NBER WORKING PAPER SERIES

DESIGNING DIFFERENCE IN DIFFERENCE STUDIES WITH STAGGERED TREATMENT ADOPTION: KEY CONCEPTS AND PRACTICAL GUIDELINES

Seth M. Freedman
Alex Hollingsworth
Kosali I. Simon
Coady Wing
Madeline Yozwiak

Working Paper 31842
http://www.nber.org/papers/w31842

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023

This paper was written for an upcoming volume of the Annual Review of Public Health in April 2024. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w31842

Designing Difference in Difference Studies With Staggered Treatment Adoption: Key Concepts and Practical Guidelines

Seth M. Freedman, Alex Hollingsworth, Kosali I. Simon, Coady Wing, and Madeline Yozwiak

NBER Working Paper No. 31842
November 2023
JEL No. I0,I1

ABSTRACT

Difference-in-Difference (DID) estimators are a valuable method for identifying causal effects in the public health researcher's toolkit. A growing methods literature points out potential problems with DID estimators when treatment is staggered in adoption and varies with time. Despite this, no practical guide exists for addressing these new critiques in public health research. We illustrate these new DID concepts with step-by-step examples, code, and a checklist. We draw insights by comparing the simple 2 × 2 DID design (single treatment group, single control group, two time periods) with more complex cases: additional treated groups, additional time periods of treatment, and with treatment effects possibly varying over time. We outline newly uncovered threats to causal interpretation of DID estimates and the solutions the literature has proposed, relying on a decomposition that shows how the more complex DID are an average of simpler 2X2 DID sub-experiments.

Seth M. Freedman
School of Public and Environmental Affairs
1315 E. 10th St.
Bloomington, IN 47405
freedmas@indiana.edu

Coady Wing
Indiana University
1315 E 10th St
Bloomington, IN 47405
cwing@indiana.edu

Alex Hollingsworth
Department of Agricultural, Environmental, and Development Economics, Department of Economics, and the
John Glenn College of Public Affairs
The Ohio State University
2120 Fyffe Road
Columbus, OH 43210
and NBER
hollingsworth.126@osu.edu

Madeline Yozwiak
Energy Justice Lab
Indiana University
myozwiak@iu.edu

Kosali I. Simon
O'Neill School of Public and Environmental Affairs
Indiana University
1315 East Tenth Street
Bloomington, IN 47405-1701
and NBER
simonkos@indiana.edu

1 Introduction

Many studies in the social and health sciences rely on comparisons where treatment exposure changes over time for some units. In the simplest case, the treatment effect is estimated via Difference-in-Differences (DID), which compares two groups (treated and untreated) across two time periods (pre-treatment and post-treatment). However, many applications involve more complex settings with multiple groups and periods and staggered treatment adoption. These designs are often analyzed using a two-way fixed effects (TWFE) regression that includes group and time-period fixed effects.[1] A recent methodological literature examines the staggered adoption case in detail, highlighting previously unrecognized sources of confounding and proposing new estimation strategies (5, 11, 16, 27).

In this review, we provide a guide to the key ideas and conclusions from this new literature and describe novel threats to internal and external validity of the results. We focus on two major themes. First, the combination of staggered adoption and time varying treatment effects can introduce confounded comparisons into the TWFE regression estimator. Second, researchers may find it helpful to conceive of a staggered adoption design as a collection of simpler difference-in-difference comparisons and to take control over the sub-experiments that contribute to their analysis. Understanding these two ideas can help researchers analyze data from staggered adoption designs more effectively. To make the ideas concrete, we use a running example of a health intervention to illustrate why previous methods are biased and how this can be accounted for through careful research design.[2] Our review focuses on methods appropriate for binary treatments and that do not involve time-varying covariates, which are widely applicable to health policy and public health research and are the most well developed in the methodological literature.

2 A Stylized Example

To describe the challenges created by staggered treatment adoption DID designs and to explain proposed solutions, we created a simulated example inspired by studies showing that the introduction of sulfa drugs reduced mortality (20, 29). Sulfa drugs were introduced in all

[1] Modern applications in public health research and health services research are quite common. See for example, Gupta et al. (18), Mullachery et al. (22), Yan et al. (34), and Kim et al. (21). More formally, we obtained all articles published by the *American Journal of Public Health* (AJPH) between 2018-2022. Using a natural language search of each article's full text, we found AJPH published an average of ten articles per year that used a DID.

[2] There are a number of existing excellent review articles of the modern econometric literature focusing on DID, including: de Chaisemartin and D'Haultfoeuille (10), Roth et al. (26), and Baker et al. (2). Our goal is to provide an accessible foundation to applied researchers, with examples and lessons relevant health and social scientists.

U.S. states in 1937. Our simulated example imagines that sulfa drugs were instead introduced in states at different times. Figure 1 shows these simulated data: mortality rates for states that "gained access" to sulfa in 1930, 1940, and 1945, along with a group of states that never gained access during the hypothetical study time period.

In this example, the treatment effect varies over time in that treatment gradually reduces mortality in subsequent years. In addition, the effect varies by the year in which treatment was introduced (i.e., timing-group) in that treatment that began in earlier periods is more effective than treatment that began in later periods.[3] We also differ the base mortality rate by timing-group, where the earlier timing groups have higher mortality rates, introducing a source of geographic heterogeneity.[4] We create the simulated data so that the introduction of sulfa drugs gradually reduces mortality. For the 1930 group, mortality rates begin to fall at a rate of 7.5 per year. Likewise, they fall by 5.0 and 2.5 in the 1940 and 1945 groups. Because treatment changes the time trend of mortality, this is an example of a time varying treatment effect, a common situation in many public health interventions. In general, there can be many sources of treatment effect heterogeneity. The treatment effect could vary with time since treatment, geography, calendar time of treatment, etc. This example is not meant to explore all possible sources of heterogeneity. Code and data for the example are available at `https://github.com/hollina/arph-did-example`.

3 Simple and Staggered DID

The staggered adoption design maps to real world settings where treatment "rolls out" differently across across geographic areas, such as states or counties. Staggered adoption obscures distinctions like treatment vs. control and pre vs. post. It helps to view the staggered adoption design as a collection of simpler 2×2 DID designs, which we refer to as "sub-experiments." This perspective suggests researchers should develop principles for actively deciding which 2×2 DIDs will contribute to the analysis and be on guard against DID comparisons that may be confounded.

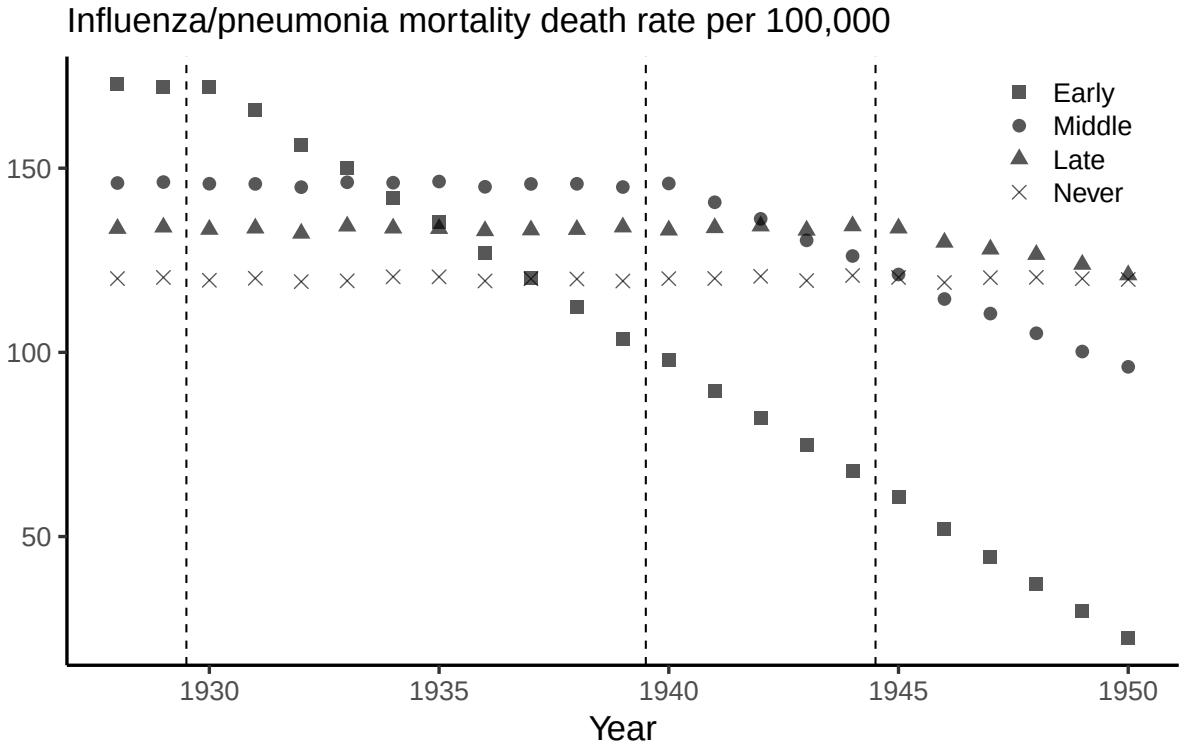
In this section, we set out notation that clarifies the "ensemble of sub-experiments" view.[5] We review core DID assumptions and show how they identify causal relationships in the 2×2 setting. Then we show how these ideas generalize using the concept of a $group \times time$ average

[3]Time here refers to event-time (i.e., time since treatment). In some literatures, this variation in treatment effect based on when treatment adoption occurred may be referred to as a "period-varying" treatment effect.

