X's are all discrete and low dimensional, this is simple and does not involve functional form restrictions: create covariate-specific strata and then, within each strata, compute the proportion of treated units, and call these estimates $\widehat{\pi}(x_k)$, x_k being a strata-indicator. When X's have continuous components, or when there are too many strata relative to the available sample size, one can adopt a flexible working model, $\pi_\omega(X_i)$, for the propensity score. A common choice for $\pi_\omega(X_i)$ is a (weighted) logistic model whose parameters can be estimated using maximum likelihood—or an alternative estimation procedure such as inverse probability tilting (Graham, Pinto and Egel, 2012). We follow this strategy in our Medicaid application and report in columns (2) and (4) of Table 6 the unweighted and weighted maximum likelihood logit coefficients from our propensity score model. Our covariates appear to explain expansion decisions better than untreated outcome trends, suggesting that an estimation approach based on propensity scores may change our $ATT(2014)$ estimate more than the RA

approach did. We then use these logit coefficients to get the fitted values for all observations, $\widehat{\pi}_\omega(X_i)$, that will serve as our estimates of $p_\omega(X_i)$.

From these fitted values, we can estimate $ATT(2)$ by

$$\widehat{ATT}_{ipw}(2) = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{w}_{\omega,D=1}(D_i) - \widehat{w}_{\omega,D=0}(D_i, X_i) \right) \Delta Y_{i,t=2}, \tag{4.8}$$

where

$$\widehat{w}_{\omega,D=1}(D) = D\omega \Big/ \frac{1}{n} \sum_{i=1}^{n} \omega_i D_i, \quad \widehat{w}_{\omega,D=0}(D, X) = \frac{\omega(1-D)\widehat{\pi}_\omega(X)}{1-\widehat{\pi}_\omega(X)} \Big/ \frac{1}{n} \sum_{i=1}^{n} \frac{\omega_i(1-D_i)\widehat{\pi}_\omega(X_i)}{1-\widehat{\pi}_\omega(X_i)}. \tag{4.9}$$

We report the $ATT(2014)$ estimates and their standard errors using this IPW DiD procedure in Table 7 (labeled as "IPW") using both unweighted and population-weighted procedures, where we use a logistic regression that is linear in covariates as our propensity score working model. We use the delta method to account for the estimation uncertainty inherited in this two-step estimation procedure when computing standard errors. The IPW DiD estimates are generally similar to the RA DiD estimates in that we obtain negative estimates that are larger in magnitude when using population weights. However, despite neither being statistically significant, the unweighted IPW estimate is less than half the size of the RA estimate. This is consistent with the broad conclusion from Table 6 that our covariates explain Medicaid expansion better than mortality trends.

IPW estimators for $ATT(2)$ tend to be noisy when fitted propensity score estimates are too close to 1 among untreated units, a condition related to the Assumption SO. The reason for this is simple: when $\widehat{\pi}_\omega(X_i)$ is close to one among units with $D_i = 0$, the (estimated) inverse probability weights $\widehat{w}_{\omega,D=0}(D, X)$ become more volatile, as one is essentially dividing by zero. A good practice when using IPW estimators is to check the plausibility of strong overlap using estimated propensity scores. Figure 1 plots the propensity scores that come from the logit models in Table 6 for the two groups of counties. Few untreated units have very high estimated propensity scores, so extreme weighting is not a significant concern.[14] In addition, propensity scores of non-expansion counties seem to lie within the support of the expansion counties' propensity scores, supporting strong overlap. Trimming high- or low-propensity score observations from the sample may be warranted when overlap is weak; for a discussion, see, for example, Crump, Hotz, Imbens and Mitnik (2009), Sasaki and Ura (2022), and Ma, Sant'Anna, Sasaki and Ura (2023).

The reliability of these two approaches to estimating conditional DiD designs centers on selecting good models for two different functions: the conditional expectation of untreated outcome changes and the true, unknown propensity score $p_\omega(X_i)$. So which approach should one pick? In practice, there are major advantages to combining both in a way that leads to estimators for the $ATT(2)$ that are more robust against model misspecification (Sant'Anna and Zhao, 2020). This is the so-called doubly robust (DR) approach, sometimes referred to as the augmented inverse

---

[14]By default, the `did` and `DRDID R` packages from Callaway and Sant'Anna (2021) and Sant'Anna and Zhao (2018) trim untreated units with propensity scores in excess of 0.995.

Figure 1: **Distribution of Propensity Scores**
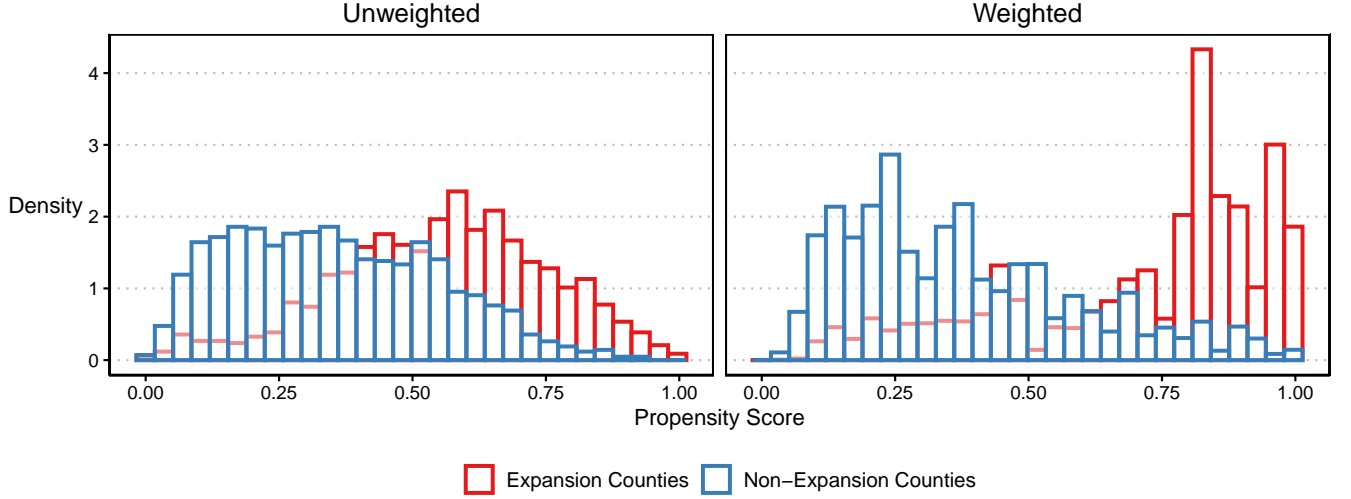
Notes: This figure shows the distribution of propensity scores using both the weighted and unweighted logit propensity score estimates in our 2×2 Medicaid example.

probability weighting approach; see, for instance, Sant'Anna and Zhao (2020) and Chang (2020).

The key idea of the DR DiD approach is to express the $ATT(2)$ in terms of both $p(X_i)$ and $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$ in a way that it gives provides some "protection" in case the working models for these functions, $\widehat{\pi}(X_i)$ and $\widehat{\mu}_{\Delta,G=0}(X_i)$, are wrong. The resulting DR estimator for the $ATT(2)$ is consistent when *either* of these nuisance working models is correctly specified. If one is exactly right, it does not matter if the other is wrong. Furthermore, if both working models are only slightly wrong, their errors will multiply, and DR will perform (asymptotically) better than either one alone.[15] Following the steps in Sant'Anna and Zhao (2020), we can express the $ATT(2)$ as

$$ATT(2) = \mathbb{E}\left[\left(w_{\omega,D=1}(D_i) - w_{\omega,D=0}(D_i, X_i) \right) \left(\Delta Y_{i,t=2} - \mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0] \right) \right], \qquad (4.10)$$

with the weights as defined in (4.7). Notice that when we omit the $\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]$ term from (4.10), we are back to the IPW estimand (4.6). When we omit the $w_{\omega,D=0}(D_i, X_i)$ term from (4.10), we are back to the regression adjustment estimand (4.2).

Constructing a DR DiD estimator for the $ATT(2)$ based on (4.6) is straightforward. Choose a flexible working model for the propensity score and compute its fitted values, $\widehat{\pi}_\omega(X_i)$. Then, a flexible working model for the outcome evolution of untreated units and compute its fitted values for all treated and untreated units, $\widehat{\mu}_{\omega,\Delta,D=0}(X_i)$, and then use the plug-in principle to estimate

---

[15]We usually arrive at such estimands by deriving the efficient influence function; see, for example, Sant'Anna and Zhao (2020) for a discussion in DiD setups. See also Seaman and Vansteelandt (2018) for an overview. When we adopt nonparametric or machine-learning-based estimators for the nuisance functions, we get a different type of double robustness, "rate double robustness," where we can trade-off precision between the different working models. For more details, see, Kennedy, Ma, McHugh and Small (2017) and Smucler, Rotnitzky and Robins (2019).

the $ATT(2)$ by

$$\widehat{ATT}_{dr}(2) = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{w}_{\omega,D=1}(D_i) - \widehat{w}_{\omega,D=0}(D_i, X_i)\Big)\Big(\Delta Y_{i,t=2} - \widehat{\mu}_{\omega,\Delta,D=0}(X_i)\Big), \tag{4.11}$$

with the estimated weights as defined in (4.9).

We report the $ATT(2014)$ estimates and standard errors using the DR DiD procedure with a linear in covariates working model for $\mu_{\omega,\Delta,D=0}(X_i)$ and (weighted) logistic regression working model that is also linear in covariates for $p_\omega(X_i)$ in Table 7 (labeled as "Doubly Robust"). Population-weighted results are fairly similar across our methods, but the unweighted DR DiD estimates are about halfway between the RA and IPW results. However, an advantage of the DR DiD procedures is that model misspecifications are arguably less of a concern.

Overall, in this section, we discussed different "forward-engineered" estimators for the $ATT(2)$ that respect conditional parallel trends, correctly incorporate covariates, and do not restrict treatment effect heterogeneity. Between RA, IPW, and DR DiD estimators, we recommend practitioners favor the *doubly robust* one compared with the RA and IPW appraoches on their own, as this procedure adds additional "protection." When strong overlap fails, though, RA DiD estimators can still work because they extrapolate the outcome model to obtain predicted counterfactual outcome changes even for treated units with covariate values not observed in the comparison group. The credibility of this extrapolation, however, rests on the accuracy of the working model outside the support of X_i in the untreated group. If this extrapolation is not reliable, one can make the DR approach more robust against weak overlap by trimming "extreme" propensity scores and performing a bias correction to ensure that the target parameter (the $ATT(2)$ in our case) remains the same. See Ma et al. (2023) for details.

4.5 Heterogeneity analysis

A different motivation for using covariates is to estimate heterogeneous treatment effects by the values of X_{it}. The identification result in equation (4.2) can be altered slightly to show that $ATT(2)$ is simply an aggregation of covariate-specific 2×2 DiD estimands:

$$ATT(2) = \mathbb{E}_\omega\Big[\mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 1] - \mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0]\Big|D_i = 1\Big].$$

Thus, Assumptions CPT and SO also imply that we can identify conditional ATT parameters:

$$\begin{aligned}ATT_{X_i}(2) &\equiv \mathbb{E}_\omega[Y_{i,t=2}(1) - Y_{i,t=2}(0)|D_i = 1, X_i]\\&= \mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 1] - \mathbb{E}_\omega[\Delta Y_{i,t=2}|X_i, D_i = 0].\end{aligned}$$

This not only demonstrates the building block structure of conditional designs, but also connects the strategies we discussed for estimating $ATT(2)$ itself to the underlying treatment effect heterogeneity by covariate values. When this heterogeneity is of interest in its own right it can also be targeted, identified, and estimated (at least under some additional conditions).

When all covariates are discrete, estimating $ATT_{X_i}(2)$ parameters is fairly straightforward. One can form saturated partitions x_k, $k = 1, \ldots, K$, and then subset the data to contain only information from the units from the specified partition k. At this stage, the analysis is analogous to the unconditional DiD setup (except that you need to repeat this step from all partitions of interest), as each $ATT_{x_k}(2)$ is identified by a (conditional on a discrete-variable) comparison of means' that is,

$$ATT_{x_k}(2) = \mathbb{E}_\omega\big[\Delta Y_{i,t=2}|D_i = 1, X_i = x_k\big] - \mathbb{E}_\omega\big[\Delta Y_{i,t=2}|D_i = 0, X_i = x_k\big].$$

One can estimate these $ATT_{x_k}(2)$'s using their sample analogs or using two-way fixed effects regression specifications analogous to (3.7). Inference is also standard, provided that each partition is sufficiently large. This type of exercise is commonly used to conduct heterogeneity analysis. For example, suppose one wants to see if the effect of Medicaid expansion on adult mortality rate varies across US census regions. Then, one would partition the data into US regions and run one (unconditional) DiD analysis for each region, provided we have treated and untreated units in each region.

When some covariates are continuous or there are so many partitions that each one contains few observations, one can still *identify* the $ATT_{X_i}(2)$'s, though they are hard to estimate unless auxiliary assumptions hold. In such cases, it is customary to identify and estimate a more aggregated conditional ATT parameter than $ATT_{X_i}(2)$. For instance, one may want to assess if the effect of Medicaid expansion on adult mortality rates is higher (or lower) in counties with an unemployment rate above the median than it is in those below. Similar partitions could be made for any other covariates. To formalize this notion of "partition specific" ATT, let $PART(X_i)$ be some user-specified partition of the covariate space such that $PART(X_i) \in \{1, 2, \ldots, K\}$, and define

$$ATT_k(2) = \mathbb{E}_\omega\big[Y_{i,t=2}(1) - Y_{i,t=2}(0)|D_i = 1, PART(X_i) = k\big].$$

Under Assumptions CPT and SO, it follows that

$$ATT_k(2) = \mathbb{E}_\omega\big[\Delta Y_{i,t=2}|D_i = 1, PART(X_i) = k\big]$$
$$- \mathbb{E}_\omega\bigg[\mathbb{E}_\omega\big[\Delta Y_{i,t=2}|D_i = 0, X_i, PART(X_i) = k\big]\Big|D_i = 1, PART(X_i) = k\bigg],$$

which implies that

$$ATT(2) = \sum_{k=1}^{K} P_\omega(PART(X_i) = k|D_i = 1)ATT_k(2).$$

Thus, for heterogeneity analysis with covariates, one can partition the data with a user-specified partition map $PART(X_i)$ and then, within each of these partitions, use arguments like the ones we used to establish (4.2) to guarantee that each partition-specific ATT is identified (which is precisely what we did to get $ATT_k(2)$ above). Regarding estimation and inference, one can use regression adjustment, inverse probability weighting, or doubly robust estimators as in Section 4.4; the difference is that these are implemented "locally" on each partition.

Obtaining the overall $ATT(2)$ is just a matter of aggregating these partition-specific ATTs using weights equal to the relative partition size among treated units. All these parameters have clear causal interpretations, can be used to answer different policy questions, and are identified under the same identification assumptions already discussed. Other types of heterogeneity analysis are also possible and even attractive. For instance, Abadie (2005) discusses how one can highlight how the ATT varies across a subset of the covariates required for conditional parallel trends to hold. For example, this would entail checking if the average effect of Medicaid expansion among expansion counties varies with the 2013 poverty rate and/or median income. One can also get the best linear approximation of these conditional ATT curves, which involves estimating fewer parameters. These heterogeneity analyses are generally more granular than the partition-based ones discussed above and do not require discretizing the data. They complement each other well.

We close this section with a remark on the choice of partition and the type of heterogeneity analysis to conduct. Heterogeneity analysis has the potential to offer policymakers and researchers novel insights about the treatment of interest and its mechanisms, opening the door for more informed policy recommendations and targeted expansions. But how should one define the subgroups in which to estimate heterogeneous effects? If researchers knew the relevant partition that policymakers care about, they could aim to estimate these particular partition-specific ATTs. But this is rarely the case. Taking no stand about heterogeneity implies reporting unit-level effects, which requires incredibly strong assumptions and could also lead to noisy estimates. On the other hand, taking a too-coarse partition may mask important types of treatment effect heterogeneity. So, it seems that a balance between these extremes is important for a policy-relevant heterogeneity analysis. Estimating conditional ATTs across one or two covariate dimensions is useful, but it may mix some other interactive effects. Another potential avenue is to go beyond averages and adapt the sorted-effects procedure proposed by Chernozhukov, Fernández-Val and Luo (2018) to DiD designs. It would also be interesting and practically relevant to extend the heterogeneity tools for experimental data discussed in Chernozhukov, Demirer, Duflo and Fernández-Val (2023) to DiD setups. In our view, this is an area in which applied econometrics practice would benefit from more thorough methodological guidance.

5   DiD designs with multiple time periods

The previous sections focused on fundamental identification issues and estimation approaches for 2×2 building blocks, but generally did not build them into anything because they targeted the only feasible $ATT(t)$, $ATT(2)$. DiD designs with multiple periods are notably different. They can target ATTs in each period after treatment to trace out dynamics, and they can produce cross-group outcome trends from *before* treatment starts to evaluate the plausibility of parallel trends. Multiple periods also admit the possibility of more complex treatment variation. We will focus on staggered-timing designs in which treatment "turns on" at different times for different units and

stays on, as the Medicaid expansion did. It is possible, however, to use DiD methods to study treatments that "turn off" or occur more than once, though some modifications are warranted; see, for example, de Chaisemartin and D'Haultfoeuille (2020, 2023a).

We need to expand the notation to define the relevant concepts with multiple periods. When treatment can happen only at one point in time (so we still have only two treatment groups with D_i equal to one or zero), the only changes we need to make are to acknowledge that time runs from $t = 1, 2, \ldots, T$, and that the map between potential outcomes and observed outcomes is

$$Y_{i,t} = D_i Y_{i,t}(1) + (1 - D_i) Y_{i,t}(0).$$

By Assumption NA, we have that $Y_{i,t} = Y_{i,t}(0)$ for all pre-treatment periods, and that, in post-treatment periods, we observe $Y_{i,t}(0)$ if the group remains untreated by $t = T$, and $Y_{i,t}(1)$ for groups treated at the unique treatment date, $t = g$.

5.1 Simple event studies $(\mathbf{2 \times T})$

The term "event study" refers to estimating and reporting effects across a range of time periods before and after treatment. A design with one treatment timing group and multiple time periods $(2 \times T)$ is the simplest case in which to discuss event studies. We thus expand our analysis of the 2014 Medicaid expansion group to include data from 2009 to 2019. We report population-weighted results for brevity. Figure 2 plots the time series of the weighted mortality rates for the 2014 and post-2019 expansion counties. It is the analog of Table 2 in the sense that it presents the raw data elements necessary to construct event study estimates. The treatment year 2014 naturally divides the x-axis into two windows: post-treatment (2014-2019) and pre-treatment (2009-2013). An event study constructs DiD-type estimates in both windows, but they have different interpretations.

5.1.1 Event study estimates in the post-treatment periods

The target parameters in a simple event study are the $ATT(t)$'s with the same definition as in (3.5); there are just more of them to identify, estimate, and interpret than in a 2×2 design. The $ATT(t)$s in the post-treatment period, $t \geqslant g$, reflect treatment effect dynamics. For example, economic models that view health as a stock imply that $ATT(t)$ may grow if Medicaid stimulates health investments. Furthermore, people and institutions may take time to adjust their behavior after Medicaid expands, suggesting a dynamic treatment effect. Event study parameters answer these kinds of subtle questions.

