Nothing to See Here? A non-inferiority approach to parallel trends

Alyssa Bilinski, PhD*
Departments of Health Services, Policy, and Practice & Biostatistics
Brown University

Laura Hatfield, PhD
Statistics and Data Science Department
National Opinion Research Center

Abstract

Difference-in-differences is a popular method for observational health policy evaluation. It relies on a causal assumption that in the absence of intervention, treatment groups' outcomes would have evolved in parallel to those of comparison groups. Researchers frequently look for parallel trends in the pre-intervention period to bolster confidence in this assumption. The popular "parallel trends test" evaluates a null hypothesis of parallel trends and, failing to find evidence against the null, concludes that the assumption holds. This tightly controls the probability of falsely concluding that trends are not parallel but may have low power to detect non-parallel trends. When used as a screening step, it can also introduce bias in treatment effect estimates. We propose a non-inferiority/equivalence approach that tightly controls the probability of missing large violations of parallel trends measured on the scale of the treatment effect. Our framework nests several common use cases, including linear trend tests and event studies. We show that our approach may induce no or minimal bias when used as a screening step under commonly-assumed error structures, and absent violations, can offer a higher-power alternative to testing treatment effects in more flexible models. We illustrate our ideas by re-considering a study of the impact of the Affordable Care Act's dependent coverage provision.

Keywords: Causal inference; controlled pre/post designs; equivalence tests; longitudinal data; quasi-experimental designs; sensitivity analyses; statistical power.

1 Introduction

Difference-in-differences (DiD) is a popular design for impact evaluation in observational settings with longitudinal data in fields ranging from health policy and economics to education and political science.[1-5] To conduct causal inference with DiD, we assume that in the absence of an intervention, the treated groups' average potential outcomes would have evolved in parallel with those of the comparison groups (the "parallel trends assumption").

Most authors employing DiD investigate whether outcomes were parallel prior to the intervention, reasoning that parallel trends in the pre-intervention period increase confidence in the untestable counterfactual parallel trends assumption.[6] We reviewed all 51 DiD papers published in the *Journal of the American Medical Association* (*JAMA*) and *JAMA Internal Medicine* during 2018-2022 and observed two common approaches to evaluating pre-intervention trends (Table S1). First, about half ($n = 27, 53\%$) plotted outcomes in treatment and comparison groups over time and concluded that they looked parallel. For instance, one paper noted, "Visual inspection...revealed small, non-significant differences in pre-intervention episode spending for episodes at [treated] versus comparison hospitals."[7]

Second, following *JAMA* guidance, many papers conducted statistical tests for parallel trends.[8] The most common approach ($n = 18, 35\%$) was to fit a linear regression model, test whether a pre-intervention linear time slope differed between treated and comparison groups, and if $p > 0.05$ for the test, conclude that trends were parallel. For instance, one paper stated, "We directly examined for this possibility by fitting a model containing a treatment indicator, a continuous time variable, the interaction of these 2 variables, and all patient- and hospital-level covariates, restricted to the pre-regulation period...We considered parallel trends as being present if the interaction term from this model was not significant."[9]

A related approach ($n = 9, 18\%$) was to fit an "event study" regression with coefficients for the differential difference between treated and comparison units in each period relative to an omitted reference period. If trends are indeed parallel prior to the intervention, coefficients in the pre-period should all be close to zero and lack a trend. Thus, authors typically evaluate the collection of point estimates and their 95% point-wise confidence intervals. For instance, one paper wrote, "We tested the parallel trends assumption ... through event study analyses, which is recommended when evaluating health policies. ... Small and statistically non-significant estimates before adoption suggest that the parallel trends assumption was satisfied."[10]

Although testing attempts to make evaluation of pre-intervention parallel trends more formal and systematic than visual inspection, there are two key problems with these conventional tests of pre-intervention trends. First, they test a null hypothesis of no difference, thus tightly controlling the probability of falsely declaring non-parallel trends (Type I error) but failing to control the probability of missing non-parallel trends (Type II error). Previous authors have argued that these tests are often under-powered, leading to unwarranted confidence in the parallel trends assumption by conflating low power with no violation of parallel trends.[11;12] Conversely, when the sample size is large, these tests may flag even trivial trend differences as statistically significant.[11;13] Although DiD assumes that counterfactual trends would have been parallel, in practice, we may tolerate trend differences that are "small enough," as determined by context-specific knowledge[14].

Second, using a test for parallel trends as a screening step (i.e., to decide whether to present DiD results) can distort the treatment effect estimates in studies that pass such a test. Previous work has noted that when trends are truly divergent, a testing step may disproportionately admit cases where random error makes the pre-period trend difference unusually small, thereby exacerbating bias in event study coefficients.[13;15]

Our paper makes three main contributions to the literature on evaluating and communicating DiD's sensitivity to the parallel trends assumption. First, we propose a non-inferiority/equivalence approach to testing for parallel trends in the pre-intervention period. This evaluates evidence against the null hypothesis that there are meaningful differences in trends, thus addressing the Type I/Type II error problem.[16;17] Other authors have proposed non-inferiority/equivalence approaches for event studies[18] and for balance and placebo tests in other designs outside of DiD.[19;20] We present a general framework that shifts focus from measuring violations to measuring their impact on the treatment effect: comparing treatment effect estimates from a reduced model (which assumes parallel trends) and an expanded model (which allows for non-parallel trends). This accommodates a range of common expanded models, including both linear time trends[21–23] and event studies.[13;18] It also allows us to specify our non-inferiority/equivalence threshold on the scale of the treatment effect and bound potential bias in our treatment effect estimator. We therefore connect to other sensitivity analysis methods that use observed differences in pre-intervention outcome evolution to construct sensitivity bounds on treatment effect estimates.[14;24] Our approach also shares conceptual underpinnings with e-values, which ask, "How big would a violation of the causal assumption need to be to meaningfully change my effect estimate?"[25]

Second, we extend prior work by establishing conditions under which our testing procedure, if used as a screening step, might introduce bias.[13;18] Leveraging past literature on model selection,[26;27] we show that with i.i.d. normal errors, unit-heteroskedastic normal errors, or time-invariant error correlation, our procedure does not add bias to reduced model treatment effect estimators. Under other error structures (e.g., autocorrelation), we describe conditions under which test-induced bias is small in magnitude.

Third, we examine the power of our procedure and show that when trends are indeed parallel or nearly parallel, we have high power to pass a non-inferiority test when our threshold is anchored to a treatment effect for which the overall study is well-powered.

The rest of the paper proceeds as follows. In Section 2, we introduce our notation, target estimand, estimation procedure, and framework for evaluating pre-intervention parallel trends. We then present a non-inferiority/equivalence testing framework, formulated in terms of reduced and expanded models, and demonstrate how this nests several common approaches, including event study formulations. In Section 3, we characterize conditions under which our estimator introduces no or small bias if employed as a screening step and consider the power of non-inferiority and equivalence tests to detect meaningful violations of parallel trends. Section 4 demonstrates the performance of our approach in a simulation study. Section 5 applies our ideas to re-analyze a study of the Affordable Care Act's effect on dependent insurance coverage rates of young people and presents an empirical simulation based on that application. We conclude in Section 6 with a summary of our findings and recommendations for practice.

