Title

Advances in Difference-in-differences Methods for Policy Evaluation Research.

Journal

Epidemiology, 35(5)

Authors

Wang, Guangyi
Hamad, Rita
White, Justin

Publication Date

2024-09-01

DOI

10.1097/EDE.0000000000001755

Peer reviewed

OPEN

Advances in Difference-in-differences Methods for Policy Evaluation Research

Guangyi Wang,[a,b] *Rita Hamad,*[b] *and Justin S. White*[a,c]

Abstract: Difference-in-differences (DiD) is a powerful, quasi-experimental research design widely used in longitudinal policy evaluations with health outcomes. However, DiD designs face several challenges to ensuring reliable causal inference, such as when policy settings are more complex. Recent economics literature has revealed that DiD estimators may exhibit bias when heterogeneous treatment effects, a common consequence of staggered policy implementation, are present. To deepen our understanding of these advancements in epidemiology, in this methodologic primer, we start by presenting an overview of DiD methods. We then summarize fundamental problems associated with DiD designs with heterogeneous treatment effects and provide guidance on recently proposed heterogeneity-robust DiD estimators, which are increasingly being implemented by epidemiologists. We also extend the discussion to violations of the parallel trends assumption, which has received less attention. Last, we present results from a simulation study that compares the performance of several DiD estimators under different scenarios to enhance understanding and application of these methods.

Keywords: Difference-in-Differences designs; Health policy evaluation; Heterogeneity-robust estimators; Quasi-experimental methods; Staggered policy implementation

(Epidemiology 2024;35: 628–637)

Submitted September 25, 2023; accepted May 21, 2024

From the [a]Philip R. Lee Institute for Health Policy Studies, University of California San Francisco (UCSF), San Francisco, CA; [b]Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA; and [c]Department of Health Law, Policy & Management, Boston University, Boston, MA.

Supported by an NIH grant (R01HL151638).

The authors report no conflicts of interest.

The simulation code (including the data generating process), is included in the Supplemental Material.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Justin White, Department of Health Law, Policy & Management, Boston University School of Public Health, 715 Albany Street, Room 249West, Boston, MA 02118. E-mail: juswhite@bu.edu.

ISSN: 1044-3983/24/355-628637

DOI: 10.1097/EDE.0000000000001755

INTRODUCTION

Difference-in-differences (DiD) is a quasi-experimental approach well suited to analyzing the effects of policies using longitudinal data.[1,2] The approach has deep roots in epidemiology, such that Ignaz Semmelweis's 1861 publication on antiseptic hand-washing in Hungarian maternity wards and John Snow's 1855 examination of the cholera outbreak in London prefigured the modern use of DiD over a century later.[3,4] DiD has become a widely used approach for assessing the health effects of programs, interventions, and policies. For example, studies have extensively employed the approach to investigate the health impacts of the Medicaid expansion, paid family leave laws, revisions of food and nutrition programs, and policy expansions during the COVID-19 pandemic.[1,2,5–9] To reduce confounding and support causal inference, DiD compares changes in outcomes over time between a "treated" group exposed to a policy change and a "comparator" group not exposed to the change. A fundamental underlying assumption, known as the parallel trends assumption, is that the treated and comparison groups would have had parallel outcome trends in the absence of the policy change. Any differential trends observed between the treated and comparator groups after the policy change are attributed to the effect of the policy change.

DiD takes different forms depending on the setting. The canonical DiD model compares changes in outcomes between two groups (treated and comparator) over two time periods, before and after treatment. This 2×2 DiD model is easily extended to multiple groups and multiple time periods, typically by including fixed effects (i.e., indicator variables) for group and time in a specification commonly referred to as generalized DiD or two-way fixed effects DiD regression. The two-way fixed effects specification has been a mainstay of policy evaluation in public health, economics, and other social sciences.[10,11]

Recent work has shown that the two-way fixed effects design requires an additional assumption to generate unbiased DiD estimates. Specifically, it requires that treatment effects are constant across treated groups and over time. This assumption can easily be violated as treatment effects can differ by group characteristics, by calendar time (i.e., when the group is treated), and dynamically (i.e., treatment effects vary over time).[12,13] Heterogeneous treatment effects threaten the validity of many research designs, potentially biasing two-way fixed effects estimates.

Given the growing popularity of the two-way fixed effects design in health research, it is important for epidemiologists to be aware of the potential limitations of the method and common solutions. Building on recent methodologic reviews in the econometrics literature,[14,15] we aim to provide theoretical foundations to help epidemiologists understand the fundamental problems associated with the two-way fixed effects design in the presence of heterogeneous treatment effects. We describe recent methodological advances that provide several alternative, heterogeneity-robust DiD estimators. We also discuss parallel trend assumption violations, which have received less attention than heterogeneous treatment effects. Finally, we conduct a simulation study that compares the performance of two-way fixed effects and new approaches under different scenarios to enhance the understanding and application of these methods in policy evaluations with health outcomes.

THE BASIC DID DESIGN

The simplest form of DiD involves two groups and two periods, with the policy implemented in only one group at one point during the study period. In the first "pre" period, both groups are unexposed. In the second "post" period, one group is exposed to the policy while the other is not. To ground our discussion, consider an evaluation of paid family leave laws that allow working parents to take time off work after the birth of a child. We might wish to understand the effects of paid family leave laws on the health outcomes of parents and their children. A randomized experiment would be difficult to undertake due to ethical concerns and cost. Thus, we turn to real-world data. California was the first US state to implement a paid family leave policy in 2004. By comparing trends in outcomes between California and states without a paid family leave policy, researchers have evaluated the effects of California's policy on a variety of outcomes, such as breastfeeding and maternal and child health.[5,16–18]

Researchers commonly use a regression framework to estimate a DiD model:

$$Y_{g,t} = \kappa + \alpha TREAT_g + \beta POST_t + \delta\,(TREAT_g \cdot POST_t) + \varepsilon_{g,t}, \quad (1)$$

where outcome $Y_{g,t}$ is regressed on $TREAT_g$, a binary indicator for whether an observation is in the treated group (i.e., California), and $POST_t$, a binary indicator of whether the observation falls in the postperiod (i.e., after paid family leave implementation in 2004), as well as an interaction term $TREAT_g \cdot POST_t$, a constant term κ, and regression residual $\varepsilon_{g,t}$. In the absence of the policy change, the potential outcome for the treated group is determined by a time-invariant group effect, and a time effect that is common to both groups . δ is the coefficient of interest, representing the average treatment effect on the treated (ATT), identified for the 2 × 2 DiD as the regression coefficient on the interaction term (i.e., $TREAT_g \cdot POST_t$), as shown in Table 1. $Y_{g,t}$ can consist of aggregated data (e.g., by state) or disaggregated data (e.g., by individual). In the case of disaggregated data, $Y_{g,t}$ is the average outcome of individuals within cell (g, t), and the regression is weighted by the number of individuals within that cell. In practice, the above model is often further modified to include time-varying confounders (e.g., state-level characteristics, individual-level characteristics when using individual-level data), although recent research notes that care must be taken in how covariates enter the model (discussed later).[19] The eAppendix; http://links.lww.com/EDE/C157 provides further details about identifying the DiD effect with three key assumptions.