[4]This is consistent with evidence from Hollingsworth et al. (19) who show that areas with modern medical facilities benefited from sulfa drugs more than areas without such facilities.

[5]Our notation is is similar to recent work by Goodman-Bacon (16) and Callaway and Sant'Anna (5), both of which view *timing group* $\times$ *time* treatment effects as a key building block for understanding the staggered adoption design.

Figure 1: Stylized Example: Pneumonia/Influenza Mortality and Sulfa Drugs



Notes: This plot presents a stylized example inspired by Jayachandran et al. (20). Each point is annual death rate per 100,000 population from influenza and pneumonia for a given group. Data are constructed so that there are four groups: the early treated group sees an annual decline beginning in 1930, the middle treated group sees the decline beginning in 1940, the late treated group sees the decline in 1945, and the never treated group has a constant death rate. There is also treatment effect heterogeneity. The annual decline is largest for the early treated group (an additional 7.5 decline in the death rate in each year following treatment) and smallest (additional decline of 2.5 per year) for the latest treated group. The middle treated group sees a reduction of an additional 5 deaths per 100,000 in year each following treatment.

treatment effect as an organizing concept.

3.1 Design and Assumptions

We use $i = 1...N$ to index individual observations, $s = 1...S$ to index the collection of groups, and $t = T_1...T_T$ to index calendar time periods. We focus on situations where treatment exposures occur at the *group* $\times$ *time* level, and where treatments remain in place until the end of the study period. Let A_s represent the calendar period when units in group s are first exposed to treatment, and set $A_s = \infty$ for groups that never adopt treatment during the study period. In our simulated Sulfa drugs data, A_s defines four groups that adopt in 1930, 1940, 1950, and ∞ (never exposed). Accordingly $D_{st} = 1(t \geq A_s)$ is a binary treatment variable indicating whether treatment is active in group s in period t.

We represent causal relationships using potential outcomes. $Y_{ist}(0)$ represents the outcome person i from group s would experience in calendar period t under a hypothetical scenario in which the person's group is never exposed to treatment. $Y_{ist}(a)$ represents the outcome that the same person would experience at t if she were first exposed to treatment in calendar period a. The causal effect of adopting treatment in period a compared to never adopting treatment is $\beta_{ist}(a) = Y_{ist}(a) - Y_{ist}(0)$. Notice the subscripts on $\beta_{ist}(a)$, which emphasize that the treatment effect can differ across units, groups, and time periods. Often researchers will be interested in averages of this effect. In the DID setting, the likely average of interest is the average treatment effect on the treated (ATT) evaluated at a particular calendar date. In our notation, this is $ATT(a, t) = E[\beta_{ist}(a)|A_s = a]$. Finally, the realized outcome depends on the adoption date so that $Y_{ist} = Y_{ist}(0) + \sum_{a=T_1}^{T_T} \beta_{ist}(a) \times 1(A_s = a)$. This is the untreated outcome plus the treatment effect if the treatment is actually in place.

The DID design provides a way to recover causal relationships under two main assumptions: the non-anticipation assumption, and the common trends assumption.

Assumption 1. *No Anticipation: The average causal effect of adopting treatment in period a is equal to zero for all calendar periods prior to period a. For periods $t < a$*

$$E\left[Y_{ist}(a) - Y_{ist}(0)|A_s = a\right] = 0$$

Assumption 2. *Common Trends: In the absence of treatment exposure, the average change across post-treatment time periods would be the same in the treatment group ($A_s = a$) and the comparison group ($A_s > a$). For periods $t > a$*

$$E\left[Y_{ist}(0) - Y_{ist-1}(0)|A_s = a\right] = E\left[Y_{ist}(0) - Y_{ist-1}(0)|A_s > a\right]$$

Assumption 1 is a version of the *strict exogeneity* assumption, familiar from panel data models. In the DID literature, people sometimes invoke Assumption 1 by saying there are no pre-trends. The assumption could fail – for example – if treatment exposure occurs in response to volatility in the outcome variable, or if behavior changes due to expectations of future treatment. In our simulated sulfa drugs example, the assumption holds by construction. But in the real world, the no anticipation assumption would fail if states with unusually high mortality in one year were more likely to gain access to sulfa drugs in subsequent years.

We state Assumption 2 (common trends) in terms of a comparison of a treatment group made up of all units that adopt in period a, and a control group made up of all units from groups that have not yet adopted treatment, including both never treated groups and groups that adopt at a later time. In practice, researchers may choose to work with a specialized common trends assumption that only relies on a never treated comparison group. In Assumption 2, $E\left[Y_{ist}(0) - Y_{ist-1}(0)|A_s = a\right]$ represents the time trend that the treated group would have experienced in the absence of treatment exposure. This is a counterfactual that we cannot observe directly. The common trend assumption implies that the counterfactual trend is equal to the observed trend in the control group. The word "trend" does not imply a linear trend over multiple periods: it is simply a change between two periods. In the Sulfa drugs example, common trends implies that the the 1930 adoption group would have continued on the same time trend as the never treated states if not for the new drugs.

Neither of these two assumptions is fully testable because both depend on counterfactual quantities. However, good applied studies present ancillary analysis that partially test or probe the credibility of these assumptions using event studies and related methods (Wing et al. (31)).

3.2 2×2 DID

The 2×2 DID design has two periods ($t = [1, 2]$) and two groups ($s = [1, 2]$). The first group is never treated so $A_1 = \infty$. The second group has $A_2 = 2$, meaning it is first exposed in period 2. Periods 1 and 2 are the "pre-" and "post-" periods, respectively. The "difference-in-differences" estimator is the difference between the expected pre-post change in realized outcomes in the treatment group and control group. Combining the estimator with Assumptions 1 and 2 gives:

$$
\begin{aligned}
\Delta_{DID} &= E[Y_{i22} - Y_{i21}|A_s = 2] - E[Y_{i12} - Y_{i11}|A_s = \infty] \\
&= E[Y_{i22}(2) - Y_{i21}(2)|A_s = 2] - E[Y_{i12}(0) - Y_{i11}(0)|A_s = \infty] \\
&= E[Y_{i22}(2) - Y_{i21}(0)|A_s = 2] - E[Y_{i12}(0) - Y_{i11}(0)|A_s = \infty] \\
&= E[\beta_{i22}(2)|A_s = 2] - \{E[Y_{i22}(0) - Y_{i21}(0)|A_s = 2] - E[Y_{i12}(0) - Y_{i11}(0)|A_s = \infty]\} \\
&= E[\beta_{i22}(2)|A_s = 2] \\
&= ATT(2,2)
\end{aligned}
$$

The second equality substitutes potential outcomes, and the third imposes the no-anticipation assumption. In the fourth line we re-express $Y_{i22}(2)$ in terms of the untreated outcome and treatment effect. The fifth line imposes common trends, and shows that the DID estimator identifies the average treatment effect for the treated group in the post-period, which is equivalent to $ATT(2,2)$ in the final line.

Adding structure makes the DID more intuitive. Write the untreated outcome as $Y_{ist}(0) = c_s + b_t + e_{ist}$. Then the treated outcome is $Y_{ist}(a) = Y_{ist}(0) + \beta_{ist}(a)$. Clearly, the time trend is $E[Y_{is2}(0) - Y_{is1}(0)] = b_2 - b_1$ in both groups: that's what common trend looks like in this case. The group difference is $E[Y_{i2t}(0) - Y_{i1t}(0)] = c_2 - c_1$ in both periods, showing that the DID assumptions do not require that the groups are "comparable". What matters is that disparities do not change over time.

The 2×2 DID can also be formed using regressions. Define $Treat_s = 1(A_s = 2)$ and $Post_t = 1(t = 2)$ and then estimate an OLS regression $Y_{ist} = \beta_0 + \beta_1 Treat_s + \beta_2 Post_t + \beta_3(Treat_s \times Post_t) + e_{ist}$. In the 2×2 case $\beta_3 = \Delta_{DID}$. Alternatively, we could estimate a linear TWFE regression model $Y_{ist} = \beta_{FE} D_{st} + c_s + b_t + e_{ist}$, where c_s and b_t represent unobserved fixed effects and $\beta_{FE} = \Delta_{DID}$.