Identification of each $ATT(t)$ follows from the same arguments outlined in section 3. Note, however, that an event study analysis requires parallel trends in *every* post-period, as in the following assumption.

Assumption PT-ES ($2 \times T$ Parallel Trends)**.** The average change of $Y_{i,t}(0)$ from $Y_{i,t=g-1}(0)$ is

Figure 2: **County Mortality Trends by Expansion Decision**
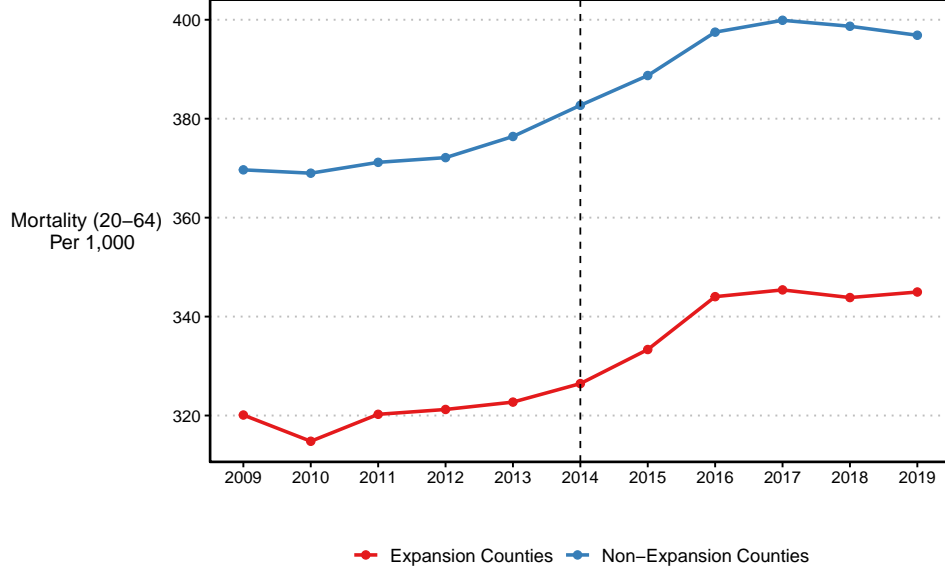
Notes: This figure shows county population-weighted average mortality rates for adults ages 20-64 by expansion category from 2009 to 2019.

the same between treated and comparison units for all post-treatment periods $t \geqslant g$; that is,

$$\mathbb{E}_\omega[Y_{i,t}(0) - Y_{i,t=g-1}(0)|D_i = 1] = \mathbb{E}_\omega[Y_{i,t}(0) - Y_{i,t=g-1}(0)|D_i = 0] \quad \forall t \geqslant g. \tag{5.1}$$

Assumption PT-ES suggests that learning about long-run effects requires stronger assumptions than learning about short-run effects. That is, to identify the average effect of Medicaid expansion in 2019 among expansion counties, Assumption PT-ES requires parallel trends to hold in every year from 2014 to 2019. On the other hand, if we are interested in learning only about short-run effects—say effects up until 2015—we would require it to hold from 2014 and 2015 only.

If Assumption PT-ES holds (as well as no-anticipation), then each $ATT(t)$ is identified by a DiD comparison between period $g-1$ and t, as in (3.5), and the $ATT(t)$s can be estimated by the familiar comparison of four sample averages:

$$\widehat{ATT}(t) = (\overline{Y}_{\omega,D=1,t} - \overline{Y}_{\omega,D=1,t=g-1}) - (\overline{Y}_{\omega,D=0,t} - \overline{Y}_{\omega,D=0,t=g-1}). \tag{5.2}$$

Figure 3 plots (weighted) event study estimates for the 2014 Medicaid expansion group. $\widehat{ATT}(t)$'s lie to the right of the vertical dashed line. The estimate for event-time 0 (that is, $t = g = 2014$) equals the 2×2 results from Table 2 (-2.6). The other estimates have the same DiD form: comparisons of cross-group changes in different post-2014 years (t) but always relative to 2013 ($g-1$). The point estimates do not suggest large mortality effects from Medicaid expansion among expansion counties.

Figure 3: **2 × T Event Study**
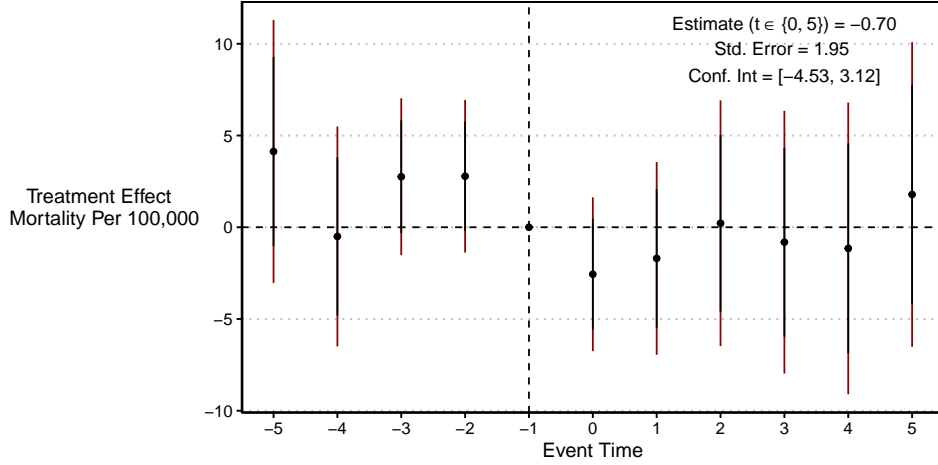
Notes: This figure shows the event study estimates in the $2 \times T$ case, comparing counties that expanded in 2014 to counties that had not yet expanded by 2019. It uses the unconditional estimator from Callaway and Sant'Anna (2021). The outcome variable is the crude mortality rate for adults ages 20-64, and the standard errors are clustered at the county level. The point estimate is reported by the circles, and both point-wise (black) and simultaneous (red) confidence intervals are reported with the vertical lines (see section 5.1.3).

5.1.2 Event study estimates in the pre-periods

Multiple time periods also allow for falsification/placebo tests based on DiD-type comparisons between *pre*-treatment periods. The no-anticipation assumption implies that all $ATT(t)$'s before time g are equal to zero, which means that a 2×2 estimand between periods $t = g-k$ and $t = g-1$, $k > 1$, equals a difference in weighted average *untreated* potential outcome trends (as they are all from pre-treatment periods):

$$\tau_{-k} = \mathbb{E}_\omega[Y_{i,t=g-k}(0) - Y_{i,t=g-1}(0)|D_i = 1] - \mathbb{E}_\omega[Y_{i,t=g-k}(0) - Y_{i,t=g-1}(0)|D_i = 0]$$
$$= \mathbb{E}_\omega[Y_{i,t=g-k} - Y_{i,t=g-1}|D_i = 1] - \mathbb{E}_\omega[Y_{i,t=g-k} - Y_{i,t=g-1}|D_i = 0].$$

The τ_{-k} terms are usually called "differential trends" or "pre-trends," and they appear to the left of the vertical dashed line in Figure 3. The credibility of event study analyses rests in large part on finding small estimates of τ_{-k}, but Figure 3 illustrates how challenging it can be to draw conclusions from informally looking at these pre-trends. No individual τ_{-k} is statistically significant, so we fail to reject the null hypothesis that the pre-trend estimates equal zero (individually or jointly). This kind of result is often interpreted to mean that parallel trends holds. But the τ_{-k}'s also tend to be positive with a mean of about 2.3, which is larger in magnitude than all post-period point estimates except one. Sometimes, this kind of result is used to argue that parallel trends fails. Do these results support or refute parallel trends?

Recent methodological work suggests three practical lessons about using pre-trend estimates to assess parallel trends; examples include Bilinski and Hatfield (2018), Manski and Pepper (2018), Kahn-Lang and Lang (2020), Roth (2022), Rambachan and Roth (2023), Dette and Schumann (2024), and Freyaldenhoven, Hansen, Pérez-Pérez and Shapiro (2024). The most fundamental, which will tend to shape language more than practice, is that Assumption PT-ES is not testable,

as it only makes restrictions on untreated potential outcomes in post-treatment periods, $t \geqslant g$. Under no-anticipation, pre-trends do measure differences in untreated outcome trends between treated and untreated units, but they necessarily measure them in the "wrong" periods $t < g$. This does not mean parallel pre-trends are not informative; it just means they are not the *same* as the parallel trends assumption in Assumption PT-ES.

When specifically are observed pre-trends uninformative about the parallel trends assumptions required for identification? One case is when pre-trends are measured too far before treatment starts. The conditions or shocks that jointly shape treatment decisions and untreated outcomes may differ many periods before treatment, but this need not imply that they generate bias in the periods after treatment. A second case, discussed in Section 3.2, is when units select into treatment based on time-varying pre-treatment unobservables. As treatment selection depends on pre-treatment unobservables, non-parallel pre-trends may appear as a consequence of the selection mechanism, though parallel trends may still be plausible. Ghanem et al. (2022) show that in this case, a *necessary* condition for parallel trends is that the untreated potential outcomes satisfy a martingale property. When this martingale property does not hold, parallel trends cannot hold either, *regardless* of the shape of pre-trends. In such situations, it is worthwhile to assess the potential bias components of the DiD estimator using the "selection-aware" benchmarking tools presented in Ghanem et al. (2022), as well as to resort to their theory-based templates for justifying parallel trends based on selection mechanisms. Ultimately, how informative pre-trends are for the plausibility of Assumption PT-ES is case-specific, though we would of course always prefer to have parallel pre-trends.

When Assumption PT-ES holds in all periods (pre- and post-treatment), it is worth noting that one can construct estimators for the $ATT(t)$ that are more precise than those in (5.2); see, e.g., Borusyak et al. (2024), Gardner (2021), Harmon (2024), Marcus and Sant'Anna (2021), Wooldridge (2021), and Chen, Sant'Anna and Xie (2024). In such cases, though, one would *require* parallel trends to hold pre-treatment periods, parallel pre-trends becomes directly testable (as the model is overidentified), and τ_{-k} could be used to assess its plausibility directly. Other types of overidentification tests could also be used in this case to assess the validity of the DiD model directly; see Marcus and Sant'Anna (2021) and Chen et al. (2024) for a discussion.

The second lesson is that statistical precision shapes the usefulness of pre-trend estimates. The hypothesis tests for parallel pre-trends in Figure 3 are low-powered to detect practically important violations (Roth, 2022, and Freyaldenhoven et al., 2024). The τ_{-k} estimates fail to rule out flat pre-trends, but as we shall see below, they also fail to rule out large pre-trends that would indicate serious bias in the $\widehat{ATT}(t)$'s. They simply do not say very much.[16] Roth (2022) discusses how conditioning the analysis on these kinds of low-powered tests can exacerbate biases and should be

---

[16]Another possibility is that very precisely estimated pre-trends are distinguishable from zero yet also rule out even small violations. The magnitude of the violations also matters: a precisely estimated but small violation of parallel trends in pre-treatment periods is "better" than an imprecisely potentially large estimated pre-trend that does not rule out zero.

interpreted with care.

The third lesson is that researchers can make better use of pre-trend estimates by taking a stand on the size of plausible and/or problematic parallel trends violations. For example, Bilinski and Hatfield (2018) propose selecting a value for differential trends that *would* fully explain the estimated treatment effects and then using it instead of zero as the null hypothesis when conducting statistical tests of the estimated pre-trends; see also Dette and Schumann (2024). Rambachan and Roth (2023) develop inference methods for an approach that bounds the $ATT(t)$'s under assumptions about the maximum size of parallel trends violations based on pre-treatment periods (see also Manski and Pepper, 2018). One can either select a magnitude using contextual knowledge or set it equal to a multiple of the largest period-to-period pre-trend estimate. One then constructs an identified set that contains $ATT(t)$ not under parallel trends but under the weaker assumption that parallel trends is not violated by more than this pre-determined maximum in each post-period. Importantly, Rambachan and Roth (2023) also show how to use the estimated covariance matrix of the event study estimates to construct confidence intervals around the set.[17] Given the existence of these methods, which are theoretically grounded in how to consider violations of pre-trends, we caution producers (and consumers) of DiD work against the common practice of using only a simple "eye-test" for whether pre-trends differ substantially from zero.

In our application, Rambachan and Roth (2023)'s method underscores how little information the pre-trend estimates convey. The largest one-period pre-trend in Figure 3 is between event-time -5 and -4, when outcomes fall by roughly four deaths more in the expansion group versus the non-expansion group. If we assume that parallel trend violations are no bigger than this, the identified set for $ATT(2014)$ is $-2.6 \pm 4 = [-6.6, 1.4]$, and given the size of the pre-period standard errors, we obtain a robust confidence interval of $[-11.1, 5.1]$, which spans implausibly large effects in both directions. Assuming smaller parallel trend violations would shrink this interval, but applying the method to subsequent event-times widens it. In general, Rambachan and Roth (2023) provide a flexible method to use information about potential parallel trends violations drawn either from external knowledge or the τ_{-k} estimates, as well as the precision of the pre-trend estimates, to gauge (statistically) the robustness of the $\widehat{ATT}(t)$'s.

When pre-trends suggest that Assumption PT-ES fails, a way forward is to assume that it holds only after conditioning on covariates and proceeding similarly to Section 4.4; we discuss this path in detail in Section 5.1.4. Alternatively, one can attempt to parametrically model the violations of parallel trends. Usually this is done by including unit-specific linear trends; see, e.g., Mora and Reggio (2019), Wooldridge (2021, Section 7), Lee and Wooldridge (2023), and Freyaldenhoven et al. (2024). We note, however, that the practice of using unit-specific linear trends deviates from the standard DiD procedures: it relies on alternative identification assumptions involving an explicit

---

[17]In cases where a measured variable accounts for the pre-trends, Freyaldenhoven, Hansen and Shapiro (2019) develop two-stage-least-squares estimators that recover $ATT(t)$ parameters by extrapolating the pre-period relationship between outcomes and the covariate into the post-period.

parametric model for unit-specific trends. We also note that sensitivity analysis procedures that do not rely on such models are available—Rambachan and Roth (2023) is one example—and we encourage practitioners to consider them.

5.1.3 Estimation and aggregating across time in event-studies

The link between a $2 \times T$ event study and a series of 2×2 DiD building blocks makes estimation simple. The point estimates in Figure 3 are the $ATT(t)$ estimates based on (5.2).[18] An equivalent way to obtain all the $\widehat{ATT}(t)$'s in one step is to run a TWFE regression with time fixed effects, θ_t, unit fixed effects, η_i, and a set of interactions between the treatment group dummy and the time dummies. Omitting the treatment interaction for $t = g - 1$ avoids multicollinearity and fixes $g - 1$ as the baseline period for all β_t estimates, which matches Assumption PT-ES. This generalizes the TWFE regression equation for a single $ATT(t)$ in (3.7) to

$$Y_{i,t} = \theta_t + \eta_i + \sum_{k=1}^{g-2} \beta_k \left(\mathbf{1}\{G_i = g\} \cdot \mathbf{1}\{t = k\} \right) + \sum_{k=g}^{T} \beta_k \left(\mathbf{1}\{G_i = g\} \cdot \mathbf{1}\{t = k\} \right) + \varepsilon_{i,t}. \tag{5.3}$$

This regression produces identical estimates to those obtained "by hand" via (5.2): $\widehat{\beta}_t = \widehat{ATT}(t)$. It also generates (point-wise) confidence intervals based on clustered standard errors, as discussed in Section 3.3. An additional issue in event study inference, however, involves the fact that we are now estimating many treatment effect parameters. Thus, when we compare across event study estimates, we are conducting many hypothesis tests, and the usual normal critical values used to construct confidence intervals do not account for these. Asymptotically correct inferences about the entire event study curve require "inflating" critical values to perform a multiple hypothesis test adjustment. In Figure 3, the thick black bars represent the standard pointwise confidence intervals from clustered standard errors at the county level, while the red line shows uniform confidence bands that cover the 95% confidence interval for the entire treatment path of the event study coefficients after accounting for multiple testing. These are produced by default in the Callaway and Sant'Anna (2021) statistical packages, using a multiplier bootstrap procedure to compute critical values of the sup-t test statistic. Alternatively, one can construct these using the estimated variance-covariate matrix of all $\widehat{\beta}_t$'s paired with the Montiel Olea and Plagborg-Moller (2018) simulation procedure. Alternative bootstrap procedures, such as the nonparametric bootstrap, the multiplier bootstrap, and the weighted/Bayesian bootstrap, can also be used to compute the sup-t critical values that account for multiple testing.[19]

---

[18]We estimate these with the `did` R package from Callaway and Sant'Anna (2021).

[19]The sup-t critical value governs the width of the uniform confidence band that yields simultaneous coverage probabilities for a given confidence level (Montiel Olea and Plagborg-Moller, 2018). In our context, its main idea is constructing asymptotically valid critical values for the entire event-study trajectory based on the maximum of all t-statics (one for each event-time considered). This procedure avoids the conservativeness of multiple-testing corrections such as Bonferroni's. See Montiel Olea and Plagborg-Moller (2018) for a general discussion.

A final estimation issue arises when targeting aggregations of the $ATT(t)$'s. For example, the average treatment effect in the post-period, $ATT_{\text{avg}} = \frac{1}{T-(g-1)} \sum_{t=g}^{T} ATT(t)$, is a convenient scalar measure that improves statistical precision, especially when $ATT(t)$ is relatively constant. The easiest way to get an estimate of $\widehat{ATT}_{\text{avg}}$ is just to construct it from the 2×2 $\widehat{ATT}(t)$ building block estimates. Standard post-estimation commands achieve this if the event study estimates come from a regression, and newer DiD packages report this parameter automatically. A common shortcut, however, is to run a second regression that replaces the event study dummies with the treatment status dummy. $D_{i,t} = \mathbf{1}\{D_i = 1\} \times \mathbf{1}\{t \geqslant g\}$:

$$Y_{i,t} = \theta_t + \eta_i + \beta^{OLS} D_{i,t} + \varepsilon_{i,t}. \tag{5.4}$$

Unfortunately, the (weighted) least squares estimator $\widehat{\beta}^{OLS}$ does not generally equal the $\widehat{ATT}_{\text{avg}}$. The reason is that $\widehat{\beta}^{OLS}$ is equivalent to first collapsing the multiple-periods data to averages in the post- and pre-periods and then estimating a 2×2 DiD on the resulting means. This, in turn, is the same as subtracting the average pre-period τ_{-k} estimate (including the zero at time $g - 1$) from the average post-period $\widehat{ATT}(t)$ estimate. This implies that interpreting both β_ts from (5.3) and β^{OLS} from (5.4) requires Assumption PT-ES to hold in every time period, not just in the post-treatment periods.[20] To the extent that the gap in mean outcomes over the whole pre-period differs from the gap in outcomes in period $t = g - 1$, the two summary parameters will not be equal. In Figure 3, the two summary parameters are quite different: $\widehat{ATT}_{\text{avg}}$ equals -0.70 while $\widehat{\beta}^{OLS}$ equals -2.53.[21]

5.1.4 Covariates in event studies

Another advantage of seeing event studies as collections of 2×2 building blocks is that all the tools for incorporating covariates from Section 4 immediately apply to each event study estimate. In fact, the only difference is that instead of using "short-differences," $\Delta Y_{i,t=2}$, one would now use "long-differences," $Y_{i,t} - Y_{i,t=g-1}$. This would imply that using the regression-adjustment procedure would require estimating a working model for $\mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|D_i = 0, X_i]$ for each time t. The propensity score working model used to construct the IPW DiD estimate (4.6), on the other hand, is exactly the same as in a 2×2 analysis of the same groups. Since the DR DiD estimation procedure builds on both RA and IPW procedures, it would involve estimating different outcome-regression working models for each time t. We also note that the potential pitfalls of controlling for covariates in a TWFE specification still apply with multiple periods and actually become more complex (Caetano and Callaway, 2024).