TABLE 1. Calculating the 2 × 2 DiD Estimand (δ) in a Regression Framework

	Comparator	Treated	Difference
Preperiod	κ	$\kappa + \alpha$	α
Postperiod	$\kappa + \beta$	$\kappa + \alpha + \beta + \delta$	$\alpha + \delta$
Difference	β	$\beta + \delta$	δ

The notation is drawn from Equation 1. Each cell represents the value of the outcome variable for the row-column combination. For example, the comparator group in the policy period has an outcome value of κ, and the treated group in the pre-policy period has an outcome value of $\kappa + \alpha$. The treated-comparator difference in the pre-period is $(\kappa + \alpha) - \kappa = \alpha$.

EXTENDING DID TO MORE COMPLEX POLICY INTERVENTIONS

Although a simple 2 × 2 DiD design is intuitive, policy settings are often more complex, as when an intervention is implemented in multiple groups at multiple time points (known as a staggered design), and thus there is no single pre- or postperiod. Returning to the paid family leave example, we note that several states have passed a paid family leave law since California's law went into effect, with states adopting the policy at different times (e.g., New Jersey in 2009, New York in 2018).[20]

A popular approach to accommodate a staggered rollout is a generalized DiD design, an extension of 2 × 2 DiD. In practice, the implementation is carried out using the regression specification in Equation (2), commonly referred to as a two-way fixed effects estimator. Two-way fixed effects DiD can be decomposed into the estimation of many pairwise 2 × 2 DiDs, and the total estimated effect is a variance-weighted average of all 2 × 2 DiDs.[12] The key underlying assumption is an extension of the canonical parallel trends assumption. The simplest extension, which is also the strongest, requires the parallel trends assumption to hold for all periods and all treated groups. In the paid family leave example, this assumption implies that, if there had been no paid family leave law, health outcomes of parents and their children would have evolved in parallel on average for each state that passed a paid family leave law compared with states without such a law:

$$Y_{g,t} = \alpha_g + \beta_t + \delta D_{g,t} + \varepsilon_{g,t}, \quad (2)$$

where $Y_{g,t}$ is the outcome of group g at period t. $D_{g,t}$ indicates the treatment status in the group g at period t (e.g., whether the policy

has been implemented in a given state at time t), equivalent to the $TREAT_g \cdot POST_t$ interaction term in the 2×2 case. α_g and γ_t are fixed effects for group and time (e.g., indicator variables for state and year). δ is the key coefficient of interest, representing the estimated policy effect under the assumptions of DiD. Note that Eq. (2) does not include separate terms for the treatment indicator, postperiod group indicator, or constant term used in the 2×2 DiD in Eq. (1) because they are absorbed by the group and time-fixed effects. The group-fixed effects account for all observed or unobserved group-specific, time-invariant factors, and the time-fixed effects account for all observed or unobserved period-specific factors that are common to all groups. Researchers often estimate a version of Eq. (2) that adjusts for observed time-varying confounders that may differentially affect the treated group and interpret estimates in light of any residual confounding. However, care must be taken in how covariates enter the model (see the *Conditional Parallel Trends Assumption* section).

In many cases, a treatment effect varies with time since exposure. To explore dynamic treatment effects, an event-study DiD specification may be used. It allows for examining anticipation effects and phase-in effects in a single regression. In the event-study regression, Eq. (2) is replaced by a set of indicator variables measuring the time relative to treatment for the group g, often referred to as event-time dummies. We first generate a centered time variable s, such that the event occurs at $s = 0$. We then estimate the following equation:

$$Y_{g,t} = \alpha_g + \beta_t + \sum_{s=-q}^{-2} \gamma_s D_{g,s} + \sum_{s=0}^{M} \delta_s D_{g,s} + \varepsilon_{g,t}, \quad (3)$$

where δ_s measures the treatment "lags," the period-by-period differences between the treated group g and the comparator group that occur s periods after the event, from time 0 (the immediate effect) to time M, and γ_s measures the treatment "leads," the effects that occur s periods before the event. By convention, the reference (or omitted) period is set to $s = -1$, one period before the event, and all coefficients are interpreted relative to this base period. While the lag coefficients δ_s are of main interest, showing how DiD effects grow or fade over time, the pattern of lead coefficients γ_s is also of substantive interest, as small and statistically insignificant coefficients indicate that the outcome trends between the treated and comparator groups were similar before the event ("pretrends"), providing support for the parallel trends assumption. Event-study estimates are often presented graphically, with centered time on the horizontal axis and the DiD estimates and confidence intervals on the vertical axis.[21] Recently, more sophisticated tests of parallel pretrends have been proposed,[22,23] to address potential problems associated with using event studies to test pretrends (discussed later).[24]

KEY LIMITATIONS OF TWO-WAY FIXED EFFECTS DID

Recent advances in the economics literature have shown that two-way fixed effects DiD designs may generate biased DiD estimators under heterogeneous treatment effects, that is, when effects vary across groups or time.[12,25] In the following section, we provide a short summary of the theoretical foundation of the problem.

