Statistics in Medicine

WILEY

Statistics in Medicine

REVIEW ARTICLE

Difference-in-Differences for Health Policy and Practice: A Review of Modern Methods

Shuo Feng[1] | Ishani Ganguli[2] | Youjin Lee[1] | John Poe[3] | Andrew Ryan[4] | Alyssa Bilinski[1,4]

[1]Department of Biostatistics, Brown University School of Public Health, Providence, Rhode Island, USA | [2]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA | [3]Hessian LLC, Cabot, Arkansas, USA | [4]Department of Health Services, Policy & Practice, Brown University School of Public Health, Providence, Rhode Island, USA

Correspondence: Alyssa Bilinski (alyssa_bilinski@brown.edu)

Received: 17 July 2024 | **Revised:** 15 August 2025 | **Accepted:** 18 August 2025

Funding: This work was supported by the Centers for Disease Control and Prevention through the Council of State and Territorial Epidemiologists (NU38OT000297-02, S.F. and A.B.), the National Institute of General Medical Sciences (1R35GM155224, S.F. and A.B.), the National Institute of Diabetes and Digestive and Kidney Disease (R01DK136515, Y.L.), and the National Institute on Aging (K23AG068240, I.G. and 2P01AG027296-16, A.B.).

ABSTRACT

Difference-in-differences (DiD) is a popular observational causal inference method in health policy, employed to evaluate the real-world impact of policies and programs. To estimate treatment effects, DiD relies on a "parallel trends assumption" that treatment and comparison groups would have had parallel trajectories on average in the absence of an intervention. Recent years have seen both growing use of DiD in health policy and medicine and rapid advancements in DiD methods. To support DiD implementation in these fields, this paper reviews and synthesizes best practices and recent innovations. We provide recommendations to practitioners in four areas: (1) assessing causal assumptions; (2) adjusting for covariates and other approaches to relax causal assumptions; (3) accounting for staggered treatment timing; and (4) conducting robust inference, especially when normal-based clustered standard errors are inappropriate. For each, we explain challenges and common pitfalls in traditional DiD and recommend methods to address these. We explore current treatment of these topics through a focused literature review of medical DiD studies.

1 | Introduction

Difference-in-differences (DiD) is a popular method for observational causal inference in health policy. DiD evaluates the impact of policies and programs using time-series data by comparing outcome trajectories between treatment and non-experimental comparison groups [1]. Its popularity partly stems from its flexibility and broad applicability: DiD allows for treatment effect estimation even when there exist no comparison groups that are exactly comparable to the treatment groups. It instead relies on the assumption that average differences between treatment and comparison groups would have been stable over time absent intervention (a counterfactual "parallel trends assumption"). Examples of DiD applications in health policy include analyzing the impact of Medicaid expansion [2, 3] and other insurance programs [4–7] on insurance coverage and health, same-sex marriage on mental health [8, 9], closures of automobile assembly plants on opioid overdoses [10], sweetened beverage taxes on soda consumption [11], and restrictions on sales of flavored tobacco products on youth smoking [12].

Growing use of DiD in health policy and medicine represents an important expansion in observational causal inference in these fields. Historically, medical researchers relied primarily

on adjustment for observed confounders (e.g., regression adjustment, stratification, propensity score weighting, or matching) and epidemiological designs (e.g., case control studies) to address selection bias when estimating treatment effects in non-randomized settings. However, in recent years, institutions [13] and journals [14] have expressed increased interest in econometric causal inference methods, including difference-in-differences, that can adjust for some unobserved confounders and, in certain cases, allow for rigorous causal interpretations of results from observational studies. Researchers have likewise expanded the use of DiD designs, with PubMed search results for "difference-in-differences" increasing from 2 in 2000 (0.4 per 100 000 PubMed entries) to 23 in 2010 (2.4/100 000) and 841 in 2024 (48/100 000) [15, 16]. Broader DiD adoption has been accompanied by significant methodological advancements over the past decade, which have refined the approach but rendered its implementation and interpretation more challenging [17–21].

In this paper, we synthesize best practices and recent innovations for medical and health policy researchers. Although previous reviews have summarized DiD methods for economics audiences [17–19], we both adapt key points from prior reviews (see in particular Roth et al. [17]) and develop new material specifically for a medical audience, extending existing tutorials in the health literature [1, 22–23]. Compared to researchers in other fields, those in health policy are diverse in disciplinary backgrounds and often implement DiD with the objective of guiding time-sensitive decisions despite limited data. To increase accessibility to audiences from different disciplines, we include both text descriptions (designed to stand alone) and mathematical notation. We also explore how DiD has been implemented in this literature and emphasize opportunities for future applications.

In the following sections, we begin by providing an overview of DiD assumptions and estimation. We then summarize methods for: (1) evaluation of DiD's causal assumptions [20, 24–27], (2) covariate adjustment and other techniques to relax causal assumptions [28–32], (3) staggered treatment timing [18, 33–37], and (4) inference [38–42]. For each, we detail recent advancements that can address common pitfalls in DiD implementation (Table 1) and describe applications in recent *JAMA* Network papers.

2 | DiD Assumptions and Estimation

DiD requires time-series data for treated and comparison units, with units observed both before and after the start of an intervention. The key identifying assumption in DiD is the "parallel trends assumption," which requires that on average, outcomes for treatment and comparison groups would have had parallel trajectories in the absence of the intervention. DiD also assumes no anticipation (i.e., no effect in treatment groups prior to the intervention) and no spillovers (e.g., no treatment effect in comparison groups) [17, 44]. Combined, these assumptions allow us to estimate what would have occurred in treatment groups, absent intervention (Figure 1, dotted lines), and use this to estimate the average treatment effect on the treated (ATT). In this section, we define assumptions and estimators traditionally used in DiD and preview recent advancements.

2.1 | Traditional DiD Assumptions and Estimation

Traditionally, DiD assumes we have N_1 treated units and N_0 comparison units, and treatment begins for all treated units at the same time period T^*. Let $Y_{i,t}(0)$ and $Y_{i,t}(1)$ denote the untreated and treated potential outcomes of unit i at time period t, respectively. Let D_i be a binary indicator for the treatment status of unit i. The causal quantity of interest is the expected population-level difference between treated and untreated potential outcomes among treated units post-intervention, that is, the average treatment effect on the treated:

$$ATT = \mathbb{E}\left[Y_{i,t}(1) - Y_{i,t}(0) | D_i = 1, t \geq T^*\right]$$

Note that the untreated potential outcome is unobserved for treated units. To identify the ATT, we therefore require three assumptions:

First, Assumption 1 requires that, on average, the untreated potential outcomes in treatment and comparison groups would have followed parallel trajectories.

Assumption 1. (Parallel trends):

$$\begin{aligned}&\mathbb{E}\left[Y_{i,t}(0) | D_i = 1, t \geq T^*\right] - \mathbb{E}\left[Y_{i,t}(0) | D_i = 1, t < T^*\right] \\ &\quad = \mathbb{E}\left[Y_{i,t}(0) | D_i = 0, t \geq T^*\right] - \mathbb{E}\left[Y_{i,t}(0) | D_i = 0, t < T^*\right]\end{aligned}$$

Second, the no-anticipation assumption states that the treatment has no effect on the treated units prior to the intervention.

Assumption 2. (No anticipation):

$$Y_{i,t}(1) = Y_{i,t}(0), \quad i \in \left\{i : D_i = 1\right\}, t < T^*$$

Third, the Stable Unit Treatment Value Assumption (SUTVA) requires that the potential outcomes of one unit do not depend on the treatment status of any other units, which implies no spillover effects.

Assumption 3. (Stable Unit Treatment Value Assumption [SUTVA]): For any time t, the potential outcome of unit i, $Y_{i,t}(d)$, does not depend on the treatment status of any other unit $j \neq i$, for $d \in \{0, 1\}$ with no hidden levels of treatment.

Together, no anticipation and SUTVA allow us to link the potential outcomes to the observed outcomes as follows, which is also sometimes referred to as consistency [44]:

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

Under Assumptions 1–3, we can write the average post-intervention untreated potential outcomes in the treated group as:

$$\begin{aligned}&\mathbb{E}\left[Y_{i,t}(0) | D_i = 1, t \geq T^*\right] \\ &= \mathbb{E}\left[Y_{i,t}(0) | D_i = 1, t < T^*\right] + \mathbb{E}\left[Y_{i,t}(0) | D_i = 1, t \geq T^*\right] \\ &\quad - \mathbb{E}\left[Y_{i,t}(0) | D_i = 1, t < T^*\right]\end{aligned}$$

TABLE 1 | Pitfalls and recommendations for health policy DiD.

	Pitfalls	Recommendations	Literature Review
Evaluating causal assumptions	1. Lack study-specific explanation or contextual evidence for the validity of causal assumptions.	1. Provide context-specific theory when justifying causal assumptions, ideally prior to seeing the outcome data.	Example: One study argued that a delayed increase in COVID-19 cases and deaths following the admission of COVID-19-positive patients was consistent with epidemiologic expectations [43].
	2. Rely on under-powered pre-trend tests (i.e., $p > 0.05$ implies parallel trends).	2. Explore and justify the observed level differences, trend trajectories, and effect timing. Consider non-inferiority tests and event study plots to diagnose violations of causal assumptions.	67% of papers ($n = 28$) used traditional tests. 43% of papers ($n = 18$) reported an event study plot. None discussed non-inferiority tests or calculated test power based on pre-intervention data.
Relaxing causal assumptions	3. Treat all covariates as confounders.	3. Understand what is—and is not—a confounder in DiD. For example, there is no need to adjust for covariates that only affect outcome levels. However, covariates that drive differential trends may be confounders, and researchers may also want to adjust for time-varying sample composition, provided it is unrelated to treatment. Beware covariates that have time-varying values within units (e.g., blood pressure), due to risk of reverse causality and lack of coherent notion of parallel trends.	Among 34 papers that adjusted for covariates, 79% ($n = 27$) directly added covariates into their regression specification, while the others conducted matching or weighting ($n = 5$) and doubly robust ($n = 2$). 82% ($n = 28$) included only baseline covariates, while the other six adjusted for both baseline and time-varying covariates.
	4. Directly add covariates into TWFE regression specification.	4. Be cautious when adding covariates directly into regression specifications. Instead, we encourage the use of a saturated specification (e.g., group-time effects) to allow differential trends by covariate value, or alternatively, consider DiD-specific adjustment techniques such as trajectory modeling and propensity score weighting.	81% of papers ($n = 34$) adjusted for covariates. Of these, 15% ($n = 5$) used recent DiD estimators (e.g., matching, trajectory modeling).
	5. Ignore the impact of potential violations of causal assumptions, leaving the study's conclusions untested for robustness.	5. Consider estimators that bound the impact of causal assumption violations. Use sensitivity analyses to communicate to what extent the conclusions are robust under potential violations of parallel trends, no anticipation, or no spillover assumptions.	None of the papers used a bounding method to explore the impact of causal assumption violations.