3.3 Two Group Event Studies

Most empirical applications include data from more than just a single pre-and post time period. A basic event study has two groups ($s = [1,2]$) but multiple time periods $t = T_1 ... T_T$. In our 2×2 example, group 1 (never treated) has $A_1 = \infty$. Group 2 is treated at $A_2 = a$, with $a > T_1$. The pre-period runs from T_1 to $a - 1$ and the post period runs from $t = a$ to T_T. By adding periods to the 2×2 design, the event study makes it possible to learn more about time varying treatment effects, and also enables some partial tests of the core DID assumptions.

We use $ATT(a, t*) = E[Y_{ist*}(a) - Y_{ist*}(0)|A_s = a]$ to represent the average effect of adopting

in period a on outcomes experienced in calendar period $t*$ among units in timing group $A_s = a$. With this notation, $ATT(a, a)$ represents the immediate effect, and $ATT(a, a+k)$ represents the effect k periods after initial adoption.

Under Assumptions 1 (no anticipation) and 2 (common trends), $ATT(a, a+k)$ is identified for each $k = 0...T_T - a$. The DID estimator of $ATT(a, a+k)$ is:

$$\begin{aligned}
\Delta_{ES}^{a+k} &= E[Y_{i2,a+k} - Y_{i2,a-1}|A_s = a] - E[Y_{i1,a+k} - Y_{i1,a-1}|A_s = \infty] \\
&= E[\beta_{i2,a+k}(a)|A_s = 2] - \{E[Y_{i2,a+k}(0) - Y_{i2,a-1}(0)|A_s = 2] \\
&\quad -E[Y_{i1,a+k}(0) - Y_{i1,a-1}(0)|A_s = \infty]\} \\
&= E[\beta_{i2,a+k}(a)|A_s = a]
\end{aligned}$$

The logic is the same as the 2×2 case: realized outcomes are replaced with potential outcomes, and the no anticipation and common trend assumption are imposed. By definition $E[\beta_{i2,a+k}|A_s = a] = ATT(a, a+k)$, which is the average causal effect k periods after treatment adoption. Applying the estimator repeatedly for different choices of k traces out the treatment effect in event time. Interestingly, this sequence of DID estimators works by comparing a focal post-period $(a + k)$ with a fixed pre-period $(a - 1)$, which is simply the last period before the treated group adopts treatment.

In addition to tracing out time varying ATTs in the post-period, the event study also provides a way to partially test the identifying assumptions. Specifically, if we slightly strengthen the common trends assumption to hold for all periods rather than only the post-treatment periods, then under the combination of no-anticipation and all period common trends, we expect that $ATT(a, a-h) = 0$ for $h = 1...T_1 + a - 1$. These pre-period ATTs can be estimated using $\Delta_{ES}^{a-h} = E[Y_{i2,a-h} - Y_{i2,a-1}|A_s = a] - E[Y_{i1,a-h} - Y_{i1,a-1}|A_s = \infty]$. Rejecting the null that these pre-period DIDs are equal to zero implies that the common trend + no anticipation assumptions is not met.

In practice, it is convenient to estimate these event study DIDs using a single linear regression:

$$Y_{ist} = \sum_{h=1}^{a-2} \alpha_h 1[A_s = a] \times 1[t = h] + \sum_{k=a}^{T_T} \beta_k 1[A_s = a] \times 1[t = k] + c_s + b_t + e_{ist}$$

In this specification, each $\beta_k = \Delta_{ES}^{a+k} = ATT(a, a+k)$ and each $\alpha_k = \Delta_{ES}^{a-h} = ATT(a, a-h)$. The regression model estimates the full set of post-period and pre-period DIDs in one pass

through the data, and provides a simple platform for estimating standard errors and performing hypothesis tests. In applied work, graphs of the pre-period and post-period coefficients from the event study regression are very common. Under the null hypothesis implied by the identifying assumptions, the collection of pre-period coefficients – the α_h – should be equal to zero, and the post-period coefficients will trace out the pattern of time varying treatment effects.

3.4 Staggered Adoption Designs

The staggered adoption design expands the event study to allow for multiple groups with different treatment adoption dates. Often researchers are interested in the causal effects of a state law that has been adopted in a set of states at different times. In our simulated example, different states gain access to sulfa drugs in 1930, 1940, and 1945.

Compared to the simple 2×2 and basic event study designs, staggered adoption muddies the definition of treated and control groups and pre- and post- time periods. Until recently, the typical approach was to combine the staggered adoption DID design with a statistical model that allows for group and period fixed effects. The workhorse specification is the two-way fixed effects model:

$$Y_{ist} = \beta_{FE} D_{st} + c_s + b_t + e_{ist}.$$

Our earlier review covers the TWFE model in detail (31), and an important advantage of TWFE estimator is that it allows social and health science researchers to draw on their existing and often extensive experience with panel data statistical models. However, it is important to understand how to interpret the TWFE estimator in light of the emerging literature on the staggered adoption design. A first point is that, in the 2×2 DID setting, β_{FE} is identical to the Δ_{DID} parameter. A second point is that when treatment effects are *constant*, the TWFE estimator is consistent for the homogeneous constant effect parameter. However, the constant treatment effect assumption is restrictive. It requires that the causal effect of the treatment is does not differ across units, groups, and time periods. In that constant effects scenarios, $\beta_{ist}(a) = \beta$. Thus, under the right conditions, the TWFE model is a useful and convenient platform for analyzing data from a staggered adoption design. In particular, if treatment effects are constant or at least not very heterogeneous, then the TWFE model provides a convenient modeling framework.

In many cases, however, researchers using the TWFE model are not intentionally asserting a constant treatment effects assumption implied by the model. Applied researchers often explore treatment effect heterogeneity by augmenting the basic model to include interaction

terms allowing treatment effects to vary across observed sub-populations, or to vary over time using modified event study specifications. These techniques are often helpful, but we think in practice most applied researchers seem to view the β_{FE} coefficient not as an estimate of a true constant treatment effect but as "some kind of average" of underlying heterogeneous effects. Recent work by Goodman-Bacon (16) and de Chaisemartin and D'Haultfœuille (11) provides a clearer account of how the β_{FE} parameter represents a variance of treatment weighted combination of underlying heterogeneous effects. However, these studies also highlight conditions under which the summary measure may be confounded by the interaction of treatment effect heterogeneity and staggered adoption.

At a broad level, the recent literature shifts the focus away from matters of statistical modelling and towards the research design itself. The *group* $\times$ *time* treatment effect — the $ATT(a,t)$ — is a key building block for interpreting the staggered adoption design (6). The group-time ATT has the same meaning in the staggered adoption case as it did in the 2×2 and event study cases. The difference is that the staggered design distinguishes between *multiple* $ATT(a,t)$ parameters because there are more adoption groups and periods.

In principle, the staggered adoption design makes it possible to identify a collection of different $ATT(a,t)$ parameters using the same no anticipation and common trends assumptions used in the simpler designs. The trick is to apply the standard DID estimator to the correct combination of periods and groups. For a generic $ATT(a,t)$ effect the DID comparison is:

$$\begin{aligned}
\Delta_{SA}^{a,a+k} &= E[Y_{is,a+k} - Y_{is,a-1} | A_s = a] - E[Y_{is,a+k} - Y_{is,a-1} | A_s > a+k] \\
&= E[\beta_{is,a+k}(a) | A_s = a] - \{E[Y_{is,a+k}(0) - Y_{is,a-1}(0) | A_s = a] \\
&\quad - E[Y_{is,a+k}(0) - Y_{is,a-1}(0) | A_s > a+k]\} \\
&= E[\beta_{is,a+k}(a) | A_s = a] \\
&= ATT(a, a+k)
\end{aligned}$$

Once again the logic of the derivation parallels the 2×2 and basic event study case. The only difference is that the treatment group is defined by conditioning on the value of the adoption date, and the control group is defined as all groups that adopt after focal post-period $(a+k)$. This definition uses all feasible control observations. In practice, researchers may choose to work with a more specialized subsets of the feasible set of controls, such as the set of never treated controls or perhaps a set of controls that do not adopt until some specified period in event or calendar time. Constraints like this might be desirable in applications where researchers wish to estimate a sequence of post adoption ATTs for each timing group and want to ensure that variation in treatment effects from one period to the next do not

arise because of changes in the composition of the control group.

In addition to estimating a sequence of time varying treatment effects for each adoption group, the staggered adoption design also allows researchers to construct partial tests of the common trend and no-anticipation assumption using the same approach described for event study designs. Specifically, DID comparisons can be formed to estimate the pre-adoption effects — $ATT(a, a-h)$ — which should each be equal to zero under the null hypothesis that the strong common trend and no-anticipation assumptions are valid.

With a balanced control group, the sequence of pre- and post-treatment $ATT(a,t)$ parameters can be estimated using an event study regression with the sample limited to observations from the specified adoption group, the balanced (clean) control group, and the relevant calendar time periods. The specification is identical to the basic event study specification.