2 Testing Framework

2.1 DiD setup, assumptions, and estimator

We begin with a canonical DiD setup, in which we observe $i = 1, \ldots, n$ units during time periods, $t = 1, \ldots, T$, where units are in two groups: a treated group of n_1 units (i.e., the set $\mathcal{N}_1$) drawn from a treated population and a comparison group of n_0 units drawn from a comparison population ($\mathcal{N}_0$). At time T_1, an intervention begins for the treated group only. Let $G_i = \mathbb{I}(i \in \mathcal{N}_1)$ be an indicator of the treated group. Let $Y_{it}(d)$ denote the potential outcome for unit i at time t under treatment condition d, where $d = 0$ indicates no treatment and $d = 1$ indicates treatment. The distinction between the (actual) treatment group and (hypothetical) treatment condition enables

us to consider counterfactual outcomes, such as the untreated potential outcome of a unit in the treated group during the post-intervention period, $Y_{it}(0)$ for $i \in \mathcal{N}_1, t \geq T_1$, where $t \geq T_1$ includes $\{T_1, \ldots, T\}$. Let y_{it} be the realized outcome for unit i at time t.

Our causal target quantity is the average effect of treatment on the treated (ATT) over all the post-intervention periods,

$$ATT = \frac{1}{T - T_1 + 1} \sum_{t=T_1}^{T} \mathbb{E}\left(Y_{it}(1) - Y_{it}(0)|G_i = 1\right) .$$

To identify this, we assume that the expected untreated potential outcomes of the two groups would have evolved in parallel, absent an intervention. Formally,

$$\underbrace{\frac{1}{T_1 - 1} \sum_{t=1}^{T_1 - 1} \left(\mathbb{E}(Y_{it}(0) \mid G_i = 1) - \mathbb{E}(Y_{it}(0) \mid G_i = 0)\right)}_{\text{average pre-intervention difference}} = \underbrace{\frac{1}{T - T_1 + 1} \sum_{t=T_1}^{T} \left(\mathbb{E}(Y_{it}(0) \mid G_i = 1) - \mathbb{E}(Y_{it}(0) \mid G_i = 0)\right)}_{\text{average post-intervention difference, absent treatment}} .$$

We combine this with other standard DiD assumptions: 1) no anticipation, which requires that units do not respond to the intervention before it begins, i.e., $y_{it} = Y_{it}(0) \ \forall \ i$ when $t < T_1$; and 2) the stable unit treatment value assumption, which rules out interference or spillovers and hidden levels of treatment.[2]

Then, we can re-write the identified ATT in terms of observable quantities,

$$ATT = \frac{1}{T - T_1 + 1} \left[\sum_{t=T_1}^{T} \mathbb{E}(y_{it} \mid G_i = 1) - \mathbb{E}(y_{it} \mid G_i = 0) \right] - \frac{1}{T_1 - 1} \left[\sum_{t=1}^{T_1 - 1} \mathbb{E}(y_{it} \mid G_i = 1) - \mathbb{E}(y_{it} \mid G_i = 0) \right] .$$

To estimate this, we could plug in sample averages for each of the expectation terms above. However, in practice, it is more common to estimate the ATT using regression, particularly using a two-way fixed effects (TWFE) estimation approach,

$$y_{it} = \beta G_i \mathbb{I}(t \geq T_1) + \alpha_i + \gamma_t + \epsilon_{it} , \tag{1}$$

where α_i is a unit fixed effect, γ_t is a time fixed effect, and ϵ_{it} is idiosyncratic mean-zero error. The coefficient β from this model corresponds to the ATT identified above.[2]

In this paper, we use a TWFE model that includes a treatment effect at *each* post-intervention

time,

$$y_{it} = \sum_{k=T_1}^{T} \beta_k G_i \mathbb{I}(t = k) + \alpha_i + \gamma_t + \epsilon_{it} . \tag{2}$$

The coefficients β_k capture the difference between treated and comparison groups at each post-intervention time relative to the average difference in the pre-intervention period. The *average* of these coefficients, $\beta = \frac{1}{T-T_1+1} \sum_{k=T_1}^{T} \beta_k$, corresponds to the ATT.[2] The ATTs from Eq. (1) and (2) are equivalent in the balanced panel characterized here and used throughout this paper, but by using the latter, we ensure that expanded models introduced in the next section will identify differential pre-trends using only pre-intervention data.[28]

Recent research has highlighted that the correspondence between the ATT identified by parallel trends and the β coefficient from a TWFE model does not hold when treatment timing is staggered and treatment effects are heterogeneous[29–34] or we condition parallel trends on covariates and include them in the model.[35–37] Nonetheless, we introduce our approach in the simple case where the ATT *can* be estimated with TWFE, then discuss some extensions in later sections (defining additional variables as needed). Next, we turn to assessing whether the parallel trends assumption required for DiD seems plausible in a particular context.

2.2 Traditional versus non-inferiority/equivalence parallel trends tests

As noted in our literature review, biomedical researchers often investigate differential trends by specifying a model that includes a linear time trend difference between the treated and comparison groups and fitting this model only to pre-intervention data. For example, if our analytic model is Eq. (2), we might fit the following model to pre-intervention data:

$$y_{it} = \alpha_i + \gamma_t + \theta G_i t + \epsilon_{it} , \tag{3}$$

The coefficient θ captures the differential trends between groups, and parallel trends in the pre-intervention period imply $\theta = 0$. Thus, a common practice is to test

$$H_0 : \theta = 0 \quad \text{versus} \quad H_A : \theta \neq 0 .$$

If $p > 0.05$ for this test, researchers conclude that trends are parallel and report β estimated with

the model in Eq. (2).[8]

Although the objective is to provide evidence of parallel pre-intervention trends, as discussed above, this conflates a lack of evidence against the null with evidence of parallel trends. However, the challenge of wanting to "prove the null" is not unique to tests of pre-intervention trends in DiD. Drug and device manufacturers often wish to show that their novel treatment is equivalent or non-inferior to the standard of care, and statistical tests have been developed for this purpose.[17] For these tests, we select a threshold δ that represents the maximum difference we can tolerate and then formulate a test for evidence against differences larger than this threshold. In the DiD context, we might test either of the following sets of hypotheses:

$$\text{Non-inferiority: } H_0 : \theta \geq \delta \quad \text{versus} \quad H_A : \theta < \delta$$

$$\text{Equivalence: } H_0 : |\theta| \geq \delta \quad \text{versus} \quad H_A : |\theta| < \delta .$$

This allows us to say whether "large" violations of parallel trends can be ruled out with some level of statistical certainty.[19;38]

Non-inferiority tests use the same test statistics as traditional tests but different cutoffs. For example, with a standard assumption that $\hat{\theta} \sim N\left(\theta, \sigma_{\hat{\theta}}^2\right)$, for a two-sided Wald test, we calculate the test statistic $w = \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}}$. A traditional two-sided test rejects the null if $w > z_{1-\alpha/2}$ or $w < z_{\alpha/2}$. The one-sided non-inferiority test above rejects if $w < z_\alpha + \frac{\delta}{\hat{\sigma}_{\hat{\theta}}}$, and the equivalence test if both $w < z_\alpha + \frac{\delta}{\hat{\sigma}_{\hat{\theta}}}$ and $w > z_{1-\alpha} - \frac{\delta}{\hat{\sigma}_{\hat{\theta}}}$.[17] We can therefore use the standard two-sided 95% confidence interval to understand the values of δ that would lead us to reject the null in non-inferiority tests at the 2.5% level.