TABLE 1 | (Continued)

	Pitfalls	Recommendations	Literature Review
Staggered treatment timing	6. Use static TWFE models when treatment was introduced at different times in different cohorts.	6. When treatment adoption is staggered, use saturated estimators with clean comparison groups, particularly those that estimate group-time effects. Explicitly specify the comparison groups and pay attention to the time periods in which parallel trends are required.	48% of papers ($n = 20$) analyzed an intervention with staggered treatment timings. 50% of these ($n = 10$) used an estimator that explicitly accounted for staggered treatment (e.g., Callaway and Sant'Anna or Sun and Abraham's estimator) in the main analysis.
Inference	7. Fail to cluster standard errors at the level of treatment.	7. Estimate standard errors clustered at the level of treatment assignment. When treatment is assigned at an aggregated level, analyzing individual level data is unlikely to substantially increase power.	48% of papers ($n = 20$) clustered standard errors at the level of treatment assignment.
	8. Conduct inference using normal-based clustered standard errors when the number of treated or untreated clusters is small.	8. Where assumptions for normal-based clustered standard errors may not be met, consider alternative inference methods, particularly the wild cluster bootstrap. Consider alternative inference techniques such as permutation-based methods or conformal inference.	Among the 20 papers that clustered at the level of treatment assignment, 30% ($n = 6$) had either large numbers of clusters or used methods that account for a small number of treatment and/or comparison clusters.

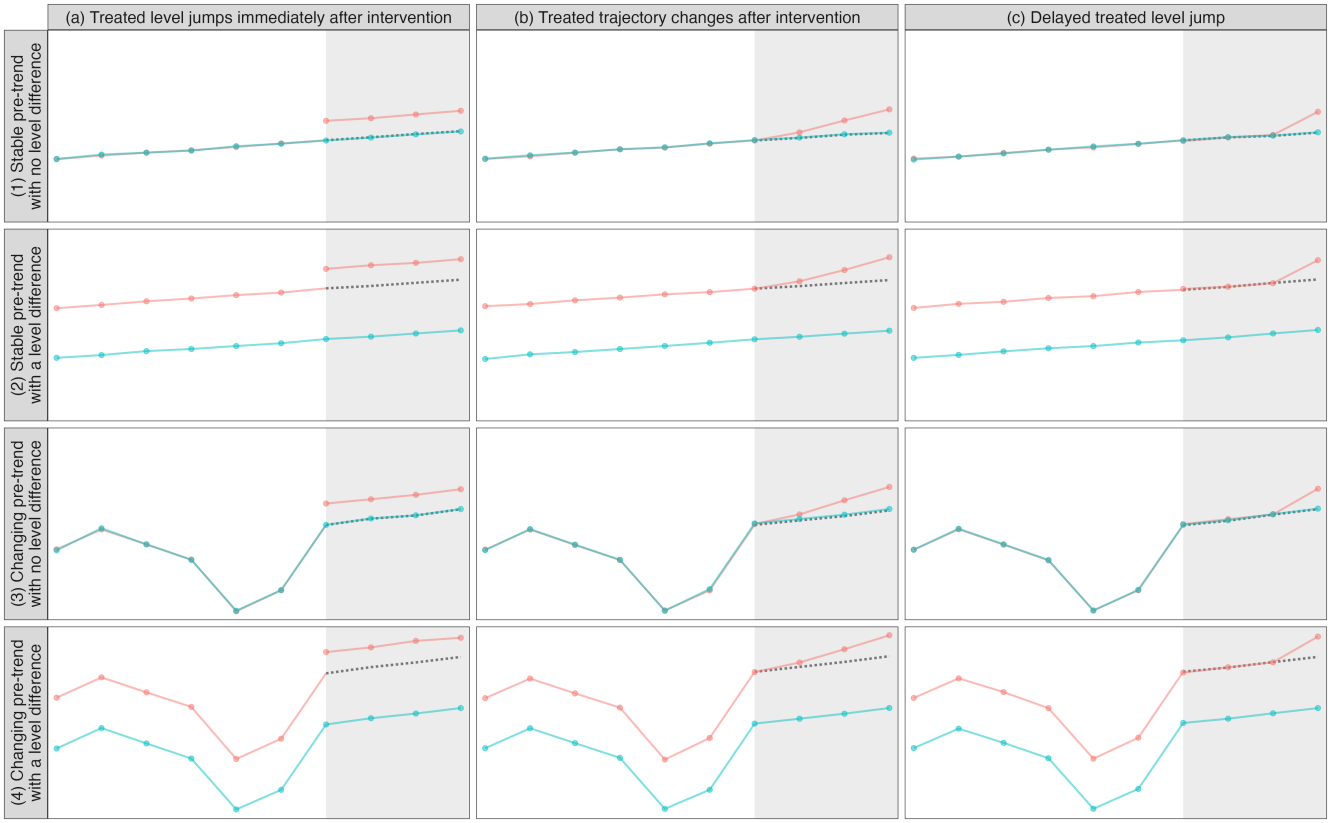

FIGURE 1 | Outcome trajectories. Red and blue solid lines represent the observed outcome trajectories over time in treatment and comparison groups, respectively. The dotted black line indicates untreated potential outcome in treatment groups under a strong parallel trends assumption. The shaded area marks the post-intervention period.

$$
\begin{aligned}
&= \mathbb{E}\left[Y_{i,t}(0)|D_i = 1, t < T^*\right] + \mathbb{E}\left[Y_{i,t}(0)|D_i = 0, t \geq T^*\right] \\
&\quad - \mathbb{E}\left[Y_{i,t}(0)|D_i = 0, t < T^*\right] \qquad \text{(Assumption 1)} \\
&= \mathbb{E}\left[Y_{i,t}(1)|D_i = 1, t < T^*\right] + \mathbb{E}\left[Y_{i,t}(0)|D_i = 0, t \geq T^*\right] \\
&\quad - \mathbb{E}\left[Y_{i,t}(0)|D_i = 0, t < T^*\right] \qquad \text{(Assumption 2)} \\
&= \mathbb{E}\left[Y_{i,t}|D_i = 1, t < T^*\right] + \mathbb{E}\left[Y_{i,t}|D_i = 0, t \geq T^*\right] \\
&\quad - \mathbb{E}\left[Y_{i,t}|D_i = 0, t < T^*\right] \qquad \text{(Assumption 3, Equation 1)}
\end{aligned}
$$

Therefore, the ATT is identified:

$$
\begin{aligned}
ATT &= \mathbb{E}\left[Y_{i,t}(1) - Y_{i,t}(0)|D_i = 1, t \geq T^*\right] \\
&= \mathbb{E}\left[Y_{i,t}|D_i = 1, t \geq T^*\right] - \mathbb{E}\left[Y_{i,t}|D_i = 1, t < T^*\right] \\
&\quad - \left(\mathbb{E}\left[Y_{i,t}|D_i = 0, t \geq T^*\right] - \mathbb{E}\left[Y_{i,t}|D_i = 0, t < T^*\right]\right)
\end{aligned}
$$

In this expression, the ATT is the difference in average treatment and comparison outcomes after versus before the intervention. Although this can be estimated from sample analogs of each term, in this simple setup, DiD is commonly implemented via a two-way fixed effects (TWFE) estimator using ordinary least squares (OLS) regression. Researchers posit the model:

$$Y_{i,t} = \eta_i + \tau_t + \delta D_i \mathbb{I}(t \geq T^*) + \epsilon_{i,t} \tag{2}$$

where $Y_{i,t}$ is the outcome for unit i at time t, for $i = 1, \ldots, N$ (the total number of units) and $t = 1, \ldots, T$ (the total number of time periods). η_i is the fixed-effect for each unit i, τ_t is the fixed-effect for each time period t, and $\epsilon_{i,t}$ is the idiosyncratic error for unit i at time t. We do not specify an intercept term, assuming this is absorbed in the fixed effect terms. Here, δ corresponds to the ATT under the previous three assumptions, and the TWFE estimator $\widehat{\delta}$ can be obtained through OLS corresponding to Equation (2). Importantly, the terms "DiD" and "TWFE" are not interchangeable; the TWFE estimator is only used in a subset of DiD analyses. Last, we note that the following specification, which includes binary indicators for treatment and post-intervention timing statuses rather than unit and time fixed effects, provides the same ATT ($\widehat{\delta}$) when estimated with OLS:

$$Y_{i,t} = \beta_1 D_i + \beta_2 \mathbb{I}(t \geq T^*) + \delta D_i \mathbb{I}(t \geq T^*) + \epsilon_{i,t}$$

2.2 | Alternative Assumptions and Saturated Estimators

The conditions under which the DiD estimator derived above is valid are well-established in the literature. In practice, however, researchers often seek to analyze more complex setups. For example, several authors have developed estimators for cases when treatment started at different calendar times for different units ("staggered adoption") [34, 35, 45–48]. Many new specifications are "saturated," estimating an ATT specific to each post-intervention time period and sometimes to specific units or groups of units. These techniques rely on different versions of the parallel trends assumption, requiring parallel trends across specific time periods and units, rather than over the average of time periods and units as in Assumption 1. For example,

Assumption 4 requires parallel trends over all time periods and adoption cohorts [35], where adoption cohorts are defined as the set of units that initiated treatment at the same calendar time.

Assumption 4. (Strong parallel trends): For every adoption cohort $E_i = e \neq e'$ that initiated treatment at time e and every pair of time periods $t \neq s$ [34]:

$$\mathbb{E}\left[Y_{i,t}(0) - Y_{i,s}(0) | E_i = e\right] = \mathbb{E}\left[Y_{i,t}(0) - Y_{i,s}(0) | E_i = e'\right]$$

It is likewise possible to assume parallel trends relative to specific pre-intervention time periods (e.g., the last pre-intervention period) [34].

With saturated specifications, researchers can estimate a set of treatment effects that can be aggregated into one overall ATT of interest through user-specified weights [34]. For example, we may estimate the following saturated specification:

$$Y_{i,t} = \eta_i + \tau_t + \sum_{e \in \mathscr{E}} \sum_{j=e}^{T} \delta_{e,j} \mathbb{I}\left(E_i = e\right) \mathbb{I}(t = j) + \epsilon_{i,t}$$

In this specification, estimate $\hat{\delta}_{e,j}$ represents the treatment effect at post-intervention time period j in cohort e, that is, $\mathbb{E}\left[Y_{i,j}(1) - Y_{i,j}(0) | E_i = e\right]$, under Assumption 4.

Saturated models, which will be discussed further throughout the text, both prevent several problems that can arise when incorporating staggered treatment timing and time-varying covariates and allow researchers to more intentionally aggregate ATTs across different groups. We will therefore argue that saturated specifications are a useful default.