---

[20]The decision to estimate each $\widehat{ATT}(t)$ relative to period $t = g - 1$ comes directly from the choice to define PT that way. If parallel trends holds in every period, one can typically form more efficient estimators than those discussed above. See Marcus and Sant'Anna (2021) for a discussion

[21]Interestingly, this quantity is almost identical to our estimate of $ATT(2014)$, but rather than representing anything reassuring, it comes from the offsetting effects of positive pre-period estimates and small post-period ones.

For completeness and ease of access, we list the RA, IPW, and DR estimands for $ATT(t)$:

$$ATT_{\mathrm{ra}}(t) = \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|D_i = 1] - \mathbb{E}_\omega\Big[\mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|X_i, D_i = 0]\Big|D_i = 1\Big],$$

$$ATT_{\mathrm{ipw}}(t) = \mathbb{E}\Big[\Big(w_{\omega,D=1}(D_i) - w_{\omega,D=0}(D_i, X_i)\Big)(Y_{i,t} - Y_{i,t=g-1})\Big],$$

$$ATT_{\mathrm{dr}}(t) = \mathbb{E}\Big[\Big(w_{\omega,D=1}(D_i) - w_{\omega,D=0}(D_i, X_i)\Big)\Big(Y_{i,t} - Y_{i,t=g-1} - \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|X_i, D_i = 0]\Big)\Big],$$

where $w_{\omega,D=1}(D_i)$ and $w_{\omega,D=0}(D_i, X_i)$ are as defined in (4.7).

Figure 4 shows weighted event study estimates using our three preferred covariate strategies: regression adjustment, inverse propensity weighting, or doubly-robust estimation. In our case, covariates do little to change the unadjusted estimates. Note, however, that Borgschulte and Vogler (2020) use an IPW estimator with different covariates selected by a lasso procedure and obtain notably stronger evidence of mortality reductions. Evidently, the exact set of covariates one conditions on matters a great deal in this analysis, and, as we have discussed earlier, one should attempt to include in X_i all the determinants of the change in untreated potential outcome or of the treatment assignment.

Figure 4: **$2 \times$ T Event Study with Covariates**
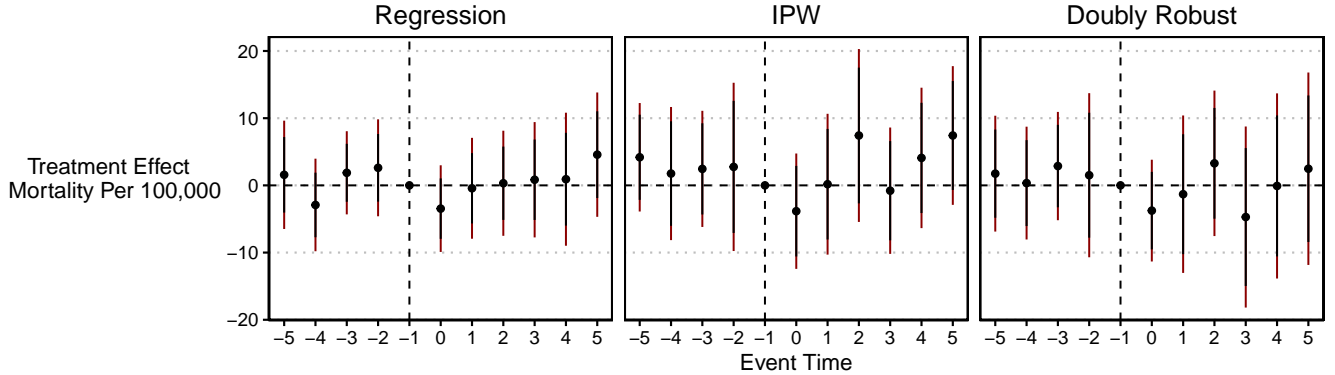
Notes: This figure shows the corresponding event study estimates with the inclusion of covariates. The outcome variable is the crude mortality rate for adults ages 20-64, and the covariates include the percentage of the county population that is female, the percentage of the county population that is white, the percentage of the county population that is Hispanic, the unemployment rate, the poverty rate, and county-level median income. The point estimate is reported by the circles, and both point-wise (black) and simultaneous (red) confidence intervals are reported with the vertical lines.

5.2 Staggered treatment adoption ($\mathbf{G}_\# \times \mathbf{T}$)

Viewing $2 \times T$ event studies as a collection of 2×2 DiD building blocks makes the jump to staggered timing designs straightforward. The key distinction is that with staggered timing, each treatment date defines a distinct treatment group, and each of these has its own set of simple event study parameters. New choices arise about the comparison units used to identify and estimate these group-specific event studies, as well as about how to aggregate the estimates across timing groups. The 2×2 structure, as well as all the tools we have developed to evaluate parallel trends (covariate balance and pre-trends) and to estimate (with or without weights and covariates), carry over.

When treatment start dates can vary across units, we need to allow the potential outcomes, and thus the target parameters and identifying assumptions, to reflect this richer notion of treatment. We therefore index potential outcomes by the time treatment begins, g: $Y_{i,t}(g)$; and use $Y_{i,t}(\infty)$ to denote never-treated potential outcomes.[22] We use G_i to denote each unit's treatment date, and with some abuse of terminology, we call units not exposed to treatment by period T the "never-treated" group.[23] Finally, we use $\mathcal{G}$ to represent the set of all treatment times (rows of Table 1 in our example). With these modifications, we can map potential outcomes to observed outcomes using a generalization of (3.1):

$$Y_{i,t} = \sum_{g \in \mathcal{G}} Y_{i,t}(g) \mathbf{1}\{G_i = g\}.$$

With multiple treatment groups, we also need to extend our notion of no-anticipation (though its empirical content is exactly the same).

Assumption NA-S (No-Anticipation with staggered treatment timing)**.** For all units i that are eventually treated and all pre-treatment periods t, $Y_{i,t}(g) = Y_{i,t}(\infty)$.

Like Assumption NA, Assumption NA-S imposes that treatment effects are zero in all pre-treatment periods as a consequence of units not acting on the potential knowledge of future treatment dates before they are actually exposed to treatment. We maintain this assumption throughout this section.

Finally, we assume that a "never-treated" group always exists in our staggered DiD setup. If all units are eventually treated, we drop all the data from when the last cohort is treated, so the last-treated cohort becomes the "never-treated" cohort, and T here denotes the number of available periods in the subset of the data that we will use in our analysis. This is essentially without loss of generality, because under standard DiD assumptions, we cannot identify any ATT for periods where all units are treated. We also dropped data from units treated in the first available period, $G_i = 1$, as such treatment group does not have any pre-treatment data, preventing us from conducting a DiD analysis.

Figure 5 plots average weighted mortality data by the year of Medicaid expansion (G_i) and time. As in Table 2 and Figure 2, these are all the means necessary to calculate a staggered DiD

---

[22]Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be s-dimensional vectors of zeros and ones, respectively, and denote the potential outcome for unit i at time t if first exposed to the treatment at time g by $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$, and denote by $Y_{i,t}(\mathbf{0}_T)$ the outcome if untreated by time $t = T$. We discussed the two-period treatment this way in section 3.1 when we defined potential outcomes as a function of the period one and period two treatment. These treatment paths define the potential outcomes we work with: $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$. Writing potential outcomes as functions of treatment paths helps with transparency regarding causal parameters of interest and the DiD design we are in. When treatment can turn on and off, writing potential outcomes in terms of the entire path becomes crucial to avoid "hidden" assumptions that rule out treatment effect heterogeneity and dynamics. Because of space constraints, we do not cover these cases in this article.

[23]In practice, "never-treated" really means "not observed to be treated by $t = T$." Given more data, units untreated at T could, in many cases, take up the treatment. In fact, this is the case with the Medicaid expansion. We use data through 2019 but include states that expanded Medicaid in 2020, 2021, and 2023 as "never treated," alongside states that have not expanded as of 2024.

---

estimate.

Figure 5: **County Mortality Trends by Expansion Decision with Staggered Timing**
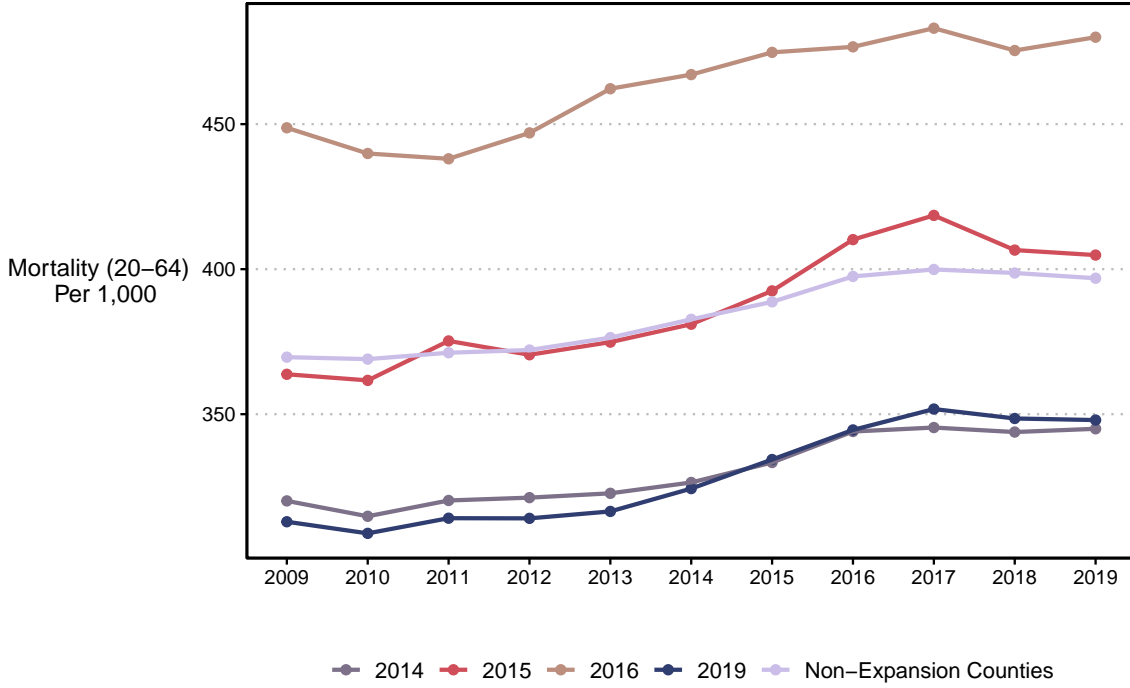
Notes: This figure shows county population-weighted average mortality rates for adults ages 20-64 by expansion year from 2009 to 2019.

5.2.1 Building block parameters with staggered adoption

Staggered treatment timing affects the structure of a DiD analysis because it changes the definition of treatment. Until now, we have used counties in the 2014 expansion states as the only treatment group represented with a single treatment dummy, D_i. But Table 1 shows that as of 2019, there were four different groups of expansion states defined by whether they expanded Medicaid in 2014, 2015, 2016, or 2019. Therefore, D_i is not rich enough to capture the relevant definition of treatment groups in staggered setups, because there are many treatment groups, not just two. Fortunately, we can use the treatment timing notation, G_i, to define ATT parameters, parallel trends assumptions, and estimators, just as we have done so far.

A simple 2×2 DiD design had one target parameter $(ATT(2))$, and a $2 \times T$ DiD design had $T-1$ of them: $T - (g-1)$ post-treatment parameters, $ATT(t), t \geqslant g$, and $g-2$ pre-trend parameters. In staggered DiD designs, each treatment group (sometimes referred to as a cohort), defined by its treatment date g, has its own set of $T-1$ event study parameters. We call these group-time average treatment effects:

$$ATT(g,t) = \mathbb{E}_\omega[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g].$$

Each $ATT(g,t)$ is the average treatment effect of starting treatment at period g relative to never-starting it, at time period t, among units that actually started treatment in period g. It is simply

a set of event study parameters for treatment timing group g. The unobserved counterfactuals are now untreated potential outcome means for each treatment group in each period, $\mathbb{E}_\omega[Y_{i,t}(\infty)|G_i = g]$.

5.2.2 Identification with staggered designs

Identifying the $ATT(g, t)$'s works exactly as in the previous sections because they are just group-specific event studies. Under no anticipation, a set of parallel trends assumptions for $t \geq g$ identifies the causal post-treatment parameters. DiD comparisons for $t < g$ represent differential pre-trends in untreated potential outcomes.

The most important way that staggered DiD changes this approach is that having access to multiple treatment groups with different treatment starting dates allows one to use alternative sets of comparison groups. For example, our Medicaid analysis so far has used counties in states that did not expand Medicaid by 2019 as the comparison group. For estimating, say, the ATT for the 2014 expansion group in 2015, counties in states that did not expand until 2016, 2017, or 2018 could also serve as comparison units. Choosing which comparison groups to use to identify an $ATT(g, t)$ is directly tied to which form(s) of parallel trends hold. Relative to simple event-studies and especially 2×2 setups, staggered timing creates many potential parallel trends assumptions. Here, we discuss three types of staggered parallel trends. The first two use either the never-treated units or any not-yet-treated groups as the comparisons for all eventually-treated groups (Callaway and Sant'Anna, 2021), and the third option assumes that parallel trends holds in all periods and in all groups, which is something that several DiD methods require; see de Chaisemartin and D'Haultfoeuille (2020); Sun and Abraham (2021); Wooldridge (2021); Borusyak et al. (2024); Harmon (2024).[24]

Assumption PT-GT-Nev (Parallel Trends based on never-treated groups)**.** For every eventually treated group g and post-treatment time period $t \geq g$,

$$\mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g] = \mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = \infty].$$

Assumption PT-GT-NYT (Parallel Trends based on not-yet-treated groups)**.** For every eventually treated group g, not-yet-treated group g' and time periods t such that $t \geq g$ and $g' > t$,

$$\mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g] = \mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g'].$$

Assumption PT-GT-all (Parallel Trends for every period and group)**.** For every treatment groups g and g' and time periods t,

$$\mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g] = \mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g'].$$

---

[24]While these three types of parallel trends are fairly intuitive choices, others are possible. For instance, Cengiz, Dube, Lindner and Zipperer (2019) use a comparison group of units treated at least $\delta + 1$ periods after time g. Thus, their parallel trends is tailored to this particular choice. As a result, all of their $ATT(g, t)$ estimates from time g to time $g + \delta$ use the same comparison group. Marcus and Sant'Anna (2021) also discuss other alternative parallel trends assumptions.

Assumption PT-GT-Nev is the analog of the PT assumptions we used in the $2 \times T$ design. It uses the never-treated units as the relevant comparison group for all eventually-treated units, and it imposes parallel trends in post-treatment periods only. In our Medicaid application, this would entail using the non-expansion counties as the comparison group for the 2014, 2015, 2016, and 2019 expansion groups. In addition, since Assumption PT-GT-Nev imposes parallel trends only for the future, the farthest we can go into pre-treatment periods is $t = g - 1$, which will serve as the only (justifiable) baseline period. More formally, under Assumption PT-GT-Nev, it is straightforward to show that for post-treatment periods,

$$ATT(g,t) = \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i = g] - \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i = \infty], \tag{5.5}$$

which shows that $ATT(g,t)$ is identified (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021). Note that (5.5) highlights that we are essentially back to a 2×2 design when it comes to learning about $ATT(g,t)$: it leverages data from only two periods, t (post) and $g - 1$ (pre), and two treatment groups, $G_i = g$ (treated) and $G_i = \infty$ (comparison).

Assumption PT-GT-NYT allows one to use not only the never-treated units but any group of units that are not-yet-treated by time t. In our Medicaid example, we could now use non-expansion counties and 2016 and 2019 expansion counties as comparison groups when estimating $ATT(2014, 2015)$. Using Assumption PT-GT-NYT, Callaway and Sant'Anna (2021) has shown that we can identify the $ATT(g,t)$ for post-treatment periods $t \geq g$ by[25]

$$ATT(g,t) = \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i = g] - \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i > \max\{g,t\}]. \tag{5.6}$$

Like equation (5.5), this follows a 2×2 set up: it leverages data from only two periods, t (post) and $g - 1$ (pre), and two treatment groups, $G_i = g$ (treated), and $G_i > t$ (comparison).[26]

Finally, Assumption PT-GT-all allows the use of any not-yet-treated units as a comparison group, as well as any pre-treatment period as a baseline period. For instance, to identify the ATT for 2015 expansion counties, we can now use the never-treated group and the 2016 and 2019 expansion groups, and use any or all of the years from 2009 to 2014 as a baseline. Using Assumption PT-GT-all, it is easy to show that, for any pre-treatment period $t_{\text{pre}} < g$ and any not-yet-treated group $g' > t$, we can identify the $ATT(g,t)$ for group g's post-treatment periods

---

[25]de Chaisemartin and D'Haultfoeuille (2020) have derived the same results but restrict attention to instantaneous treatment effects—that is, $ATT(g,g)$'s. They do allow for treatment turning on and off, but also impose that being exposed to a treatment today does not affect outcomes tomorrow (a no-carryover assumption). See de Chaisemartin and D'Haultfoeuille (2023a) for some extensions.

[26]We note that more general results are also possible. In fact, for any not-yet-treated group $g' > t$, it is easy to show that under Assumption PT-GT-NYT, for $t \geq g$, $ATT(g,t) = \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i = g] - \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i = g']$. One can then flexibly combine the different comparison groups by leveraging user-specified or efficiency-oriented weights. Chen et al. (2024) discuss how to efficiently explore all the information implied by the identification assumptions to form semiparametrically efficient DiD estimators—that is, estimators that asymptotically enjoy the shortest possible (theoretically justified) confidence intervals without making strong functional form or model-based assumptions related to error terms such as homoskedasticity and restrictions on serial dependence.

$t \geqslant g$ by

$$ATT(g,t) = \mathbb{E}_\omega[Y_{i,t} - Y_{i,t_{\mathrm{pre}}}|G_i = g] - \mathbb{E}_\omega[Y_{i,t} - Y_{i,t_{\mathrm{pre}}}|G_i = g'], \tag{5.7}$$

which again maps back to the 2×2 DiD setup, as it leverages two periods, t (post) and t_{pre} (pre), and two groups, $G_i = g$ (treated) and $G_i = g'$ (comparison). This representation makes it clear that, in practice, one can use various pre-treatment periods and not-yet-treated comparison groups to characterize the $ATT(g,t)$ under Assumption PT-GT-all. We can also combine several of these to form an $ATT(g,t)$ estimand that uses more data (Wooldridge, 2021; Gardner, 2021; Liu, Wang and Xu, 2024; Borusyak et al., 2024; Chen et al., 2024). An intuitive estimand that naturally extends (5.6) and allows us to use more pre-treatment data is given by

$$ATT(g,t) = \mathbb{E}_\omega[Y_{i,t} - \overline{Y}_{i,t \leqslant g-1}|G_i = g] - \mathbb{E}_\omega[Y_{i,t} - \overline{Y}_{i,t \leqslant g-1}|G_i > \max\{g,t\}], \tag{5.8}$$

where $\overline{Y}_{i,t \leqslant g-1} = \sum_{s=1}^{g-1} Y_{i,s}/(g-1)$ is the time average of group g pre-treatment periods for each unit i; see, for example, Callaway (2023, Section 3.2), Lee and Wooldridge (2023), and the discussion in de Chaisemartin and D'Haultfoeuille (2023b, Section 3.2.4). Although the estimand does look different from the previous one, it too resembles a 2×2 design: it effectively leverages two periods, t (post) and the average of all period $t < g$ (pre), and two groups, $G_i = g$ (treated) and $G_i > t$ (comparison). In practice, however, there is no general econometrics guarantee that estimators based on (5.8) will be more precise than estimators based on (5.6); see, for instance, Harmon (2024). This arises because it is not always optimal to weigh all pre-treatment periods equally when forming estimators for $ATT(g,t)$. In fact, as discussed in Chen et al. (2024), the optimal way (or, more formally, the semiparametric efficient way) to aggregate information across pre-treatment periods and comparison groups under Assumption PT-GT-all depends on the correlation structure of how the outcome changes over time across different comparison groups. And an effective way to leverage this information consists of constructing DiD estimators for $ATT(g,t)$'s that efficiently weigh several 2×2 DiD estimators for the $ATT(g,t)$ that use different comparison groups and different baseline periods. These efficiency weights do not depend on additional hard-to-motivate assumptions (e.g., homoskedasticity or restriction on the serial correlation), arise as a consequence of the information content of the identification assumptions, and can be transparently visualized using the tools provided Chen et al. (2024). What is perhaps most important here is that even these more complex DiD estimators closely resemble the approach we took in the 2×2 design.