An immediate question is what to do with all of these causal effect estimates? In a small staggered adoption design, where there are not many timing groups, it may be sensible to simply examine each of the effects in isolation. This provides insight into the degree of treatment effect heterogeneity, and the pre-treatment ATTs may help researchers gauge the credibility of each sub-experiment. However, this approach can be unwieldy in larger staggered adoption designs. Estimating multiple sub-group effects may also be statistically inefficient. For both reasons, it will often make sense to aggregate or average the $group \times time$ estimates into a single summary average causal effect parameter. Following Callaway and Sant'Anna (5), we can think of an abstract representation of a summary parameter as:

$$\delta_w = \sum_a \sum_t w(a,t) ATT(a,t)$$

In this expression, $w(a,t)$ is a weight that is attached to a particular $group \times time$ cell, and δ_w is the summary parameter based on that weighting scheme $w()$. In some instances, researchers may want to estimate a collection of summary effects that characterize averages of the $group \times time$ ATTs by *event time* (i.e., years relative to the date of adoption). For example, Wing et al. (30) propose a trimmed aggregate ATT weighting scheme that enforces compositional balance across event times and show how to estimate this parameter directly using a stacked DID estimator. In other cases, weights might be chosen to build a summary measure of average causal effects across all units up to a specific point in calendar time. The details of how best to construct interesting weighted summaries depends on the research question. Callaway and Sant'Anna (5) provide a detailed discussion of several options that will often be useful in applied work.

Many estimators used in applied work – including the TWFE estimator – provide a kind of automated aggregation. These regression based aggregations usually do not correspond to

any of the weighted summary concepts developed in Callaway and Sant'Anna (5), although the stacked DID estimator in Wing et al. (30) is one exception. Instead, standard regression based aggregations produce summaries of underlying heterogeneity that are a byproduct of the optimization problem underlying the regression. Goodman-Bacon (16) shows – for example – that the TWFE estimator applied to the staggered adoption design can be interpreted as a "variance of treatment" weighted average of underlying *group* × *time* average effects. These byproduct averages do not have much appeal as conceptual objects of interest, although in many situations they may not be very different from more theoretically coherent approaches, and they may also have advantages in the form of simplicity and also statistical precision.

3.5 New Understandings of Threats to Validity

The TWFE estimator is a widely used way to analyze data from a staggered adoption DID. Viewed through the lens of panel data econometric models, the TWFE estimator forms an estimate using all available "within-group" variation in the treatment variable. Purely between group and time series variation in the treatment variable is discarded in order to eliminate possible confounding from group and time unobserved effects. As we explained earlier, the TWFE estimator is equivalent to the DID estimator in the 2×2 setting and in the staggered adoption setting under a constant treatment effects assumption. The connection to the DID research design is murkier in staggered adoption designs with treatment effect heterogeneity.

In an influential paper Goodman-Bacon (16) showed that the within variation in treatment can be expressed as a collection of 2×2 DIDs. In that sense, the TWFE estimator is a weighted average of these underlying DIDs. Goodman-Bacon (16) works out the weights assigned to each underlying DID comparison, which is helpful in understanding which policy changes "drive" the overall estimate.

However, the most important contribution of the Goodman-Bacon (16) paper is the *discovery* that some of the 2×2 DID comparisons that contribute to the TWFE parameter are actually confounded even when the common trend and no-anticipation assumptions are valid as stated. Goodman-Bacon (16) categorizes the 2×2 DID comparisons that contribute to the TWFE estimator into three types:1) Treated vs Never Treated DIDs ($A_s = a$ vs $A_s = \infty$), 2) Early vs Late DIDs ($A_s = a$ vs $A_s = c > a$), and 3) Late vs Early DIDs ($A_s = a$ vs $A_s = b < a$). In each case, the actual variation comes from specific calendar time periods during which one group changes status and the other does not. These three types of comparisons are weighted together to form a single summary coefficient, β_{FE}.

The problematic DIDs involve comparisons in which a treatment group that changes treatment status between periods is compared with an "already treated" comparison group.

This type of late adopter vs. earlier adopter DID is based on within-group variation in the TWFE sense. But it can create problems if treatment effects vary with time since event. For example, consider a DID in which timing group $A_s = a$ is compared with an already treated control group with $A_s = b < a$:

$$\begin{aligned}
\Delta_{Bad}^{a,a+k} &= E[Y_{is,a+k} - Y_{is,a-1}|A_s = a] - E[Y_{is,a+k}Y_{is,a-1}|A_s = b] \\
&= E[\beta_{is,a=k}(a) + Y_{is,a+k}(a)(0) - Y_{is,a-1}(0)|A_s = a] \\
&\quad - E[(Y_{is,a+k}(0) + \beta_{is,a+k}(b)) - (Y_{is,a-1}(0)) + \beta_{is,a-1}(b))|A_s = b] \\
&= E[\beta_{is,a=k}(a)|A_s = a] + E[\beta_{is,a+k}(b) - \beta_{is,a-1}(b)|A_s = b] \\
&\quad \{E[Y_{is,a+k}(0) - Y_{is,a-1}(0)|A_s = a] - E[Y_{is,a+k}(0) - Y_{is,a-1}(0)|A_s = b]\} \\
&= ATT(a, a+k) + E[\beta_{is,a+k}(b) - \beta_{is,a-1}(b)|A_s = b]
\end{aligned}$$

The first line shows the DID comparing the later adopting group with $A_s = a$ to the early adopting group with $A_s = b < a$ before and after the later adopting group is treated. The second line substitutes the potential outcomes and imposes the no-anticipation assumption. Importantly, the realized outcomes in the early adopting control group are not "untreated outcomes" in this comparison. Since the $A_s = b$ group has already been exposed to treatment, its realized outcomes include the causal effects $\beta_{ist}(b)$. The third line re-arranges. The term in braces is the difference in the time trend in untreated outcomes in each group, which equals zero under the common trends assumption. The final line shows that $\Delta_{Bad}^{a,a+k}$ is equal to $ATT(a, a+k)$ plus a *bias term* that is driven by time varying treatment effects in the early treatment comparison group. Depending on the sign and magnitude of the bias term, $\Delta_{Bad}^{a,a+k}$ can be biased up or down compared with $ATT(a, a+k)$. Sign flips are possible, for example. The bias occurs despite the fact that both the common trend assumption and the no anticipation assumption are valid. A key point here is that in the staggered adoption design some of "within variation" does not identify a causal effect. In particular, within variation based on the comparison of later and earlier adoption groups is confounded by time varying treatment effects. In situations where treatment effects do not vary much over time, the bias term would disappear and these comparisons would identify the relevant ATT. Because the bad Late vs Early comparisons are included in the overall weighted average, it is possible that β_{FE} is not simply a weighted average of underlying $group \times time$ ATTs.

The implicit weights correspond to factors that influence the amount of variance contributed by each underlying 2×2 DID. As a result, these weights are a function of the size of the timing group and how close to the middle of the overall study window the timing group is

treated. Groups treated towards the middle of the study window experience more variation in their treatment variable and get more weight. These weights determine how important the bad Late vs Early comparisons are in the overall analysis. If the bad comparisons receive negligible weight, then this strange new source of bias may not be practically important, for example.

Figure 2a illustrates the weights associated with each type of DID comparison in our simulated sulfa drugs example. Start with the $A_s = 1940$ adoption group, which adopts near the middle of the study window. It is possible to construct three separate 2×2 DIDs in which the 1940 adoption group serves as the treated group.[6] The first and most straightforward estimator is to compare the 1940 group to the group of states that never obtained sulfa drugs before and after 1940 ($A_s = \infty$). The comparison is represented by the blue square with the label "1940 vs. never". This is a clean comparison that will not suffer from bias if the standard DID assumptions hold. Indeed, the DID in the graph nearly perfectly recovers the true average treatment effect for the 1940 treated group of -25.

A second 2×2 DID compares the 1940 treatment group to the set of states that obtained sulfa drugs in 1945. This "Early vs Late" comparison is represented by a red circle with the label "1940 vs. 1945." Since the true treatment effect gets larger over time in this example, the 1940 vs 1945 DID represents an average effect from earlier in event time than the 1940 vs Never comparison. The third 2×2 DID is a bad (Late vs Early) DID that compares the 1940 group to 1930 group. The estimate is shown by the green triangle with the label "1940 vs. 1930." The Late vs Early comparison is an example of a problematic DID comparison because the control group is already treated. And indeed the bad comparison yields a *positive* (i.e. wrong signed) treatment effect estimate of 54, implying that introducing sulfa drugs increased mortality.