3 | Literature Review

To explore DiD in the recent medical literature, we surveyed DiD analyses published in three *Journal of American Medical Association* (*JAMA*) network journals between January 2022 and November 2024: *JAMA* ($n = 13$, 31%), *JAMA Internal Medicine* ($n = 16$, 38%), and *JAMA Pediatrics* ($n = 13$, 31%) (see Supporting Information for search criteria). These papers evaluated a range of policies, with interventions related to Medicare or Medicaid as the most popular ($n = 10$, 24%) [49–58]. Other policies included private equity acquisition of hospitals [59, 60], cigarette menthol flavor bans [61, 62], and anti-discrimination legislation [63, 64]. Broadly, these DiDs leveraged the fact that many US health care policies—such as phased insurance expansion [58], elimination of asset tests [65], and antibullying policies [64]—were introduced at the state- or local-level, leaving comparable states or localities untreated. In the following sections, we explore how these papers integrated best practices.

4 | Evaluating Causal Assumptions

4.1 | Recommendation 1a: Provide Context-Specific Theory When Justifying Causal Assumptions

DiD relies on assumptions—parallel trends, no anticipation, and SUTVA—that are not directly testable because we cannot observe what would have happened absent intervention. The parallel trends assumption has been a particular focus of methodological inquiry. It invites the question of why we believe, prior to seeing the data, that outcomes in treatment and comparison groups would have had similar counterfactual trends, even if their levels differed [20]. Recent work has emphasized that researchers should draw on context-specific theory to provide evidence as to why treatment and comparison groups would have been expected to trend in parallel in the absence of an intervention [20, 27], ideally prior to seeing outcome data [17]. This might involve, for instance, highlighting similar exposure to market shocks or similar policy environments across treatment and comparison units.

Where formal theory is well-developed, researchers may be able to translate the parallel trends assumption into domain-specific conditions and use context to help select functional form or scale of the outcome, to which the parallel trends assumption is usually sensitive [17, 32, 66]. For example, if the parallel trends assumption holds in the outcome variable's original levels, it will generally not hold in log-transformed levels (for other transformations, see [67, 68]), unless there was randomization of treatment and/or no differential time trends [32]. Recent work has described epidemiological conditions required for parallel trends to hold with infectious disease outcomes under different outcome transformations and specifications, and proposed estimators for this purpose [27, 69]. Other work has adapted the parallel trends assumption to a survival analysis with time-to-event data, formalizing a version of parallel trends on hazard rates [70, 71].

Last, domain knowledge may help to inform potential violations of spillover and no-anticipation assumptions. For example, if individuals can easily avoid a soft drink tax by shopping in a neighboring jurisdiction, researchers might want to remove these jurisdictions from the set of comparison groups [72]. Similarly, if a policy is announced or enacted in legislation for an extended period prior to implementation, researchers may end the pre-intervention period prior to the time at which individuals likely anticipated its adoption [73–75].

4.2 | Recommendation 1b: Explore and Explain the Observed Level Differences, Trend Trajectories, and Effect Timing

Researchers often use pre-intervention data to help support the plausibility of causal assumptions, especially parallel trends. In theory, assumptions can be satisfied with or without a level difference between treatment and comparison groups, and with or without stable pre-intervention trends (rows of Figure 1). Likewise, there can be different post-intervention patterns in treatment effects, including: (a) level changes immediately post-intervention, (b) trajectory changes immediately post-intervention, or (c) delayed level or trajectory changes (columns of Figure 1). However, DiD designs are often thought to be most compelling when level differences between groups are small before the intervention [20], and there is a sizable change in the treated outcome shortly after the intervention (e.g., Figure 1, column a).

When there are increasing treatment effects (e.g., Figure 1, column b) as an intervention is phased in, researchers

might investigate whether these suggest continuation of a pre-intervention trend or are driven by other post-intervention policies or changes. When treatment effects are delayed (e.g., Figure 1, column c), researchers should explore contextual evidence to support the plausibility of the observed effect timing (e.g., time needed for a drug to become effective) and again assess whether other contemporaneous events may have produced the observed effect. When there is instability in pre-intervention trends (e.g., Figure 1, rows 3 and 4), such as that induced by the COVID-19 pandemic in many outcomes, researchers may need to account for the sensitivity of results to the length of pre-intervention periods, the set of appropriate comparison units, and whether these may suggest anticipation effects. They should also consider whether changes in the comparison group post-intervention may suggest spillover effects.

4.3 | Recommendation 1c: Use Non-Inferiority Tests and Event Study Plots to Diagnose Violations of Causal Assumptions

Researchers frequently employ statistical tests to evaluate evidence in favor of parallel pre-intervention trends. Traditionally, pre-trend tests have been based on a null hypothesis that there was no violation of parallel trends during pre-intervention periods, e.g., no pre-intervention slope difference. However, as several papers have noted [20, 24, 76–77], these tests can be misleading because they may not have adequate statistical power to detect a violation, even if there are non-parallel pre-intervention trend differences between treatment and comparison groups [20, 24, 76–77].

One alternative is to consider non-inferiority tests [24, 78], which specify a null hypothesis that the violation of parallel trends exceeds some threshold (instead of no violation) and proceed with DiD only if there is sufficient evidence to reject the null [17, 24]. A related approach is to report an estimate of the power of pre-trends tests against what the researcher believes to be meaningful violations of parallel trends [76]. However, conducting statistical pre-trend tests and conditioning the analysis on passing these tests may exacerbate bias in the estimated treatment effects because the observed draws of data that pass such tests are a select sample from the true underlying data-generating process [76].

Event study plots (e.g., Figure 2) provide another useful diagnostic for evaluating pre-intervention trends. These show estimated "treatment effects" at each time point, comparing the change in treatment and comparison groups relative to a reference period (e.g., the last pre-intervention time period). The pre-intervention effects can be interpreted as placebo effects that should be small in magnitude with narrow confidence intervals, without evidence of trends or anticipation effects. With a common intervention time for all units, an event study design can be specified:

$$Y_{i,t} = \eta_i + \tau_t + \sum_{j \neq T^* - 1} \delta_j D_i \mathbb{I}(t = j) + \epsilon_{i,t} \quad (3)$$

We encourage the use of event studies to visually assess pre-intervention trends between treatment and comparison groups, explore timing and trends in post-intervention treatment effects, and examine bounds of confidence intervals to rule out unlikely pre-intervention trend patterns. As with tests of slope differences, event study estimates should not be evaluated based on whether confidence intervals overlap zero [24, 76]. Further, when treatment adoption is staggered, the interpretation of event study plots depends on the specification [79], and pre-intervention points may not display secular trends unless they are all constructed relative to the same pre-intervention period (i.e., *base period* = *"universal"* in R or *long2* in Stata).

In summary, because causal assumptions are not directly testable, researchers should justify why they expect that treatment and comparison groups would have had parallel trends absent intervention even if pre-intervention levels differ. Context-specific theory may inform the plausibility of causal assumptions. Researchers may also report results from non-inferiority tests and present event study plots to visualize any pre-existing trends and the pattern of post-intervention evolution in the treated outcomes.

4.4 | Literature Review

Most studies provided little contextual support for causal assumptions or the observed patterns of treatment effects. However, some discussed this; for example, one paper argued that a delay in the impact of admitting COVID-19 patients to skilled nursing facilities on cases and deaths was to be expected, especially for the latter, due to the time needed for transmission and the length of the disease course [43].

Two-thirds of the articles we reviewed ($n = 28$, 67%) implemented traditional pre-trend tests, with a null of no violation of parallel trends. None discussed non-inferiority tests or calculated test power based on pre-intervention data.

Close to half ($n = 18$, 43%) of the studies reported an event study plot. These graphs showed different treatment effect patterns. Many papers [54, 64, 80–83] showed an immediate and persistent jump in outcome levels after the intervention. Others [10, 65] found no immediate level changes but increasing treatment effects over time and/or after a period of treatment adoption, similar to columns (b) and (c) in Figure 2. For instance, one study found an increasing treatment effect of the safer supply policy in opioid prescription over time [84].

5 | Covariate Adjustment and Relaxing Causal Assumptions

In some applications, DiD's causal assumptions can seem implausible, and researchers may seek to adapt DiD to address potential confounding. In this section, we discuss approaches for relaxing causal assumptions. We first describe how researchers may posit parallel trends only conditional on covariates and adjust estimation accordingly [85–89]. We then describe methods that derive bounds on bias induced by violations of parallel trends and/or other causal assumptions [90–92].

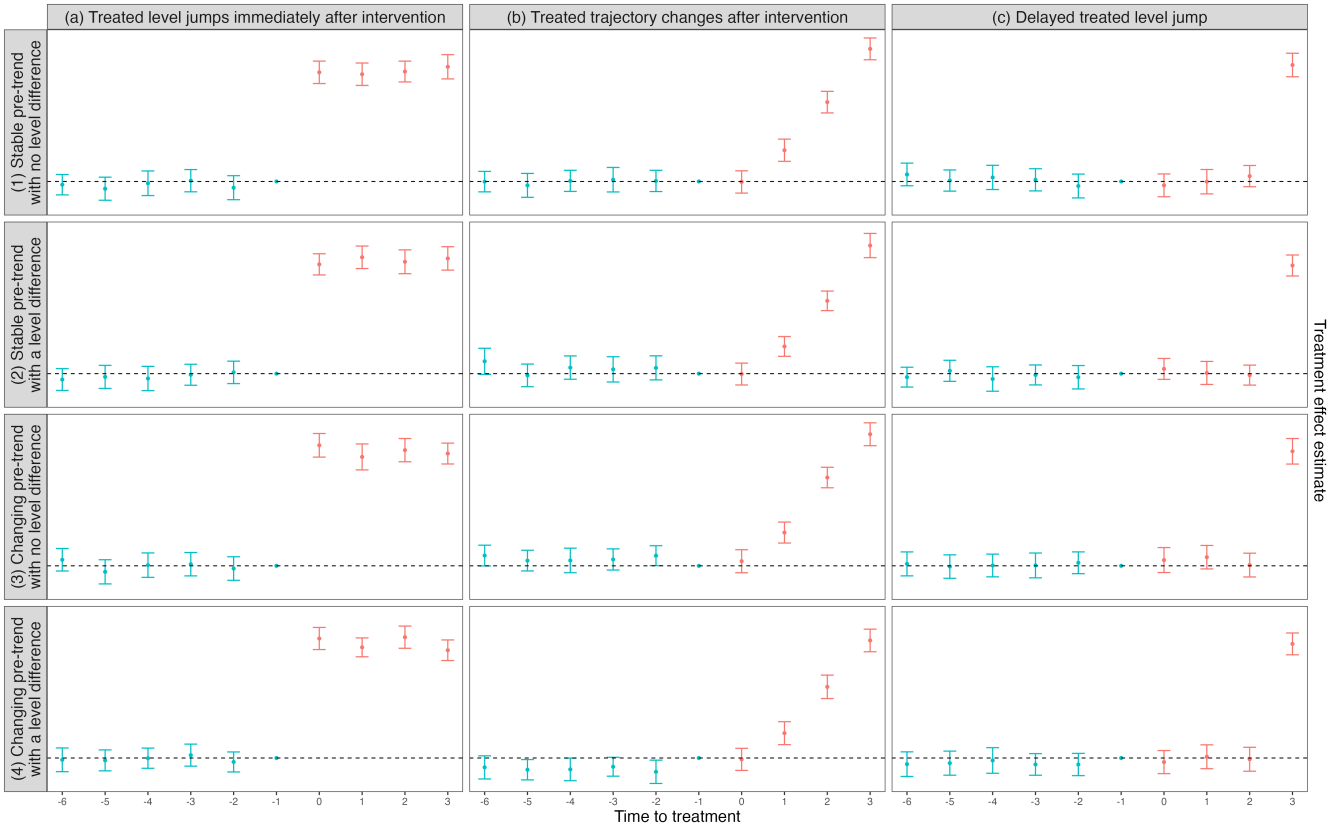

FIGURE 2 | Event study plots under different scenarios. The event study plot in each panel corresponds to the outcome trajectories of the corresponding panel in Figure 1. The dotted horizontal line represents 0, that is, no treatment effect. The vertical error bars represent 95% confidence intervals. Treatment effect estimates were generated according to Equation (3).