In the end, a natural question arises: Which parallel trends assumption should one use? This context-specific question is hard to answer, as each assumption has pros and cons. For instance, Assumption PT-GT-all leads to more precise $ATT(g,t)$ estimators because it uses data from multiple pre-treatment periods and multiple comparison groups. Given that power is important when conducting causal inference, this is appealing. On the other hand, it imposes parallel pre-trends, an assumption that is not *required* for identification of $ATT(g,t)$ and that we have not imposed in $2 \times T$ DiD designs (see our discussion of equation (5.3)). If pre-trends are not parallel,

then estimates of $ATT(g,t)$ based on Assumption PT-GT-all can be biased.

The other extreme is to make Assumption PT-GT-Nev and use only comparison groups made up of never-treated units. This avoids compositional changes in the comparison group over time, does not restrict pre-trends, and identifies all the $ATT(g,t)$'s.[27] It also avoids using as a comparison group units that may have chosen to begin treatment in a given period because of pre-treatment outcomes, which potentially violates parallel trends. For instance, states that expanded Medicaid in 2016 may have done so on the basis of county mortality rates from previous years, likely violating the parallel trends assumption for this group. On the other hand, never-treated units may have remained untreated for reasons related to trends in $Y_{i,t}(0)$. Non-expansion counties may be too different from expansion counties for them to reflect the relevant counterfactual. This could be, in part, justified after examining the differences in covariate levels and trends between treatment and control groups, as we saw in the discussion of Table 4. Also, depending on how widespread treatment is, there may be too few never-treated units to obtain precise estimates.

We view Assumption PT-GT-NYT as a middle step that uses all not-yet-treated units as a comparison group without restricting all pre-treatment trends to be parallel. It uses more information than Assumption PT-GT-Nev, which can lead to gains in precision and helps to incorporate covariates. While it uses less information than Assumption PT-GT-all, it is also less susceptible to bias from violations of parallel pre-trends. For our Medicaid application, we favor Assumption PT-GT-NYT, as we prefer not to impose parallel pre-trends from 2009 to 2014 or to rely exclusively on comparisons to the set of states that have not expanded Medicaid as of 2024. On the other hand, if parallel trends are not plausible for a particular group of eventually-treated units, perhaps owing to selection based on time-varying unobservables (Ghanem et al., 2022), it is important to remove these units from the DiD analysis to retain interpretability. Ultimately, the plausibility of each parallel trends assumption may vary across different contexts. At the very least, we strongly recommend that researchers clearly state the specific parallel trends assumption they are actually imposing in their analysis to allow readers to discuss its plausibility in a scientifically grounded manner.

5.2.3   Estimators for staggered designs without covariates

The identification results for $ATT(g,t)$ discussed in Section 5.2.2 suggest very simple and intuitive estimators for the $ATT(g,t)$. Given the estimand that comes from the chosen parallel trends assumption, the estimators replace the population (weighted) expectations with their sample analogs. The principle is the same as in the 2×2 setup of Section 3.3.

For example, under Assumption PT-GT-NYT, we can leverage (5.6) and form plug-in estima-

---

[27]Without never-treated units, we cannot estimate $ATT(g,t)$ for the last observed treatment date, which shapes the feasible target parameters. For example, in our Medicaid expansion example, without the presence of never-treated (by 2019) counties, we would not be able to estimate the treatment effects in 2019 (i.e., $ATT(g = 2015, t = 2019)$).

tors for $ATT(g,t)$ using

$$\widehat{ATT}_{\text{nyt}}(g,t) = \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i = g\}\omega_i(Y_{i,t} - Y_{i,t=g-1})}{\sum_{i=1}^{n} \mathbf{1}\{G_i = g\}\omega_i} - \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i > t\}\omega_i(Y_{i,t} - Y_{i,gt=-1})}{\sum_{i=1}^{n} \mathbf{1}\{G_i > t\}\omega_i}. \tag{5.9}$$

This simple estimator is what Callaway and Sant'Anna (2021) propose when one uses the not-yet-treated group as the comparison group.[28]

When Assumption PT-GT-Nev holds, it is straightforward to build on (5.5) and estimate $ATT(g,t)$ by

$$\widehat{ATT}_{\text{never}}(g,t) = \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i = g\}\omega_i(Y_{i,t} - Y_{i,t=g-1})}{\sum_{i=1}^{n} \mathbf{1}\{G_i = g\}\omega_i} - \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i = \infty\}\omega_i(Y_{i,t} - Y_{i,gt=-1})}{\sum_{i=1}^{n} \mathbf{1}\{G_i = \infty\}\omega_i}. \tag{5.10}$$

This estimator was proposed by Callaway and Sant'Anna (2021) and Sun and Abraham (2021) when using never-treated units as a comparison group, though Sun and Abraham (2021) arrive at this using a fully saturated regression specification and estimating the regression coefficients $\beta_{g,e}^{SA}$ with (weighted) least squares,

$$Y_{i,t} = \theta_t + \eta_i + \sum_{g \neq \infty} \sum_{e \neq -1} \beta_{g,e}^{SA} \mathbf{1}\{G_i = g\}\mathbf{1}\{G_i + e = t\} + \epsilon_{i,t}. \tag{5.11}$$

It is straightforward to show that $\beta_{g,e}^{SA} = \widehat{ATT}_{\text{never}}(g, g+e)$, emphasizing that (5.11) is just a way to contrast sample means across groups and periods that respect Assumption PT-GT-Nev.

When Assumption PT-GT-all holds instead, one can construct plug-in estimators for (5.8):

$$\widehat{ATT}_{\text{avg}}(g,t) = \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i = g\}\omega_i(Y_{i,t} - \overline{Y}_{i,t \leqslant g-1})}{\sum_{i=1}^{n} \mathbf{1}\{G_i = g\}\omega_i} - \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i > t\}\omega_i(Y_{i,t} - \overline{Y}_{i,t \leqslant g-1})}{\sum_{i=1}^{n} \mathbf{1}\{G_i > t\}\omega_i}.$$

Alternatively, Wooldridge (2021) proposed constructing estimators for $ATT(g,t)$ based on Assumption PT-GT-all, using following "extended" TWFE specification:

$$Y_{i,t} = \theta_t + \eta_i + \sum_{g \neq \infty} \sum_{s=g}^{T} \beta_{g,t}^{W} \mathbf{1}\{G_i = g\}\mathbf{1}\{s = t\} + \epsilon_{i,t}, \tag{5.12}$$

where the $\beta_{g,t}^{W}$'s are estimated using (weighted) least squares. Wooldridge (2021) shows that $\widehat{\beta}_{g,t}^{W}$ consistently estimate $ATT(g,t)$ under Assumption PT-GT-all, though we do not know the exact way that $\widehat{\beta}_{g,t}^{W}$ combines pre-treatment periods and not-yet-treated units, which is to say that we do not know the statistical estimand associated with $\beta_{g,t}^{W}$. Wooldridge (2021) also shows that $\widehat{\beta}_{g,t}^{W}$ is numerically the same as the "imputation" estimators proposed by Gardner (2021), Liu et al. (2024), and Borusyak et al. (2024) with balanced panel data (and these specifications do not have covariates).[29]

---

[28]Recently, Dube, Girardi, Jordà and Taylor (2024) show that one can also get $ATT(g,t)$ estimates that are equivalent to $\widehat{ATT}_{\text{nyt}}(g,t)$ by using local projections. One can also get similar estimators using a "stacked DiD" procedure akin to what Fadlon and Nielsen (2021), Deshpande and Li (2019), and Cengiz et al. (2019) have implemented.

[29]Imputation procedures work in two-steps. The first step uses all untreated observations to run the regressions $Y_{it} = \theta_t + \eta_i + \epsilon_{it}$ using only data from the (i,t) pairs that satisfy this criterion, and get the fitted values $\widehat{Y}_{i,t}(0) = \widehat{\theta}_t + \widehat{\eta}_i$ for all eventually-treated observations. The second step estimates $ATT(g,t)$ by

Overall, these different $ATT(g,t)$ estimators highlight that we can leverage our DiD expertise built in the 2×2 setup to estimate heterogeneous $ATT(g,t)$ parameters. The exact estimator we can use is shaped by which parallel trends assumption holds. In our Medicaid application, we report in Figure 6 our $ATT(g,t)$ estimates based on Assumption PT-GT-NYT and (5.9). As we have four sets of counties defined by their Medicaid expansion timing, we report four sets of event studies, one for each expansion group. For the 2014, 2016, and 2019 expansion groups, we find that Medicaid did not lead to significant changes in adult mortality rates. For the 2015 expansion group, adult mortality rates rose after expansion. Regarding pre-trends, Figure 6 suggests that there may be some non-negligible pre-trends for the 2016 expansion group, though these are not statistically different from zero.

In Section 4, we highlighted that unconditional assumptiosn such as Assumption PT-GT-NYT may fail in our Medicaid context. One reason is that important determinants of changes in untreated adult mortality rates are imbalanced across treatment and comparison groups. In such cases, one should interpret the results in Figure 6 with great care. In addition, it is also important to highlight that, as indicated in Table 1, the 2015, 2016, and 2019 expansion groups are relatively small and represent only 6%, 2%, and 3% of the US population, respectively. Thus, analyzing these groups separately may be "too noisy" and not representative of the overall effect for the US population. This does not mean that these $ATT(g,t)$'s are not useful. They are actually an essential part of our DiD analysis, but we may want to aggregate the cohorts to get more informative target parameters for the overall treated population. We now turn to how to aggregate the $ATT(g,t)$'s and then discuss how to incorporate covariates.

5.2.4 Aggregating group-time average treatment effects

The previous section highlighted that, in many applications, estimating all $ATT(g,t)$'s precisely and attaching policy-relevant interpretations to them may be challenging. Aggregating them into a summary treatment effect measure therefore has clear benefits: it improves precision, reduces the number of results, and yields a parameter that averages over all treated units like the $ATT(2)$ identified in 2×2 designs.

Aggregation in staggered designs involves a notion of time (either calendar time t or event-time $e = t - g$), a length of time (how many periods to aggregate across), and group weights (so larger treatment groups can "matter more" than smaller ones). Given some set of weights, it is simple

---

$\widehat{ATT}_{\text{imp}}(g,t) = \frac{\sum_{i=1}^{n} \mathbf{1}\{G_i=g\}(Y_{i,t}-\widehat{Y}_{i,t}(0))}{\sum_{i=1}^{n} \mathbf{1}\{G_i=g\}}$. See Gardner (2021), Liu et al. (2024), and Borusyak et al. (2024) for details. Note that the specification in (5.12) is similar to the Sun and Abraham (2021)'s specification (5.11), but it omits the pre-treatment event-time dummies. This is justified because (5.12) imposes parallel pre-trends (Assumption PT-GT-all) while (5.11) does not (it effectively relies on Assumption PT-GT-Nev). Thus, in general, one should not expect $\widehat{\beta}_{g,t}^{W}$ to be equal to $\widehat{\beta}_{g,t-g}^{SA}$.

Figure 6: **ATT(g,t)s for Each Expansion Group**
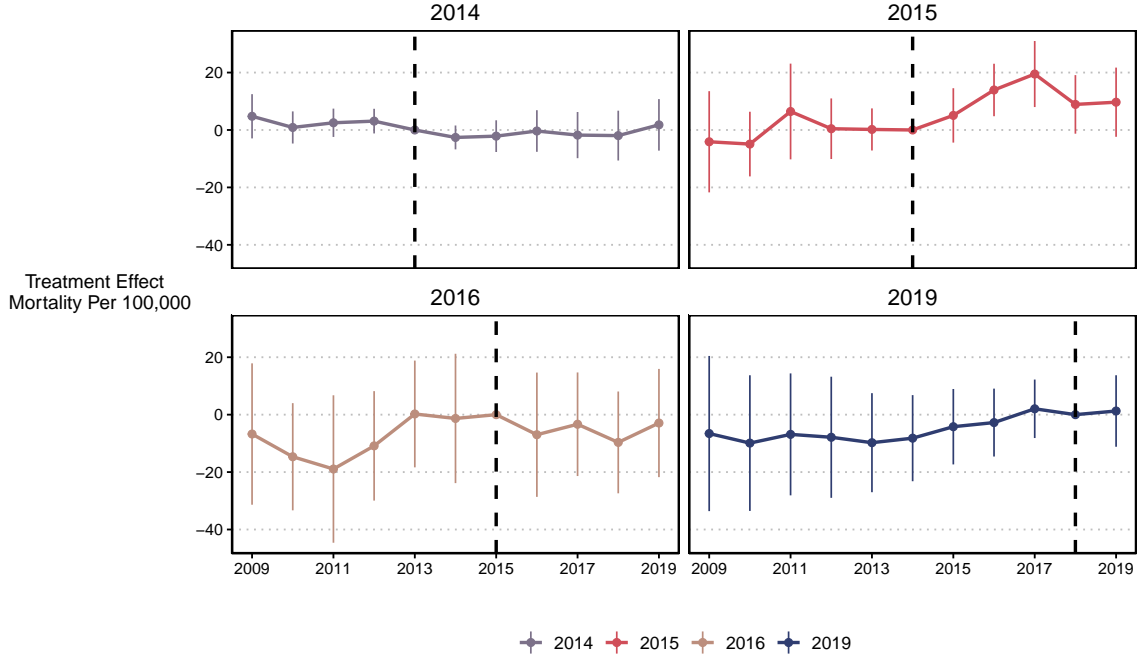
Notes: This figure shows the group-time ATT estimates ($ATT(g,t)$) in calendar time for the four groups of counties that expanded Medicaid before 2019. It uses not-yet-treated units as the comparison group and shows their uniform confidence intervals at the 95% significance level. The outcome variable is the crude mortality rate for adults ages 20-64, and standard errors are clustered at the county level. The vertical line represents the year before Medicaid expansion (i.e., g - 1) for the timing group.

to average the $ATT(g,t)$ building blocks into many kinds of summary parameters:

$$ATT_{\text{aggte}} = \sum_{g,t} w_{\omega,g,t} ATT(g,t),$$

where $w_{\omega,g,t}$ is a "generic" (ω-weighted) group and time-specific (non-negative) weight that sums up to one. Specific choices for the $w_{\omega,g,t}$ weights map to different ways to aggregate and present interpretable causal effects in this kind of complex setting.

As highlighted in Section 5.1, an appealing feature of having access to data from multiple periods is that we can assess how average treatment effects evolve with the time since treatment, or event-time $e = t - g$. Figure 6 displays event studies for each expansion group, and the aggregation question is how to combine them into a single summary event study.

A basic observation about weighting is that we can give positive weights only to the $ATT(g,t)$'s that we actually identified and estimated. Earlier treated groups have $ATT(g,t)$ estimates for later event-times by definition (and later treated groups have estimates for earlier pre-trends), so it will not be possible to include every group in an aggregated event-study parameter at every event-time. To see which $ATT(g,t)$'s will contribute to our event study, Figure 7 recenters our $ATT(g,t)$ estimates in event time instead of calendar time. That is, we plot $ATT(g, g+e)$ against e for each expansion group.

We will take "vertical" weighted averages of each available $ATT(g, g + e)$ for each event time e based on Figure 7. For instance, to estimate an aggregate event study in event time 0 (a

47

Figure 7: **ATT(g, t) in Event Time**
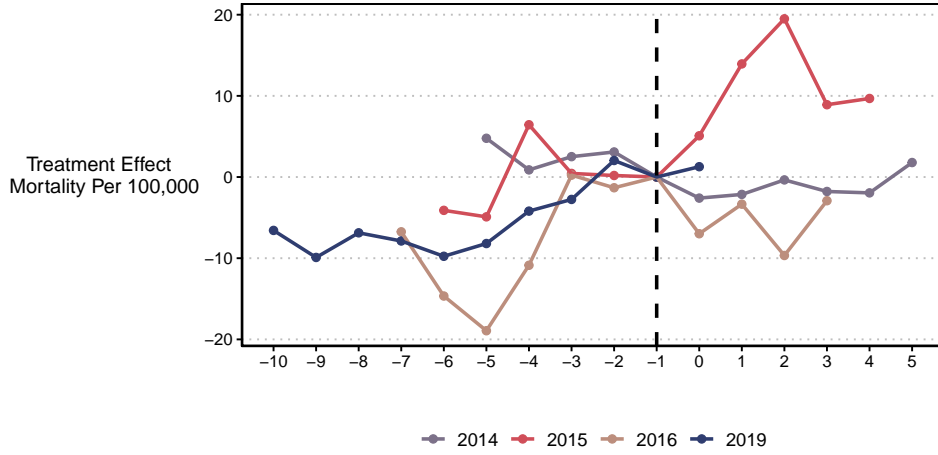
Notes: This figure shows the group-time ATT estimates ($ATT(g,t)$) in relative event time for the four treatment timing groups of counties that expanded Medicaid before 2019, using not-yet-treated units as the comparison group. The outcome variable is the crude mortality rate for adults ages 20-64.

measure of instantaneous treatment effects), we would average estimates of $ATT(2014, 2014)$, $ATT(2015, 2015)$, $ATT(2016, 2016)$ and $ATT(2019, 2019)$. When we are interested in event time 1, we would now average $ATT(2014, 2015)$, $ATT(2015, 2016)$, $ATT(2016, 2017)$. The same logic applies to other event times.

When constructing timing-group weights at a given event-time, it is also important to account for group sizes so that the resulting parameters equal sensible averages of treated units. Table 1 contains all the information necessary for this. The 2014 expansion group accounts for 80% of treated adults in the groups we consider, while the 2016 expansion group accounts for 3.5%. If we would like our aggregate event study to be a representative summary of the dynamic effects among treated counties, we should choose weights that are proportional to the treatment group size.