However, this approach introduces a new challenge: how do we decide what difference in trends we can tolerate? On the scale of the slope of the differential linear time trend, there is no clear analog of the "clinically meaningful difference" thresholds used in testing drugs and devices. Therefore, we next develop a framework for testing how differential trends impact our treatment effect estimates, which allows us to specify our threshold on the scale of the treatment effect itself.

2.3   Non-inferiority/equivalence tests on the treatment effect scale

We introduce our testing framework on the scale of the treatment effect using a pair of models: one reduced and one expanded. Suppose the reduced model is Eq. (2). One possible expanded

model is

$$y_{it} = \sum_{k=T_1}^{T} \beta_k^{(e)} G_i \mathbb{I}(t = k) + \alpha_i^{(e)} + \gamma_t^{(e)} + \theta G_i t + \epsilon_{it}^{(e)} , \tag{4}$$

where superscript (e)'s distinguish this (e)xpanded model's parameters from those of the reduced model. This model contains a linear trend difference, $\theta G_i t$, as in the model used to test for parallel pre-intervention trends in Eq. (3). However, we now have all the post-period treatment effect coefficients in the model and intend to fit it on all the data. (This will recover an unbiased treatment effect under an assumption of "parallel growth."[21])

Denote the average of the post-period coefficients from this model as $\beta^{(e)} = \frac{1}{T-T_1+1} \sum_{k=T_1}^{T} \beta_k^{(e)}$. To the extent that this differs from β derived from Eq. (2), it is because of non-zero θ; thus, $\beta - \beta^{(e)}$ measures the impact of a differential linear trend on the treatment effect. We can therefore select a threshold that quantifies the maximum difference in the treatment effect that would imply substantive equivalence. That is, we specify our hypotheses as

$$\text{Non-inferiority: } H_0 : \beta - \beta^{(e)} \geq \delta \quad \text{versus} \quad H_A : \beta - \beta^{(e)} < \delta \tag{5}$$

$$\text{Equivalence: } H_0 : \left| \beta - \beta^{(e)} \right| \geq \delta \quad \text{versus} \quad H_A : \left| \beta - \beta^{(e)} \right| < \delta . \tag{6}$$

If we reject the null using this procedure, we have established that non-parallel trends are not likely to meaningfully change our treatment effect. In this case, the difference in treatment effect estimates is simply a scaling of the differential slope $\hat{\theta}$.

Proposition 1 (Reduced vs. expanded model estimators (linear trend difference)). *The difference between ordinary least squares (OLS) ATT estimators corresponding to model specifications in Eqs. (2) and (4) is a linear transformation of the differential trends parameter estimate $\hat{\theta}$ from Eq. (4):*

$$\hat{\beta} - \hat{\beta}^{(e)} = \left(\frac{1}{T - T_1 + 1} \sum_{t=T_1}^{T} t - \frac{1}{T_1 - 1} \sum_{t=1}^{T_1 - 1} t \right) \hat{\theta} = \frac{T}{2} \hat{\theta}. \tag{7}$$

Therefore, the impact of differential linear trends depends on the magnitude of the slope difference $\hat{\theta}$ and the length of the study period. (See derivation in Appendix A.) However, a differential linear trend is not the only way to expand the model to accommodate deviations from parallel trends.

2.4 Reduced/expanded model testing framework

We next generalize our approach to accommodate a broader array of reduced and expanded model specifications. Collect the parameters of the reduced model into the p-vector $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_p \right]'$ and denote the corresponding $nT \times p$ design matrix $\mathbf{X}$. If the reduced model is correctly specified, the ATT of interest is an average of a subset $\mathcal{K}$ of these parameters where $|\mathcal{K}| = K$: $\beta = \frac{1}{K} \sum_{k \in \mathcal{K}} \beta_k$. The TWFE specification in Eq. (2) discussed above will be our main example of a reduced model in this text.

The expanded model includes the same covariates and parameters as the reduced model as well as additional terms. We denote additional parameters in the expanded model as $\boldsymbol{\theta} = \left[\theta_1, \ldots, \theta_q \right]'$, with $\mathbf{Z}$ the corresponding $nT \times q$ design matrix. We can concatenate the design matrix for the expanded model as $\mathbf{X}^{(e)} = \left[\mathbf{X} \quad \mathbf{Z} \right]$. Let $\boldsymbol{\beta}^{(e)} = \left[\beta_1^{(e)}, \ldots, \beta_p^{(e)} \right]'$ indicate parameters corresponding to shared predictors across models. The corresponding ATT of interest for this model is an average of the subset $\mathcal{K}$ of these: $\beta^{(e)} = \frac{1}{K} \sum_{k \in \mathcal{K}} \beta_k^{(e)}$. Finally, collect the outcomes and error terms into nT-vectors $\mathbf{y}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon}^{(e)}$. We can then write reduced and expanded models as

$$\text{Reduced: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{8}$$

$$\text{Expanded: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{(e)} + \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}^{(e)} \tag{9}$$

For instance, to represent the TWFE model in Eq. (2), we set $\boldsymbol{\beta} = \left[\beta_{T_1}, \ldots, \beta_T, \alpha_1, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_T \right]'$. To represent the expanded differential linear trend model in Eq. (4), let $\boldsymbol{\theta} = \theta$, with $\mathbf{Z}$ as an nT-vector of $G_i t$ values (i.e., $q = 1$ additional parameter in the expanded model). We assume that $q \geq 1$ (i.e., the expanded model adds at least one parameter) and both models are identified (i.e., $\left[\mathbf{X} \quad \mathbf{Z} \right]$ has full column rank).

Using this general framework, we avoid having to derive the equivalent of Eq. (7) for each set of models. In the following section, this framework will allow us to characterize test properties. To estimate variance and conduct tests, we assume that the expanded model is correctly specified and that the more restrictive reduced model may be correctly specified only if $\boldsymbol{\theta} = \mathbf{0}$. We briefly discuss implications of expanded model misspecification in Section 3.

2.4.1 Test statistics in reduced/expanded model framework

Assuming a Gaussian error structure, a test comparing coefficients from the reduced and expanded models has the following form.