5.1 | Recommendation 2a: Understand What Is—And Is Not—A Confounder in DiD

In many observational epidemiology studies, researchers evaluate the relationship between an outcome and a treatment of interest, adjusting for covariates. For example, they may estimate a regression corresponding to the following model:

$$Y_i = \delta D_i + \boldsymbol{\beta} \boldsymbol{X}_i + \epsilon_i$$

where δ is the treatment effect of interest, $\boldsymbol{\beta}$ is a row vector of coefficients, and $\boldsymbol{X}_i$ is a column vector of observed covariates for unit i. In this case, any variable that affects both treatment status and the outcome can be a confounder, and failing to adjust for it may bias treatment effect estimates.

By contrast, in DiD, covariates that only produce time-invariant level differences between treatment and comparison outcomes are not confounders [89], and excluding them does not bias ATT estimates (Table 2, row 1). However, covariates that drive differential trends may confound treatment effect estimates. We argue that researchers should focus on adjusting for confounders measured pre-intervention for which they believe a "conditional parallel trends assumption" (Assumption 5) holds. This implies parallel trends within populations in covariate-defined strata (e.g., parallel trends within smokers and non-smokers): [17, 34]

Assumption 5. (Conditional parallel trends): For $t_1 < T^*$, $t_2 \geq T^*$, and a vector of covariates $\boldsymbol{X}_i$,

$$\mathbb{E}\left[Y_{i,t_2}(0) - Y_{i,t_1}(0) | D_i = 1, \boldsymbol{X}_i\right] = \mathbb{E}\left[Y_{i,t_2}(0) - Y_{i,t_1}(0) | D_i = 0, \boldsymbol{X}_i\right]$$

This translates to two scenarios that merit adjustment. First, suppose researchers have cross-sectional data, meaning that different individuals are sampled at each observed time point. In this case, researchers may need to account for covariates that shift an outcome's level, even if they do not affect an individual's trajectory (Table 2, row 2, Case 1) because changing composition may drive differential trends. Second, in either cross-sectional or longitudinal/panel data (in which the same individuals are sampled at each time point), researchers may adjust for covariates that shift the outcome trend (Table 2, row 3, Case 2).

We urge caution when adjusting for confounders that themselves vary over time (e.g., blood pressure) for two reasons (Table 2, row 4) [93]. First, this invites a risk of adjusting for variables on the causal pathway if the confounder can itself be affected by treatment. Second, conditioning the parallel trends assumption on a time-varying variable makes it difficult to conceptualize a cohesive notion of underlying parallel trends—if trends are only parallel conditional on a time-varying variable, they may not be, in fact, believed to be parallel in any well-defined subgroup. As a result, in this case, other study designs may be more appropriate than DiD.

TABLE 2 | Summary of DiD confounding.

	Narrative		How does parallel trends assumption hold?	Longitudinal (panel) vs. cross-sectional data • Longitudinal: The same group of individuals is followed over time • Cross-sectional: Different individuals are sampled for treatment and comparison groups over time	Baseline covariate vs. repeated measures • Baseline: The covariate takes one value per individual and can be measured at baseline • Repeated measures: The covariate can take different values per individual	Covariate effect	Recommended adjustment method(s)	Impact of naïve adjustment
	Description	**Example**	**Unconditionally, conditionally, or else?**			**Does the covariate shift the outcome's level or its trajectory?**	**What adjustment (if any) is necessary to produce an unbiased treatment effect estimate?**	**What are the impacts of naïve regression adjustment on the treatment effect estimate?**
Longitudinal data, level differences	The same group of individuals is followed over time in treatment and comparison groups. Baseline covariates are measured once and only shift the outcome's level.	A study follows the same individuals in treatment and comparison groups over time. Covariates may be the individual's inherent characteristics (e.g., birth year) that only shift the outcome's level.	Unconditionally	Longitudinal	Baseline	Level	Not needed	No change in treatment effect (no bias).
Cross-sectional data, level differences	Different individuals are sampled for treatment and comparison groups at different time points. Baseline covariates are measured once and only shift the outcome's level.	A survey is conducted every 6 months and sample different participants. Population-level characteristics (e.g., proportion of individuals in each age group) vary between rounds of the survey. However, at the respondent level, covariates values do not vary over time.	Conditional on the baseline covariates	Cross-sectional	Baseline	Level	• Saturated model • Trajectory modeling[a] ◦ Matching • Inverse probability weighting • Doubly-robust estimators	Will recover an unbiased treatment effect in saturated specifications if covariate functional form is correctly specified.

TABLE 2 | (Continued)

	Narrative		How does parallel trends assumption hold?	Longitudinal (panel) vs. cross-sectional data • **Longitudinal: The same group of individuals is followed over time** • **Cross-sectional: Different individuals are sampled for treatment and comparison groups over time**	Baseline covariate vs. repeated measures • **Baseline: The covariate takes one value per individual and can be measured at baseline** • **Repeated measures: The covariate can take different values per individual**	Covariate effect	Recommended adjustment method(s)	Impact of naïve adjustment
	Description	**Example**	**Unconditionally, conditionally, or else?**			**Does the covariate shift the outcome's level or its trajectory?**	**What adjustment (if any) is necessary to produce an unbiased treatment effect estimate?**	**What are the impacts of naïve regression adjustment on the treatment effect estimate?**
Trajectory differences	Baseline covariates take only one value per individual or unit but shift the outcome's trajectory.	There are different outcome trajectories among rural and urban respondents in the absence of an intervention, and the comparison group has a greater proportion of urban enrollees.	Conditional on the baseline covariates	Either	Baseline	Trajectory	• Trajectory modeling[a] ◦ Matching • Inverse probability weighting • Doubly-robust estimators	May not recover an unbiased treatment effect, even with saturated specifications.
Repeated measures	Covariates are measured at multiple times. Their values vary across different measurements.	A clinical trial where the covariates include repeated measurements of the patient's vital signs such as blood pressure. These values affect the outcome.	Not recommended: it is difficult to envision how parallel trends would hold conditioning on repeated measures	Either	Repeated measures	Either	DiD may not be a reliable method due to: • risk of reverse causality • lack of cohesive notion of parallel trends	May not recover an unbiased treatment effect, even with saturated specifications.

[a] Regression adjustment in the literature.

5.2 | Recommendation 2b: Be Cautious When Adding Covariates Directly Into Regression Specifications

Mirroring methods used in other observational studies, many DiD studies add covariates directly into the TWFE specification (Equation 2) to adjust for confounders [17, 89]. However, this approach may not be appropriate. In Case 1 (cross-sectional data, covariates affect outcome's level), although this will reduce bias, adding time-varying covariates into TWFE, rather than a saturated specification, will fit covariate coefficients in part based on time heterogeneity in treatment effects [24, 94]. Therefore, if adding covariates directly into a regression specification, researchers should adopt a saturated model, estimating effects at each time period (Table 2, row 2):

$$Y_{i,t} = \eta_i + \tau_t + \sum_{j \geq T^*} \delta_j D_i \mathbb{I}(t = j) + \boldsymbol{\beta} \boldsymbol{X}_i + \epsilon_{i,t}$$

In **Case 2** (cross-sectional or longitudinal data, covariates affect outcome's trajectory), bias may persist after adding covariates to a TWFE specification both because of the above issue and because naïvely adding covariates typically does not allow *trends* to differ by covariate value (Table 2, row 3). To address this, one approach could involve adding a group-covariate interaction to a saturated model encoding some time-varying functional form: [45]

$$Y_{i,t} = \eta_i + \tau_t + \sum_{j \geq T^*} \delta_j D_i \mathbb{I}(t = j) + \boldsymbol{\beta} \boldsymbol{X}_{i,t} D_i + \epsilon_{i,t}$$

For example, allowing for differential linear time trends would correspond to setting $X_{i,t} = tX_i$.

5.3 | Recommendation 2c: Apply Adjustment Techniques That Allow Parallel Trends Conditional on Covariates

Several other techniques have been proposed to allow for unbiased or consistent treatment effect estimation assuming conditional parallel trends (Assumption 5) with potentially fewer parametric assumptions. These fall into two main categories, which we denote trajectory modeling [28, 29, 86, 88, 95] and propensity score weighting [17, 30].

Techniques in the first category involve modeling subgroup- or covariate-specific trends, using these to estimate subgroup effects, and combining them into an ATT [28, 29, 86, 89]. To do this, researchers would model the evolution in untreated outcomes from pre- to post-intervention in the comparison group, conditional on covariates, and use this fitted model to project untreated outcome evolution in the treated group. That is, we first fit a conditional expectation function (e.g., $\hat{\mathbb{E}}\left[Y_{i,t_2} - Y_{i,t_1} | D_i = 0, X_i\right]$ [17], assuming longitudinal data) using only data from the comparison group, and evaluate it given covariate distribution among treated units. We can then use this to generate an overall ATT estimate: [17, 28]

$$\widehat{ATT}_{\text{trajectory}} = \frac{1}{N_1} \sum_{i:D_i=1} \left[\left(Y_{i,t_2} - Y_{i,t_1}\right) - \hat{\mathbb{E}}\left[Y_{i,t_2} - Y_{i,t_1} | D_i = 0, \boldsymbol{X}_i\right] \right]$$

Trajectory modeling produces consistent estimates if covariate-specific conditional expectation functions are correctly specified. Matching is a special form of such trajectory modeling, in which researchers define subgroup-specific trajectories based only on comparison units with similar (or identical) baseline characteristics and compare these to their treated counterparts (i.e., DiD on a "matched" subsample of data) [86, 88, 95].