Putting these pieces together, we can formally state the exact summarized causal parameter that highlights treatment effect dynamics in terms of event time:

$$ATT_{\text{es}}(e) = \mathbb{E}_\omega \left[ATT(G, G + e) \middle| G + e \in [1, T], G \leqslant T \right]$$
$$= \sum_{g < \infty} w_{\omega, g, e}^{es} ATT(g, g + e), \tag{5.13}$$

where each weight $w_{g,e}^{es}$ gives the share of a group $G = g$ among treated units that have been exposed to treatment for exactly e periods (the groups that we have data for event time e in Figure 7), and is formally defined as

$$w_{g,e}^{es} = \mathbf{1}\{g + e \leqslant T\} P_\omega (G = g | G + e \leqslant T, G \leqslant T).$$

Note that $ATT_{\text{es}}(e)$ gives the average treatment effect among the units that have been exposed to treatment for exactly e periods, conditional on being observed having participated in the treatment for that number of periods (the condition that $G + e \in [1, T]$) and ever-participating in the

treatment by period T ($G \leqslant T$). One can also take a simple average of all available post-treatment event times, $ATT_{es}(e), e \geqslant 0$, and report an overall ATT measure. See Callaway and Sant'Anna (2021) for a discussion of alternative aggregations based on calendar time and groups.

Estimating $ATT_{es}(e)$ is straightforward and, once again, relies on the plug-in principle: we need to replace $ATT(g, g + e)$ with its sample analogs (which we already computed in Figure 6), and use the relative adult population share of expansion group g among eventually-treated units as estimates of the event study weights. Figure 8 reports our population-weighted estimates of the event-study aggregation from event times -5 to 5, which respectively correspond to 5 years before the Medicaid expansion and 5 years after the Medicaid expansion. We also report pointwise and simultaneous confidence intervals at the 95% significance level. Overall, the results suggest that Medicaid expansion has no effect on adult mortality rates among counties that eventually experience a Medicaid expansion. The pre-trends are also fairly close to zero, suggesting that our parallel trends assumption may be reasonable.

Figure 8: **G × T Event Study without Covariates**
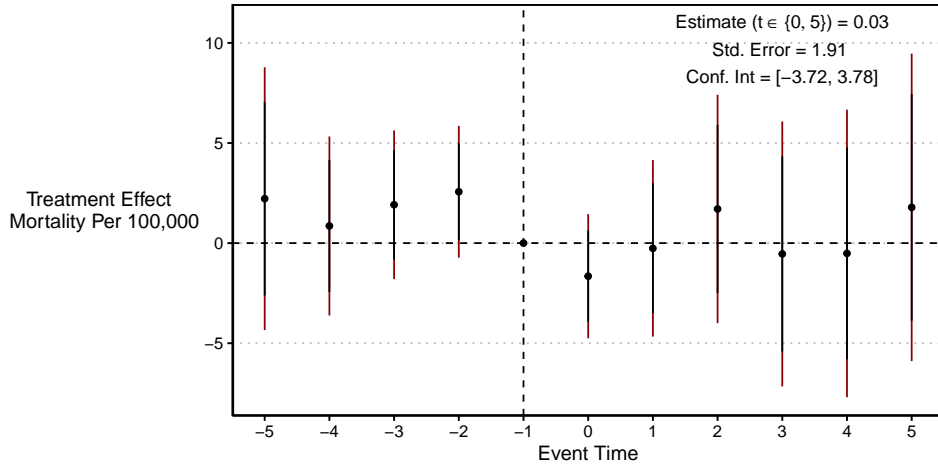
Notes: This figure shows the event study estimates with staggered treatment timing using the doubly-robust estimation method from Callaway and Sant'Anna (2021), using not-yet-treated units as the comparison group. The outcome variable is the crude mortality rate for adults ages 20-64. The point estimate is reported by the circles, and both 95% point-wise (black) and simultaneous (red) confidence intervals are reported with the vertical lines. We also report the simple average of all non-negative event times as a summary of the overall ATT (together with their standard errors and 95% confidence interval).

We conclude this section by stressing that the way we have constructed the event study parameters in Figure 8 uses all available information from Figure 6. A potential drawback of this strategy is that we do not always use the same set of groups across all event times. Practitioners usually refer to this as "imbalance in event time." For instance, the 2019 expansion group contributes only to event time $e = 0$, not $e = 1$ or later event times. When compositional changes are a concern, one can impose balance in event time and estimate a balanced event study aggregation:

$$ATT_{es,bal,[\underline{e},\overline{e}]}(e) = \mathbb{E}\left[ATT(G, G + e)\middle| G + \overline{e} \in [1, T], G + \underline{e} \in [1, T], G \leqslant T\right]$$

$$= \sum_{g \in \mathcal{G}_{treat}} w^{es,bal}_{g,[\underline{e},\overline{e}]} ATT(g, g + e), \tag{5.14}$$

where balanced event-time weights $w_{g,[\underline{e},\overline{e}]}^{es,bal}$ are given by

$$w_{g,[\underline{e},\overline{e}]}^{es,bal} = \mathbf{1}\{g+\overline{e} \leqslant T\}\mathbf{1}\{g+\underline{e} \geqslant 1\}P_\omega(G = g|G+\overline{e} \in [1,T], G+\underline{e} \in [1,T], G \leqslant T).$$

Although intimidating, $w_{g,[\underline{e},\overline{e}]}^{es,bal}$ just measures the relative size of a particular treatment group that was kept in the balanced data. We can interpret $ATT_{\text{es,bal},[\underline{e},\overline{e}]}(e)$ as the average group-time average treatment effect among units whose event time is equal to e *and is observed to participate in the treatment for at least $\overline{e}$ periods, and have at least $\underline{e}$ available pre-treatment periods* (if $\underline{e}$ is negative).[30]

5.2.5 Estimators for staggered designs with covariates

As the discussions in Section 5.2.2 made it clear, we can view the staggered DiD setups as a collection of simpler 2×2 DiD building blocks. A benefit of this interpretation is that, if parallel trends holds only after conditioning on the covariates that determine untreated potential outcome changes, we can easily leverage all the results discussed in Section 4 to identify, estimate and make inference about the $ATT(g,t)$'s, using regression adjustment, inverse probability weighting, or doubly robust methods.

Of course, to proceed in this manner, we would need to adopt conditioned on covariates versions of Assumption PT-GT-Nev, PT-GT-NYT or PT-GT-all, as well as impose an overlap condition. Given that all these are fairly similar to each other, here we state only an extension of Assumption PT-GT-NYT and a strong overlap condition that can be used for all three cases.

Assumption CPT-GT-NYT (Conditional Parallel Trends based on not-yet-treated groups)**.** For every eventually treated group g, not-yet-treated group g', time periods t such that $t \geqslant g$ and $g' > t$, and every covariate value X_i,

$$\mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g, X_i] = \mathbb{E}_\omega[Y_{i,t}(\infty) - Y_{i,t-1}(\infty)|G_i = g', X_i].$$

Assumption SO-GT (Strong overlap with staggered adoption)**.** For every group $g \in \mathcal{G}$, the conditional (weighted) probability of belonging to a treatment group g, given observed covariates X_i that are determinants of untreated potential outcome growth, is uniformly bounded away from zero and one. That is, for some $\epsilon > 0$ and for every group $g \in \mathcal{G}$, $\epsilon < P_\omega[G_i = g|X_i] < 1 - \epsilon$.

Using these identifying assumptions and building on the results in Sections 4 and 5.2.2, we can follow the arguments in Callaway and Sant'Anna (2021) and establish that the post-treatment $ATT(g,t)$'s are identified by the regression-adjusted, inverse-probability-weighted, and doubly-

---

[30]This discussion assumes that there are no "holes" in event-times for each treatment group. It is straightforward to adjust the interpretation to those more complicated cases, as the same logic can be applied.

robust estimands given by[31]

$$ATT_{\text{ra}}(g,t) = \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|G_i = g] - \mathbb{E}_\omega\Big[\mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|X_i, G_i > t]\Big|G_i = g\Big],$$

$$ATT_{\text{ipw}}(g,t) = \mathbb{E}\left[\Big(w_{\omega,G=g}(G_i) - w_{\omega,g,t}(G_i, X_i)\Big)(Y_{i,t} - Y_{i,t=g-1})\right],$$

$$ATT_{\text{dr}}(g,t) = \mathbb{E}\left[\Big(w_{\omega,G=g}(G_i) - w_{\omega,g,t}(G_i, X_i)\Big)\Big(Y_{i,t} - Y_{i,t=g-1} - \mathbb{E}_\omega[Y_{i,t} - Y_{i,t=g-1}|X_i, G_i > t]\Big)\right],$$

where $(w_{\omega,G=g}(G_i)$ and $w_{\omega,g,G>t}(G_i, X_i)$ are the analogs of the weights in (4.7) and are defined as

$$w_{\omega,G=g}(G) = \omega\mathbf{1}\{G = g\}\Big/\mathbb{E}[\omega\mathbf{1}\{G = g\}],$$

$$w_{\omega,g,t}(G, X) = \frac{\omega\mathbf{1}\{G > t\}\mathbf{1}\{G \neq g\}p_{\omega,g,t}(X)}{1 - p_{\omega,g,t}(X)}\Bigg/\mathbb{E}\left[\frac{\omega\mathbf{1}\{G > t\}\mathbf{1}\{G \neq g\}p_{\omega,g,t}(X)}{1 - p_{\omega,g,t}(X)}\right],$$

and $p_{\omega,g,t}(X) = \mathbb{E}_\omega[\mathbf{1}\{G_i = g\}|X, \mathbf{1}\{G_i = g\} + \mathbf{1}\{G_i > t\} = 1]$ denote the (weighted) probability of belonging to the group g given covariates X and that the unit belongs to either to group g—the treated group for the $ATT(g,t)$ of interest—or the not-yet-treated group $G_i > t$— the comparison group.

Estimating the $ATT(g,t)$'s follows exactly as in Section 4, and event study aggregations follow from the arguments in Section 5.2.4. Figure 9 reports event study summary estimates incorporating covariates into the Medicaid analysis. Like Figure 8, it suggests that Medicaid expansion had no effect on adult mortality among counties that expanded Medicaid by 2019. Given the point estimates and uniform confidence interval, we can be reasonably confident that the treatment effects are not greater than 6 or less than 11 deaths per 100,000 adults for the 6 year period following Medicaid expansion.

5.3 Limitations of TWFE regressions

Our framework emphasizes building an estimator from 2×2 components, each of which targets a well-defined ATT parameter under a specific parallel trends assumption. The most common estimator for staggered designs, a TWFE regression, comes instead from extending convenient estimation tools that work well in the 2×2 case. A TWFE specification that estimates a summary treatment effect parameter is:

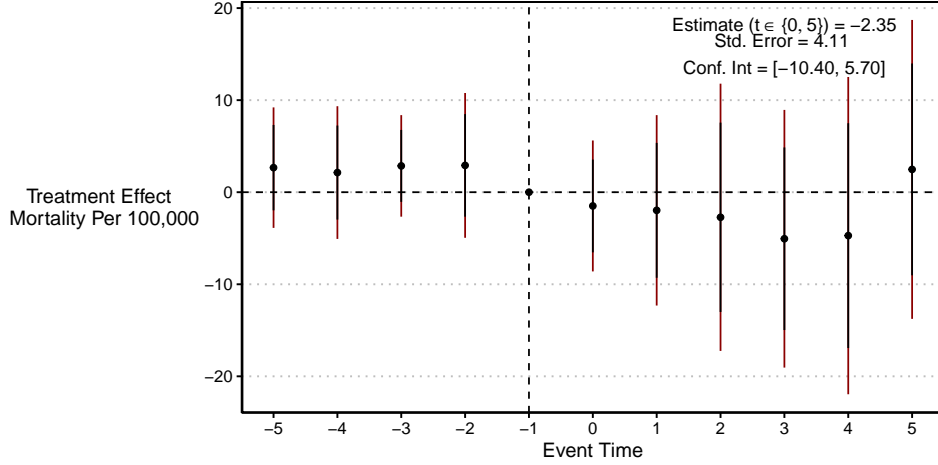$$Y_{i,t} = \theta_t + \eta_i + \beta^{twfe}D_{i,t} + e_{i,t}. \tag{5.15}$$

In this section, we abstract from weights.

A major breakthrough in recent DiD research has been to demonstrate two potentially large problems with β^{twfe} (de Chaisemartin and D'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and

---

[31]Wooldridge (2021) propose alternative estimators for the $ATT(g,t)$ that incorporate covariates and their interactions with group and time dummies into (5.12). Although Borusyak et al. (2024) and de Chaisemartin and D'Haultfoeuille (2020) also allow for covariates in their estimation procedure, they affect the outcome levels and not their changes over time. Thus, to some extent, these procedures do not build on conditional parallel trends like Assumption CPT-GT-NYT.

Figure 9: **G × T Event Study with Covariates**

Notes: This figure shows the event-study estimates with staggered treatment timing, using the doubly-robust estimation method from Callaway and Sant'Anna (2021). The outcome variable is the crude mortality rate for adults ages 20-64, and the covariates include the percentage of the county population that is female, the percentage of the county population that is white, the percentage of the county population that is Hispanic, the unemployment rate, the poverty rate, and county-level median income. The point estimate is reported by the circles, and both 95% point-wise (black) and simultaneous (red) confidence intervals are reported with the vertical lines. We also report the simple average of all non-negative event times as a summary of the overall ATT (together with their standard errors and 95% confidence interval).

Abraham, 2021; Borusyak et al., 2024). The primary issue comes from the fact that TWFE implicitly uses already-treated comparison groups. Even if PT holds for all groups and all periods, the resulting estimand can actually put negative weight on certain $ATT(g,t)$ parameters. The only way TWFE avoids the problem is if treatment effects do not change over time, a strong additional assumption.

To isolate this issue, consider a setting with two time periods and three groups: a group that enters treatment in the first period ($G_i = 1$), a group that becomes treated in the second time period ($G_i = 2$), and a never treated group ($G_i = \infty$). This is a staggered design because group 1 and group 2 are treated at different times, but because there are only two time periods, we can re-write the TWFE specification as

$$\Delta Y_{i,2} = \Delta\theta_t + \beta^{twfe}\Delta D_{i,2} + \Delta e_{i,2}.$$

Because $\Delta D_{i,2}$ takes only two values—1 for units whose treatment status increases ($G_i = 2$), and 0 for units whose treatment status does not change ($G_i = 1 \ and \ G_i = \infty$)—the TWFE estimand is the following simple comparison of means:

$$\begin{aligned} \beta^{twfe} &= \mathbb{E}[\Delta Y_{i,2}|\Delta D_{i,2} = 1] - \mathbb{E}[\Delta Y_{i,2}|\Delta D_{i,2} = 0] \\ &= \Big(\mathbb{E}[\Delta Y_{i,2}|G_i = 2] - \mathbb{E}[\Delta Y_{i,2}|G_i = \infty]\Big)(1 - w_1) + \Big(\mathbb{E}[\Delta Y_{i,2}|G_i = 2] - \mathbb{E}[\Delta Y_{i,2}|G_i = 1]\Big)w_1, (5.16) \end{aligned}$$

where $w_1 = \frac{p_1}{p_1 + p_\infty}$ and $p_g = P(G = g)$ is group g's share of units. The TWFE coefficient, in this case, is a weighted average of two DiD terms that use the already-treated units or the never-treated units as comparisons. We have already discussed how under PT between group $G_i = 2$ and never-treated units, the first term in the $\widehat\beta^{twfe}$ decomposition, $\mathbb{E}[\Delta Y_{i,2}|G_i = 2] - \mathbb{E}[\Delta Y_{i,2}|G_i = \infty]$, equals $ATT(2,2)$. But what about the second term with the already treated comparison group?

It turns out that under parallel trends and no-anticipation, if we add and subtract different terms, this type of estimand generally equals a combination of treatment effects for both groups:

$$
\begin{aligned}
\mathbb{E}[\Delta Y_{i,2}|G_i = 2] - \mathbb{E}[\Delta Y_{i,2}|G_i = 1] \;=\;& \mathbb{E}[Y_{i,2}|G_i = 2] - \mathbb{E}[Y_{i,1}|G_i = 2] \\
& - \Big(\mathbb{E}[Y_{i,2}|G_i = 1] - \mathbb{E}[Y_{i,1}|G_i = 1] \Big) \\
=\;& \mathbb{E}[Y_{i,2}(2) - Y_{i,2}(\infty)|G_i = 2] + \mathbb{E}[Y_{i,2}(\infty) - Y_{i,1}(\infty)|G_i = 2] \\
& - \Big(\mathbb{E}[Y_{i,2}(1) - Y_{i,2}(\infty)|G_i = 1] - \mathbb{E}[Y_{i,1}(1) - Y_{i,1}(\infty)|G_i = 1] \Big) \\
& - \mathbb{E}[Y_{i,2}(\infty) - Y_{i,1}(\infty)|G_i = 1] \\
=\;& ATT(2,2) - \Big(ATT(1,2) - ATT(1,1) \Big) \\
& + \mathbb{E}[Y_{i,2}(\infty) - Y_{i,1}(\infty)|G_i = 2] - \mathbb{E}[Y_{i,2}(\infty) - Y_{i,1}(\infty)|G_i = 1] \\
=\;& ATT(2,2) - \Big(ATT(1,2) - ATT(1,1) \Big),
\end{aligned}
$$

where the first equality follows from the linearity of expectations; in the second, we add and subtract averages of untreated potential outcomes and explore no-anticipation, and the observation rules that $Y_{i,t}(g)$ is observed for units with $G_i = g$; in the third equality we rearrange terms; and the last equality follows from a parallel trends for units in group $G = 1$ and $G = 2$.

The DiD estimand with an already treated comparison group thus equals:

$$
\beta^{twfe} = ATT(2,2) + ATT(1,1)w_1 - ATT(1,2)w_1. \tag{5.17}
$$

The $ATT(1,2)$ building block receives *negative* weight in the overall TWFE estimand unless there are no treatment effect dynamics ($ATT(1,1) = ATT(1,2)$). In this example, the problem is easy to fix: drop the always-treated units and target $ATT(2,2)$. With multiple periods, however, the problem is more complex.

There are two primary results regarding TWFE for general staggered timing designs. One is a decomposition of the TWFE *estimator*. Goodman-Bacon (2021) expresses the TWFE estimator as a weighted average of all possible 2×2 DiD comparisons between pairs of groups and time periods during which one group enters treatment and the other does not. The terms in his decomposition are not two-period $ATT(g,t)$-type estimators; they aggregate over the relevant pre- and post-periods just as β^{OLS} from (5.4) does. They include many comparisons to already-treated units like (5.16). This mechanical decomposition always has strictly positive weights that are larger for larger groups and for groups treated closer to the middle of the panel, which have a larger variance of $D_{i,t}$ conditional on the fixed effects. This result shows how TWFE estimators actually function and creates a clear link to the estimation approaches we outlined above for summary parameters. Both are averages of 2×2 DiD terms, but they differ in which comparisons they use and how they aggregate.

The second result is a decomposition of the TWFE *estimand*, which clarifies why the TWFE estimator is not guaranteed to identify a desirable parameter in staggered designs, even if parallel trends holds. The reason is their use of already-treated comparisons like the ones in (5.17) (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfoeuille, 2020; Borusyak et al., 2024; Imai

and Kim, 2021; Strezhnev, 2018; Sun and Abraham, 2021). This will tend to bias β^{twfe} away from the sign of ATT_{avg}, and β^{twfe} can even have the opposite sign of ATT_{avg} (Baker et al., 2022). It may appear that a more flexible regression specification could solve this problem, but Sun and Abraham (2021) show that a TWFE event study specification suffers from a similar bias when the dynamics of the $ATT(g,t)$'s differ across cohorts. Moreover, the "variance-weighting" feature of OLS means that β^{twfe} has non-intuitive weights even when $ATT(g,t) = ATT(g)$.