The bias occurs because the difference between the 1940 treated group and the 1930 treated group is smaller before 1940 than it is after 1940. The reason for this can be seen by examining Figure 1, focusing on the 1940 sulfa introduction group (circles) and the group treated ten years earlier in 1930 (triangles). Prior to 1930, both groups, the triangles and squares, share a common trend. The 1930 group then gets treated and experiences a treatment effect that grows over time. In the time after it is treated it acts as a comparison group for the middle treated 1940 group. However, these two groups are no longer on common trends, since the treatment itself has changed the trend for the 1930 treated group. During the pre-period of this 2×2 DID, which runs between 1930 and 1939, the triangles experience a flat trend, while the squares are following a steep downward trend driven by the treatment

[6] We make the graph using the bacondecomp package in R. See our online supplement at, `https://github.com/hollina/arph-did-example`.

itself in the earlier treated group.

Unfortunately for the two-way fixed effects DID estimate, this biased estimate receives a non-negligible weight (18%) when forming the combined treatment effect estimate. This bias can be seen in any of the three potential Late vs Early comparisons in our example. Because of these biased comparisons, the two-way effects estimate is quite biased relative to the average real treatment effect. This is well evidenced in Figure 2b, where the TWFE estimate is above zero at a *positive* 5.1 (p-value of .03), while other estimates of the "real" average treatment effect are well below zero, and which we discuss in more detail below.

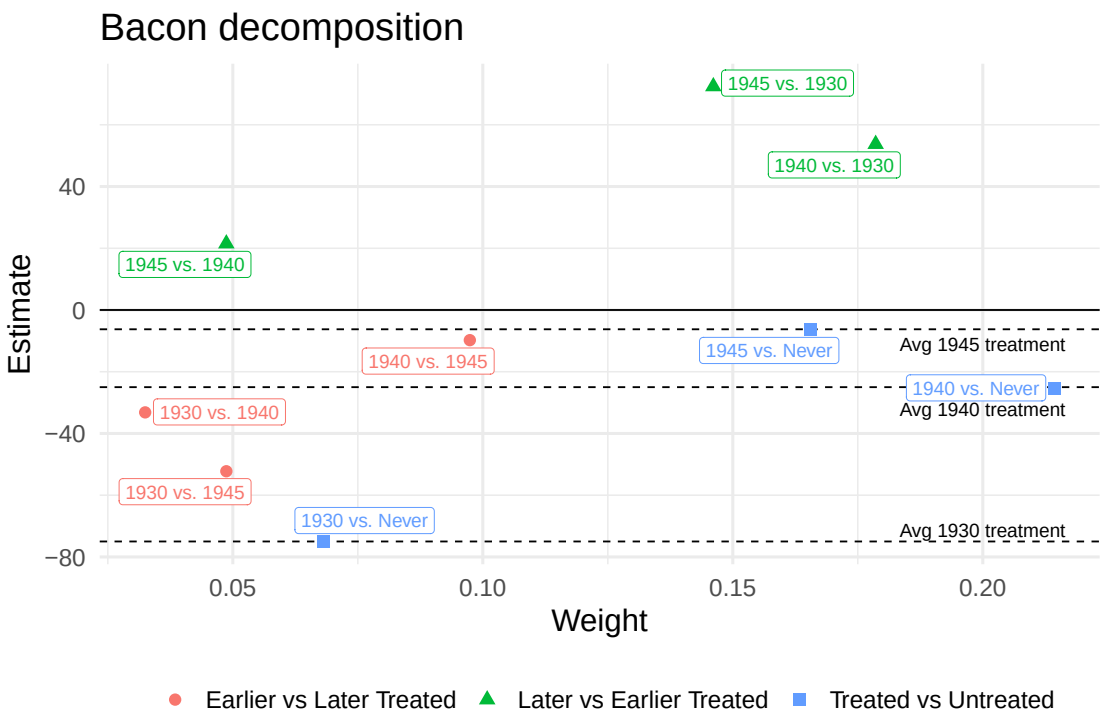
One key insight of Goodman-Bacon (16) is that we can avoid this bias by simply removing these potentially biased comparisons from our estimation strategy. This insight underlies one of the key design principles that have emerged from the new DID literature: the importance of using so-called "clean controls". The idea is that – in a staggered adoption setting – causal inferences should be based on DIDs that compare treated timing groups to never treated comparison units, to future treated comparison units, or both. Causal inference should not be built on comparisons between treated timing groups and previously treated comparison units. In the sulfa drugs example, we would exclude the comparison of the 1940 group as the treated group to the earlier exposed 1930 group. Likewise, we would exclude the comparisons based on the 1945 group as treated to the earlier exposed 1930 or 1940 group.

4 Estimation Strategies and Techniques

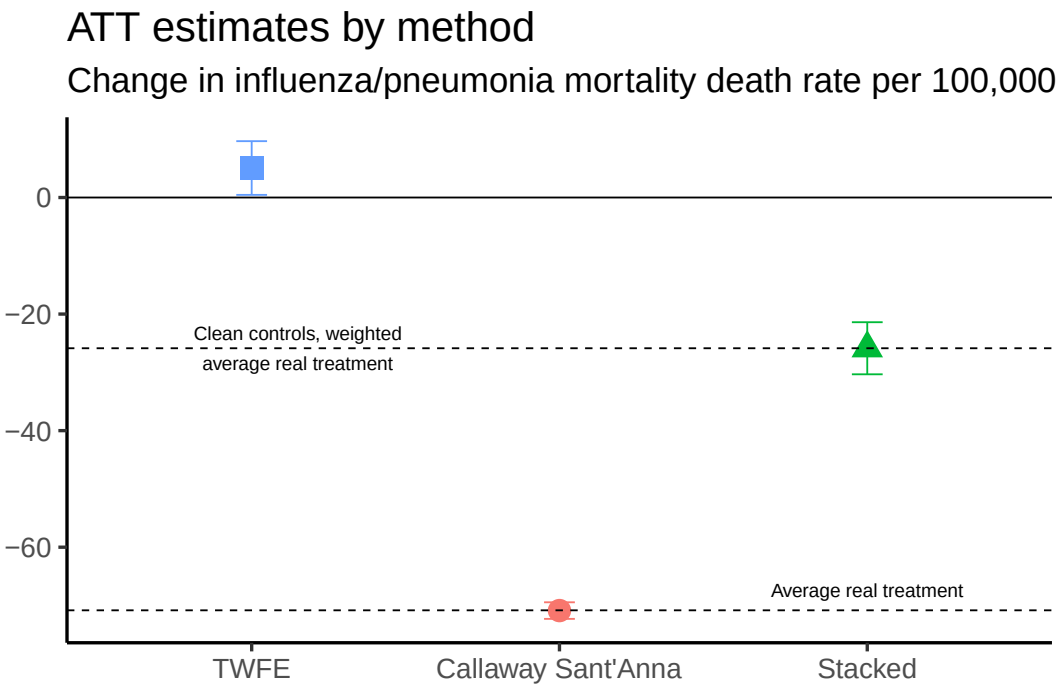
Studies based on staggered adoption DID designs are increasingly expected – by reviewers and editors at many journals – to address the concerns raised by the staggered adoption design. In this section, we describe two specific methods: (i) stacked difference-in-difference introduced in Cengiz et al. (7) and Deshpande and Li (13), and (ii) the explicit group-time approach developed by Callaway and Sant'Anna (6). Our discussion of the stacked DID emphasizes clean controls and balanced sample composition, and it aggregates results "automatically", see Wing et al. (30) for further analysis of the stacked DID estimation strategies. The Callaway and Sant'Anna (6) estimator hews closely to the *group* $\times$ *time* perspective presented in this paper but it provides explicit options for aggregating effects in flexible ways. After discussing these two methods in some detail, we provide a broad overview of several other leading approaches.

Figure 2: Staggered designs and average treatment effects

(a) Goodman-Bacon (16) decomposition of two-way fixed effects DiD



(b) Comparison of ATT across methods



Notes: In the top panel each label takes the format of "treated vs. comparison".

4.1 Stacked Estimation and Clean Controls

In the stacked DID framework, each policy adoption is viewed as a separate sub-experiment, and each sub-experiment is designed to be free from confounded DID comparisons. We describe the inclusion criteria, data structure, and estimation strategies used in the stacked DID approach.

4.1.1 Inclusion Criteria and Sub-Experiments

In a basic event study, there are two groups and possibly multiple periods pre and post. In the staggered adoption setting, the maximum length of time before treatment adoption is $A_s - T_1$ and the maximum length of time after treatment adoption is $T_T - A_s$. Thus, the length of the feasible pre and post periods varies across adoption groups.

To implement a stacked analysis, we impose a fixed event time window that will be used across all sub-experiments. Let κ_a be the length of the pre-treatment period and κ_b to be the length of the post-treatment period. The κ parameters are a design choice with practical implications. A shorter event time window may allow more policy events to be studied. A longer window allows treatment effects that vary with time since treatment to be studied, perhaps for a smaller subset of adoption events. Let $\Omega_A = \{A_s | T_1 + \kappa_a \leq A_s \leq T - \kappa_b\}$ represent the set of policy changes that are feasible to study given a choice of κ_a and κ_b. Use $d \in \Omega_A$ to index the admissible sub-experiments.