Proposition 2 (Reduced vs. expanded model estimators (Gaussian errors)). *Assume reduced and expanded models as in Eq. (8) and Eq. (9) and that the expanded model is correctly specified, with* $\boldsymbol{\epsilon}^{(e)} \sim N(0, \boldsymbol{\Omega})$, *where* $\boldsymbol{\Omega}$ *is an* $nT \times nT$ *matrix. Let* $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$ *and denote* $\mathbf{V}^{(e)}$ *analogously for the expanded model. The difference between OLS estimators* $\hat{\beta}_k$ *and* $\hat{\beta}_k^{(e)}$ *is:*

$$\hat{\beta}_k - \hat{\beta}_k^{(e)} \sim N\left(\beta_k - \beta_k^{(e)}, \boldsymbol{\Sigma}_{k,k} + \boldsymbol{\Sigma}_{k,k}^{(e)} - 2\boldsymbol{\Sigma}_{k,k}^*\right),$$

where $\boldsymbol{\Sigma} = \mathbf{V}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\mathbf{V}$, $\boldsymbol{\Sigma}^{(e)} = \mathbf{V}^{(e)}\mathbf{X}^{(e)\prime}\boldsymbol{\Omega}\mathbf{X}^{(e)}\mathbf{V}^{(e)}$, $\boldsymbol{\Sigma}^* = \mathbf{V}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}^{(e)}\mathbf{V}^{(e)}$, *and* $\mathbf{A}_{k,k}$ *indicates the entry in the kth row and kth column of the matrix* $\mathbf{A}$.

The proof, given in Appendix B, uses model comparison methods developed in other contexts to derive the covariance of the difference in these coefficients across model specifications.[26;27] The result allows us to define the following procedure, using OLS and standard normal-based variance estimators:

1. Estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{(e)}$ using OLS and $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Sigma}}^{(e)}$, $\hat{\boldsymbol{\Sigma}}^*$ under chosen error assumptions (using residuals from the expanded model to construct the latter three quantities).

2. Create a linear combination of parameter estimates, $\hat{\beta} - \hat{\beta}^{(e)}$, and the corresponding standard error of the difference.

3. Conduct a non-inferiority Wald test on $\hat{\beta} - \hat{\beta}^{(e)}$, the difference in ATT estimates between reduced and expanded models, using the hypotheses defined in Eq. (5) (for non-inferiority) or Eq. (6) (for equivalence).

We walk through detailed implementation of this procedure, accounting for heteroskedastic errors, clustering, and survey weights, in Appendix B and provide R code for readers (`https://github.com/laura-hatfield/NonInfParTren/`).

We can gain some intuition about this procedure by considering the special case of independent and identically distributed errors.

Proposition 3 (Reduced vs. expanded model estimators (*i.i.d.* errors)). *Assume reduced and expanded models as in Eq. (8) and Eq. (9) and that the expanded model is correctly specified, with* $\epsilon_{it}^{(e)} \overset{i.i.d.}{\sim} N\left(0, \sigma_{(e)}^2\right)$. *Recall that* $\beta = \frac{1}{K}\sum_{k \in \mathcal{K}} \beta_k$ *and* $\beta^{(e)} = \frac{1}{K}\sum_{k \in \mathcal{K}} \beta_k^{(e)}$ *are the parameters of interest. The difference between the corresponding OLS ATT estimators is:*

$$\hat{\beta} - \hat{\beta}^{(e)} \sim N\left(\beta - \beta^{(e)}, \sigma_{\hat{\beta}^{(e)}}^2 - \sigma_{\hat{\beta}}^2\right), \tag{10}$$

where $\sigma_{\hat{\beta}^{(e)}}^2$ *is the variance of* $\hat{\beta}^{(e)}$ *(corresponding to the expanded model), and* $\sigma_{\hat{\beta}}^2$ *is the variance of* $\hat{\beta}$ *(corresponding to the potentially misspecified reduced model but defined using common error variance,* $\sigma_{(e)}^2$*).*

The proof, in Appendix B, illustrates how the variance of the difference between $\hat{\beta}$ and $\hat{\beta}^{(e)}$ is less than the variance of $\hat{\beta}^{(e)}$. This occurs because $Var(\hat{\beta}) = Cov\left(\hat{\beta}, \hat{\beta}^{(e)}\right)$, and thus testing the difference between treatment effect estimates subtracts off the shared component of the two models, thereby reducing the variance compared to estimating the treatment effect in the expanded model.

In the special i.i.d. case, the lack of covariance between the difference $\hat{\beta} - \hat{\beta}^{(e)}$ and the estimated coefficient from the reduced model $\hat{\beta}$ implies that conditioning on $\hat{\beta} - \hat{\beta}^{(e)}$ will not bias the reduced model estimator. We build on this in the following section, discussing the potential of a non-inferiority/equivalence test to induce bias in ATT estimators. We also discuss the relationship between the power of the overall study and the power of these non-inferiority/equivalence tests.

Overall, this test procedure resembles a Hausman specification test, except that the Hausman test uses a null hypothesis of a correct reduced model specification (rather than a null that only the expanded model is correctly specified).[26;39]

2.5 Functional form selection

Within this reduced/expanded framework, researchers have broad latitude to select functional form. Table 1 gives several examples, using the TWFE model in Eq. (2) as the reduced model and several possible expanded models.

2.5.1 Linear time trends

Three specifications in Table 1 add differential linear time trends. The first is simply Eq. (4), which allows treated groups to have a differential linear time trend with slope θ. A second model allows each unit its own linear trend, θ_i. (In a balanced panel, this produces the same overall ATT estimate as the first.) The third allows differential linear time trends in groups of units defined by something other than treatment; units with the same value of a covariate (e.g., rural counties) have slope $\theta_{\ell(i)}$ for groups $\ell = 1, \ldots, L$.

2.5.2 Event studies

Another popular approach to testing parallel trends involves event study models, also shown in Table 1. Adding to our reduced model treatment group-specific time fixed effects θ_k for $k = 1, ..., T_1 - 2$ at each pre-intervention time relative to the final pre-intervention time yields the conventional event study specification:

$$\textbf{Expanded: } y_{it} = \sum_{k=T_1}^{T} \beta_k^{(e)} G_i \mathbb{I}(t = k) + \alpha_i + \gamma_t^{(e)} + \sum_{k=1}^{T_1-2} \theta_k G_i \mathbb{I}(t = k) + \epsilon_{it}^{(e)} . \qquad (11)$$

If the parallel trends assumption holds exactly, all the θ_k will be zero. As noted in our literature review, it is common to conclude that trends are parallel if the confidence intervals all cover zero. Joint F-tests may offer further formalization of this test, evaluating a collection of θ_k.