While TWFE remains common, it has well-understood, potentially serious, and easily remedied problems, and we do not recommend using it. In many cases, especially those with many untreated units or minimal treatment effect dynamics, TWFE estimates may be similar to those derived from the theoretically grounded estimators discussed above. The only way to be sure, however, is to estimate both. In that case, it is unclear why a researcher would not report the estimates from a procedure motivated by a desirable target parameter and a credible PT assumption.

6 Conclusion

The starting point of this paper was a 2×2 DiD design that researchers have been using for almost 200 years. The end point was a design with five treatment groups, 11 years of data, six covariates, three types of parallel trends assumptions, and four estimation techniques. Our fundamental message is that without understanding how complex designs are built up from simpler ones, it is exceedingly difficult to navigate all the empirical tools now available for DiD designs. This lesson applies not only to the design details we considered here—weighting, covariates, and staggered designs—but to any DiD design.

The forward-engineering philosophy we followed in this paper suggests a set of steps that researchers can follow in any DiD study:

Step 1. *Define target parameters.* Adopt a potential outcomes notation that fits the study's specific setting and use it to define causal target parameters that answer the study's motivating question. Building block causal parameters usually aggregate across units using (conditional) weighted averages, and summary target parameters aggregate across the building blocks. This step fixes the study's goals in terms of causal quantities and facilitates comparisons with related studies.

Step 2. *State (formally) the identification assumptions.* DiD studies leverage parallel trends assumptions, but they also rely on no-anticipation and, in some cases, overlap conditions, or more. Be explicit about which form of these assumptions is required for identification in the study. Engage with the theoretical arguments necessary for them to hold and generate appropriate empirical evidence, such as pre-treatment differential trends, that can falsify or (indirectly) support their plausibility.

Step 3. *Determine the appropriate estimation method.* In some DiD designs, estimation is as simple as replacing population expectations with sample means. In others, such as conditional DiD designs, estimation involves choosing econometric techniques (e.g., a regression adjustment, inverse probability weighting, or doubly robust procedure) to map theoretical quantities to estimable sample quantities. Each of these strategies relies on additional modeling restrictions that should be stated clearly.

Step 4. *Discuss sources of uncertainty.* Statistical inference procedures for DiD designs stem from basic assumptions about where randomness comes from in a given design. Some researchers may adopt a sampling approach to inference, whereas others may be more comfortable with a design-based perspective. It is important to discuss what variables of the model are being treated as fixed and what variables are considered random, as well as to use inference techniques that are compatible with the model structure and assumptions.

Step 5. *Estimate.* Steps 1-4 provide a specific structure for using data to estimate the causal parameters of interest.

Step 6. *Conduct sensitivity analysis.* A clear statement of the identification and estimation assumptions also facilitates a clear statement of what violations of those assumptions might mean. No study is robust to all the ways its assumptions may fail, but a good study should be robust against likely violations of plausible magnitudes. Combine context-specific knowledge about how the assumptions from Step 2 might be violated, and by how much, with the structure of the estimator from Step 3 to evaluate how much the DiD estimates vary if the key identification assumptions are not exactly true.

Step 7. *Conduct heterogeneity analysis.* Sometimes aggregate parameters defined in Step 1 mask important heterogeneity, in which case the forward-engineering approach simply suggests targeting sub-group parameters as well. This can include variation in parameters over time, between groups of units with different characteristics, or across different sources of treatment variation. Be clear about which types of heterogeneous effects are relevant and how they are identified and estimated.

Step 8. *Keep learning.* DiD is not the only or the best research design in all settings; it is just one of many causal inference techniques. If the assumptions required for a DiD analysis appear implausible *ex-ante* or are refuted by evidence or non-robustness in practice, then explore different designs. If existing DiD methods do not provide enough guidance, then use a forward engineering approach to deduce what advances would help.

Some researchers may still prefer to use standard regression tools to conduct DiD studies. The properties and pitfalls of some popular regression specifications are now well understood, and one can easily explain how this choice fits with (and perhaps satisfies) the steps above. But using

simple regressions in any DiD-type setting is an implicit choice to reverse-engineer a research design from the statistical method, rather than forward-engineer a reliable estimator from a substantive question and transparent assumptions. Ultimately, important questions and credible identification strategies should guide DiD analyses (regression-based or not), not the other way around.

Although this paper is by no means an exhaustive guide to DiD practice, the eight steps above are a rigorous framework for tackling all the DiD topics that we did not cover. The Appendix briefly discusses DiD methods for (a) treatments that turn on and off over time, (b) continuous and multi-valued treatments, (c) triple differences, (d) distributional parameters, and (e) repeated cross-section or unbalanced panel data. While each of these designs differs from what we covered in the main text, a forward-engineering approach that moves from defining parameters and assumptions, to settling on estimation and inference techniques, to probing robustness, applies equally to all of them. While the specifics of any given DiD analysis may change across research questions, treatment variables, econometric techniques, and data structures, the principles by which one can conduct reliable and transparent causal inference stay the same.[32]

---

[32]There are other DiD topics of interest that we do not cover, including fuzzy DiD and instrumented DiD designs (de Chaisemartin and D'Haultfoeuille, 2018; Miyaji, 2024), nonlinear DiD models (Wooldridge, 2023; Tchetgen Tchetgen, Park and Richardson, 2024), issues related to few clusters (Roth et al., 2023, Section 5), and situations with multiple treatments (de Chaisemartin and D'Haultfoeuille, 2023a; Yanagi, 2023). We also do not cover some methods that address violations of parallel trends (Freyaldenhoven et al., 2019; Arkhangelsky, Athey, Hirshberg, Imbens and Wager, 2021; Callaway and Karami, 2023; Imbens, Kallus and Mao, 2021), nor do we examine setups that impose as-good-as-random treatment timing (Athey and Imbens, 2022; Roth and Sant'Anna, 2023a; Arkhangelsky, Imbens, Lei and Luo, 2024).

List of Acronyms

Acronym	Definition
2×2	Two-Group Two-Time-Periods DiD
$2 \times T$	Two-Group T-Time-Periods DiD
ACA	Affordable Care Act
ATT	Average Treatment Effect on the Treated
CPT	Conditional Parallel Trends
CPT-GT-NYT	Conditional Parallel Trends Based on Not-Yet-Treated Groups
DiD	Difference-in-Differences
DR	Doubly Robust
ETWFE	Extended Two-Way Fixed Effects
IPW	Inverse Probability Weighted
NA	No Anticipation
NA-S	No Anticipation with Staggered Treatment Timing
OLS	Ordinary Least Squares
PT	Parallel Trends
PT-ES	Parallel Trends Event Study
PT-GT-all	Parallel Trends for Every Period and Group
PT-GT-Nev	Parallel Trends Based on Never-Treated Groups
PT-GT-NYT	Parallel Trends Based on Not-Yet-Treated Groups
RA	Regression Adjustment
SO	Strong Overlap
SO-GT	Strong Overlap With Staggered Adoption
TWFE	Two-Way Fixed Effects

A   Some additional DiD-related procedures

This section discusses some important DiD-related topics that we did not cover in our main text. These discussions are short by design, and we focus on providing the main ideas related to challenges and solutions specific to the problem. We abstract from weights and use $\mathbb{E}[\cdot|\cdot]$ to denote (conditional) expectations.

A.1   Setups with treatment turning on and off

Our main text focuses on setups where treatment remains in place from the period it begins until the end of the sample period, but in practice, some treatments turn on and off over time. This is the setting tackled by de Chaisemartin and D'Haultfoeuille (2020, 2023a), Imai, Kim and Wang (2023), and Liu et al. (2024).

To tackle this problem from first principles, we need to augment the potential outcomes to reflect the richer notion of treatment *sequences*. Following Robins (1986), let $Y_{i,t}(\mathbf{d})$ denote the potential outcome for unit i at time t if this unit received the T-dimensional treatment sequence $\mathbf{d} \in \{0,1\}^T$. For simplicity, let's say that $T = 3$ and that no unit is treated in the first period. In this case, we have four treatment sequences (or histories), which define four potential outcomes for each unit: $Y_{i,t}(0,0,0), Y_{i,t}(0,0,1), Y_{i,t}(0,1,0)$ and $Y_{i,t}(0,1,1)$. We then define treatment groups by treatment sequences: $G = \mathbf{d}_0 \equiv (0,0,0)$ (never-treated), $G = \mathbf{d}_1 \equiv (0,0,1)$ (treated in the third period), $G = \mathbf{d}_2 \equiv (0,1,1)$ (treated in the second and third period), and $G = \mathbf{d}_3 \equiv (0,1,0)$ (treated only in the second period). In general, we would have as many groups as we have different (realized) treatment sequences. Recall that in a staggered timing design with an absorbing treatment, treatment timing fully characterizes a treatment sequence.

Once potential outcomes and groups are well-defined, one can move to parameters of interest. We proceed similarly to the staggered treatment setup in Section 5.2 and consider group-and-time specific ATTs as building blocks, except that groups are now based on more complex treatment sequences. Let $\mathbf{0}$ denote a T-dimensional vector of zeros. One intuitive building block parameter on which to base a DiD analysis is

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

the average treatment effect at time period t of being exposed to treatment sequence $\mathbf{d}$ instead of never being exposed to treatment, among units that received treatment sequence $\mathbf{d}$.[33]

Next, one needs to establish identification for the parameters, and propose appropriate estimators and inference procedures. Following similar arguments to those in Section 5.2, a DiD approach to this problem would involve imposing a parallel trends assumption (potentially conditional on covariates) and a no-anticipation assumption to establish that the $ATT(\mathbf{d}, t)$'s are identified. If

---

[33]One could also adopt alternative building blocks not discussed here, such as the average effect of treatment lasting one period longer or a treatment spell of a given length beginning one period later.

each treatment group is sufficiently large, one could proceed in a similar fashion as the staggered setup, comparing average outcome paths for a given sequence with the average outcome path for never-treated (or not-yet-treated) units. One can also aggregate these different $ATT(\mathbf{d}, t)$ to form different summary parameters.

In practice, however, it is often the case that the number of treatment groups is large and each group is small. This essentially creates a "curse of dimensionality" problem: there are too many building block parameters defined for too-small groups to be estimated reliably. In such cases, additional assumptions that limit treatment effect dynamics (or how past treatments affect future outcomes) are often imposed, and different aggregated summary parameters are usually targeted. We provide a brief overview of several different solutions that have been proposed to address this issue.

de Chaisemartin and D'Haultfoeuille (2020) impose a "no-carryover" assumption that implies that past treatments do not affect future outcomes; which is to say that treatment effects in a given period last only during that period. With such an assumption (in addition to parallel trends and a no-anticipation assumption), they propose DiD estimators for an instantaneous average treatment effects parameter by comparing currently treated units with untreated units. Imai et al. (2023) adopt a similar approach, though they impose a limited-carryover assumption where treatments may last for ℓ periods (with ℓ specified by the researcher). They then propose estimators for an average treatment effect of switching into treatment in period t among units that experience the policy change in period t, and share the same treatment history over the previous k periods; see Liu et al. (2024) for a related procedure. Finally, de Chaisemartin and D'Haultfoeuille (2023a) avoid making assumptions related to carryover effects and extend the DiD framework in de Chaisemartin and D'Haultfoeuille (2020) to allow for treatment effect dynamics. The way they proceed is to first "staggerize" treatment sequences according to first-time of treatment exposure, compute a staggered DiD procedure for this "intention-to-treat" type parameter, and normalize them by a DiD estimate based on the number of treated periods. A potential challenge with deChaisemartin2023b-intertemporal-treatments's (2023a) approach is the interpretability of their proposed summary parameter, though we should acknowledge that this is a complex setup.

One important takeaway is that comparing these DiD procedures that allow for treatments to turn on and off may be challenging, as they target different causal parameters of interest, and practitioners should be aware of the different assumptions and limitations. We refer the reader to de Chaisemartin and D'Haultfoeuille (2023b) and Liu et al. (2024) for additional discussions on these types of DiD estimators.

A.2 DiD setups with continuous or multi-valued treatments

Our paper focuses on binary treatments, but many treatments take multiple values or are even continuous. A number of recent papers have studied this particular type of treatment de-

sign. These include Callaway, Goodman-Bacon and Sant'Anna (2021, 2024); de Chaisemartin, D'Haultfoeuille, Pasquier and Vazquez-Bare (2024a); and de Chaisemartin, D'Haultfuille and Vazquez-Bare (2024b). Here we focus on a two-period setting in which no unit is treated in period one and some units receive a treatment with varying intensities (or doses) in period two. Most of the key results that distinguish multi-valued from binary treatments are evident with two periods (Callaway et al., 2024).

We now need to define potential outcomes that reflect varying treatment intensity. We denote $Y_{i,t}(0, d)$ as potential outcomes for unit i in period t if they are untreated in period one and receive treatment dosage d in period two. As we focus on setups where all units are untreated in period one, we simplify notation and index potential outcomes by treatment intensity in period two; that is, $Y_{it}(d) = Y_{i,t}(0, d)$. An important feature is that d is not restricted to $\{0, 1\}$ and can take on richer treatment intensities instead. We denote the treatment dosage for unit i as D_i in period two and stress that in this context, our notion of the treatment group is tied to units' treatment dosage: groups are defined by their treatment dosage in period 2.

A multi-valued treatment defines several different types of causal parameters that may be of interest. For instance, dose-specific average treatment effect parameters such as

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

reflect the average effect of dose d relative to no treatment. Here $ATT(d|d')$ is the average treatment effect for units that experienced dose d'; when $d' = d$, it is the ATT among units that received dose d. On the right side, $ATE(d)$ is defined analogously, except that it is the effect on the overall population. Of course, one can also aggregate these dose-specific parameters to form more precisely estimable summary quantities; see, e.g., Callaway et al. (2021).

The two treatment effect parameters above provide average treatment effects in levels, and so one reason why they vary could be because d itself varies. To account for the differences in d, one may be interested in "per-dosage" effects:

$$ATT_{\mathrm{pd}}(d|d') = \frac{ATT(d|d')}{d} \quad \text{and} \quad ATE_{\mathrm{pd}}(d) = \frac{ATE(d)}{d}.$$

One can also aggregate these parameters across dosages to analyze $\mathbb{E}[ATT_{\mathrm{pd}}(D|D)|D > 0]$, an average treatment effect among treated (or, more generally, among switchers). One can also consider weighted averages of these to learn about $\mathbb{E}[ATT(D|D)|D > 0]/\mathbb{E}[D|D > 0]$; see de Chaisemartin et al. (2024a) for a general discussion about such target parameters.

Finally, researchers are often interested in the causal effect of a marginal increment in the dose. This notion is the average causal response (ACR), similar to Angrist and Imbens (1995), defined as follows (when the dose is absolutely continuous):

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

Here $ACRT(d|d)$ equals the derivative of the average potential outcome in period two for units

that received dose d evaluated at d—this is equivalent to the derivative of $ATT(l|d)$ with respect to l, evaluated at $l = d$. We can interpret $ACR(d)$ analogously.[34]

The relevant questions pertain to (a) what assumptions are needed to impose to identify these parameters, (b) how to estimate and make inferences about these parameters of interest once identification is established, (c) how to summarize treatment effect heterogeneity across doses to generate interpretable aggregated causal parameters, and (d) whether traditional regression specifications based on TWFE recover a sensible and easy-to-understand causal parameter of interest. These questions are addressed in detail by Callaway et al. (2021) and de Chaisemartin et al. (2024a).

Callaway et al. (2021) highlight how, when no units are treated in period one, identification and estimation of $ATT(d|d)$'s (or their functionals) follows the binary case. They propose flexible nonparametric estimators for the $ATT(d|d)$ curve—the relationship between outcome changes (minus the average change for untreated units) and the dose d, making it possible to visualize and make inference about treatment effect heterogeneity across dosages. They also propose estimators that aggregate across dosage values and can be more precisely estimated. The identification of causal response parameters or ATE-type parameters, however, requires a stronger version of parallel trends that holds for potential outcomes at non-zero treatment doses. Under these strong parallel trends and no anticipation assumptions, they discuss estimation and inference procedures for the ACR curves and their summary measures.[35]

de Chaisemartin et al. (2024a) consider the setup where units are already exposed to different levels of treatment in period one. They discuss how one can identify causal quantities that generalize $ATT_{\mathrm{pd}}(d|d')$ to this more complex setup when (a) a sizable number of units do not change treatment dosage over time (stayers), and (ii) there is no-carryover from past treatment to future outcomes. They propose estimation and inference procedures for aggregated parameters akin to $\mathbb{E}[ATT_{\mathrm{pd}}(D|D)|D > 0]$ and $\mathbb{E}[ATT(D|D)|D > 0]/\mathbb{E}[D|D > 0]$.

Lastly, these papers target different causal parameters, put more emphasis on different DiD designs, and, therefore, should be viewed as complements rather than substitutes. In our view, DiD with continuous treatment is another area in which more methodological research is warranted. See Callaway et al. (2021) and de Chaisemartin et al. (2024a) for a more thorough discussion of many other cases.

---

[34]For discrete treatments, ACR's are defined in a similar way but with a slightly different notation to accommodate the discreteness of d: $ACRT(d_j|d_k) = \mathbb{E}[Y_{t=2}(d_j) - Y_{t=2}(d_{j-1})|D = d_k]$, and $ACR(d_j) = \mathbb{E}[Y_{t=2}(d_j) - Y_{t=2}(d_{j-1})]$.

[35]Interestingly, they also show that commonly used TWFE regression specifications are too rigid to lead to easy-to-interpret causal parameters of interest. In fact, they show that one can provide several different decompositions of the TWFE treatment coefficient depending on the specific causal parameter being used as a building block for the analysis, though every decomposition considered by them has some issues related to negative-weighting, additional "bias" terms, or non-interpretable weights that can distort inference. They emphasize that all this can be easily resolved by adopting the forward-engineering approach.

A.3 Triple differences

The causal interpretation of DiD estimates depends on the plausibility of their identification assumptions, which involve a no-anticipation and a parallel trends condition. In some applications, however, these assumptions may not hold—for example, when the trends of average untreated outcomes among men and women vary across treatment groups. In these cases, a common empirical practice is to attempt to model these violations of parallel trends directly or to conduct sensitivity analysis (Freyaldenhoven et al., 2024; Rambachan and Roth, 2023). In some specific treatment designs in which treatment is rolled out to different units or groups (e.g., states), but is targeted to a specific subset (partition) of the population (e.g., women), it is possible to relax DiD-type parallel trends so that partition-specific and group-specific violations of parallel trends are allowed. Such setups are often referred to as "triple differences" (DDD). Since its introduction by Gruber (1994), DDD has become very popular among empirical researchers—see Olden and Møen (2022) for documentation. In this section, we provide a brief overview of the target parameters and identifying assumptions in DDD. We also highlight that, contrary to conventional wisdom, DDD procedures cannot generally be expressed as the difference between two DiD, especially when parallel trends assumptions only hold after conditioning on covariates or when treatment adoption is staggered. This discussion borrows heavily from Ortiz-Villavicencio and Sant'Anna (2025).