Alternatively, researchers can employ inverse probability weighting (IPW) to recover the ATT assuming conditional parallel trends. This involves fitting a propensity score model, $\hat{p}(X_i)$, that predicts treatment status from covariates and using the estimated propensity scores to define weights that account for selection into treatment [17, 30]. With longitudinal data, the ATT is then identified by weighting each unit's observed outcome change using its propensity scores: [30]

$$ATT_{\text{IPW}} = \frac{\mathbb{E}\left[\left(D_i - \frac{(1-D_i)p(\boldsymbol{X}_i)}{1-p(\boldsymbol{X}_i)}\right)\left(Y_{i,t_2} - Y_{i,t_1}\right)\right]}{\mathbb{E}\left[D_i\right]}$$

This approach provides a consistent estimator for the ATT if the propensity score model is correctly specified.

Doubly robust (DR) estimators combine trajectory modeling and propensity score estimation and produce consistent estimates if either the trajectory model or the propensity score model is correctly specified: [29]

$$ATT_{\text{DR}} = \mathbb{E}\left[\left(\frac{D_i}{\mathbb{E}\left[D_i\right]} - \frac{\frac{(1-D_i)p(\boldsymbol{X}_i)}{1-p(\boldsymbol{X}_i)}}{\mathbb{E}\left[\frac{(1-D_i)p(\boldsymbol{X}_i)}{1-p(\boldsymbol{X}_i)}\right]}\right)\left(\left(Y_{i,t_2} - Y_{i,t_1}\right) - \mathbb{E}\left[Y_{i,t_2} - Y_{i,t_1} | D_i = 0, \boldsymbol{X}_i\right]\right)\right]$$

Other innovations in confounding adjustment methods include Bayesian approaches for trajectory modeling [96, 97], machine learning to estimate weights for IPW and doubly robust estimators [98–100], and IPW approaches driven by latent variables or marginal structure models to facilitate a better understanding of selection into treatment [100–102]. Under different assumptions, these techniques may allow for treatment effect estimation assuming conditional parallel trends.

5.4 | Recommendation 2d: Consider Estimators That Bound the Impact of Causal Assumption Violations

Beyond methods that directly model potential confounders, researchers may use sensitivity analyses to quantify how sensitive substantive conclusions are to violations of causal assumptions. As above, most work in this area has focused on sensitivity analyses for parallel trends: for example, Rambachan and Roth [91] proposed an approach to bound how large violations of parallel trends must be to meaningfully shift effect estimates and developed methods to construct valid confidence intervals for the treatment effect under certain violations. Keele et al. [103] likewise

developed a sensitivity analysis method based on matching to assess how strong unobserved confounders must be to alter or reverse their conclusions. Ye et al. [92] proposed bounding bias by choosing two comparison groups with offsetting biases, assuming that the treated group's counterfactual path would have been sandwiched or "bracketed" between those of the chosen comparison groups. DiD estimates based on these two groups then provide lower and upper bounds on the treatment effect. These approaches allow researchers to evaluate and communicate the robustness of conclusions across reasonable assumptions even lacking data on potential confounders (see examples [104, 105]).

Sensitivity analyses can also be helpful when researchers suspect violations of the no anticipation assumption. Although the common practice is to end the pre-intervention period early before any impact of anticipation effects could happen [75], researchers may consider methods proposed by Chen et al. [73] to bound violation impact under different assumptions about the magnitude of anticipation effects.

In summary, when faced with violations of parallel trends, researchers might carefully consider potential confounders and use trajectory modeling, propensity score weighting, or combined doubly robust estimators to account for them (Table 2, rows 2 and 3). Sensitivity analyses like bounding can be helpful to quantify the robustness of substantive conclusions when causal assumptions are violated.

5.5 | Literature Review

Among papers we reviewed, most (34 of 42, 81%) adjusted for covariates in their analyses, with 27 (79%) directly adding the covariates into regression specifications. Among them, nine studies (33%) employed a saturated model. Five studies (15%) conducted matching or weighting based on the covariates, and two used a doubly robust estimator. None of the papers used a bounding method to explore the impact of parallel trends violations. Twenty-eight out of these 34 studies that adjusted for covariates (82%) included only baseline covariates in their model specifications, while the other six (18%) adjusted for both baseline and time-varying covariates.

6 | DiD With Staggered Treatment Timing

Researchers frequently apply DiD to settings in which units initiate treatment at different calendar times. In this case, conducting DiD with a static TWFE specification (Equation 2) can lead to a biased ATT estimate unless treatment effects are homogeneous both over time and across all treatment adoption cohorts (Table 3, row 1) [17]. This occurs because the treatment effect estimate produced by TWFE is a weighted average of possible 2×2 comparisons over all groups and treatment times, including using early treated units as comparison units after treatment [36]. When the treatment effect is changing over time, this can even flip the sign of the treatment effect [17, 34, 36]. Furthermore, these 2×2 comparisons are combined into a treatment effect estimate based on weights designed to minimize variance. When treatment effects differ across adoption cohorts, this treatment effect estimand may no longer reflect the ATT of interest.

6.1 | Recommendation 3: When Treatment Adoption Is Staggered, Use Saturated Estimators With Clean Comparison Groups, Particularly Those That Estimate Group-Time Effects

Many saturated models have been developed to address issues related to staggered treatment adoption [34, 35, 45–48]. These models can invoke different versions of the parallel trends assumption, and researchers often need to specify the time periods and units over which the parallel trends assumption is believed to hold. In this section, we summarize three common techniques, in the order of increasing flexibility: dynamic TWFE, stacked DiD, and group-time estimators, and describe the conditions under which each of them performs well.

6.1.1 | Dynamic TWFE

One simple extension to the static TWFE specification (Equation 2) is the dynamic TWFE specification, which partially addresses these problems. In dynamic TWFE, the single treatment status indicator is replaced by multiple indicators, estimating a treatment effect for each post-intervention time period relative to treatment initiation, where time since treatment for unit i and time t is denoted $R_{i,t} = t - E_i$:

$$Y_{i,t} = \eta_i + \tau_t + \sum_{r \geq 0} \delta_r D_i \mathbb{I}\left(R_{i,t} = r\right) + \epsilon_{i,t}$$

The dynamic TWFE specification yields a sensible estimand for the ATT if heterogeneity only exists in the number of time periods since treatment, but it still requires the sequence of treatment effects to be the same across all cohorts regardless of calendar adoption time (Table 3, row 2) [35, 46]. For example, dynamic TWFE allows treatment effects to be different in year 2 compared to year 3 following treatment adoption, but the year 2 effect is assumed to be the same across all adoption cohorts [17]. This may not be plausible if early adopters had a different experience of the policy (e.g., a pilot phase) or if there was selection into treatment.

6.1.2 | Stacked DiD

Nevertheless, there are other methods that allow for treatment effect heterogeneity both across cohorts and over time. For example, in stacked regression, researchers can create a new dataset by stacking each treated cohort with the group of comparison units that were not yet treated over a pre-selected time horizon (e.g., four time periods before and two time periods after the treated group received the treatment) [48]. This approach allows both time and cohort heterogeneity, requiring parallel trends over only the horizon of time that the researcher specifies. The overall ATT is implicitly a weighted average of treatment effects in each cohort, with weights reflecing the variance of each cohort-specific treatment effect [47].

6.1.3 | Group-Time Estimators

For further flexibility, researchers may generate a treatment effect estimate for each cohort at each post-intervention time (often called a "cohort-time" or "group-time" treatment effect [34, 35, 45]) and aggregate as desired. There is growing consensus that

TABLE 3 | Heterogeneous treatment effects and staggered treatment timing estimators.

	Narrative		Is static TWFE biased?	Heterogenous treatment effects by time since treatment	Heterogenous treatment effects by cohort	Comparison group	
Label	**Description**	**Example**	**Yes/no**	**Are treatment effects different for different lengths of time since treatment?**	**Are treatment effects different across different cohorts?**	**What is typically used as the comparison group?**	**Recommended estimators**
Homogeneous treatment effects across units and time periods	Treatment effects are the same over time and across cohorts.	The effect of insurance expansion is the same regardless of county or how long ago the county implemented the policy.	No	No	No	Weighted average of both not-yet-treated and any never-treated groups	• Static TWFE
Heterogeneous treatment effects across time periods only	Treatment effects can differ according to length of time since treatment but must be the same across adoption cohorts.	The effect of insurance expansion is stronger in the counties that have implemented the policy for a longer time, but it is the same for all counties at 1 month post-intervention despite the timing of adoption.	Yes	Yes	No	Weighted average of both not-yet-treated and any never-treated groups	• Dynamic TWFE
Heterogenous treatment effects across units and time periods	Treatment effects can vary across different cohorts and over time.	The effect of insurance expansion can depend on both the timing of adoption or how long the counties have implemented the policy. Either not-yet-treated counties, never-treated counties, or a combination of both may be used as the comparison group.	Yes	Yes	Yes	Varies	• Imputation method by Borusyak, Jaravel, and Spiess [46] (comparison group: average of all never-treated and not-yet treated) • Stacked regression by Cengiz et al. [48] (comparison group: never-treated or not-yet-treated) • Sun and Abraham [35] group-time estimator (comparison group: never-treated or last-treated) • Callaway and Sant'Anna [34] group-time estimator (comparison group: never-treated or not yet treated)

researchers should default to these saturated models, including estimators proposed by Callaway and Sant'Anna, Sun and Abraham, and Borusyak, Jaravel, and Spiess (Table 3, row 3) [34, 35, 46, 106].

To estimate a group-time average treatment effect

$$ATT(e,t) = \mathbb{E}\left[Y_{i,t}(1) - Y_{i,t}(0) | E_i = e\right]$$

Callaway and Sant'Anna [34] proposed to estimate a treatment effect for each cohort e at each post-intervention time period t using only the subset of data containing time periods t and $e-1$ for units in adoption cohort e and those never-treated (or not-yet-treated at t):

$$Y_{i,t} = \eta_i + \tau_t + \delta_{i,t}\mathbb{I}(E_i = e) + \epsilon_{i,t}$$

They then described several approaches to aggregate individual group-time average treatment effects into an overall ATT estimate [34, 106].

Other group-time estimators differ in the choice of comparison groups and pre-intervention time periods [17, 107–108]. The Callaway and Sant'Anna [34] estimator uses only the last pre-treatment period for reference, which invokes a weaker version of the parallel trends assumption (i.e., assuming that counterfactual trends would have been parallel between the last pre-intervention and the post-intervention periods) but may have correspondingly lower power [108] than estimators that use all pre-intervention periods as reference (e.g., [46]). In practice, software to implement these estimators also often allows for user customization. For example, the Sun and Abraham [35] estimator, motivated by strong parallel trends (Assumption 4), can be implemented with software (R fixest) that allows users to pick one or multiple reference periods and either never-treated or last-treated units as the comparison group [109].