However, to adopt a non-inferiority testing framework on the scale of the treatment effect, we must clarify how we believe the pre-intervention coefficients inform us about parallel outcome evolution into the post-intervention period, absent intervention.[14;18] Models that add differential linear trends extrapolate pre-intervention trend differences into the post-intervention period, which is why their impact on the treatment effect depends on both the magnitude of differential slopes and the length of the study period (Proposition 1). The situation is more complicated for event study models. If we believe the pre-period θ_k represent transient shocks, we might choose δ to be the largest treatment effect impact we can tolerate and test $H_0 : \theta_k \geq \delta$. For instance, Rambachan and Roth[14] use the maximum pre-period differential change to construct sensitivity bounds, and Dette and Schumann[18] suggest formulating a non-inferiority test on the maximum or mean of the pre-intervention coefficients. The reasoning is that we are looking for evidence against shocks large enough to substantially impact our treatment effect, if they were to strike in the post-intervention

period. If we instead believe the pre-period θ_k represent single-period differential trends, we might formulate our test based on extrapolating these trend differences into the post-intervention period (as in Proposition 1 and Eq. (7)). In this case, we are looking for evidence against differential trends large enough to impact our treatment effect, if they were to persist into the post-intervention period.

Tests on specific pre-intervention coefficients and tests on extrapolations into the post-intervention period can both be cast in terms of reduced vs expanded model effect estimates as in Proposition 2. For example, suppose we want to test the average pre-intervention coefficient $\frac{1}{T_1-2}\sum_{k=1}^{T_1-2}\theta_k$.[18] If we test $H_0 : \beta - \beta^{(e)} \geq \delta$, where β is from the reduced model in Eq. (2) and $\beta^{(e)}$ is from the expanded event study model in Eq. (11), this is equivalent to evaluating $H_0 : \frac{1}{T_1-1}\sum_{k=1}^{T_1-2}\hat{\theta}_k \leq -\delta$, which can be scaled to conduct our test of interest (see Appendix C for derivation). In this case, $\beta^{(e)}$ is estimated relative only to the reference period, rather than to the average of all pre-intervention periods. The choice of δ determines whether and how a trend difference is assumed to be extrapolated into the post-intervention period. In Appendix C, we also show other tests of event study coefficients can be estimated in the form required by Proposition 2. This crosswalk will be useful for understanding the implications of conditioning estimates on event study non-inferiority tests in the following section.

2.5.3 Other specifications

As alternative expanded specifications, we could add differential time fixed effects for units with the same covariate value/cluster membership or some combination of differential linear time trends and differential time fixed effects.[40] Another strategy assumes that trends are parallel only for units with the same covariate values (i.e., conditional parallel trends).[29] To the extent that these can be encoded in an expanded model using regression, they are further examples of our reduced/expanded model testing framework. However, in addition to regression-based estimators for conditional parallel trends,[37] other popular techniques use propensity score weighting or doubly robust estimators; the latter are beyond the scope of this paper.[4;41]

3 Post-test bias and power

3.1 Post-selection estimation and inference

Do we abandon our DiD study if we fail a test of parallel trends? Usually not. Rather, we may selectively report results if trends are "parallel enough" or present expanded model results if not. However, previous work has highlighted that distortions may arise from using tests of parallel pre-intervention trends as such a screening step, specifically when presenting effect estimates from event study models.[13] We build on these results to show that non-inferiority/equivalence tests, when used as a screening step, may introduce no or minimal bias in the reduced model estimator but substantial bias in the expanded model estimator. In the following sections, we show these results in detail.

3.1.1 Reduced model

Distortions in the reduced model estimator depend on the covariance between the reduced model treatment effect estimator and the difference between reduced and expanded model treatment effect estimators. We first consider the case of no covariance (and thus, no distortion), then expand to more general cases.

Given that the expanded model may also be misspecified, we introduce an explicit correct model specification. Following the pattern above, let $\mathbf{fl} = \left[\gamma_1, \ldots, \gamma_r\right]'$ denote the additional parameters and $\mathbf{W}$ their corresponding design matrix. Then the correct model is,

$$\textbf{Correct:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{(w)} + \mathbf{Z}\boldsymbol{\theta}^{(w)} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{(w)} \, , \tag{12}$$

which we use to define the "no covariance condition."

Assumption 1 (No covariance condition)**.** Assume that the correct model specification follows Eq. (12), with $\boldsymbol{\epsilon}^{(w)} \sim N(0, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is an $nT \times nT$ matrix, and that the reduced and expanded models are specified as in Eq. (8) and Eq. (9), with $\hat{\beta}$ and $\hat{\beta}^{(e)}$ denoting corresponding OLS ATT estimators. Further assume that the combination of error structure and model specifications yields $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$.

Assumption 1 holds straightforwardly in the i.i.d. case when the expanded model is correctly specified, following Proposition 3. In Appendix D, we also show that the expanded model need not

be correctly specified for Assumption 1 to hold in the i.i.d case. We also characterize conditions under which Assumption 1 holds in the generalized error setup from Proposition 2. For example, Assumption 1 holds for linear and event study tests if errors are independent across units, but heteroskedastic (i.e., $\epsilon_{it} \sim N(0, \sigma_i^2)$ with all ϵ_{it} independent) or error correlation is constant within units. In these cases as well, the expanded model need not be correctly specified for Assumption 1 to hold. When Assumption 1 holds, conditioning on testing does not induce bias in the reduced model estimator.

Proposition 4 (Reduced model test-induced distortions under Assumption 1). *Under Assumption 1, if we conduct a non-inferiority test with a threshold δ, rejecting the null if $\hat{\beta} - \hat{\beta}^{(e)} < \delta^*$, where $\delta^* = z_\alpha \sigma_{\hat{\beta} - \hat{\beta}^{(e)}} + \delta$, then there is no distortion in the reduced model ATT induced by conditioning on the test result:*

$$\mathbb{E}\left(\hat{\beta} \middle| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) - \mathbb{E}\left(\hat{\beta}\right) = 0$$

and likewise,

$$Var\left(\hat{\beta} \middle| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) - Var\left(\hat{\beta}\right) = 0.$$

The proof, given in Appendix E, derives from the assumption that the covariance between $\hat{\beta}$ and $\hat{\beta} - \hat{\beta}^{(e)}$ is 0, and for multivariate normally distributed variables, zero covariance implies independence. In Appendix E, we extend this result to show that it holds when we formulate the test as an equivalence test (Corollary 2).

Note that this does not imply that $\hat{\beta}$ is unbiased, only that conditioning whether to report $\hat{\beta}$ on a non-inferiority test does not add bias, and that results of the non-inferiority test still bound its bias. Therefore, if we pass a non-inferiority test, it is appropriate to present $\hat{\beta}$ from the reduced model and the results of the test of $\hat{\beta} - \hat{\beta}^{(e)}$ as a bound on the bias. These results do not require the threshold to be pre-specified or explicit, and testing does not introduce distortions in variance.

However, there may be test-induced bias in other circumstances, particularly when errors are autocorrelated over time. We formalize this in the following proposition.