We start our analysis by discussing potential outcomes and treatment design. As we focus on binary treatments (with potential staggered adoption), the potential outcome is the same as discussed in the main text, with $Y_{i,t}(g)$ denoting the potential outcome for unit i in time t if first exposed to treatment in period g. In DDD setups, a unit i is *exposed* to treatment in period t if (i) it belongs to a group (e.g., state) that enabled treatment in period g and t is a post-treatment period, $t \geq g$, and (ii) it belongs to the subset of the population that qualifies (or is *eligible*) for treatment (e.g., women). Let $S \in \mathcal{S} \subseteq \{2, ..., T\} \cup \{\infty\}$ denote the time each group (e.g., state) enables the policy/treatment, with the notion that $S = \infty$ if the policy is not enabled in the observed time frame. We also denote the partition of the population that (eventually) qualifies for the treatment by Q with $Q_i = 1$ if unit i is (eventually) eligible for treatment and $Q_i = 0$ otherwise. With these notations, we can define the treatment groups G_i according to the first time a unit i is *exposed* to treatment; that is, $G_i = S_i$ if $Q_i = 1$ and $G_i = \infty$ if $Q_i = 0$.[36]

Similar to standard DiD designs, DDD is interested in the $ATT(g,t)$-type parameters discussed in Section 5.2.1. Given the particular structure of the DDD problem, we can write $ATT(g,t)$'s as

$$ATT(g,t) \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|S_i = g, Q_i = 1],$$

to stress that it measures the average treatment effect at time period t of first being exposed to treatment in period g versus not being exposed to treatment, among units that are actually exposed to treatment in period g, i.e., units that are in groups that the policy was first enabled

---

[36] Note that when all units are eligible for treatment, we have $G_i = S_i$, which gets us back to a (staggered) DiD setup.

in period g and that qualify for treatment. One can also analyze aggregations of these $ATT(g, t)$ parameters to form causal summary parameters that can be more precisely estimated and highlight treatment effect heterogeneity in some specific directions. This would follow the exact same steps as we discussed in Section 5.2.4, once again highlighting the importance of our forward-engineering approach.

Identifying these causal parameters involves a no-anticipation assumption and a (conditional) parallel trends assumption. Assumption NA-S can be recycled here, as DDD has the same empirical content as DiD when it comes to no-anticipation. The parallel trends assumption, though, needs to be adjusted as an empirical appeal of DDD is to allow for violations of Assumption PT-GT-all, or the other PT variations discussed in Section 5.2. Here, we consider a variation of Assumption PT-GT-all that holds only after conditioning on covariates and allows for some partition-specific and group-specific non-parallel trends.

Assumption DDD-PT-GT-all (DDD-Parallel Trends for every period and group)**.** For every group s and s' and time periods t, with probability one,

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s, Q = 1, X\right] \quad - \quad \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s, Q = 0, X\right]$$

$$=$$

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s', Q = 1, X\right] \quad - \quad \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty)|S = s', Q = 0, X\right].$$

When there are only two periods, $t = 1, 2$, and two groups, $S \in \{2, \infty\}$, and covariates play no role in terms of identification—that is, Assumption DDD-PT-GT-all holds without X (or, equivalently, with $X = 1$ for all units)—Olden and Møen (2022) show that one can identify $ATT(2, 2)$ as the difference of two DiD estimands:

$$
\begin{aligned}
ATT(2, 2) \quad = \quad & \mathbb{E}\left[Y_{t=1} - Y_{t=1}|S = 2, Q = 1\right] - \mathbb{E}\left[\left[Y_{t=1} - Y_{t=1}|S = 2, Q = 0\right]\right. \\
& - \left(\mathbb{E}\left[Y_{t=1} - Y_{t=1}|S = \infty, Q = 1\right] - \mathbb{E}\left[\left[Y_{t=1} - Y_{t=1}|S = \infty, Q = 0\right]\right).
\end{aligned}
$$

Estimation and inference would be straightforward, as one could use the analogy principle or a two-way fixed effects regression with triple interactions—see Olden and Møen (2022) for details.

Ortiz-Villavicencio and Sant'Anna (2025) show that DDD estimands cannot be written as the difference of two DiD estimands when covariates are important for identification, or when treatment adoption is staggered over time and one wants to use not-yet-treated units as a comparison group (as is commonly done in DiD setups). They show how ignoring these considerations and proceeding as if DDD were indeed just a difference of two DiDs can lead to severely biased estimates for the $ATT(g, t)$'s. Ortiz-Villavicencio and Sant'Anna (2025) also show how one can avoid these issues by adopting a forward-engineering approach to the DDD problem. They propose regression-adjusted, inverse probability weighting, and doubly robust estimators for DDD setups that can reliably recover $ATT(g, t)$ and their associated summary parameters under mild assumptions. The paper discusses using multiple comparison groups to generate more precise estimates than simply using

a single comparison group. Relatedly, Strezhnev (2023) discusses several limitations of common two-way fixed effects regression specifications commonly used for DDD analysis.

Sometimes researchers use the term "triple differences" to mean different things and often use different identification assumptions to estimate these different quantities. Caron (2025) discusses using a triple difference strategy to estimate treatment effect heterogeneity. We recommend that practitioners be transparent about target parameters, research designs, and identification assumptions to allow the research community to understand the goals and the differences between DDD procedures.

A.4   Distributional DiD procedures

Our paper focuses on learning about *average* treatment effects in various DiD setups. However, approaches that embrace heterogeneity can also target quantities that describe heterogeneity other than average treatment effect parameters. In some settings, researchers may want more information about the distributional impacts of treatment participation. For instance, if a policymaker faces two different labor market programs with very similar average effects on earnings, they may prefer the one that potentially has a higher impact on the lower tail of the income distribution. Difference-in-Differences-type strategies can also be used to identify, estimate, and make inferences about various distributional features of the outcome of interest. This area has received a substantial amount of methodological consideration by econometricians in recent years; see Athey and Imbens (2006), Bonhomme and Sauder (2011), Callaway, Li and Oka (2018), Callaway and Li (2019), Roth and Sant'Anna (2023b), Ghanem, Kédagni and Mourifié (2023), Fernández-Val, Meier, van Vuuren and Vella (2024b), and references therein. For some empirical literature using distributional DiD procedures, see Meyer, Viscusi and Durbin (1995), Finkelstein and McKnight (2008), and Cengiz et al. (2019), among many others.

An analysis of distributional quantities does not require different potential outcomes notation from Section 5.2; it just targets functionals of the potential outcome distributions other than their means. The first thing to notice is that there are several types of distributional causal parameters in the treated group that one may care about. The unique feature of them is that they are all functionals of $F_{Y_t(g)|G=g}(y) = \mathbb{P}(Y_t(g) \leqslant y | G = g)$ and $F_{Y_t(\infty)|G=g}(y) \equiv \mathbb{P}(Y_t(\infty) \leqslant y | G = g)$. Examples of such functionals include distributional treatment effects in time period t among units first treated in period g (denominated in probability units),

$$DTT(y|g,t) = F_{Y_t(g)|G=g}(y) - F_{Y_t(\infty)|G=g}(y),$$

quantile treatment effects in time period t among units first treated in period g (denominated in outcome units),

$$QTT(\tau|g,t) = F^{-1}_{Y_t(g)|G=g}(\tau) - F^{-1}_{Y_t(\infty)|G=g}(\tau),$$

where $F^{-1}_{Y_t(g)|G=g}(\tau) = inf\{y : F_{Y_t(g)|G=g}(y) \geqslant \tau\}$ denotes the τ-quantile of $Y_t(g)$ among units in

group $G = g$, and $F_{Y_t(\infty)|G=g}^{-1}(\tau)$ is defined analogously. Other functionals related to inequality measures can also be obtained; see Firpo and Pinto (2016) for a discussion on this topic.

To make inferences about these different causal parameters, one needs to identify $F_{Y_t(\infty)|G=g}(y)$ and $F_{Y_t(g)|G=g}(y)$. Identification of $F_{Y_t(g)|G=g}(y)$ is usually non-controversial, as we can use data from units in group $G = g$ to learn about the distribution of $Y_t(g)$. The main challenge is related to how to learn the counterfactual distribution $F_{Y_t(\infty)|G=g}(y)$ from the data. This is where different DiD-type procedures differ, as each paper in this literature relies on different and often non-nested identification assumptions that allow them to identify $F_{Y_t(\infty)|G=g}(y)$. Given the space constraints, we do not provide explicit and detailed discussion about how these different DiD-related distributional procedures function. However, all distributional DiD estimators share our forward-engineering approach; they clearly state their identification assumptions and target parameters and then provide estimators that recover well-defined causal quantities. We also note that most distributional DiD methodological papers focus on two-period and two-group setups. However, it is straightforward to build similar arguments to those in Section 5.2 to extend the designs to more general settings, which is again another benefit of the forward-engineering approach to causal inference.

We close this section by noting that there exist other types of distributional parameters of interest related to the distribution *of* the treatment effects in period t among the units in group g, $\mathbb{P}(Y_t(g) - Y_t(\infty) \leqslant y|G = g)$. In general, such causal quantities cannot be point identified, as discussed in Heckman, Smith and Clements (1997b), Fan and Yu (2012) and Callaway (2021). However, often one can still partially identify such policy-relevant parameters under different restrictions. We refer the reader to Callaway (2021) for a more detailed discussion of this topic.

A.5   Repeated cross-sections and unbalanced panel data

An appealing feature of DiD procedures is that, although helpful, a balanced panel is not a requirement for DiD analyses, which can also be deployed with repeated cross-sectional data or unbalanced panels. Indeed, as discussed in Section 3.3 and made explicit in equation 3.6, the 2×2 building block in unconditional DiD analyses involves only averages that are group and time specific, and does not require the same unit to be observed in all periods. As discussed in Callaway and Sant'Anna (2021), the same applies to unconditional staggered adoption setups, and one need not enforce a balanced panel even within each subset of the data used to estimate the $ATT(g,t)$ building blocks. One caveat is that the interpretation of the parameter of interest may change, which we discuss more below.

When covariates are available and play an important role in the plausibility of the identification assumptions, the differences between DiD with a balanced panel and repeated cross-sections (or unbalanced panel) are subtle, can be practically important, and are often not discussed in methodological papers. The gist of the problem relates to potential compositional changes over

time. Most DiD papers that rigorously discuss repeated cross-section setups, including Abadie (2005), Sant'Anna and Zhao (2020), and Callaway and Sant'Anna (2021), rule out compositional changes by assuming that the joint distribution of covariates and treatment groups is invariant over time, a stationarity-type assumption. However, this may not be warranted in empirical applications, and erroneously imposing this additional assumption can lead to biases (Hong, 2013; Sant'Anna and Xu, 2023). On the other hand, when this stationarity assumption is justified and correctly used, the gains in power when conducting inference for DiD parameters can be noticeable (Sant'Anna and Xu, 2023). In what follows, we use the 2×2 setup to explain how compositional changes can complicate the analysis and why ruling it out leads to a gain in precision.

To see how issues related to compositional changes affect the analysis, let us first assume that there are no compositional changes and that the stationarity assumption is valid. In this case, the average treatment effect on the treated in period two (post-treatment) can be written as

$$
\begin{aligned}
ATT(2) &\equiv \mathbb{E}[Y_{i,t=2}(1)|D_i = 1] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1] \\
&= \mathbb{E}[Y_{i,t=2}(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1, T_{i,t=2} = 1] \\
&= \mathbb{E}[Y_i(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_i(0)|D_i = 1, T_{i,t=2} = 1],
\end{aligned}
\tag{A.1}
$$

where $T_{i,t}$ is an indicator if unit i is observed in period t, $Y_i(d) = T_{i,t=2} \, Y_{i,t=2}(d) + T_{i,t=1} \, Y_{i,t=1}(d)$ is the potential outcome for unit i, $D_i = 1\{G_i = 2\}$ is a treatment group dummy that equals one if a unit is first treated in period two and zero if it is untreated in both periods. We also set X_i to be a vector of (pre-treatment) covariates. Note that even here, we already use the stationarity condition that the joint distribution of (D_i, X_i) is invariant to $T_{i,t=2}$ to move from the first to the second line and establish (A.1).

To identify $ATT(2)$ it is often constructive to first establish the identification of its conditional-on-covariates analog; that is, the conditional ATT in period two among units with covariates X_i, $ATT_{X_i}(2)$. This is exactly how we proceeded in Section 4.2. Under the stationarity condition, and similarly to (A.1), we can express this quantity as[37]

$$
\begin{aligned}
ATT_{X_i}(2) &\equiv \mathbb{E}[Y_{i,t=2}(1)|D_i = 1, X_i] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1, X_i] \\
&= \mathbb{E}[Y_i(1)|D_i = 1, X_i, T_{i,t=2} = 1] - \mathbb{E}[Y_i(0)|D_i = 1, X_i, T_{i,t=2} = 1].
\end{aligned}
\tag{A.2}
$$

Next, we have to establish the identification of this quantity. As expected, we will again use conditional parallel trends, no-anticipation, and overlap assumptions. The no-anticipation condition used here is the same as the one in the main text. The conditional parallel trends and overlap assumptions need to be modified, as we now work with multiple partitions of the data depending on treatment status and the period a unit is observed. In this sense, we modify Assumptions CPT and SO to the following related, but different, assumptions. These modifications are warranted regardless of whether compositional changes are present; this step is instead tied to

---

[37]To guarantee that all the conditional expectations in (A.2) are well-defined, we need an overlap condition that guarantees that $P(T_{i,t=2} = 1, D_i = 1|X_i) > 0$. We discuss this below.

data structure.[38]

Assumption CPT-RCS (2×2 Conditional Parallel Trends with repeated cross-sections). We assume that, with probability one,

$$
\mathbb{E}[Y_{i,t=2}(0)|X_i, D_i = 1, T_{i,t=2} = 1] \quad - \quad \mathbb{E}_\omega[Y_{i,t=1}(0)|X_i, D_i = 1, T_{i,t=1} = 1]
$$
$$
= \tag{A.3}
$$
$$
\mathbb{E}[Y_{i,t=2}(0)|X_i, D_i = 0, T_{i,t=2} = 1] \quad - \quad \mathbb{E}[Y_{i,t=1}(0)|X_i, D_i = 0, T_{i,t=1} = 1].
$$

Assumption SO-RCS (Strong overlap with repeated cross-sections). For some $\epsilon > 0$ and every $(d, s) \in \{0, 1\} \times \{0, 1\}$, $\epsilon < P[D_i = d, T_{i,t=2} = s|X_i] < 1 - \epsilon$.

With this modification, we can now show that when Assumptions NA, CPT-RCS, and SO hold, the conditional ATT parameter $ATT_{X_i}(2)$ is identified by[39]

$$
ATT_{X_i}(2) = (\mathbb{E}[Y_i|D_i = 1, T_{i,t=2} = 1, X_i] - \mathbb{E}[Y_i|D_i = 1, T_{i,t=1} = 1, X_i])
$$
$$
- (\mathbb{E}[Y_i|D_i = 0, T_{i,t=2} = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, T_{i,t=1} = 1, X_i]). \tag{A.4}
$$

This step provides a methodological justification to estimate the $ATT_{X_i}(2)$'s using four conditional expectations that use only the available data. In addition, it highlights that, under the stationarity assumption, once we learn the $ATT_{X_i}(2)$'s, we can aggregate them using the covariate distribution of treated units *available from both time periods* to get the $ATT(2)$. More formally, $ATT(2)$ is identified by

$$
ATT(2) = \mathbb{E}[ATT_{X_i}(2)|D_i = 1]
$$
$$
= \mathbb{E}[ATT_{X_i}(2)|D_i = 1, T_{i,t=2} = 1]P(T_{i,t=2} = 1|D_i = 1)
$$
$$
+ \mathbb{E}[ATT_{X_i}(2)|D_i = 1, T_{i,t=1} = 1]P(T_{i,t=1} = 1|D_i = 1),
$$

where $ATT_{X_i}(2)$ is given by (A.4). This is the second place where ruling out compositional changes and imposing the stationarity assumption helps: it allows you to use covariates from treated units in the entire dataset to recover $ATT(2)$. The fact that you can pool data across all periods to learn about $ATT(2)$ translates to gains in power, as formally discussed by Sant'Anna and Zhao (2020) and Sant'Anna and Xu (2023). The third place where the stationarity condition affects the analysis is in the characterization of how "the most precise" (regular and asymptotically linear) estimator for the $ATT(2)$ should look. This point relates to the semi-parametric efficiency bound and the construction of efficient (and doubly robust) estimators. As these points are slightly more technical, we refer the readers to Sant'Anna and Zhao (2020) and Sant'Anna and Xu (2023) for more details.

---

[38]In setups where we rule out compositional changes and impose the stationarity condition that the joint distribution of (D_i, X_i) is invariant to $T_{i,t=2}$, we may not need to modify Assumption CPT. We do it here for transparency purposes.

[39]We also require the assumption that the pooled repeated cross-section data $\{Y_i, D_i, X_i, T_{i,t=2}, T_{i,t=1}\}_{i=1}^n$ is *iid*, though this is fairly standard and uncontroversial; see, for instance, Abadie (2005) and Sant'Anna and Zhao (2020, Assumption 1). We maintain this condition as an assumption throughout this section.

Overall, ruling out compositional changes over time allows for the pooling of information across the entire dataset, which has an impact on the definition of target parameters and leads to more precise inference procedures. But what happens when this condition fails? How does this affect the analysis?

First, this matters for the definition of the treatment effect of interest. In setups where the sampling varies across periods, we do not have a single notion of $ATT(2)$. Instead, we need to accommodate the fact that the $ATT(2)$ may vary across units sampled from different periods. Thus, when we do not rule out compositional changes, we must be explicit about the treated subpopulation that we are interested in. It is common to focus on the average treatment effect in period two among treated units *that are also sampled in period two*, that is,[40]

$$
\begin{aligned}
ATT(2|T_{t=2} = 1) &\equiv \mathbb{E}[Y_{i,t=2}(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_{i,t=2}(0)|D_i = 1, T_{i,t=2} = 1] \\
&= \mathbb{E}[Y_i(1)|D_i = 1, T_{i,t=2} = 1] - \mathbb{E}[Y_i(0)|D_i = 1, T_{i,t=2} = 1].
\end{aligned}
\tag{A.5}
$$

Although (A.5) has the same statistical estimand as (A.1)—that is, the formulas on the right-hand side of the equation coincide—it has a very different interpretation. It is the $ATT(2)$ among units sampled in period 2, and is not an "overall" $ATT(2)$. One may think this difference is merely cosmetic, but as discussed below, this has implications for constructing estimands when covariates are important for identification. Where covariates do not play an important role, it is simply a matter of changing the interpretation of your reported estimates (which also applies to unconditional staggered setups, to be clear).

When covariates do play an important role, however, and when Assumptions CPT-RCS, NA and SO-RCS hold, the conditional ATT parameter among units sampled in period two, $ATT_{X_i}(2|T_{t=2})$ is identified by

$$
\begin{aligned}
ATT_{X_i}(2|T_{t=2}) &= (\mathbb{E}[Y_i|D_i = 1, T_{i,t=2} = 1, X_i] - \mathbb{E}[Y_i|D_i = 1, T_{i,t=1} = 1, X_i]) \\
&\quad - (\mathbb{E}[Y_i|D_i = 0, T_{i,t=2} = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, T_{i,t=1} = 1, X_i]),
\end{aligned}
\tag{A.6}
$$

which, in turn, implies that $ATT(2|T_{t=2} = 1)$ is identified by

$$
ATT(2|T_{t=2} = 1) = \mathbb{E}[ATT_{X_i}(2|T_{t=2})|D_i = 1, T_{i,t=2} = 1].
\tag{A.7}
$$

Several remarks are worth making. First, the statistical estimand in (A.6) is the same as when one rules out compositional changes as in (A.4), suggesting that, once again, what changes in this step is the interpretation. However, these interpretative issues have direct consequences on the appropriate method for aggregating across covariate values. As clearly stated in (A.7), in the presence of potential compositional changes, one is not allowed to pool information across periods to identify (and also estimate and make inference) about $ATT(2|T_{t=2} = 1)$. As discussed

---

[40]One may also be interested in the ATT in period two among treated units that are sampled in period one. The arguments required to establish (point) identification of this parameter differ from those we use here. A main challenge is that we do not observe $\mathbb{E}[Y_{i,t=2}(1)|D_i = 1, T_{i,t=1} = 1]$, and the parallel trends assumption we leverage does not involve treated potential outcomes.

in Sant'Anna and Xu (2023), ignoring these issues and pooling data from all periods in the presence of compositional changes leads to a bias that is important to be aware of. We refer the reader to Sant'Anna and Xu (2023) for a discussion related to unbalanced panels and also on a discussion about doubly robust and semiparametric efficient DiD estimators under compositional changes. We are not aware of any papers that formally extend the discussion in Sant'Anna and Xu (2023) to staggered DiD designs. Still, this extension is surely possible by following our forward-engineering approach to DiD.