Saturated group-time estimators provide researchers with the flexibility to aggregate treatment effects meaningfully into an overall ATT, as specified by the user, for example, larger weights in the first few post-treatment periods if the aim is to evaluate the short-term effect of an intervention. Overall, these alternatives provide well-defined causal parameters with transparent weights over the 2×2 comparisons and avoid using treated units as comparison groups [17]. In a large empirical reanalysis [21], estimators that account for staggered treatment timing were found to sometimes produce substantially smaller treatment effect estimates. These estimators also often produced larger standard errors and would require more power to reject the null hypothesis than TWFE [21].

6.2 | Literature Review

In our literature review, 20 (48%) of the 42 studies had a staggered treatment design; however, about half of them ($n=9$, 45%) used a static TWFE model as the main estimator. One study did not clearly specify its estimator. Of the remaining studies that implemented estimators accounting for the staggered treatment ($n=10$, 50% of all studies with staggered treatment), Callaway and Sant'Anna ($n=3$, 30%) and stacked DiD ($n=3$, 30%) were the most popular methods. The remaining adjustment methods included dynamic TWFE ($n=2$), two-stage DiD proposed by Gardner [47] ($n=1$), and the d'Chaisemartin and D'Haultfoeuille estimator [33] ($n=1$).

7 | DiD With Robust Inference

7.1 | Recommendation 4: Where Assumptions for Normal-Based Clustered Standard Errors May Not Be Met, Consider Alternative Inference Methods, Particularly the Wild Cluster Bootstrap

In DiD studies, treatment generally occurs at an aggregate level (e.g., state), but data may be collected either at the same aggregated or a more granular level. For example, we may have patient-level data, while the intervention was implemented at the hospital level, or we may have county-level data while the policy was implemented by state. In all scenarios, it is recommended to estimate the standard errors clustered at the level of treatment assignment [17, 110–111]. When treatment is assigned at an aggregate level, analyzing individual-level data is unlikely to substantially increase power.

Table 4 summarizes inference methods for DiD studies. Normal-based clustered robust inference (e.g., *cluster robust* in Stata) is one of the most popular methods for clustering standard errors. While normal-based inference accounts for error correlation within clusters, it requires a substantial number of both treated and untreated clusters (e.g., at least 25 or 30 clusters in each class), which may not exist in many DiD applications (Table 4, row 1) [17, 114]. If there are only a few clusters in the data, normal-based clustered standard errors are typically anti-conservative [114, 115], leading to inappropriately small p values and an inflated Type I error rate.

Several alternative methods have been proposed to conduct inference with a small number of clusters. One popular approach is the wild cluster bootstrap (or wild score bootstrap for generalized linear models) [41]. In each bootstrap replicate of these methods, we multiply the estimated residuals (or scores) in each cluster by a draw of an independent random variable with mean 0 and variance 1 [39, 116]. Simulations suggest that these methods can perform well with about five clusters, although estimates may be conservative when the proportion of treated units is small (Table 4, row 2) [39, 114, 116]. However, despite strong simulation performance, the wild cluster bootstrap technically invokes an assumption of a balanced proportion of treated units among clusters, which may not be plausible in many DiD applications where the treatment is assigned at the cluster level (see Canay et al. [117]).

Other methods rely on different assumptions (Table 4, rows 3 and 4) [115]. Conley and Taber proposed an approach to estimate the distribution of errors in the treated clusters using that from the untreated clusters [40]. This approach may be powerful when there are only a few treated clusters but many untreated ones from which to estimate the error distribution. However, it requires that errors in the treated units have the same distribution as those in the untreated units, which may be violated under treatment effect heterogeneity. Alternatively, conformal

TABLE 4 | Summary of DiD inference methods.

	Examples of popular implementation	Data notes: Consider • Number of clusters • Size of each cluster • Number of time periods	Heterogeneous treatment effects: Does method perform well when different clusters have different treatment effects?	Other
Normal-based clustered standard error	• Sandwich estimators for standard errors: ◦ "vcovCL(cluster = clustvar)" in R ◦ "vce(cluster clustvar)" in Stata	• Requires large number of treated and untreated clusters	Yes	• Anti-conservative with few treated or comparison units
Wild bootstrap	• Wild cluster bootstrap (OLS) [41] • Wild score bootstrap (GLM) [41]	• Can work well with about five large clusters • Assumes the proportion of treated clusters is not small (less sensitive in some simulations)	Yes	• Estimates will be conservative if the proportion of treated clusters is too small
Permutation	• Hageman (2022) [112]: bound on the maximal tolerance of heterogeneity between clusters		• Untreated potential outcomes cannot have any cluster-specific heterogeneity	• Assumes the untreated potential outcomes for treated clusters could be inferred using any single untreated cluster, as cluster size grows large
	• Fisher randomization test (FRT) [113]		Yes	• Assumes random treatment assignment and exchangeability between treated and untreated clusters • Tests sharp null: no effect for any units
Other methods	• Conley and Taber (2011) [40]: learn the distribution of errors in treated clusters using that from untreated clusters	• Requires large number of untreated clusters	• Yes, but heterogeneity in treatment effects often violates the assumption that errors in treated clusters share the same distribution as those in untreated clusters	• Assumes treated and untreated clusters have the same distribution of errors
	• Chernozhukov, Wüthrich, and Zhu (2021) [42]: conformal inference	• Requires large number of time periods	• Yes, provided that the proportion of treated units in each cluster remains the same over time	• Assumes strong parallel trends assumption over many time periods

inference can perform well when there are many time periods available. However, it assumes that the proportion of treated units in each cluster to vary little over time and imposes a strong parallel trends assumption over all pre- and post-treatment periods [42].

Taken together, researchers should cluster standard errors at the level of treatment assignment and be wary of normal-based cluster inference if there are insufficient numbers of treatment or comparison clusters and apply alternatives.

7.2 | Literature Review

Among papers reviewed, normal-based standard errors were by far the most common method ($n = 40$, 95%) for conducting inference. Two papers implemented bootstrap algorithms for inference, one with traditional bootstrap [118] and another one with wild cluster bootstrap [119]. 48% of studies ($n = 20$) clustered standard errors at the level of treatment assignment. Among them, 30% ($n = 6$) had either large numbers of clusters or used methods that account for a small number of treatment and/or comparison clusters. Five papers (12%) studied an intervention that was assigned at an individual patient or physician level.

8 | Conclusion

DiD is one of the most popular observational causal inference tools in health policy and medical research. It can allow for rigorous causal interpretations in certain observational studies. When an intervention is introduced first in a few cities or states, DiD can support evaluation and learning from early adopters before scaling to additional locations.

This paper identifies the key challenges in DiD studies and compiles recommendations that have been proposed in recent literature to support analyses of US health care policies. We argue that researchers should start by evaluating causal assumptions based on context, visualizing data in plots, using statistical tools with adequate power, and selecting an outcome functional form guided by domain-specific knowledge. They can also relax assumptions, for example by assuming parallel trends only in subgroups of the data or conditional on covariates. When treated units initiate treatment at different times, we recommend that researchers default to group-time treatment effect estimation methods that perform well with heterogeneous treatment effects across time and cohorts. Finally, normal-based clustered inference requires both many treated and many untreated clusters. When this is not the case, researchers should opt for alternatives that require weaker assumptions.

Overall, we are optimistic that these recent innovations can strengthen DiD practice and associated policy recommendations.

Acknowledgments

The authors gratefully acknowledge feedback from Jason Buxbaum, Travis Donahoe, Jonathan Roth, and Pedro Sant'Anna. This work was supported by the Centers for Disease Control and Prevention through the Council of State and Territorial Epidemiologists (NU38OT000297-02, S.F. and A.B.), the National Institute of General Medical Sciences (1R35GM155224, S.F. and A.B.), the National Institute of Diabetes and Digestive and Kidney Disease (R01DK136515, Y.L.), and the National Institute on Aging (K23AG068240, I.G. and 2P01AG027296-16, A.B.).

Conflicts of Interest

Ishani Ganguli is an Associate Editor at JAMA Internal Medicine. The other authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. E. C. Caniglia and E. J. Murray, "Difference-in-Difference in the Time of Cholera: A Gentle Introduction for Epidemiologists," *Current Epidemiology Reports* 7, no. 4 (2020): 203–211.

2. S. Miller and L. R. Wherry, "Health and Access to Care During the First 2 Years of the ACA Medicaid Expansions," *New England Journal of Medicine* 376, no. 10 (2017): 947–956.

3. S. A. M. Khatana, A. Bhatla, A. S. Nathan, et al., "Association of Medicaid Expansion With Cardiovascular Mortality," *JAMA Cardiology* 4, no. 7 (2019): 671–679.

4. A. Mulcahy, K. Harris, K. Finegold, A. Kellermann, L. Edelman, and B. D. Sommers, "Insurance Coverage of Emergency Care for Young Adults Under Health Reform," *New England Journal of Medicine* 368, no. 22 (2013): 2105–2112.

5. J. M. McWilliams, L. A. Hatfield, B. E. Landon, P. Hamed, and M. E. Chernew, "Medicare Spending After 3 Years of the Medicare Shared Savings Program," *New England Journal of Medicine* 379, no. 12 (2018): 1139–1149.

6. K. E. Joynt Maddox, E. J. Orav, J. Zheng, and A. M. Epstein, "Evaluation of Medicare's Bundled Payments Initiative for Medical Conditions," *New England Journal of Medicine* 379, no. 3 (2018): 260–269.

7. M. L. Barnett, A. Wilcock, J. M. McWilliams, et al., "Two-Year Evaluation of Mandatory Bundled Payments for Joint Replacement," *New England Journal of Medicine* 380, no. 3 (2019): 252–262.

8. J. Raifman, E. Moscoe, S. B. Austin, and M. McConnell, "Difference-In-Differences Analysis of the Association Between State Same-Sex Marriage Policies and Adolescent Suicide Attempts," *JAMA Pediatrics* 171, no. 4 (2017): 350–356.

9. J. Raifman, E. Moscoe, S. B. Austin, M. L. Hatzenbuehler, and S. Galea, "Association of State Laws Permitting Denial of Services to Same-Sex Couples With Mental Distress in Sexual Minority Adults: A Difference-In-Difference-in-Differences Analysis," *JAMA Psychiatry* 75, no. 7 (2018): 671–677.

10. A. S. Venkataramani, E. F. Bair, R. L. O'Brien, and A. C. Tsai, "Association Between Automotive Assembly Plant Closures and Opioid Overdose Mortality in the United States: A Difference-In-Differences Analysis," *JAMA Internal Medicine* 180, no. 2 (2020): 254–262.

11. J. Cawley, D. Frisvold, A. Hill, and D. Jones, "The Impact of the Philadelphia Beverage Tax on Purchases and Consumption by Adults and Children," *Journal of Health Economics* 67 (2019): 102225.

12. A. S. Friedman, "A Difference-in-Differences Analysis of Youth Smoking and a Ban on Sales of Flavored Tobacco Products in San Francisco, California," *JAMA Pediatrics* 175, no. 8 (2021): 863–865.