"Negative Weights" and "Forbidden Comparisons"

Under the parallel trends assumption, de Chaisemartin and D'Haultfœuille[25] show that the two-way fixed effects DiD estimator δ in Eq. (2) can be written as the weighted average of treatment effects from multiple 2×2 DiDs:[14,25]

$$\mathrm{E}[\delta] = \mathrm{E}\left[\sum_{(g,t):D_{g,t}\neq 0} W_{g,t}\tau_{g,t}\right], \quad (4)$$

where $\tau_{g,t}$ represents the estimated ATT of group g at period t. Each group-time ATT is the difference in potential outcomes with and without treatment: $\tau_{g,t} = Y_{g,t}(1) - Y_{g,t}(0)$. $W_{g,t}$ are weights summing to 1 and are proportional to and of the same sign as the following equation:

$$D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}, \quad (5)$$

where $D_{g,\cdot}$ is the average proportion of the time group g is treated across periods, $D_{\cdot,t}$ is the average treatment status at period t across groups, and $D_{\cdot,\cdot}$ is the average treatment status across groups and periods.

Eq. (5) is decreasing in $D_{g,\cdot}$ and $D_{\cdot,t}$ (i.e., smaller $W_{g,t}$ in Eq. (4) with larger $D_{g,\cdot}$ and $D_{\cdot,t}$), meaning that $\mathrm{E}[\delta]$ is down-weighted for groups that receive treatment for longer periods and for time periods when more groups are treated. Eq.(4) and Eq.(5) together imply that, under heterogeneous treatment effects (e.g., when treatment effects are larger among early-treated groups or larger in later periods), $\mathrm{E}[\delta]$ will not be equal to the average treatment effect across all cells, resulting in a biased ATT (if defined as the simple average of all group-time ATTs). More importantly, Eq. (5) implies that some weights may be negative, for example, when $D_{g,\cdot} + D_{\cdot,t}$ is large. In this case, even if the treatment effect is positive for each (g,t), $\mathrm{E}[\delta]$ could be negative, causing the DiD to have a flipped sign. For example, if the effect for an early-treated group (treated in period 2) is 10% and 40% in periods 2 and 3, respectively, and the effect for a late-treated group (treated in period 3) is 10% in period 3, the two-way fixed effects estimator may assign weights of +1 and −½ to the effect for the early-treated group in periods 2 and 3, respectively, and weight of +½ to the effect for the late-treated group in period 3, then $\mathrm{E}[\delta] = 1 \times 10\% - {}^{1}/_{2} \times 40\% + {}^{1}/_{2} \times 10\% = -5\%$. Hence, the estimated DiD is negative, despite the effect being positive for both early- and late-treated groups. For additional details, please refer to de Chaisemartin and D'Haultfœuille.[14]

More generally, if treatment effects vary across time or groups, it is possible for $\tau_{g,t}$ to receive lower or even negative weights in Eq. (5) for some combinations of g and t, generating biased DiD estimates. Returning to the PFL case,

TABLE 2. Summary of Key Estimation Strategies of HTE-robust Approaches

	Summary
Group-time estimator approach	Proposed by Callaway and Sant'Anna[13], this approach first identifies the $ATT_{g,t}$, comparing the outcome evolution of group g (units treated in period g) from the last pretreatment period (i.e., $g-1$) to period t, against the outcome evolution of a control group (not-yet treated or never-treated) over the same time periods. These identified parameters then can be extended to estimate more aggregated ATTs of interest. de Chaisemartin and D'Haultfœuille's approach operates similarly, except employing different weighting for the estimated treatment effect across groups.
Imputation approach	Proposed by Borusyak et al[26], Gardner[31], and Liu et al[32], this approach fits a TWFE regression using only observations for units and periods not-yet-treated to impute a counterfactual outcome for each treated unit in the absence of treatment. The individual treatment effects are then aggregated to an overall ATT. Whereas Callaway and Sant'Anna[13] only uses the outcome of the last pretreated period as a baseline, this approach uses average outcomes across all pretreatment periods as a baseline. The choice of baseline period creates tradeoffs, as stated in the Discussion.
Regression approach	Sun and Abraham[27], Cengiz et al[28], and Wooldridge[29] propose regression-based methods. Sun and Abraham uses a fully saturated model, running a regression with leads and lags of the treatment (event-study specification) interacted with group indicators. Wooldridge uses a similar regression, but with group indicators interacted with time periods. Cengiz constructs cohort-specific data sets for each treated unit, which includes the respective cohort and all never-treated units; these cohort-specific data sets are stacked to compute an overall ATT using stack-unit and stack-year fixed effects. While Sun-Abraham and Cengiz generate event-study estimators and Wooldridge generates $ATT_{g,t}$, both allow for aggregation to a more aggregated ATT of interest.

ATT indicates average treatment effect on the treated; TWFE, two-way fixed effects.

treatment effects in California (early-treated group) in later periods may receive lower or even negative weights. The larger the treatment effects in California in later periods, the more likely that the DiD estimator will be underestimated and even have an incorrect sign.

Goodman-Bacon[12] offers helpful intuition to understand this issue. Two-way fixed effects' weighting of 2×2 DiDs makes both "clean" comparisons between treated and not-yet-treated groups as well as "forbidden" comparisons between groups treated in a later period (late-treated) and already treated groups (early-treated).[12] Under heterogeneous treatment effects, the forbidden comparisons may lead to negative weighting problems and a biased DiD estimator (see eAppendix; http://links.lww.com/EDE/C157 for illustration using the PFL example). Goodman-Bacon[12] proposes a diagnostic to assess how much weight is placed on each 2×2 comparison, including those involving forbidden comparisons.[12]

Event-study designs similarly face contamination from forbidden comparisons under heterogeneous treatment effects (see eAppendix; http://links.lww.com/EDE/C157).

ALTERNATIVE HETEROGENEOUS TREATMENT EFFECTS-ROBUST DID ESTIMATORS

In recent years, several alternatives to two-way fixed effects DiD have been proposed to allow for heterogeneous treatment effects. Popular estimators include those proposed by Callaway and Sant'Anna[13] Borusyak et al[26] Sun and Abraham[27] de Chaisemartin and D'Haultfœuille[25] Cengiz et al.[28] and Wooldridge.[29]

Key Estimation Strategies

Several of these alternative methods use similar strategies. The first step involves targeting treatment effect parameters (i.e., group-time treatment effects, $ATT_{g,t}$) using only "clean" comparisons and avoiding "forbidden" comparisons. The second step involves aggregating treatment effect parameters to final target parameters of interest, such as an overall ATT, using an appropriate weighting approach. Note that there are several ways to summarize $ATT_{g,t}$'s into an overall ATT (see eAppendix; http://links.lww.com/EDE/C157).[13,30] Throughout the remainder of this article, the overall ATT is denoted by the simple option, representing an average of all $ATT_{g,t}$'s weighted by the group size of each $ATT_{g,t}$.