We close this section by highlighting that, in practice, it is possible to test for compositional changes by comparing the estimates from estimators that impose it and those that do not. Sant'Anna and Xu (2023) discuss Hausman-type tests in the two-period setting, though one can extend those to more general setups. We also highlight that when it comes to DiD setups with staggered adoption, some equivalence results discussed in Section 5.2 no longer hold with repeated cross-sections or unbalanced panel data. For instance, the Sun and Abraham (2021) regression-based strategy to estimate $ATT(g,t)$'s using (5.11) no longer coincides with Callaway and Sant'Anna's (2021) estimators using the analog of (5.10):

$$\widehat{ATT}_{\text{never}}(g,t) = (\overline{Y}_{G=g,t} - \overline{Y}_{G=g,t=g-1}) - (\overline{Y}_{G=\infty,t} - \overline{Y}_{G=\infty,t=g-1}),$$

where $\overline{Y}_{G=a,t=s}$ is the sample mean of Y among units that belong in group $G = a$ and are observed in period $t = s$. In fact, it is unclear exactly what estimand is being recovered when one uses (5.11) with an unbalanced panel. If one replaces unit fixed effects with treatment group dummies in (5.11), such equivalence is restored, though we suspect that many practitioners do not use this alternative specification. In general, we caution against extrapolating from a well-motivated regression specification that was studied under one specific setup to another related but inherited different framework. This practice has led to many issues in DiD, which can be fully avoided by adopting the forward-engineering approach discussed in this paper.

References

Abadie, Alberto, "Semiparametric Difference-in-Difference Estimators," *The Review of Economic Studies*, 2005, *72*, 1–19.

_ , **Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge**, "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, January 2020, *88* (0), 265–296.

_ , _ , _ , **and Jeffrey Wooldridge**, "When Should You Adjust Standard Errors for Clustering?," *The Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Abbring, Jaap H. and Gerard J. van den Berg, "The nonparametric identification of treatment effects in duration models," *Econometrica*, 2003, *71* (5), 1491–1517.

Angrist, Joshua D., "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 1998, *66* (2), 249–288.

_ **and Guido W. Imbens**, "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 1995, *90* (430), 431–442.

Arkhangelsky, Dmitry, Guido W. Imbens, Lihua Lei, and Xiaoman Luo, "Design-Robust Two-Way-Fixed-Effects Regression For Panel Data," *Quantitative Economics*, 2024, *15* (4).

_ , **Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager**, "Synthetic Difference-in-Differences," *American Economic Review*, 2021, *111* (12), 4088–4118.

Aronow, P. M. and Cyrus Samii, "Does regression produce representative estimates of causal effects?," *American Journal of Political Science*, 2015, *60* (1), 250–267.

Ashenfelter, Orley C. and David Card, "Using the longitudinal structure of earnings to estimate the effect of training programs," *The Review of Economics and Statistics*, 1985, *67* (4), 648–660.

Athey, Susan and Guido Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79.

_ **and Guido W. Imbens**, "Identification and Inference in Nonlinear Difference in Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

Austin, Peter C, "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples," *Statistics in Medicine*, 2009, *28* (25), 3083–3107.

Baker, Andrew C, David F Larcker, and Charles CY Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, February 2004, *119* (1), 249–275.

Bilinski, Alyssa and Laura A Hatfield, "Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions," 2018. Working Paper.

Black, Bernard, Alex Hollingsworth, Leticia Nunes, and Kosali Simon, "Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion," *Journal of Public Economics*, 2022, *213*, 104713.

Bonhomme, Stéphane and Ulrich Sauder, "Recovering Distributions in Difference-in-Differences Models: a Comparison of Selective and Comprehensive Schooling," *Review of Economics and Statistics*, 2011, *93* (May), 479–494.

Borgschulte, Mark and Jacob Vogler, "Did the ACA Medicaid expansion save lives?," *Journal of Health Economics*, 2020, *72*, 102333.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Caetano, Carolina and Brantly Callaway, "Difference-in-Differences when Parallel Trends Holds Conditional on Covariates," *arXiv:2406.15288*, 2024.

_ , _ , **Stroud Payne, and Hugo Sant'Anna Rodrigues**, "Difference in differences with time-varying covariates," 2022. Working Paper.

Callaway, Brantly, "Bounds on distributional treatment effect parameters using panel data with an application on job displacement," *Journal of Econometrics*, 2021, *222* (2), 861–881.

__ , "Difference-in-Differences for Policy Evaluation," in Klaus F. Zimmermann, ed., *Handbook of Labor, Human Resources and Population Economics*, Cham: Springer International Publishing, 2023, pp. 1–61.

__ **and Pedro HC Sant'Anna**, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

__ **and Sonia Karami**, "Treatment effects in interactive fixed effects models with a small number of time periods," *Journal of Econometrics*, 2023, *233* (1), 184–208.

__ **and Tong Li**, "Quantile Treatment Effects in Difference in Differences Models with Panel Data," *Quantitative Economics*, 2019, *10* (4), 1579–1618.

__ , **Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna**, "Difference-in-Differences with a Continuous Treatment," *arXiv:2107.02637 [econ]*, 2021.

__ , __ , **and** __ , "Event Studies with a Continuous Treatment," *AEA Papers and Proceedings*, May 2024, *114*, 601–605.

__ , **Tong Li, and Tatsushi Oka**, "Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with Only Two Time Periods," *Journal of Econometrics*, 2018, *206* (2), 395–413.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, August 2008, *90* (3), 414–427.

Caron, Laura, "Triple Difference Designs with Heterogeneous Treatment Effects," *arXiv:2502.19620*, 2025.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1405–1454.

Centers for Disease Control and Prevention, "Vital Statistics Data," `https://www.cdc.gov/nchs/nvss/index.htm` 2024. Accessed: 2024-09-17.

Chabé-Ferret, Sylvain, "Analysis of the bias of Matching and Difference-in-Difference under alternative earnings and selection processes," *Journal of Econometrics*, 2015, *185* (1), 110–123.

Chang, Neng-Chieh, "Double/debiased machine learning for difference-in-differences," *Econometrics Journal*, 2020, *23*, 177–191.

Chen, Xiaohong, Pedro H. C. Sant'Anna, and Haitian Xie, "Efficient Difference-in-Differences and Event Study Estimators," *Working Paper*, 2024.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo, "The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages," *Econometrica*, 2018, *86* (6), 1911–1938.

__ , **Mert Demirer, Esther Duflo, and Iván Fernández-Val**, "Fisher-Schultz Lecture: Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India," *arXiv: 1712.04802*, 2023, pp. 1–81.

Conley, Timothy and Christopher Taber, "Inference with "Difference in Differences" with a Small Number of Policy Changes," *Review of Economics and Statistics*, February 2011, *93* (1), 113–125.

Crump, Richard K, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik, "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 2009, *96* (1), 187–199.

Currie, Janet and Hannes Schwandt, "Mortality Inequality: The Good News from a County-Level Approach," *Journal of Economic Perspectives*, May 2016, *30* (2), 2952.

_ , **Henrik Kleven, and Esmée Zwiers**, "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, 2020, *110*, 42–48.

de Chaisemartin, Clément and Xavier D'Haultfoeuille, "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, 2018, *85* (2), 999–1028.

_ **and** _ , "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

_ **and** _ , "Difference-in-Differences Estimators of Intertemporal Treatment Effects," 2023. Working Paper.

_ **and** _ , "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey," *Econometrics Journal*, 2023, *Forthcoming.*

_ , _ , **Félix Pasquier, and Gonzalo Vazquez-Bare**, "Difference-in-Differences for Continuous Treatments and Instruments with Stayers," *arXiv:2201.06898*, 2024.

_ , **Xavier D'Haultfuille, and Gonzalo Vazquez-Bare**, "Difference-in-Difference Estimators with Continuous Treatments and No Stayers," *AEA Papers and Proceedings*, May 2024, *114*, 610–613.

Deshpande, Manasi and Yue Li, "Who Is Screened Out? Application Costs and the Targeting of Disability Programs," *American Economic Journal: Economic Policy*, November 2019, *11* (4), 213–48.

Dette, Holger and Martin Schumann, "Testing for Equivalence of Pre-Trends in Difference-in-Differences Estimation," *Journal of Business & Economic Statistics*, 2024, *42* (4).

DiNardo, John and David S. Lee, "Chapter 5 - Program Evaluation and Research Designs," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 4, Elsevier, 2011, pp. 463–536.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and other Panel Data," *Review of Economics and Statistics*, 2007, *89* (2), 221–233.

Dube, Arindrajit, Daniele Girardi, Òscar Jordà, and Alan M Taylor, "A Local Projections Approach to Difference-in-Differences," Working Paper 31184, National Bureau of Economic Research 2024.

Fadlon, Itzik and Torben Heien Nielsen, "Family Labor Supply Responses to Severe Health Shocks: Evidence from Danish Administrative Records," *American Economic Journal: Applied Economics*, July 2021, *13* (3), 1–30.

Fan, Yanqin and Zhentao Yu, "Partial Identification of Distributional and Quantile Treatment Effects in Difference-in-Differences Models," *Economics Letters*, 2012, *115* (3), 511–515.

Fernández-Val, Iván, Jonas Meier, Aico van Vuuren, and Francis Vella, "Distribution Regression Difference-in-Differences," *arXiv preprint arXiv:2409.02311*, 2024.

_ , _ , _ , **and** _ , "Distribution Regression Difference-In-Differences," *arXiv:2409.00123*, 2024.

Finkelstein, Amy and Robin McKnight, "What Did Medicare Do? The Initial Impact of Medicare on Mortality and Out-of-Pocket Medical Spending," *Journal of Public Economics*, 2008, *92* (7), 1644–1668.

Firpo, Sergio and Cristine Pinto, "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures," *Journal of Applied Econometrics*, 2016, *31* (3), 457–486.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M Shapiro, "Pre-Event Trends in the Panel Event-Study Design," *American Economic Review*, 2019, *109* (9), 3307–3338.

__ , __ , **Jorge Pérez-Pérez, and Jesse M. Shapiro**, "Visualization, identification, and estimation in the linear panel event-study design," in "Advances in Economics and Econometrics: Twelfth World Congress" 2024. Forthcoming.

Gardner, John, "Two-stage differences in differences," *Working Paper*, 2021.

Ghanem, Dalia, Désiré Kédagni, and Ismael Mourifié, "Evaluating the Impact of Regulatory Policies on Social Welfare in Difference-in-Difference Settings," *arXiv:2306.04494*, 2023.

__ , **Pedro H. C. Sant'Anna, and Kaspar Wüthrich**, "Selection and parallel trends," *arXiv:2203.09001*, 2022.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, "Contamination Bias in Linear Regressions," *American Economic Review*, 2024, *Forthcoming.*

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

Graham, Bryan, Cristine Pinto, and Daniel Egel, "Inverse Probability Tilting for Moment Condition Models with Missing Data," *The Review of Economic Studies*, 2012, *79* (3), 1053–1079.

Gruber, Jonathan, "The Incidence of Mandated Maternity Benefits," *American Economic Review*, June 1994, *84* (3), 622–641.

Harmon, Nikolaj A., "Difference-in-Differences and Efficient Estimation of Treatment Effects," *Working Paper*, 2024.

Heckman, James and Richard Robb, "Alternative methods for evaluating the impact of interventions," in James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge: Cambridge University Press, 1985, pp. 156–246.

Heckman, James J., "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *The Quarterly Journal of Economics*, 2000, *115* (1), 45–97.

__ , **Hidehiko Ichimura, and Petra Todd**, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 1997, *64* (4), 605–654.

__ , **Jeffrey Smith, and Nancy Clements**, "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 1997, *64* (4), 487–535.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart, "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political Analysis*, 2007, *15* (3), 199–236.

Hong, Seung-Hyun, "Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes," *Journal of Applied Econometrics*, 2013, *28* (2), 297–324.

Imai, Kosuke and In Song Kim, "On the use of two-way fixed effects regression models for causal inference with panel data," *Political Analysis*, 2021, *29* (3), 405–415.

__ , __ , **and Erik H. Wang**, "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data," *American Journal of Political Science*, 2023, *67* (3), 587–605.

Imbens, Guido, Nathan Kallus, and Xiaojie Mao, "Controlling for Unmeasured Confounding in Panel Data Using Minimal Bridge Functions: From Two-Way Fixed Effects to Factor Models," *arXiv:2108.03849*, 2021.

Imbens, Guido W and Donald B Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge university press, 2015.

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business and Economic Statistics*, 2020, *38* (3), 613–620.

Kennedy, Edward H, Zongming Ma, Matthew D McHugh, and Dylan S Small, "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2017, *79* (4), 1229–1245.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

Lechner, Michael, "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 2011, *4*, 165–224.

Lee, Soo Jeong and Jeffrey M. Wooldridge, "A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data," *Working Paper*, 2023. Available at SSRN: `http://dx.doi.org/10.2139/ssrn.4516518`.

Liu, Licheng, Ye Wang, and Yiqing Xu, "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data," *American Journal of Political Science*, 2024, *68* (1), 160–176.

Ma, Yukun, Pedro H. C. Sant'Anna, Yuya Sasaki, and Takuya Ura, "Doubly Robust Estimators with Weak Overlap," *arXiv:2304.08974*, 2023.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, 2015, *124*, 1–17.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

Marcus, Michelle and Pedro H. C. Sant'Anna, "The role of parallel trends in event study settings: An application to environmental economics," *Journal of the Association of Environmental and Resource Economists*, 2021, *8* (2), 235–275.

Marx, Philip, Elie Tamer, and Xun Tang, "Parallel Trends and Dynamic Choices," *Journal of Political Economy Microeconomics*, 2024, *2* (1), 129–171.

Meyer, Bruce, W. Kip Viscusi, and David Durbin, "Workers Compensation and Injury Duration: Evidence from a Natural Experiment," *The American Economic Review*, 1995, *85* (3), 322–340.

Miller, Sarah, Norman Johnson, and Laura R Wherry, "Medicaid and mortality: new evidence from linked survey and administrative data," *The Quarterly Journal of Economics*, 2021, *136* (3), 1783–1829.

Miyaji, Sho, "Instrumented Difference-in-Differences with Heterogeneous Treatment Effects," *arXiv:2405.12083*, 2024.

Mogstad, Magne and Alexander Torgovitsky, "Chapter 1 - Instrumental variables with unobserved heterogeneity in treatment effects," in Christian Dustmann and Thomas Lemieux, eds., *Handbook of Labor Economics*, Vol. 5, Elsevier, 2024, pp. 1–114.

Mora, Ricardo and Iliana Reggio, "Alternative diff-in-diffs estimators with several pretreatment periods," *Econometric Reviews*, 2019, *38* (5), 465–486.

Olden, Andreas and Jarle Møen, "The triple difference estimator," *The Econometrics Journal*, 2022, *25* (3), 531–553.

Olea, José Luis Montiel and Mikkel Plagborg-Moller, "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 2018, *33* (7), 943–964.

Ortiz-Villavicencio, Marcelo and Pedro H. C. Sant'Anna, "Better Understanding Triple Differences Estimators," 2025. Working paper.

Poirier, Alexandre and Tymon Sloczynski, "Quantifying the Internal Validity of Weighted Estimands," *arXiv:2404.14603*, 2024.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 2023, *90* (5), 2555–2591.

_ **and** _ , "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602*, 2024.

Robins, James, "A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 1986, *7*, 1393–1512.

Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, April 1983, *70* (1), 41–55.

Roth, Jonathan, "Pretest with caution: Event-study estimates after testing for parallel trends," *American Economic Review: Insights*, 2022, *4* (3), 305–322.

_ **and Pedro H. C. Sant'Anna**, "Efficient Estimation for Staggered Rollout Designs," *Journal of Political Economy Microeconomics*, 2023, *1* (4), 669–709.

__ **and** __ , "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2023, *91* (2), 737–747.

__ , __ , **Alyssa Bilinski, and John Poe**, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *Journal of Econometrics*, 2023, *235* (2), 2218–2244.

Rubin, Donald, "Estimating Causal Effects of Treatments in Randominzed and Nonrandomized Studies," *Journal of Educational Psychology*, 1974, *66* (5), 688–701.

Sant'Anna, Pedro H. C. and Jun B. Zhao, "Doubly Robust Difference-in-Differences Estimators," November 29 2018. Unpublished Manuscript.

__ **and Jun Zhao**, "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 2020, *Forthcoming.*

__ **and Qi Xu**, "Difference-in-Differences with Compositional Changes," *arXiv:2304.14256*, 2023.

Sasaki, Yuya and Takuya Ura, "Estimation and inference for moments of ratios with robustness against large trimming bias," *Econometric Theory*, 2022, *38* (1), 66–112.

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

Semmelweis, Ignaz, *Etiology, Concept and Prophylaxis of Childbed Fever*, The University of Wisconsin Press, 1983.

Sloczynski, Tymon, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *The Review of Economics and Statistics*, 2022, *104* (3), 501–509.

Smucler, Ezequiel, Andrea Rotnitzky, and James M. Robins, "A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts," *arXiv:1904.03737*, 2019.

Snow, John, *On the Mode of Communication of Cholera*, London: John Churchill, 1855.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge, "What Are We Weighting For?," *The Journal of Human Resources*, 2015, *50* (2), 301–316.

Sommers, Benjamin Daniel and Arnold M Epstein, "US governors and the Medicaid expansionno quick resolution in sight," *New England Journal of Medicine*, 2013.

Strezhnev, Anton, "Semiparametric weighting estimators for multi-period differencein-differences designs," 2018. Working Paper.

__ , "Decomposing Triple-Differences Regression under Staggered Adoption," *arXiv:2307.02735*, 2023.

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

Tchetgen, Eric J. Tchetgen, Chana Park, and David B. Richardson, "Universal Difference-in-Differences for Causal Inference in Epidemiology," *Epidemiology*, 2024, *35* (1), 16–22.

Wooldridge, Jeffrey M., "Cluster-Sample Methods in Applied Econometrics," *American Economic Review P&P*, 2003, *93* (2), 133–138.

__ , "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators," 2021. Working Paper.

__ , "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data," *Econometrics Journal*, 2023, *Forthcoming.*

Wyse, Angela and Bruce Meyer, "Saved by Medicaid: New Evidence on Health Insurance and Mortality from the Universe of Low-Income Adults," *Working Paper*, 2024.

Yanagi, Takahide, "An Effective Treatment Approach to Difference-in-Differences with General Treatment Patterns," *arXiv:2212.13226*, 2023.