Proposition 5 (Reduced model test-induced bias). *Assume setup and reduced and expanded model estimators as in Proposition 2. If we conduct a non-inferiority test with a threshold δ, rejecting the null if $\hat{\beta} - \hat{\beta}^{(e)} < \delta^*$, where $\delta^* = z_\alpha \sigma_{\hat{\beta} - \hat{\beta}^{(e)}} + \delta$, then the expectation of the reduced model ATT*

estimator conditional on passing the test may differ from its unconditional expectation:

$$\mathbb{E}\left(\hat{\beta}\middle|\hat{\beta}-\hat{\beta}^{(e)}<\delta^*\right)-\mathbb{E}\left(\hat{\beta}\right)=-\frac{Cov\left(\hat{\beta},\hat{\beta}-\hat{\beta}^{(e)}\right)}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\frac{\phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}{\Phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)},$$

where ϕ and Φ are the probability density function and cumulative distribution function of a standard normal, respectively.

Proposition 5 indicates that test-induced distortions depend on the covariance between the reduced model estimator and the difference estimator as well as the choice of threshold (proof in Appendix E). Still, even with error autocorrelation producing non-zero $Cov\left(\hat{\beta},\hat{\beta}-\hat{\beta}^{(e)}\right)$, bias may be small if the covariance term is small and/or thresholds are not too strict. It may also be conservative under positive error autocorrelation with positive treatment effects. We will explore this further via simulation in Section 4.

Last, note that conditioning on the test outcome leads to a smaller variance for $\hat{\beta}$. The variance of a truncated distribution is necessarily smaller than that of the original distribution. Although the truncation applies to the distribution of $\hat{\beta}-\hat{\beta}^{(e)}$, the joint normal relationship between this and $\hat{\beta}$ implies that the truncation also reduces the conditional variance of $\hat{\beta}$, following from standard properties of the truncated multivariate normal distribution[13].

3.1.2 Expanded model

By contrast, testing may introduce bias in the expanded model estimator, even when it does not induce bias in the reduced model estimator. This occurs in part because, although the formula of the bias appears similar to that above, the covariance between $\hat{\beta}^{(e)}$ and $\hat{\beta}-\hat{\beta}^{(e)}$ may be sufficiently large to produce meaningful distortions, even when the covariance between $\hat{\beta}$ and $\hat{\beta}-\hat{\beta}^{(e)}$ is 0. Formally, we quantify test-induced bias:

Proposition 6 (Expanded model test-induced bias). *Assume setup and reduced and expanded model estimators as in Proposition 2. If we conduct a non-inferiority test with a threshold δ, rejecting the null if $\hat{\beta}-\hat{\beta}^{(e)}<\delta^*$, where $\delta^*=z_\alpha\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}+\delta$, then the expectation of the expanded*

model ATT estimator conditional on passing the test may differ from its unconditional expectation:

$$\mathbb{E}\left(\hat{\beta}^{(e)}\Big|\hat{\beta}-\hat{\beta}^{(e)}<\delta^*\right) - \mathbb{E}\left(\hat{\beta}^{(e)}\right) = -\frac{Cov\left(\hat{\beta}^{(e)},\hat{\beta}-\hat{\beta}^{(e)}\right)}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\frac{\phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}{\Phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}.$$

The proof is provided in Appendix E, and we also illustrate this phenomenon in simulations in Section 4. Similar to the prior proposition, the expression above characterizes the case when expanded model estimates (e.g., an event study) are presented after passing a test. In this case, bias declines with a larger magnitude of δ. An analogous distortion occurs if expanded model results are reported only after failing a test (e.g., adding differential linear trends), but with an opposite impact of threshold stringency. (Presenting expanded model results after failing a test would also necessitate assessing plausibility of a different identifying assumption; we leave further consideration of this to future work.)

This result suggests that although an expanded model may reduce bias by adjusting for non-parallel trends, these reductions in bias may be offset by test-induced bias. Indeed, we will see such scenarios in our simulation study (Section 4). Overall, we may be wary of reporting the results from the expanded model after conducting a test. As above, when distortions in variance occur from a testing procedure, using unconditional variance would remain conservative.

3.2   Power of non-inferiority tests

Recall that we pass a traditional test when we have high uncertainty (i.e., low power). Thus, we may worry that switching to a non-inferiority framework will make passing too difficult. In this section, we characterize the power of our tests. We consider the power of a non-inferiority test when trends truly are parallel (i.e., $\boldsymbol{\theta} = \mathbf{0}$) and compare this test's power to that of a treatment effect test under the expanded model in the simple case of i.i.d. errors. We will explore power in more complex scenarios in simulations in Section 4.

Proposition 7 (Non-inferiority difference-in-differences power). *Assume setup and reduced and expanded model estimators as in Proposition 3. If a non-inferiority test has power p evaluated under an asymptotic normal approximation to rule out violations of parallel trends equal to or larger than β_k^* (i.e., to reject $H_0 : \beta_k - \beta_k^{(e)} \geq \beta_k^*$, with $\beta_k^* > 0$) in a Wald test at level α, and*

assuming no violation exists ($\boldsymbol{\theta} = \mathbf{0}$), then $p > p_e$, where p_e is the power to detect $\beta_k^{(e)} = \beta_k^*$ (likewise evaluated) in a one-sided Wald test at level α in an expanded model.

In other words, the power of a non-inferiority test with a given threshold will exceed power to detect an effect as large as the threshold with an expanded model (proof in Appendix F). We also show in Appendix F that this may exceed power to detect an effect in the reduced model. This does not imply that non-inferiority tests and tests of treatment effects are equivalent tasks.

Rather, this comparison helps contextualize power and illustrates that non-inferiority tests may still provide useful information even when the expanded model is under-powered for a treatment effect the size of the chosen threshold.

Power declines as the threshold becomes stricter, as we are trying to rule out smaller (harder to detect) violations. It is also lower for equivalence formulations (see Appendix F).[17;42] Last, power to rule out trend differences exceeding a threshold decreases as the true trends diverge, and violations approach the threshold value.

4   Simulations

4.1   Methods

We conducted a simulation study to demonstrate the empirical performance of non-inferiority and equivalence tests. Using Eq. (4), we generated observations for 60 units at $T = 25$ time points, assigning half to treatment beginning at $T_1 = 21$ (i.e., 5 treated periods). We drew unit fixed effects from a standard normal distribution and errors from normal distributions with two different variance structures. The first (denoted "Independent, heteroskedastic") assumed errors were independent across the 25 time points but each unit had its own variance parameter. The second (denoted "AR(1), $\rho = 0.2$") assumed errors were independent across units but followed a first-order autoregressive process with correlation parameter $\rho = 0.2$ across time within each unit.

We assumed a constant treatment effect and included scenarios either with no violation of parallel trends (i.e., $\theta = 0$) or a linear violation of parallel trends that increased the expected value of the reduced model treatment effect by 50%. For ease of interpretation, we scaled the treatment effect and violation in terms of "TX80," the effect size that a TWFE estimator had 80% power to detect under no violation of parallel trends.