13. World Health Organization, "Evaluating the Impact of Interventions Addressing Health Behaviour: Considerations and Tools for Policy-Makers," in *Evaluating the Impact of Interventions Addressing Health Behaviour: Considerations and Tools for Policy-Makers* (2024).

14. I. J. Dahabreh and K. Bibbins-Domingo, "Causal Inference About the Effects of Interventions From Observational Studies in Medical Journals," *Journal of the American Medical Association* 331, no. 21 (2024): 1845–1853.

15. Medicine NLo, "Difference-in-Differences" — PubMed Search Results 2025, https://pubmed.ncbi.nlm.nih.gov/?term=%22difference-in-differences%22&filter=years.2008-2025.

16. E. Sperr, "About PubMed by Year," 2016, https://esperr.github.io/pubmed-by-year/about.html.

17. J. Roth, P. H. Sant'Anna, A. Bilinski, and J. Poe, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *Journal of Econometrics* 235, no. 2 (2023): 2218–2244.

18. C. De Chaisemartin and X. d'Haultfoeuille, "Two-Way Fixed Effects and Differences-In-Differences With Heterogeneous Treatment Effects: A Survey," *Econometrics Journal* 26, no. 3 (2023): C1–C30.

19. D. Arkhangelsky and G. Imbens, "Causal Models for Longitudinal and Panel Data: A Survey," *Econometrics Journal* 27 (2024): utae014.

20. A. Kahn-Lang and K. Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business & Economic Statistics* 38, no. 3 (2020): 613–620.

21. A. Chiu, X. Lan, Z. Liu, and Y. Xu, "Causal Panel Analysis Under Parallel Trends: Lessons From a Large Reanalysis Study," *arXiv* 6 (2025): 230915983.

22. N. A. Haber, E. Clarke-Deelder, J. A. Salomon, A. Feller, and E. A. Stuart, "Impact Evaluation of Coronavirus Disease 2019 Policy: A Guide to Common Design Issues," *American Journal of Epidemiology* 190, no. 11 (2021): 2474–2486.

23. S. Rothbard, J. C. Etheridge, and E. J. Murray, "A Tutorial on Applying the Difference-in-Differences Method to Health Data," *Current Epidemiology Reports* 11 (2023): 85.

24. A. Bilinski and L. A. Hatfield, "Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions," *arXiv* 5 (2018): 180503273.

25. C. E. Fry and L. A. Hatfield, "Birds of a Feather Flock Together: Comparing Controlled Pre-Post Designs," *Health Services Research* 56, no. 5 (2021): 942–952.

26. J. M. Wooldridge, "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-In-Differences Estimators," *SSRN Electronic Journal* (2021): 3906345.

27. S. Feng and A. Bilinski, "Parallel Trends in an Unparalleled Pandemic: Difference-In-Differences for Infectious Disease Policy Evaluation," *medRxiv* 1 (2024): 2024.

28. J. J. Heckman, H. Ichimura, J. A. Smith, and P. E. Todd, "Characterizing Selection Bias Using Experimental Data," 1998.

29. J. J. Heckman, H. Ichimura, and P. E. Todd, "Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Programme," *Review of Economic Studies* 64, no. 4 (1997): 605–654.

30. A. Abadie, "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies* 72, no. 1 (2005): 1–19.

31. P. H. Sant'Anna and J. Zhao, "Doubly Robust Difference-In-Differences Estimators," *Journal of Econometrics* 219, no. 1 (2020): 101–122.

32. J. Roth and P. H. Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica* 91, no. 2 (2023): 737–747.

33. C. De Chaisemartin and X. d'Haultfoeuille, "Two-Way Fixed Effects Estimators With Heterogeneous Treatment Effects," *American Economic Review* 110, no. 9 (2020): 2964–2996.

34. B. Callaway and P. H. Sant'Anna, "Difference-in-Differences With Multiple Time Periods," *Journal of Econometrics* 225, no. 2 (2021): 200–230.

35. L. Sun and S. Abraham, "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects," *Journal of Econometrics* 225, no. 2 (2021): 175–199.

36. A. Goodman-Bacon, "Difference-in-Differences With Variation in Treatment Timing," *Journal of Econometrics* 225, no. 2 (2021): 254–277.

37. C. Wing, M. Yozwiak, A. Hollingsworth, S. Freedman, and K. Simon, "Designing Difference-In-Difference Studies With Staggered Treatment Adoption: Key Concepts and Practical Guidelines," *Annual Review of Public Health* 45, no. 1 (2024): 485–505.

38. S. G. Donald and K. Lang, "Inference With Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics* 89, no. 2 (2007): 221–233.

39. A. C. Cameron, J. B. Gelbach, and D. L. Miller, "Bootstrap-Based Improvements for Inference With Clustered Errors," *Review of Economics and Statistics* 90, no. 3 (2008): 414–427.

40. T. G. Conley and C. R. Taber, "Inference With "Difference in Differences" With a Small Number of Policy Changes," *Review of Economics and Statistics* 93, no. 1 (2011): 113–125.

41. D. Roodman, M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb, "Fast and Wild: Bootstrap Inference in Stata Using Boottest," *Stata Journal: Promoting Communications on Statistics and Stata* 19, no. 1 (2019): 4–60.

42. V. Chernozhukov, K. Wüthrich, and Y. Zhu, "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association* 116, no. 536 (2021): 1849–1864.

43. B. E. McGarry, A. D. Gandhi, M. A. Chughtai, J. Yin, and M. L. Barnett, "Clinical Outcomes After Admission of Patients With COVID-19 to Skilled Nursing Facilities," *JAMA Internal Medicine* 184, no. 7 (2024): 799–808.

44. M. Lechner, "The Estimation of Causal Effects by Difference-In-Difference Methods," *Foundations and Trends in Econometrics* 4, no. 3 (2011): 165–224.

45. P. Deb, E. C. Norton, J. M. Wooldridge, and J. E. Zabel, *A Flexible, Heterogeneous Treatment Effects Difference-in-Differences Estimator for Repeated Cross-Sections* (National Bureau of Economic Research, 2024).

46. K. Borusyak, X. Jaravel, and J. Spiess, "Revisiting Event-Study Designs: Robust and Efficient Estimation," *Review of Economic Studies* 91 (2024): rdae007.

47. J. Gardner, "Two-Stage Differences in Differences," *arXiv* 1 (2022): 220705943.

48. D. Cengiz, A. Dube, A. Lindner, and B. Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," *Quarterly Journal of Economics* 134, no. 3 (2019): 1405–1454.

49. J. N. Mafi, M. Craff, S. Vangala, et al., "Trends in US Ambulatory Care Patterns During the COVID-19 Pandemic, 2019-2021," *Journal of the American Medical Association* 327, no. 3 (2022): 237–247.

50. S. A. Shashikumar, B. Gulseren, N. L. Berlin, J. M. Hollingsworth, K. E. Joynt Maddox, and A. M. Ryan, "Association of Hospital Participation in Bundled Payments for Care Improvement Advanced With Medicare Spending and Hospital Incentive Payments," *Journal of the American Medical Association* 328, no. 16 (2022): 1616–1623.

51. M. W. Steenland, L. E. Pace, and J. L. Cohen, "Association of Medicaid Reimbursement for Immediate Postpartum Long-Acting Reversible Contraception With Infant Birth Outcomes," *JAMA Pediatrics* 176, no. 3 (2022): 296–303.

52. P. Singh, N. Fu, S. Dale, et al., "The Comprehensive Primary Care Plus Model and Health Care Spending, Service Use, and Quality," *Journal of the American Medical Association* 331, no. 2 (2024): 132–146.

53. R. Myerson, D. M. Qato, D. P. Goldman, and J. A. Romley, "Insulin Fills by Medicare Enrollees and Out-of-Pocket Caps Under the Inflation Reduction Act," *JAMA* 330, no. 7 (2023): 660–662.

54. S. Matta, P. Chatterjee, and A. S. Venkataramani, "Changes in Health Care Workers' Economic Outcomes Following Medicaid Expansion," *Journal of the American Medical Association* 331, no. 8 (2024): 687–695.

55. D. M. Qato, J. A. Romley, R. Myerson, D. Goldman, and A. M. Fendrick, "Shingles Vaccination in Medicare Part D After Inflation Reduction Act Elimination of Cost Sharing," *Journal of the American Medical Association* 331, no. 23 (2024): 2043–2045.

56. J. Markowski, J. Wallace, M. Schlesinger, and C. D. Ndumele, "Alternative Payment Models and Performance in Federally Qualified Health Centers," *JAMA Internal Medicine* 184, no. 9 (2024): 1065–1073.

57. K. I. Duan, E. Obara, E. S. Wong, et al., "Supplemental Oxygen Use, Outcomes, and Spending in Patients With COPD in the Medicare Competitive Bidding Program," *JAMA Internal Medicine* 184 (2024): 1457–1465.

58. K. H. Geissler, M. S. Shieh, A. S. Ash, P. K. Lindenauer, J. A. Krishnan, and S. L. Goff, "Medicaid Accountable Care Organizations and Disparities in Pediatric Asthma Care," *JAMA Pediatrics* 178, no. 11 (2024): 1208–1215.

59. S. Kannan, J. D. Bruch, and Z. Song, "Changes in Hospital Adverse Events and Patient Outcomes Associated With Private Equity Acquisition," *Journal of the American Medical Association* 330, no. 24 (2023): 2365–2375.

60. E. Schrier, H. E. M. Schwartz, D. U. Himmelstein, et al., "Hospital Assets Before and After Private Equity Acquisition," *Journal of the American Medical Association* 332, no. 8 (2024): 669–671.

61. S. Asare, A. Majmundar, J. L. Westmaas, et al., "Association of Cigarette Sales With Comprehensive Menthol Flavor ban in Massachusetts," *JAMA Internal Medicine* 182, no. 2 (2022): 231–234.

62. S. Asare, A. Majmundar, Z. Xue, A. Jemal, and N. Nargis, "Association of Comprehensive Menthol Flavor Ban With Current Cigarette Smoking in Massachusetts From 2017 to 2021," *JAMA Internal Medicine* 183, no. 4 (2023): 383–386.

63. A. Schoenbrunner, A. Beckmeyer, N. Kunnath, et al., "Association Between California's State Insurance Gender Nondiscrimination Act and Utilization of Gender-Affirming Surgery," *Journal of the American Medical Association* 329, no. 10 (2023): 819–826.

64. Y. Liang, D. I. Rees, J. J. Sabia, and C. Smiley, "Association Between State Antibullying Policies and Suicidal Behaviors Among Lesbian, Gay, Bisexual, and Questioning Youth," *JAMA Pediatrics* 177, no. 5 (2023): 534–536.

65. A. E. Austin, M. E. Shanahan, M. Frank, et al., "Association of State Expansion of Supplemental Nutrition Assistance Program Eligibility With Rates of Child Protective Services-Investigated Reports," *JAMA Pediatrics* 177, no. 3 (2023): 294–302.