We build a separate data set for each sub-experiment $d \in \Omega_A$. The observations included in a given sub-experiment are determined by three inclusion criteria:

IC 1. *Homogeneous Treatment Timing:* Treatment adoption dates are homogeneous and non-staggered.

IC 2. *Clean Controls:* The control group consists of units that are not exposed to treatment during the event study period running from $d - \kappa_a$ to $d + \kappa_b$.

IC 3. *Admissible Calendar Periods:* All observations on treated and control units come from calendar time periods that fall inside the event window so that $d - \kappa_a \leq t \leq d + \kappa_b$.

Under the homogeneous treatment timing condition, let $T_{sd} = 1(A_s = d)$ indicate that an observation from group s is a member of the treatment group in sub-experiment d. Under the clean controls condition, define $C_{sd} = 1(A_s > d + \kappa_b)$ to indicate group s is a valid clean control for sub-experiment d. Finally, under the Admissible Calendar Periods condition, let $M_{td} = 1(d - \kappa_a \leq t \leq d + \kappa_b)$ indicates that calendar period t falls inside the event window for sub-experiment d. Putting the three rules together implies that $I_{istd} = M_{td}(T_{sd} + C_{sd})$ is

a binary inclusion variable indicating whether observation i from group s in calendar period t belongs in sub-experiment d.

Applying the inclusion rule to the raw data repeatedly for each sub-experiment yields a collection of sub-experimental data sets, each centered around a specific policy change and including data only on clean controls and treated units for the appropriate calendar time periods. The sub-experimental data sets are then vertically concatenated into a single "stacked" analytic dataset. Note that some units will appear as control observations in multiple sub-experimental data sets.

4.1.2 Stacked Estimation

To estimate an event study regression based on this stacked data, we use Y_{ised} to represent the observed outcome for unit i from state s in *event time period* $e = t - d$ in sub-experiment d. Then the following regression model is fit to the stacked data:

$$Y_{ised} = \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\beta_e^{stacked} (D_{sd} \times 1[e = h]) \right] + a_{sd} + b_{de} + \epsilon_{ised} \tag{1}$$

In this regression a_{sd} and b_{ed} are a set of $group \times sub-experiment$ fixed effects and $event-time \times sub-experiment$ fixed effects. This method uses what looks like a typical TWFE regression estimate, but because of the structure of the data, it only incorporates clean controls. One way to think about this regression is as a way of estimating all of the $ATT(a,t)$ parameters and then immediately aggregating them into a single set of event time parameters, $\beta_e^{stacked}$. Versions of the fixed effects specification given above has been used in applied work by Cengiz et al. (7) and Deshpande and Li (13), although their stacking procedure is not fully balanced as we propose here. Although free of confounding from late vs early adoption comparisons, the aggregation produced by these models are still based on implicit variance of treatment weights, which may not be intuitive or appealing way to summarize the underlying group-time ATTs. Wing et al. (30) show how to construct sample weights that ensure that the stacked event study specification corresponds to a coherent aggregate ATT parameter. In the examples Wing et al consider in their paper, the correctly weighted estimator often produces estimates that are very close to the estimates from the fixed effect specification given above.

Figure 2b shows the stacked DID estimator for our sulfa drugs example. Relative to the TWFE DID estimate, we see the expected negative effect. In addition, the estimate is almost identical to one conception of the "real" effect. This comes from taking the weights and

estimates from *only* those 2x2 DID estimates in Figure 2a that come from clean estimates and do not suffer from confounding, with weights re-normalized to sum to 1. This is in essence what the stacked DID estimator is accomplishing, and because both methods are based on OLS regressions, they are both variance weighted.

4.2 Callaway and Sant'Anna and Aggregation

Callaway and Sant'Anna (5) develop an approach that is explicitly organized around the $group \times time$ ATT parameter, and the notation we use in this paper draws on their approach. The Callaway and Sant'Anna (5) paper makes two larger contributions beyond the basic framework. First, they develop strategies for incorporating time invariant baseline covariates into the analysis of a staggered adoption DID design using inverse propensity score weights, and or regression adjustment methods. Second, they discuss strategies for aggregating $group \times time$ ATT parameters.

Although there are many ways that $ATT(a,t)$ parameters could be aggregated, we think that applied researchers will often be interested in examining dynamic treatment effects using what Callaway and Sant'Anna (5) call a "balanced event study" aggregation. Using κ_a and κ_b to represent an event study window of interest, the balanced event study aggregate ATT at a specific event time q periods away from treatment adoption is:

$$\delta_q^{\kappa_a,\kappa_b} = \sum_a 1[T_1 + \kappa_a \leq a \leq T_T - \kappa_b] \times ATT(a, a+q) \times Pr(A_s = a | T_1 + \kappa_a \leq a \leq T_T - \kappa_b)$$

$\delta_q^{\kappa_a,\kappa_b}$ is an average of $ATT(a, a+q)$ parameters across groups with different values of $a \in A_s$. In practice, the idea is to estimate a family of $\delta_q^{\kappa_a,\kappa_b}$ parameters for values of $q = -\kappa_a ... 0 ... \kappa_b$ and then plot these parameters in an event study graph. The summation cycles over each of the treatment adoption dates, $a \in \Omega_A$. The first term *trims* out any adoption dates a that occurs outside the $\kappa - window$, ensuring an common composition across event times. The third term is the weight assigned to the surviving $ATT(a, a+q)$and the weight is the sample size of the adoption group relative to the total sample size of all admissible adoption groups.

In Figure 2b we show the dynamic aggregate estimate from Callaway and Sant'Anna (6) that averages the average treatment effects across all timing groups that have different lengths of exposure. This estimator is almost identical to the second "real" treatment effect, which constructs the timing group specific treatment effect by subtracting the imposed treatment effect for each group in event-time and then takes the simple average across timing-groups

for the entire post-treatment period.

4.3 Comparison of estimation techniques

Table 1 provides a side-by-side summary of six of the leading estimation techniques, highlighting attributes that are relevant for applied researchers when selecting a method.[7] Below, we give a short explanation of key characteristics and practical implications of each method.

4.3.1 Overview of Estimators

At a basic level, each estimator provides a different way to ensure that later treated groups are not being compared to earlier treated groups. Like Callaway and Sant'Anna (6), many of them explicitly estimate dynamic treatment effects akin to event studies, for each timing group, and then provide ways to aggregate these group-time estimates into a summary estimate more comparable to a simple DID coefficient.

Sun and Abraham (27) and Wooldridge (33) use ordinary least squares regressions, but include a large set of interaction terms to carefully ensure that treated units are only being compared to clean controls. Gardner (15) takes a different approach by setting up a two step regression-based method that first estimates the group- and time-fixed effects using only untreated observations before estimating treatment effects. Finally, de Chaisemartin and D'Haultfœuille (11) build an estimator focused on comparing changes in treated units just before and after treatment, to units who do not experience a change at that time.

4.3.2 Do you need a specific software package to implement?

As a practical matter, regression based estimators, such as those proposed in Gardner (15), Wooldridge (33), Sun and Abraham (28), and stacked DID, are fairly straightforward to implement using standard commands in statistical software such as R or Stata, though specialized software packages are available in most cases. On the other hand, the methods proposed by Callaway and Sant'Anna (6) and de Chaisemartin and D'Haultfœuille (11) may be better suited towards utilizing a package. While these techniques are based on simple comparisons for each timing group, they involve multiple steps and may be less intuitive to code since they do not use regressions.

[7] For additional new estimators, see the complementary reviews in Roth et al. (26) and de Chaisemartin and D'Haultfoeuille (10). Notable methods we do not cover include estimators from Dube et al. (14), Borusyak et al. (3), and de Chaisemartin and D'Haultfoeuille (9).

4.3.3 What are the primary outputs? How do they relate to a well-known estimand?

Researchers familiar with the canonical 2×2 DID design or two-way fixed effects regression will be used to estimating a single parameter that summarizes the treatment effect, or a set of event study coefficients. Three of the new estimators are similar in this regard, by outputting a single, (weighted) average effect: stacked DID, Gardner (15), and de Chaisemartin and D'Haultfœuille (11). However, a potential disadvantage of these approaches is that the single effect may not directly correspond to an easily-interpretable parameter, such as the average treatment effect on the treated (ATT). For example, stacked DID results in a variance weighted estimate, since it is using OLS methods. Gardner (15) provides multiple options for weighting estimates that correspond to different weighted averages of ATTs. de Chaisemartin and D'Haultfœuille (11) estimate a single ATT, interpreted as an average treatment effect in the first period in which a group changes treatment status.