For each scenario, we fit the reduced model in Eq. (2) and compared it to two expanded models. One included a linear trend difference, as in Eq. (4), and the other used an event study specification, as in Eq. (11). We estimated the treatment effect using all three models and tested for a violation by comparing the reduced vs. expanded model results with both non-inferiority and equivalence tests. For these tests, we considered one threshold that was more lax (100% of TX80) and one that was stricter (20% of TX80). Across 100,000 replications, we present results in terms of empirical power/Type I error (percentage of treatment effect tests and non-inferiority/equivalence tests meeting statistical significance across different violation scenarios) and bias (both in estimated treatment effects and incremental bias from conditioning on tests).

4.2   Results

The results of the simulation study are summarized in Table 2. Non-inferiority tests strictly controlled the probability of missing a violation that exceeded the threshold (i.e., making a Type I error). In scenarios with a violation of 50% TX80 and a threshold of 20% TX80, the probability of passing a non-inferiority or equivalence test remained below 5%. (See the "Power" results in the columns labeled "NI" and "EQ" in scenarios marked with double daggers.)

When the true violation was smaller than the threshold (as in the scenarios with a violation of 50% TX80 and a threshold of 100% TX80 or no violation), power depended on the strictness of the threshold. For example, in the scenario with no violation and independent, heteroskedastic errors, the probability of passing a non-inferiority test based on a linear expanded model declined from 88% at the more lax threshold (100% TX80) to 12% at the stricter threshold (20% TX80). However, at the more lax threshold, the power of the non-inferiority test was greater than that of the expanded model (52%) to detect the treatment effect. Such differences were more pronounced for expanded models with linear differential trends than event study models.

For models with independent, heteroskedastic errors, there was no test-induced bias in the reduced model treatment effect estimator. However, conditional on passing a test, there could be substantial test-induced bias in results from the expanded model (range: 8-157%), which worsened with a stricter threshold. There was also test-induced bias in the expanded model estimator conditional on not passing, though opposite in sign and larger at more lenient thresholds. In several scenarios, test-induced bias exceeded misspecification bias in the reduced model estimator, meaning that after conditioning on the test result, the expanded model estimator was more biased

than the reduced model estimator.

With autocorrelated errors, we observed some test-induced bias in the reduced model treatment effect estimator conditional on passing a test, ranging from 0 to -2% of TX80 for linear models and -3 to -9% for event study models and increasing with stricter thresholds. Nevertheless, test-induced bias in the reduced model estimator was still smaller in magnitude than that in the expanded model estimator conditional on passing (range 8%-128%). Conditional on not passing, there was less test-induced bias in the expanded model estimator only when the violation substantially exceeded the threshold (and thus the test rarely passed, meaning the conditioning event almost always occurred).

5 Application to the ACA Dependent Coverage Provision

We applied our approach to re-analyze the impact of the United States (US) 2010 Patient Protection and Affordable Care Act (ACA) on young adults' health insurance coverage. The ACA was enacted March 23, 2010 by the US Congress. Beginning September 23, 2010, it required commercial insurers to offer health insurance coverage to dependents up to age 26 on a parent or guardian's plan. Previously, the federal rules only required coverage on a parent's plan up to age 18. As a result, nearly a third of people aged 19-25 went without health insurance.[43] Several studies have assessed the impact of this provision using DiD, comparing coverage among people newly eligible to join parents' plans to coverage among other young people not affected by the policy change.[44–47]

5.1 Methods

5.1.1 Data

We used replication code and data provided by authors of one of these studies.[44] From their extract of the Survey of Income and Program Participation (SIPP), we used monthly self-reported insurance coverage in the following categories: any insurance, dependent coverage on a parent's plan, employer-sponsored insurance for oneself, privately purchased individual insurance, and government insurance (e.g., Medicare, Medicaid). As in the original analysis, we compared coverage among people aged 19-25 years (the treated group) to that of people aged 16-18 and 27-29 years (the comparison groups), omitting people aged 26 years because of their ambiguous treatment status. The study period was Aug 2008 through Nov 2011.

5.1.2 Models

As in the original authors' specification, we first fit a linear probability model for each insurance type separately,

$$\textbf{Reduced: } y_{it} = \sum_{k=20}^{40} \beta_k G_i \mathbb{I}(t=k) + \mathbf{X}_{it}\boldsymbol{\beta}_X + \alpha_{s(i)} + \gamma_t + \epsilon_{it} \, , \tag{13}$$

where y_{it} was a binary indicator for person i having coverage at month t, G_i was a binary indicator for the treated group (i.e., age 19-25), $\mathbf{X}_{it}$ was a row vector of covariates, and $s(i)$ indexed the ith person's state. The vector of covariates $\mathbf{X}_{it}$ contained age (categorical, encoding treatment status), gender (binary), race/ethnicity (categorical), marital status (binary), student status (binary), household income as a proportion of the federal poverty line (continuous), and squared household income as a proportion of the federal poverty line.

We next fit an expanded model with a differential linear time trend in the treated group $\theta G_i t$:

$$\textbf{Expanded: } y_{it} = \sum_{k=20}^{40} \beta_k^{(e)} G_i \mathbb{I}(t=k) + \mathbf{X}_{it}\boldsymbol{\beta}_X^{(e)} + \alpha_{s(i)}^{(e)} + \gamma_t^{(e)} + \theta G_i t + \epsilon_{it}^{(e)} \tag{14}$$

From both the reduced and expanded models, we focused on implementation effects, $\beta = \frac{1}{14}\sum_{k=27}^{40} \beta_k$ and $\beta^{(e)} = \frac{1}{14}\sum_{k=27}^{40} \beta_k^{(e)}$, which average the coefficients from implementation (Oct 2010) to the end of the study (Nov 2011). To determine a non-inferiority threshold for the difference in the estimated effects from the two models, we use estimated treatment effects from an earlier analysis of the ACA's dependent coverage provision.[47] That study's outcomes were slightly different, but the significant results ranged from a 2.1 percentage point decrease in "private, self or spouse" coverage to a 5.3 percentage point increase in "private, non-spouse dependent". Thus, we sought to rule out changes greater than or equal to $|2.1|$ or $|5.3|$. For the stricter threshold, we also considered a non-inferiority benchmark, denoted 2.1^*, which adjusted non-inferiority tests based on the sign of the expected treatment effect (i.e., ruling out violations > 2.1 for any health insurance and dependent coverage and < -2.1 for others).

Following the original authors, we used normal-based robust standard errors, clustered at the state level, and weighted regression with person weights from SIPP for all the models. In tests of the difference between the treatment effects from the reduced and expanded models, we implemented the inference described in Proposition 2.