Heterogeneous treatment effect-robust methods vary in implementation and can be classified into three general approaches: a group-time estimator approach, an imputation approach, and a regression-based approach. Table 2 offers a high-level overview of these approaches, providing a basic understanding of the methods and their intricacies. Table 3 outlines the implementation of various estimators, the corresponding Stata and R packages, and potential advantages and disadvantages.

Differences in the Parallel Trends Assumption and Covariate Adjustment

Callaway and Sant'Anna[13] and Sun and Abraham[27] estimators impose a weaker parallel trends assumption, which only requires parallel trends from the last pretreated period until the last time period. Borusyak et al[26] and Wooldridge[29] estimators impose a stronger parallel trends assumption, which requires parallel trends in all time periods. Different methods handle covariates differently, as discussed next.

THE CONDITIONAL PARALLEL TRENDS ASSUMPTION

Parallel trends assumption violations remain a major concern in practice for DiD designs. Typically, it is more plausible to assume that the assumption holds conditional on certain observed

TABLE 3. Summary of Novel Heterogeneity-robust Difference-in-differences Estimators

Manuscript and Stata/R Packages	Method	Advantages/Disadvantages
Callaway and Sant'Anna (2021)[13] Stata csdid hdidregress ra hdidregress ipw hdidregress aipw R Did	1. Uses never-treated or not-yet-treated observations as comparator. 2. Estimates every feasible 2 × 2 DiD available in the selected sample ($ATT_{g,t}$) and allows for doubly-robust estimation with inverse probability weighting (IPW) to reduce bias from confounding.	1. Requires a weaker parallel trends assumption. 2. Can generate more aggregated treatment effects, that is, the overall ATT, event-study estimators, and group/cohort ATTs. 3. Easy to generate event-study results and plot figures.
Borusyak et al[26] Stata did_imputation R didimputation	1. Uses never-treated or not-yet-treated observations as comparator. 2. Explicitly imputes potential outcomes for the treated group using the comparator, calculating individual treatment effects that can be aggregated.	1. More efficient than other methods under some assumptions. 2. Imposes a stronger parallel trend assumption. 3. More precise under parallel trends assumption, but less precise if the assumption is violated. 4. Easy to generate event-study results and plot figures. 5. Ready to be generalized to more complicated specifications, for example, triple-differences, adding group-specific linear trends. 6. Computationally fast.
Sun and Abraham (2021)[27] Stata eventstudyinteract staggered R fixest with sunab() staggered	1. Uses either never-treated observations as comparator, or the last-treated groups as comparator if there are no never-treated, rather than not-yet-treated observations. 2. Unlike the other methods, focuses on estimating dynamic effects separately for each cohort, which can then be aggregated.	1. Generates event-study estimators. The staggered package can be used to generate more aggregated ATTs. 2. Computationally fast.
de Chaisemartin and D'Haultfœuille (2020)[25] Stata did_multiplegt R DIDmultiplegt	1. Uses two types of DiD estimators: a. Compares the outcome evolution of groups switching from untreated to treated and groups untreated in two periods ("switchers in"). b. Compares the outcome evolution of groups switching from treated to untreated and groups treated in two periods ("switchers out").	1. Can be used beyond staggered design, i.e., allows for treatment switching in and out rather than only switching in. 2. Requires an additional parallel trends assumption, i.e., parallel trends for the second type of DiDs (i.e., "switchers out"). 3. Computationally very slow.
Cengiz et al. (2019)[28] Stata stackedev	1. Uses not-yet-treated or never-treated observations as comparator. 2. Creates event-specific datasets ("stacks"), with each stack including observations from units that receive treatment at the same time and observations never treated or treated far enough in the future (i.e., treated after examination window).	1. Only generates event-study results. There is currently no package to generate a more aggregated ATT. 2. Does not provide a way to "weight and sum" event-specific treatment effects; each event is weighted equally
Wooldridge (2021)[29] Stata jwdid hdidregress twfe wooldid R etwfe	1. Uses never-treated or not-yet-treated observations as comparator. 2. Similar to Sun and Abraham, but includes interactions between treatment-time cohorts and time-specific effects, rather than interacting dynamic effects with cohorts, to select valid comparisons.	1. Like Callaway-Sant-Anna, $ATT_{g,t}$ may be aggregated as needed to obtain average effects, calendar effects, and cohort-specific effects. 2. Can accommodate nonlinear models such as logit and Poisson. 3. Uses additional information from pretreatment periods that may improve precision. 4. May be more biased than other methods if assumptions violated.

ATT indicates average treatment effect on the treated; DiD, Difference-in-differences.

covariates. Of note, only covariates that vary by treatment status and are associated with outcome trends are considered confounders in DiD (i.e., time-varying confounders).[33] Table 1 in Zeldow and Hatfield[33] provides a comprehensive classification of potential scenarios involving confounders in DiD designs. Under the conditional parallel trends assumption, the outcome trend in each treated unit, given a vector of covariates X, will be parallel to comparator units with the same values of X. Consequently, we can infer the conditional ATT for a treated unit with X, and then identify the overall unconditional ATT by aggregating all conditional ATTs.

Researchers commonly adjust for covariates in DiD regressions to satisfy the conditional parallel trends assumption. However, several limitations have been highlighted in the literature. A prominent issue under discussion is the "bad control" problem.[19,33] If a time-varying covariate is affected by treatment status, bias can arise because the time-varying covariate may serve both as a confounder and a mediator, and thus the ATT is a combination of the direct effect of treatment and the indirect effect of treatment via the covariate. However, simply excluding it from the regression can lead to violations of the parallel trends assumption. Caetano et al[19] propose a solution to let the parallel trends assumption be conditioned on the untreated potential value of these covariates. Statistical code to implement this procedure is still under development. See eAppendix; http://links.lww.com/EDE/C157 for more discussion about the conditional parallel trends assumption (e.g., other limitations with covariate adjustment, how heterogeneous treatment effects-robust estimators handle covariates).