Alternatively, another class of methods explicitly estimate all group-time ATTs, giving the researcher flexibility to aggregate if desired (6, 28, 33). As mentioned in Section 4.2, this approach adds a degree of complexity but gives the researcher explicit control over the averaging process. For example, Callaway and Sant'Anna (6) provide a method to aggregate group-time ATTs into a single average that is comparable to the ATT estimated in a 2×2 DID.

4.3.4 What observations are included in the control group?

All methods ensure that only clean observations that have not been treated in earlier periods are included in the control group. Most methods provide flexibility to include either never treated units, not yet treated units or both in the control group. Gardner (15) and Wooldridge (33) are the exceptions in that they must use both.[8]

Why might one control group be preferred over another? Choosing to only use never-treated observations as the control group may be more robust to violations of the no anticipation assumption, since anticipation effects could impact the trend in soon to be treated observations. However, as Wooldridge (33) notes, estimators that use all untreated observations as controls may have a slight advantage in precision because more of the data is being used.

[8]In the case when all units are eventually treated, note that Wooldridge (33) provides guidance on how to use the not-yet treated and last-to-be treated as the control group.

4.3.5 What guidance exists to calculate standard errors?

The estimators in Gardner (15) and Wooldridge (33) are based on well-worn statistical methods, which include formulas for the asymptotic variance-covariance matrices. For applied work, this means any regression software can correctly calculate standard errors off-the-shelf, without corrections or additional packages.

In contrast, a straightforward reason to prefer a method's package is to correctly calculate standard errors. For example, de Chaisemartin and D'Haultfœuille (11) derive a novel asymptotic distribution for the variance-covariance matrix of their estimator, which can be implemented using their Stata package. Similarly, standard errors in both Callaway and Sant'Anna (6) and Sun and Abraham (28) are calculated via bootstrapping.

Among the techniques highlighted here, the least formal econometric theory exists on how to properly calculate standard errors within the stacked DID. Because the estimation relies on a single regression command, after the data has been reshaped, a standard approach is to simply cluster errors at the group-level, as would be common in a standard two-way fixed effects regression.

It is worth noting that the expected efficiency (statistical precision) of the new estimators remains a somewhat open question. There may also be an overall bias-efficiency trade-off between new estimators and the two-way fixed effects estimator.[9] Intuitively, ensuring clean controls often comes at a cost of excluding some of the data, and this insight guides some of practical advice in Section 5.[10]

4.3.6 What is the role, if any, for covariates?

While we have focused here on new sources of confounding due to time varying treatment effects and staggered adoption, the standard common trends assumption is still crucial to any DID design. Acknowledging this, most of the new estimators allow for a weaker identifying assumption that states trends are parallel, conditional on a set of time-invariant, observed covariates (6, 9, 15, 33).

What about *time-varying* covariates? To include time varying covariates in a DID design, researchers must assume that the covariates are not affected by the treatment and that the covariates do not impact the effect of the treatment. These are strong assumptions that may not hold in many settings.[11] Readers are referred to Gardner (15), Caetano et al. (4) and

[9] See de Chaisemartin and D'Haultfœuille (11) for an excellent discussion.

[10] We also note a separate estimator by Athey and Imbens (1) that views uncertainty as design-based, rather than sampling based, that may be appropriate in settings when the timing of treatment across groups is plausibly random.

[11] Deeper discussion of this point is available in Wooldridge (32), Pei et al. (23), and, Caetano et al. (4).

de Chaisemartin and D'Haultfoeuille (9) for early work in this space.

4.4 Other considerations

Most of the new estimators are designed for binary absorbing treatments, meaning that treatment takes on the same value for all treated groups and once a group is treated, it remains treated for the duration of the data. However, there are exceptions. A key contribution of the body of work by De Chaisemartin and D'Haultfoeuille is an explicit consideration of a broader class of treatment variables (9, 11, 12). For example, the authors have developed methods that are robust when treatments turn on and off (non-absorbing); when treatment compliance or receipt varies within groups ('fuzzy'); and when the treatment variable represents a different levels of exposure.

5 Practical Advice

In this section, we provide a check-list of suggested best practices for applied researchers confronted with a staggered adoption study design.

5.1 Explore the raw data

In some ways the new literature provides more structure to carrying out the same types of data exploration that researchers have typically performed in DID projects. One of the ways that most researchers figrst begin to probe the key assumptions of no anticipation and common trends has been to plot the raw data over calendar time among treatment and control units. Under new understandings of DID, this should still be the first step, but with the added advice to separate the data into timing groups to make a figure akin to Figure 1.

5.2 Assess if and describe why bias is likely to be an issue in your setting

With this figure in hand, researchers can begin to assess the scope for potential bias in their DID design. First, researchers can visually compare trends between timing groups (and an untreated group if applicable) in pre-treatment years to assess no-anticipation and common trends. Researchers also get a sense of whether or not any potential treatment effects change over time by visualizing whether treated groups experience a different trend after treatment.[12]

[12]Note that we recommend plotting the raw data in calendar time, not an event study. Coefficients in an event study can be affected by the same bias due to heterogeneous effects as the two-way fixed effects model (Sun and Abraham (28)).

It is also worth noting that even if adoption is staggered and treatment effects are time varying, the severity of confounding also depends on how much the adoptions spread over time. If adoption is staggered only over few and short periods, the problems pointed out here would be smaller than if adoption is spread out over many years. An emerging best practice is to also plot the adoption timeline or to include a table with adopting groups and period of adoption (as in Table 1 of Cook et al. (8)). This allows readers to see how the pattern of adoption may affect the simple two-way fixed effects estimates.

5.3 Use an empirical diagnostic method to explore the possible bias

In addition to visualizing the raw data and describing the extent of staggering, more formal diagnostics can help understand the potential extent of confounding in a TWFE regression analysis. For example, researchers can use the method in Goodman-Bacon (17), Figure 2a, to decompose all embedded two-by-two DID comparisons, their treatment effect estimates, and their contribution (weight) to the overall two-way fixed effect coefficient. This allows the author to identify how much weight "bad" controls contribute to the overall estimate, the overall magnitude of the biased comparisons relative to the aggregated two-way fixed effects estimate, and the specific comparisons that may be problematic. Similarly, Sun and Abraham (28) show how event study coefficients can be similarly "decomposed" into their underlying group-time ATTs and associated weights, and de Chaisemartin and D'Haultfœuille (11) propose a ratio statistic that assesses how robust the two-way fixed effects estimate is to treatment effect heterogeneity.

5.4 Plan to estimate your treatment effect with at least two techniques

One approach researchers can take is to consider estimating the treatment effect using at least two specifications: the simple two-way fixed effects regression, and one of the new estimators, such as those summarized in Table 1. While these new estimators are robust to heterogeneous effects and staggered adoption, there may be tradeoffs in some contexts. For one, there may be a bias-efficiency trade-off, as discussed in Section 4. In addition, readers may prefer to at least see the specification they are more accustomed to in order to help contextualize the results from new estimators.

Applied researchers are likely familiar with estimating regressions using alternate specifications as a robustness check. Our advice on how to place multiple estimators within a paper depends on the results of diagnostics. If visualizing the raw data and other diagnostics suggest that confounding due to time varying treatment effects is likely to be large, we recommend that an estimator robust to this confounding be considered the "main" estimate

and TWFE estimates potentially be included with explicit acknowledgment that they are likely confounded. On the other hand, if potential confounding seems likely to be small and authors feel the TWFE specification will be more intuitive to readers, then it could be included as the main estimate with at least one alternative specification from the more robust methods.

Because the robust methods are new within the econometric literature, their behavior and relative performance is still being established. Without a clear "best" robust estimator, at this point in time, researchers can use the characteristics of their study's setting, the desired parameters they wish to estimate, and their discretion to select the specific robust estimators to use. (See Section 4 for a detailed comparison).

5.5 Consider new robustness checks to test for pre-trends

Empirical researchers often include an event study as partial evidence of whether the "No Anticipation" and "Common Trends" assumptions from Section 3 hold (31). However, recent work has noted that many event studies are under-powered in practice and unlikely to detect meaningful violations (24, 25). Two alternative approaches have been proposed. First, researchers can use the methods in Roth (25) to determine the smallest trend that can be detected in the pre-period, with a given power, for their sample. Second, scholars can estimate bounds on their treatment effect in the post-period, given a range of possible trends in the pre-period (24).

6 Conclusion

In recent years, a rapidly developing methods literature has identified and provided solutions to overlooked problems that occur when using standard two-way fixed effects models to estimate a DID study design. In this review, we have outlined two key takeaways from this literature for public health researchers. First, when treated groups adopt treatment at different times (staggered adoption) and when the effect of treatments change over time (time-varying treatment effects), the standard regression approach with group and time fixed effects can be biased. This bias occurs because part of the estimate is based on comparing treated units to previously treated units, whose time trend itself has been impacted by their earlier treatment. Second, when adoption is staggered, the researcher really has multiple sub-experiments. Therefore, researchers must choose what weights to apply to each sub-experiment to aggregate them into a single summary estimate.