66. N. A. Haber, E. Clarke-Deelder, A. Feller, et al., "Problems With Evidence Assessment in COVID-19 Health Policy Impact Evaluation: A Systematic Review of Study Design and Evidence Strength," *BMJ Open* 12, no. 1 (2022): e053820.

67. J. M. Wooldridge, "Simple Approaches to Nonlinear Difference-in-Differences With Panel Data," *Econometrics Journal* 26, no. 3 (2023): C31–C66.

68. E. J. Tchetgen Tchetgen, C. Park, and D. B. Richardson, "Universal Difference-In-Differences for Causal Inference in Epidemiology," *Epidemiology* 35, no. 1 (2024): 16–22.

69. B. Callaway and T. Li, "Policy Evaluation During a Pandemic," *Journal of Econometrics* 236, no. 1 (2023): 105454.

70. B. Deaner and H. Ku, "Causal Duration Analysis With Diff-in-Diff," *arXiv* 1 (2024): 240505220.

71. S. Sandoval-Olascoaga, A. S. Venkataramani, and M. C. Arcaya, "Eviction Moratoria Expiration and COVID-19 Infection Risk Across Strata of Health and Socioeconomic Status in the United States," *JAMA Network Open* 4, no. 8 (2021): e2129041.

72. G. Hettinger, C. Roberto, Y. Lee, and N. Mitra, "Estimation of Policy-Relevant Causal Effects in the Presence of Interference With an Application to the Philadelphia Beverage Tax," *arXiv* 2 (2023): 230106697.

73. Z. Chen, A. Denteh, and D. Kédagni, "Anticipation Effects in Difference-in-Differences Models," 2025.

74. A. Alpert, "The Anticipatory Effects of Medicare Part D on Drug Utilization," *Journal of Health Economics* 49 (2016): 28–45.

75. Y. A. Antwi, A. S. Moriya, and K. Simon, "Effects of Federal Policy to Insure Young Adults: Evidence From the 2010 Affordable Care Act's Dependent-Coverage Mandate," *American Economic Journal: Economic Policy* 5, no. 4 (2013): 1–28.

76. J. Roth, "Pretest With Caution: Event-Study Estimates After Testing for Parallel Trends," *American Economic Review: Insights* 4, no. 3 (2022): 305–322.

77. S. Freyaldenhoven, C. Hansen, and J. M. Shapiro, "Pre-Event Trends in the Panel Event-Study Design," *American Economic Review* 109, no. 9 (2019): 3307–3338.

78. H. Dette and M. Schumann, "Testing for Equivalence of Pre-Trends in Difference-In-Differences Estimation," *Journal of Business & Economic Statistics* 42 (2024): 1289–1301.

79. J. Roth, "Interpreting Event-Studies From Recent Difference-in-Differences Methods," *arXiv* 1 (2024): 240112309.

80. K. E. Anderson, M. Sahu, M. J. DiStefano, C. V. Asche, and T. J. Mattingly, "Pharmacy Closures and Anticonvulsant Medication Prescription Fills," *Journal of the American Medical Association* 332 (2024): 1847–1849.

81. N. C. Apathy, A. J. Holmgren, and D. A. Cross, "Physician EHR Time and Visit Volume Following Adoption of Team-Based Documentation Support," *JAMA Internal Medicine* 184, no. 10 (2024): 1212–1221.

82. A. La Forgia, A. M. Bond, R. T. Braun, et al., "Association of Physician Management Companies and Private Equity Investment With Commercial Health Care Prices Paid to Anesthesia Practitioners," *JAMA Internal Medicine* 182, no. 4 (2022): 396–404.

83. S. Adkins, N. Talmor, M. H. White, C. Dutton, and A. L. O'Donoghue, "Association Between Restricted Abortion Access and Child Entries Into the Foster Care System," *JAMA Pediatrics* 178, no. 1 (2024): 37–44.

84. H. V. Nguyen, S. Mital, S. Bugden, and E. E. McGinty, "British Columbia's Safer Opioid Supply Policy and Opioid Outcomes," *JAMA Internal Medicine* 184, no. 3 (2024): 256–264.

85. E. A. Stuart, H. A. Huskamp, K. Duckworth, et al., "Using Propensity Scores in Difference-in-Differences Models to Estimate the Effects of a Policy Change," *Health Services and Outcomes Research Methodology* 14 (2014): 166–182.

86. A. M. Ryan, J. F. Burgess, Jr., and J. B. Dimick, "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences," *Health Services Research* 50, no. 4 (2015): 1211–1235.

87. S. O'Neill, N. Kreif, R. Grieve, M. Sutton, and J. S. Sekhon, "Estimating Causal Effects: Considering Three Alternatives to Difference-in-Differences Estimation," *Health Services and Outcomes Research Methodology* 16 (2016): 1–21.

88. J. R. Daw and L. A. Hatfield, "Matching and Regression to the Mean in Difference-in-Differences Analysis," *Health Services Research* 53, no. 6 (2018): 4138–4156.

89. B. Zeldow and L. A. Hatfield, "Confounding and Regression Adjustment in Difference-in-Differences Studies," *Health Services Research* 56, no. 5 (2021): 932–941.

90. C. F. Manski and J. V. Pepper, "How do Right-to-Carry Laws Affect Crime Rates? Coping With Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics* 100, no. 2 (2018): 232–244.

91. A. Rambachan and J. Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies* 90, no. 5 (2023): 2555–2591.

92. T. Ye, L. Keele, R. Hasegawa, and D. S. Small, "A Negative Correlation Strategy for Bracketing in Difference-in-Differences," *Journal of the American Statistical Association* 119 (2023): 1–2268.

93. C. Caetano, B. Callaway, S. Payne, and H. S. A. Rodrigues, "Difference-In-Differences With Time-Varying Covariates in the Parallel Trends Assumption," *arXiv* 2 (2022): 220202903.

94. J. Wolfers, "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results," *American Economic Review* 96, no. 5 (2006): 1802–1820.

95. A. M. Ryan, "Well-Balanced or Too Matchy-Matchy? The Controversy Over Matching in Difference-In-Differences," *Health Services Research* 53, no. 6 (2018): 4106–4110.

96. X. Pang, L. Liu, and Y. Xu, "A Bayesian Alternative to Synthetic Control for Comparative Case Studies," *Political Analysis* 30, no. 2 (2022): 269–288.

97. T. L. Schell, M. Cefalu, B. A. Griffin, R. Smart, and A. R. Morral, "Changes in Firearm Mortality Following the Implementation of State Laws Regulating Firearm Access and Use," *Proceedings of the National Academy of Sciences of the United States of America* 117, no. 26 (2020): 14906–14910.

98. N. C. Chang, "Double/Debiased Machine Learning for Difference-in-Differences Models," *Econometrics Journal* 23, no. 2 (2020): 177–191.

99. T. Blakely, J. Lynch, K. Simons, R. Bentley, and S. Rose, "Reflection on Modern Methods: When Worlds Collide—Prediction, Machine Learning and Causal Inference," *International Journal of Epidemiology* 49, no. 6 (2020): 2058–2064.

100. R. Bentley, E. Baker, K. Simons, J. A. Simpson, and T. Blakely, "The Impact of Social Housing on Mental Health: Longitudinal Analyses Using Marginal Structural Models and Machine Learning-Generated Weights," *International Journal of Epidemiology* 47, no. 5 (2018): 1414–1422.

101. T. Shinozaki and E. Suzuki, "Understanding Marginal Structural Models for Time-Varying Exposures: Pitfalls and Tips," *Journal of Epidemiology* 30, no. 9 (2020): 377–389.

102. S. Gruber, R. W. Logan, I. Jarrin, S. Monge, and M. A. Hernan, "Ensemble Learning of Inverse Probability Weights for Marginal Structural Modeling in Large Observational Datasets," *Statistics in Medicine* 34, no. 1 (2015): 106–117.

103. L. J. Keele, D. S. Small, J. Y. Hsu, and C. B. Fogarty, "Patterns of Effects and Sensitivity Analysis for Differences-In-Differences," *arXiv* 2 (2019): 190101869.

104. I. Ganguli, J. Souza, J. M. McWilliams, and A. Mehrotra, "Association of Medicare's Annual Wellness Visit With Cancer Screening, Referrals, Utilization, and Spending," *Health Affairs* 38, no. 11 (2019): 1927–1935.

105. I. Ganguli, C. Lim, N. Daley, D. Cutler, M. Rosenthal, and A. Mehrotra, "Telemedicine Adoption and Low-Value Care Use and Spending Among Fee-For-Service Medicare Beneficiaries," *JAMA Internal Medicine* 185 (2025): 440.

106. P. Deb, E. C. Norton, J. M. Wooldridge, and J. E. Zabel, "Aggregating Average Treatment Effects on the Treated in Difference-in-Differences Models," Working Paper, 2025.

107. A. C. Baker, D. F. Larcker, and C. C. Wang, "How Much Should We Trust Staggered Difference-In-Differences Estimates?," *Journal of Financial Economics* 144, no. 2 (2022): 370–395.

108. F. M. Hollenbach and B. Egerod, "How Many Is Enough? Sample Size in Staggered Difference-in-Differences Designs," 2024, Center for Open Science.

109. L. Berge, "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm," CREA Discussion Papers (2018).

110. A. Rambachan and J. Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv* 7 (2020): 200800602.

111. A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *Quarterly Journal of Economics* 138, no. 1 (2023): 1–35.

112. A. Hagemann, "Inference With a Single Treated Cluster," *arXiv* 1 (2020): 201004076.

113. R. A. Fisher, "The Design of Experiments," in *The Design of Experiments*, vol. xi (Oliver & Boyd, 1935), 251.

114. S. Rokicki, J. Cohen, G. Fink, J. A. Salomon, and M. B. Landrum, "Inference With Difference-In-Differences With a Small Number of Groups: A Review, Simulation Study, and Empirical Application Using SHARE Data," *Medical Care* 56, no. 1 (2018): 97–105.

115. L. Alvarez, B. Ferman, and K. Wüthrich, "Inference With Few Treated Units," *arXiv* 2 (2025): 250419841.

116. J. G. MacKinnon and M. D. Webb, "The Wild Bootstrap for Few (Treated) Clusters," *Econometrics Journal* 21, no. 2 (2018): 114–135.

117. I. A. Canay, A. Santos, and A. M. Shaikh, "The Wild Bootstrap With a "Small" Number of "Large" Clusters," *Review of Economics and Statistics* 103, no. 2 (2021): 346–363.

118. J. Alten, D. S. Cooper, D. Klugman, et al., "Preventing Cardiac Arrest in the Pediatric Cardiac Intensive Care Unit Through Multicenter Collaboration," *JAMA Pediatrics* 176, no. 10 (2022): 1027–1036.

119. D. M. Qato, J. S. Guadamuz, and R. Myerson, "Changes in Emergency Contraceptive Fills After Massachusetts' Statewide Standing Order," *Journal of the American Medical Association* 332, no. 6 (2024): 504–506.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.