VIOLATIONS OF THE PARALLEL TRENDS ASSUMPTION

The conditional parallel trends assumption may still be violated due to residual unobserved time-varying confounding. Although theoretically untestable, researchers commonly check differences in outcome trends in the treated and comparison groups during the pretreatment periods, both visually and using quantitative tests. Evidence of parallel "pretrends" is used to infer the plausibility of the parallel trends assumption. One conventional approach to test pretrends involves assessing the pretreatment coefficients from event-study DiD designs (as described earlier). With a staggered design, it is recommended to employ the heterogeneity-robust methods described above.

However, recent literature argues that the traditional parallel trends test often has low power, causing underrejection of the null hypothesis that there are no differences between trends in treated and comparison groups (i.e., type II error).[24,34] Moreover, even if pretrends are parallel, it does not guarantee that post-treatment trends would be parallel in the absence of treatment. Therefore, traditional parallel trends assumption testing is not a reliable indicator of parallel trends assumption violations.

Several articles have proposed alternative approaches to detect pretreatment violations of parallel trends, including tools to conduct power analyses and "noninferiority" approaches testing the likelihood of rejecting the null hypothesis of a large pretrend.[15,24,34] While these approaches offer valuable insights, they do not provide guidance for ATT inference when parallel trends may be violated and do not address post-treatment differences in trends. To address these issues, Rambachan and Roth[22] propose an approach for robust inference and sensitivity analysis regarding parallel trends assumption violations (see eAppendix; http://links.lww.com/EDE/C157).

SIMULATION STUDY

We conducted a Monte Carlo simulation study to test the performance of the two-way fixed effects estimator and four heterogeneous treatment effects-robust estimators under different scenarios. We considered three main scenarios: (1) constant or dynamic treatment effects (i.e., whether treatment effects vary over time), (2) homogeneous or heterogenous treatment effects (i.e., whether treatment effects are homogeneous across groups, random across groups, or larger among earlier-treated groups), and (3) with or without parallel trends assumption violations. Table 4 provides a list of scenarios, and the eAppendix; http://links.lww.com/EDE/C157 details our data generating process.

In each scenario, the simulation had 500 runs for each DiD estimator. We compared the performance of DiD estimators in terms of percent bias and root mean squared error (RMSE).

Figure 1 presents the distribution of bias (%), and Table 5 summarizes the mean bias (%) and RMSE under no parallel trends assumption violations. With constant, homogeneous effects (scenarios 1a), two-way fixed effects had the lowest bias (−0.01%) and RMSE (0.12). With constant, heterogeneous effects, two-way fixed effects had a small bias when effects varied randomly (2.38%; scenarios 1b) and a slightly larger bias under the "large-first" setting (−9.49%; scenarios 1c). The Callaway and Sant'Anna[13], Borusyak et al[26], Sun and Abraham[27], and Wooldridge[29] estimators were generally robust across these scenarios, with small bias and RMSEs. When treatment effects were dynamic (scenarios 2a–2c), two-way fixed effects had notably high bias (−66.40% to −78.62%). The Callaway and Sant'Anna[13] estimator had the lowest bias (0.64–4.14%) with dynamic effects. The Borusyak et al[26], Sun and Abraham[27], and Woodridge[29] estimators had slightly higher bias (around −6%). Borusyak et al[26] and Woodridge[29] estimators generally had lower RMSEs than Callaway and Sant'Anna[13] and Sun and Abraham[27] estimators in all scenarios, perhaps reflective of precision gains from using all pretreatment periods for comparisons.

Figure 2 and Table 5 display results under parallel trends assumption violations (scenarios 3a–3f). Compared with the scenarios with no parallel trends assumption violations, we observed large increases in bias (%) and RMSEs across all estimators, suggesting lower accuracy and precision in the models' predictions compared with the true effect. Two-way

TABLE 4. Summary of Scenarios

#	Description	Dynamic Effects Over Time?	Heterogeneous Effects Across groups?	Parallel Trends Assumption Violation?
1a	Constant, homogeneous effects without PTA violation			
1b	Constant, heterogeneous (at random) effects without PTA violation		X	
1c	Constant, heterogeneous (large first) effects without PTA violation		X	
2a	Dynamic (linear trend), homogeneous effects without PTA violation	X		
2b	Dynamic (linear trend), heterogeneous (at random) effects without PTA violation	X	X	
2c	Dynamic (linear trend), heterogeneous (large first) effects without PTA violation	X	X	
3a	Constant, homogeneous effects with PTA violation			X
3b	Constant, heterogeneous (at random) effects with PTA violation		X	X
3c	Constant, heterogeneous (large first) effects with PTA violation		X	X
3d	Dynamic (linear trend), homogeneous effects with PTA violation	X		X
3e	Dynamic (linear trend), heterogeneous (at random) effects with PTA violation	X	X	X
3f	Dynamic (linear trend), heterogeneous (large first) effects with PTA violation	X	X	X

Scenarios 1a–1c compares the performance of DiD estimators under constant effects under the PTA. Scenarios 2a–2c compares the performance under dynamic effects under the PTA. Scenarios 3a–3f compare the performance under constant and dynamic effects when the PTA is violated.
PTA indicates parallel trends assumption.