With these two points in mind, we describe some of the various estimators that are robust

to the bias introduced by treatment effect heterogeneity. We also provide some guidance on best practices for diagnosing potential bias, choosing an appropriate estimator, and robustness checks. Finally, we illustrate these problems and solutions with an example public health intervention. With these fundamental points and our check-list in mind, public health DID research can be strengthened by implementing new methods when studying interventions that occur at different times for different groups.

References

[1] Athey, S. and G. W. Imbens (2022, January). Design-based analysis in Difference-In-Differences settings with staggered adoption. *Journal of Econometrics 226*(1), 62–79.

[2] Baker, A. C., D. F. Larcker, and C. C. Y. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395.

[3] Borusyak, K., X. Jaravel, and J. Spiess (2022, April). Revisiting Event Study Designs: Robust and Efficient Estimation.

[4] Caetano, C., B. Callaway, S. Payne, and H. S. Rodrigues (2022, February). Difference in Differences with Time-Varying Covariates. arXiv:2202.02903 [econ].

[5] Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

[6] Callaway, B. and P. H. C. Sant'Anna (2021, December). Difference-in-Differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

[7] Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019, August). The Effect of Minimum Wages on Low-Wage Jobs*. *The Quarterly Journal of Economics 134*(3), 1405–1454.

[8] Cook, A. C., G. Leung, and R. A. Smith (2020, March). Marijuana Decriminalization, Medical Marijuana Laws, and Fatal Traffic Crashes in US Cities, 2010–2017. *American Journal of Public Health 110*(3), 363–369. Publisher: American Public Health Association.

[9] de Chaisemartin, C. and X. D'Haultfoeuille (2022a, March). Difference-in-Differences Estimators of Intertemporal Treatment Effects.

[10] de Chaisemartin, C. and X. D'Haultfoeuille (2022b). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. pp. 30.

[11] de Chaisemartin, C. and X. D'Haultfœuille (2020, September). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

[12] de Chaisemartin, C. and X. D'HaultfŒuille (2018, April). Fuzzy Differences-in-Differences. *The Review of Economic Studies 85*(2), 999–1028.

[13] Deshpande, M. and Y. Li (2019, November). Who Is Screened Out? Application Costs and the Targeting of Disability Programs. *American Economic Journal: Economic Policy 11*(4), 213–248.

[14] Dube, A., D. Girardi, Jordà, and A. M. Taylor (2023, April). A Local Projections Approach to Difference-in-Differences Event Studies.

[15] Gardner, J. (2022, July). Two-stage differences in differences. arXiv:2207.05943 [econ].

[16] Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research.

[17] Goodman-Bacon, A. (2021, December). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

[18] Gupta, S., L. Montenovo, T. Nguyen, F. Lozano-Rojas, I. Schmutte, K. Simon, B. A. Weinberg, and C. Wing (2023). Effects of social distancing policy on labor market outcomes. *Contemporary Economic Policy 41*(1), 166–193.

[19] Hollingsworth, A., K. Karbownik, M. A. Thomasson, and A. Wray (2022, November). The gift of a lifetime: The hospital, modern medicine, and mortality. Working Paper 30663, National Bureau of Economic Research.

[20] Jayachandran, S., A. Lleras-Muney, and K. V. Smith (2010, April). Modern medicine and the twentieth century decline in mortality: Evidence on the impact of sulfa drugs. *American Economic Journal: Applied Economics 2*(2), 118–46.

[21] Kim, N.-H., H. W. Elani, and I. Kawachi (2023). Did dental insurance expansion improve dental care needs among korean adults? difference in difference analysis. *Journal of Epidemiology 33*(2), 101–108.

[22] Mullachery, P. H., D. A. Quistberg, M. Lazo, K. Indvik, C. Perez-Ferrer, N. López-Olmedo, M. A. Colchero, and U. Bilal (2022). Evaluation of the national sobriety checkpoints program in mexico: a difference-in-difference approach with variation in timing of program adoption. *Injury epidemiology 9*(1), 32.

[23] Pei, Z., J.-S. Pischke, and H. Schwandt (2019, April). Poorly Measured Confounders are More Useful on the Left than on the Right. *Journal of Business & Economic Statistics 37*(2), 205–216. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07350015.2018.1462710.

[24] Rambachan, A. and J. Roth (2022). A More Credible Approach to Parallel Trends.

[25] Roth, J. (2022, September). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights 4*(3), 305–322.

[26] Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2022). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. pp. 54.

[27] Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

[28] Sun, L. and S. Abraham (2021, December). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

[29] Thomasson, M. A. and J. Treber (2008, January). From home to hospital: The evolution of childbirth in the United States, 1928–1940. *Explorations in Economic History 45*(1), 76–99.

[30] Wing, C., S. Freedman, A. Hollingsworth, and K. Simon (2023). Stacked difference in differences. Working paper.

[31] Wing, C., K. Simon, and R. A. Bello-Gomez (2018, April). Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. *Annual Review of Public Health 39*(1), 453–469.

[32] Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statis-*

tics 87(2), 385–390.

[33] Wooldridge, J. M. (2021, August). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators.

[34] Yan, Y., M. Yoshihama, J. S. Hong, and F. Jia (2023). Substance use among asian american adults in 2016–2020: A difference-in-difference analysis of a national survey on drug use and health data. *American Journal of Public Health 113*(6), 671–679.

	Stacked Difference-in-Difference	Gardner (2021)	Callaway & Sant'Anna (2021)	Sun & Abraham (2021)	Woolridge (2021)	De Chaisemartin & D'Haultfoeuille (2020)
Synopsis	Create sub-experiments with clean controls and stack. Run regression on reshaped data.	Two-stage DiD. Estimate group and period fixed effects in untreated sample. Use to determine treatment effect in full sample.	Decide control group. Calculate all group-time treatment effects. Flexible aggregation.	Estimate a 'saturated' event study with separate effects for each group and period. Reweight and aggregate.	Estimate all group-time treatment effects using a specific pooled OLS or fixed effects specification. Aggregate.	Use units who switch into or out of treatment to identify an average treatment effect.
Do you need a software package to implement?	A package is available (*STACKEDEV, Stata*) but not necessary. Researchers can implement directly using regressions in standard statistical software.	A package is available (*did2s, R*) but not necessary. Researchers can implement directly using regressions in standard statistical software.	A package is available (*did, R*) and recommended.	A package is available (*eventstudyinteract*, *Stata*) but not necessary. Researchers can implement directly using regressions in standard statistical software.	A package is not available. Researchers can implement directly using regressions in standard statistical software.	A package is available (*did_multiplegt*, *Stata*) and recommended.
What are the primary outputs?	Single aggregated treatment effect or even study estimates	Single aggregated treatment effect or event study estimates	Individual treatment effects by group and period (*group-time ATTs*)	Event study coefficients (*which can be interpreted as group-time ATTs in the post-period*)	Individual treatment effects by group and period (*group-time ATTs*)	Single aggregated treatment effect
How do the outputs relate to a well-known estimand—such as, the average treatment effect on the treated (ATT) calculated in a two-by-two DiD?	The estimate is interpreted as an average of group-time ATTs, where each is weighted by variance.	The estimate is interpreted as an average of group-time ATTs, where each is given equal weight.*	The group-time ATTs can be aggregated to a single ATT that is analogous to the canonical DiD.*	The event study coefficients in the post-period can be aggregated to a single ATT, weighted by the share of each group by period.	The group-time ATTs can be aggregated to a single ATT, where each is given equal weight.*	The estimate is interpreted as the average treatment effect (ATE) of all switching cells.
What observations are included in the control group?	Researcher can specify their desired inclusion criteria for control observations.	Both never-treated and not-yet-treated are included in the control observations.	Researcher can specify never-treated or not-yet treated as the control observations.	Researcher can specify never-treated or last-to-be treated as the control observations.	Both never-treated and not-yet-treated are included in the control observations. Last-to-be treated can be used if all groups are eventually treated.	Units whose treatment status does not change between two periods are used as the control observations.
What guidance exists to calculate standard errors?	No formal theory. Standard practice is to cluster at the group level.	Exact asymptotic distribution	Bootstrap	Bootstrap	Exact asymptotic distribution	Novel asymptotic distribution
What is the role (if any) for time invariant covariates?	Time invariant covariates can be included to allow for conditional parallel trends.	Time invariant covariates can be included to allow for conditional parallel trends.	Time invariant covariates can be included to allow for conditional parallel trends.	Not an explicit part of the method.	Time invariant covariates can be included to allow for conditional parallel trends.	Covariates require an assumption that the treatment effect is homogeneous.

Notes: (*) Gardner (2021), Woolridge (2021) and Callaway & Sant'Anna (2021), in particular, provide explicit guidance on multiple ways to aggregate or re-weight their estimators.

Table 1: Comparison of select estimators that are robust to treatment effect heterogeneity under staggered adoption