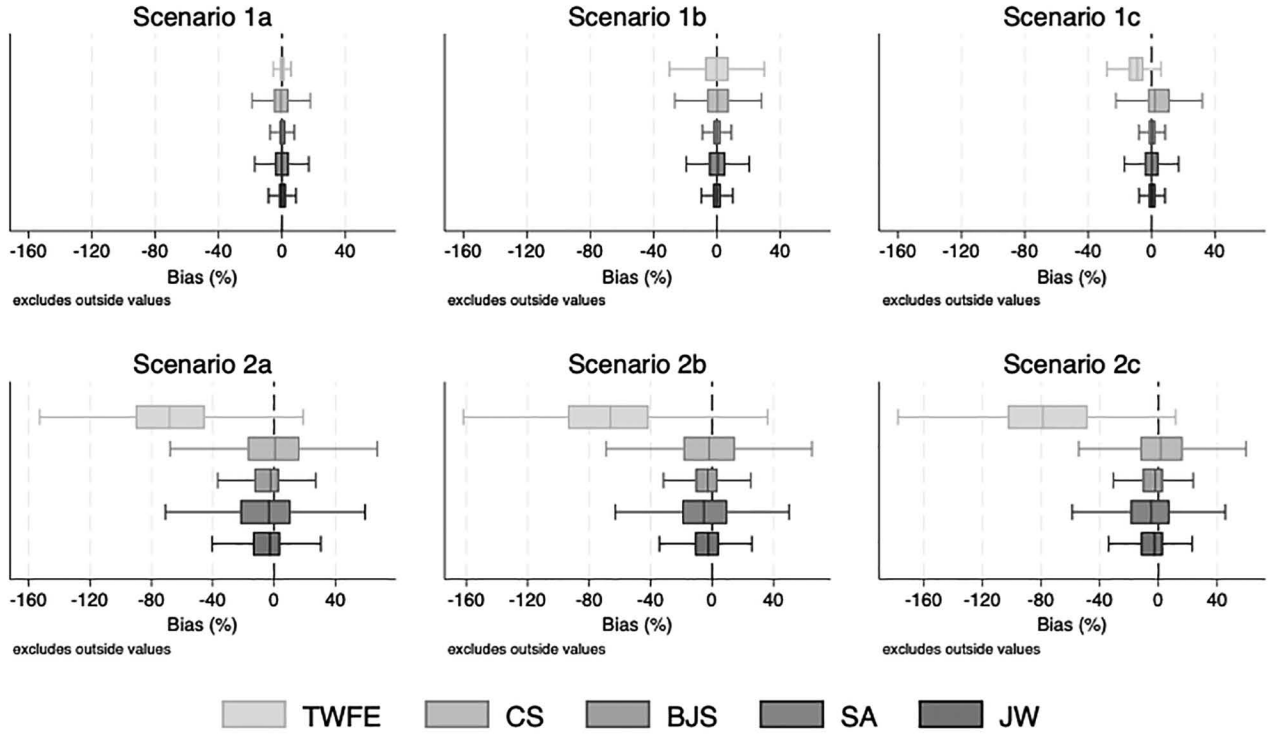

FIGURE 1. Monte Carlo simulation results for scenarios 1 and 2 (no PTA violations). Scenarios 1a–1c have constant effects, and scenarios 2a–2c have dynamic (linear trend) effects. Scenarios 1a and 2a have homogeneous effects across groups; scenarios 1b and 2b have heterogeneous (at random) effects across groups; and scenarios 1c and 2c have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4. BJS Indicates Borusyak-Jaravel-Spiess; CS, Callaway-Sant'Anna; JW, Wooldridge; PTA, parallel trends assumption; SA, Sun-Abraham; TWFE, Two-way fixed effects.

fixed effects exhibited somewhat better performance than heterogeneous treatment effects-robust estimators under constant effects but poorer performance when treatment effects grow linearly over time. Among heterogeneous treatment effects-robust estimators, Callaway and Sant'Anna[13] and Sun and Abraham[27] estimators outperformed Borusyak et al[26] and Wooldridge[29] estimators, aligning with the hypothesis that estimators reliant on a weaker parallel trends assumption

TABLE 5. Monte Carlo Simulation Estimates of the ATT

		No PTA violation (Scenarios 1 and 2)		PTA violation (Scenario 3)	
Scenario	**Methods**	**Bias (%)**	**RMSE**	**Bias (%)**	**RMSE**
Constant, homogeneous effects			Scenario 1a		Scenario 3a
	TWFE	−0.01	0.12	−8.54	0.63
	CS	−1.56	0.77	−21.53	1.21
	BJS	−0.38	0.26	−24.31	1.24
	SA	0.65	0.39	−20.23	1.15
	JW	−0.16	0.30	−24.11	1.18
Constant, random the			Scenario 1b		Scenario 3b
	TWFE	2.38	0.90	−4.95	1.07
	CS	0.04	1.45	−17.70	1.75
	BJS	0.62	0.54	−20.25	1.33
	SA	0.95	0.66	−16.93	1.26
	JW	0.55	0.56	−19.62	1.25
Constant, large-first the			Scenario 1c		Scenario 3c
	TWFE	−9.49	0.74	−15.00	1.05
	CS	7.19	1.57	−8.99	1.07
	BJS	−0.34	0.59	−17.74	1.34
	SA	0.03	0.51	−14.89	1.20
	JW	0.05	0.41	−17.09	1.20
Dynamic, homogeneous the			Scenario 2a		Scenario 3d
	TWFE	−69.24	1.77	−96.65	2.00
	CS	1.42	0.62	−57.29	1.83
	BJS	−6.51	0.40	−81.65	1.36
	SA	−6.60	0.54	−64.01	1.28
	JW	−7.58	0.47	−79.76	1.35
Dynamic, random the			Scenario 2b		Scenario 3e
	TWFE	−66.40	1.78	−87.78	1.96
	CS	0.64	0.86	−50.17	1.24
	BJS	−5.11	0.45	−67.37	1.33
	SA	−6.57	0.54	−55.51	1.25
	JW	−5.43	0.47	−66.58	1.35
Dynamic, large-first the			Scenario 2c		Scenario 3f
	TWFE	−78.62	2.28	−95.64	2.30
	CS	4.14	0.72	−37.05	1.12
	BJS	−6.05	0.45	−59.77	1.38
	SA	−6.16	0.58	−48.97	1.27
	JW	−6.34	0.58	−58.77	1.32

Each scenario is listed in Table 4.

BJS Indicates Borusyak-Jaravel-Spiess; CS, Callaway-Sant'Anna; JW, Wooldridge; PTA, parallel trends assumption; RMSE, root mean squared error; SA, Sun-Abraham; TWFE, Two-way fixed effects.

perform better under parallel trends assumption violations (details in the *Discussion* section).

The simulation results remain robust when we increase the number of units, time periods, and simulation runs (see eAppendix; http://links.lww.com/EDE/C157).

DISCUSSION

The DiD design is a powerful approach for investigating causal relationships, particularly in policy evaluation research. While heterogeneous treatment effects-robust DiD estimators are effective in reducing bias compared with traditional two-way fixed effects DiD methods, epidemiologists have only recently begun adopting them. This article discussed key issues associated with two-way fixed effects estimators in the presence of heterogeneous treatment effects, recent methodologic advancements, as well as the parallel trends assumption, which remains a major concern in practice for DiD.

We also examined the performance of two-way fixed effects estimators and several other heterogeneous treatment effects-robust estimators that have received attention recently, adding to several recent articles conducting similar simulation studies.[30,35] Our results indicated that two-way fixed effects are the most efficient option when treatment effects remained constant across groups and time, but its performance diminished

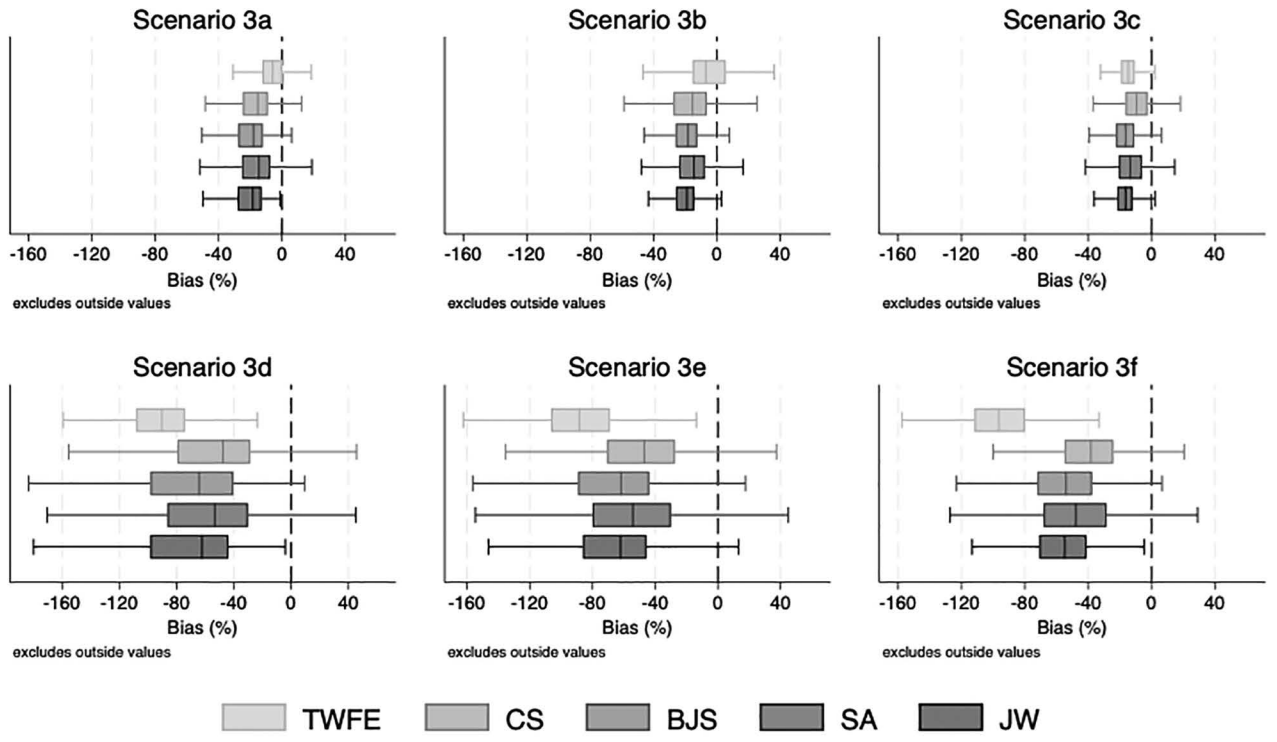

FIGURE 2. Monte Carlo simulation result for scenario 3 (with PTA violations). Scenarios 3a–3c have constant effects, and Scenarios 3d,e have dynamic (linear trend) effects. Scenarios 3a and 3d have homogeneous effects across groups; Scenarios 3b and 3e have heterogeneous (at random) effects across groups; and scenarios 3c and 3f have heterogeneous (large first) effects across groups. Each scenario is listed in Table 4. BJS Indicates Borusyak-Jaravel-Spiess; CS, Callaway-Sant'Anna; JW, Wooldridge; PTA, parallel trends assumption; SA, Sun-Abraham; TWFE, Two-way fixed effects.

notably under dynamic treatment effects. Meanwhile, heterogeneous treatment effects-robust estimators exhibited more robust results across all scenarios under the parallel trends assumption. Thus, we recommend that researchers test the robustness of two-way fixed effects results using a heterogeneous treatment effects-robust estimator for staggered design studies. Conducting the Goodman-Bacon[12] diagnostic will also further understand the extent and nature of heterogeneous treatment effects in a particular study setting and the comparisons receiving greater weight in two-way fixed effects.

Our results align with the conclusion in the literature that heterogeneous treatment effects-robust methods often yield similar estimates.[14,15] Nonetheless, we recommend researchers to carefully evaluate which method to employ based on the specific context of their study, including factors discussed below.

The preferred method depends on the reliability of the parallel trends assumption. Callaway and Sant'Anna's[13] and Sun and Abraham's[27] approaches rely on weaker assumptions of parallel trends, using only the last pretreated period's outcome for baseline comparisons. Borusyak et al[26] and Wooldridge's[29] approaches rely on stronger assumptions of parallel trends, using the average outcome of all pretreatment periods as a baseline. Tradeoffs exist between efficiency and the strength of the identifying assumption. If parallel trends hold for all groups and all periods, estimators with strong assumptions of parallel trends are more precise. However, they are also more biased when strong assumptions of parallel trends are violated, especially with early and increasing trend discrepancies between groups. However, if the failure of parallel trends is due to anticipation effects emerging just before treatment (i.e., outcomes change in anticipation of a future treatment), estimators with a strong parallel trends assumption are less biased than those with weaker assumptions. In summary, if the validity of parallel trends over longer time horizons is a concern, Callaway and Sant'Anna's[13] and Sun and Abraham's[27] approaches may be preferred.

The choice of comparator is an important design choice. Selecting both not-yet-treated and never-treated units as the comparator group increases statistical power, whereas selecting not-yet-treated units only may reduce confounding bias by closely resembling treated units. However, not-yet-treated units should be included with caution, particularly when anticipatory effects are present because this can violate the parallel trends assumption and introduce bias. Sensitivity analyses may help to determine whether estimates are robust to comparator choice. Ultimately, the decision will depend on the application.

The computation time for estimating heterogeneous treatment effects-robust DiD estimators is another factor to consider. Some estimators, for example, Borusyak et al[26], Sun and Abraham[27], and Wooldridge[29], are quick to compute. However, other estimators, for example, those proposed by de Chaisemartin and D'Haultfœuille[25] and Callaway and Sant'Anna[13], require more time due to complex weighting methods like doubly-robust estimation. de Chaisemartin and D'Haultfœuille's technique may be less desirable for large sample sizes due to its substantially slower computation speed. However, it does possess the advantage of accommodating treatments that switch on and off, while most other methods allow for nonreversible treatment only.

Concerning the parallel trends assumption, we recommend event-study plots for visually assessing pretrends. Then, it is crucial to thoroughly consider the relationships among covariates, treatments, and outcomes over time, within a causal framework and insights from prior literature, before employing advanced techniques to control these time-varying confounders. Nevertheless, given the challenge of assessing the validity of the parallel trends assumption, conducting sensitivity tests (e.g., honest DiD[22]) is important to ensure a more robust inference.

Finally, it is important to note that the field of DiD estimation is rapidly evolving, with ongoing developments incorporating more complex designs, such as continuous treatments and triple differences.[26,36,37] While beyond the scope of this article, these advancements provide methodological rigor of policy evaluations for increased translational impact.

REFERENCES

1. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health*. 2018;39:453–469.
2. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA*. 2014;312:2401–2402.
3. Coleman T. Causality in the time of cholera: John Snow as a prototype for causal inference. Available at SSRN 3262234; 2019;
4. Semmelweis IF. *The cause, concept and prophylaxis of childbed fever*. vol 5. Williams & Wilkins; 1941.
5. Hamad R, Modrek S, White JS. Paid family leave effects on breastfeeding: a quasi-experimental study of US policies. *Am J Public Health*. 2019;109:164–166.
6. Baicker K, Taubman SL, Allen HL, et al; Oregon Health Study Group. The Oregon experiment—effects of Medicaid on clinical outcomes. *N Engl J Med*. 2013;368:1713–1722.
7. Lipton BJ, Decker SL. The effect of health insurance coverage on medical care utilization and health outcomes: evidence from Medicaid adult vision benefits. *J Health Econ*. 2015;44:320–332.
8. Hamad R, Collin DF, Baer RJ, Jelliffe-Pawlowski LL. Association of revised WIC food package with perinatal and birth outcomes: a quasi-experimental study. *JAMA Pediatr*. 2019;173:845–852.
9. Batra A, Jackson K, Hamad R. Effects of the 2021 expanded child tax credit on adults' mental health: a Quasi-Experimental Study: study examines the effects of the expanded child tax credit on mental health among low-income adults with children and racial and ethnic subgroups. *Health Aff (Millwood)*. 2023;42:74–82.
10. Ryan AM, Burgess JF Jr, Dimick JB. Why we should not be indifferent to specification choices for difference-in-differences. *Health Services Res*. 2015;50:1211–1235.
11. Currie J, Kleven H, Zwiers E. Technology and big data are changing economics: mining text to track methods. AEA Pap Proc. 2020;110:42–48.
12. Goodman-Bacon A. Difference-in-differences with variation in treatment timing. *J Econometrics*. 2021;225:254–277.
13. Callaway B, Sant'Anna PH. Difference-in-differences with multiple time periods. *J Econometrics*. 2021;225:200–230.
14. De Chaisemartin C, D'Haultfœuille X. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econom J*. 2023;26:C1–C30.
15. Roth J, Sant'Anna PH, Bilinski A, Poe J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J Econometrics*. 2023;235:2218–2244.
16. Lee BC, Modrek S, White JS, Batra A, Collin DF, Hamad R. The effect of California's paid family leave policy on parent health: a quasi-experimental study. *Soc Sci Med*. 2020;251:112915.
17. Irish AM, White JS, Modrek S, Hamad R. Paid family leave and mental health in the US: a quasi-experimental study of state policies. *Am J Prev Med*. 2021;61:182–191.
18. Van Niel MS, Bhatia R, Riano NS, et al. The impact of paid maternity leave on the mental and physical health of mothers and children: a review of the literature and policy implications. *Harv Rev Psychiatry*. 2020;28:113–126.
19. Caetano C, Callaway B, Payne S, Rodrigues HSA. Difference in differences with time-varying covariates. *arXiv preprint arXiv*:220202903; 2022.
20. Rossin-Slater M. *Maternity and family leave policy*; 2017.
21. Freyaldenhoven S, Hansen C, Pérez JP, Shapiro JM. *Visualization, identification, and estimation in the linear panel event-study design*; 2021.
22. Rambachan A, Roth J. A more credible approach to parallel trends. *Rev Econ Stud*. 2023;90:2555–2591.
23. Roth J, Sant'Anna PH. When is parallel trends sensitive to functional form? *Econometrica*. 2023;91:737–747.
24. Roth J. Pretest with caution: event-study estimates after testing for parallel trends. *American Economic Review: Insights*. 2022;4:305–322.
25. De Chaisemartin C, D'Haultfoeuille X. Two-way fixed effects estimators with heterogeneous treatment effects. *Amer Econ Rev*. 2020;110:2964–2996.
26. Borusyak K, Jaravel X, Spiess J. Revisiting event-study designs: robust and efficient estimation. *Rev Econ Stud*. 2024:rdae007.
27. Sun L, Abraham S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J Econometrics*. 2021;225:175–199.
28. Cengiz D, Dube A, Lindner A, Zipperer B. The effect of minimum wages on low-wage jobs. *Q J Econ*. 2019;134:1405–1454.
29. Wooldridge JM. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. Available at SSRN 3906345; 2021;
30. Riddell CA, Goin DE. Guide for comparing estimators of policy change effects on health. *Epidemiology*. 2023;34:e21–e22.
31. Gardner J. Two-stage differences in differences. arXiv preprint arXiv: 220705943; 2022.
32. Liu L, Wang Y, Xu Y. A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data. Am J Political Sci. 2024;68:160–176.
33. Zeldow B, Hatfield LA. Confounding and regression adjustment in difference-in-differences studies. *Health Services Res*. 2021;56:932–941.
34. Bilinski A, Hatfield LA. Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions. *arXiv preprint arXiv*:180503273; 2019.
35. Baker AC, Larcker DF, Wang CC. How much should we trust staggered difference-in-differences estimates? *J Finan Econ*. 2022;144:370–395.
36. Callaway B, Goodman-Bacon A, Sant'Anna PH. Difference-in-differences with a continuous treatment. *arXiv preprint arXiv*:210702637; 2021;
37. De Chaisemartin C, d'Haultfoeuille X. *Difference-in-differences estimators of intertemporal treatment effects*. *National Bureau of Economic Research*; 2022.No. w29873.