Finally, we also replicated the original authors' tests for parallel pre-intervention trends by fitting a model with differential linear trends to data from the pre-period only, that is, from Aug 2008 to just before enactment in Feb 2010:

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta}_X + \alpha_{s(i)} + \gamma_t + \theta G_i t + \epsilon_{it} \tag{15}$$

Two differences from the authors' original model merit mention. First, the original model used interactions between treatment group and three time periods (pre-ACA, enactment, and implementation), whereas we saturated the model with coefficients for the treated group in each post-enactment month β_k and added corresponding month-year fixed effects (therefore omitting linear time trends, which had minimal impact per Tables S3 and S4). We used this saturated specification to avoid fitting a pre-intervention trend to time heterogeneity in treatment effects; as the authors only fit differential trends by treatment status to pre-intervention data, they avoided this concern. The authors' original specification also included an interaction term between state-month unemployment (continuous) and treatment group. However, we omitted this variable because of its collinearity with differential linear time trends in our reduced vs expanded testing strategy (see Appendix G).

5.1.3 Empirical simulations

To understand the magnitude of violations we could have ruled out, as well as potential distortions introduced by testing, we also generated simulations based on the SIPP data. We first fit the expanded model in Eq. (14) on the any health insurance outcome. Setting $\theta = 0$, we used the fitted model to generate predicted values, $\hat{p}_{it}$, and residuals, $\hat{u}_{it}$. We considered two data-generating processes:

1. *Normal with heteroskedastic errors across state clusters:* We estimated sample residual variance, $\hat{\sigma}_s$, by state and then simulated synthetic datasets as follows:

$$y_{it}^{sim} = \hat{p}_{it} + \theta G_i t + \epsilon_{it}^{sim},$$
$$\epsilon_{it}^{sim} \overset{ind}{\sim} N\left(0, \hat{\sigma}_{s(i)}^2\right)$$

   where $s(i)$ indicated the ith person's state. We considered θ corresponding to either no

violation or a linear violation approximately a third of the treatment effect.

2. *Cluster-resampling wild bootstrap:* For simulations that preserved intra-cluster correlation, both across individuals and over time, we bootstrapped the data set by sampling with replacement at the state cluster level. For each state draw, we then set the values of the component units:

$$y_{it}^{sim} = \hat{p}_{it} + \theta G_i t + q_{s(i)} \hat{u}_{s(i)t},$$

where q_s was a cluster-level random variable taking 1 with probability 0.5 and -1 with probability 0.5 (i.e., randomly flipping the signs of the errors by cluster). We considered the same θ as above.

We ran 15,000 replications of each scenario, each time generating data from the specified data-generating process and fitting the reduced (Eq. (13)) and expanded (linear trend difference, Eq. (14)) models. Following the estimation and testing workflow of our earlier simulations, we estimated the treatment effect in each and tested for the violation (reduced vs. expanded) using both non-inferiority and equivalence tests, with thresholds of $\{1^*, 2.1^*, |5|\}$, with 1^* added for an even more stringent threshold. We calculated power/Type I error of non-inferiority/equivalence tests and bias (in estimated treatment effects and incremental bias from conditioning on tests).

5.2 Results

5.2.1 Parallel trends tests

Akosa Antwi and co-authors plotted the proportion with any insurance in treated and comparison groups over time (Figure 1 of their paper) and reported "generally a similar pattern prior to the ACA passage." [44] To extend the visual investigation of pre-trends, we plotted differences in coverage between treated and comparison groups for each of the insurance coverage outcomes (Figures S2 and S3). Table S2 replicates the pre-intervention trend tests from the original paper's Appendix Table A1, showing no statistically significant non-parallel trends. However, absent the interaction between unemployment and treatment group, there were statistically significant differential trends in the dependent coverage outcome in the pre-period (Figure S4).

5.2.2 Treatment effects

The original authors found statistically significant increases in any (+3.2%) and dependent (+7.0%) coverage, significant decreases in employer (−3.1%) and individual (−0.8%) coverage, and a small, non-significant decrease in government coverage (−0.3%) (see their Table 2).[44] We replicated those results in Tables S3 and S4.

Our model formulation yielded very similar results, shown in Table 3 in rows with Model type "Original". Adding differential linear trends to this model (rows with Model type "+ trend") substantially reduced the estimated impacts on dependent, employer, and individual outcomes, and made the nearly zero effect on government coverage more positive (though not statistically significant). Applying our non-inferiority approach, all outcomes except dependent coverage passed at the most generous |5.3| threshold. Both any health insurance and individual coverage passed at the 2.1* threshold, and at the strictest |2.1| threshold, only individual coverage passed. With the inclusion of the unemployment/treatment interaction, the treatment effect estimates from the expanded model had substantially higher variance, and only 3 of the prior 7 tests passed (Table 3).

5.3 Empirical simulation results

In simulations, we found high power to rule out violations greater than |5| (88-100%) and low-to-moderate power to rule out those greater than 2.1 (21-60%) (Table 4). Non-inferiority tests controlled type I error at approximately 5% for simulations where violations exceeded the threshold. We observed $\leq$—1%— incremental test-induced bias in reduced model treatment effect estimators for scenarios with independent errors and $\leq$—6%— in bootstrap-based simulations that preserved the cluster correlation structure. For expanded model estimators, test-induced bias could be substantial, up to 80%.

6 Discussion

Conventional guidance in the medical literature suggests testing a null hypothesis of no differential pre-intervention trend when conducting DiD.[8] This practice obscures meaningful violations by tightly controlling Type I error, the probability of incorrectly detecting violations, rather than Type II error, the probability of missing them. By contrast, our non-inferiority/equivalence

framework informs researchers about the magnitude of violations (and their impacts on treatment effects) that can be ruled out. Our general framework for testing reduced versus expanded models enables flexible relaxations of the parallel trends assumption, allowing us to specify the non-inferiority/equivalence threshold on the scale of the treatment effect itself. We can even avoid committing to a single threshold by using the relationship between hypothesis tests and confidence intervals to rule out values outside a 95% confidence interval. We characterized conditions under which our procedure, if used as a pre-screening step, may introduce no or minimal bias in the reduced model ATT estimator. Finally, we showed that our strategy may have higher power than tests of treatment effects in the expanded model under no violation.

6.1 Implementing our framework

For applied DiD studies, we therefore recommend the following implementation of our framework:

1. Specify models.

   a. The reduced model should encode plausible causal assumptions, such as parallel trends, and encode a treatment effect β.

   b. The expanded model should encode plausible relaxations of the assumptions and encode an analogous treatment effect $\beta^{(e)}$.

   Table 1 suggests a variety of expanded models using familiar models from the literature.

2. Perform a non-inferiority/equivalence test.

   a. Using the study's power and notes below as a guide, choose a threshold δ for the largest change in the treatment effect that would still imply substantive equivalence. Alternatively, construct a 95% confidence interval around $\hat{\beta} - \hat{\beta}^{(e)}$ and present values outside the range as "ruled out" impacts on the treatment effect.

   b. Using results from Proposition 2 or 3, conduct a test that accounts for across-model dependence in the parameters, or examine the bounds of the corresponding confidence interval.

      i. If the test rejects the null (or the ruled-out range is sufficient), present $\hat{\beta}$ and the 95% CI on $\hat{\beta} - \hat{\beta}^{(e)}$ as a bound on the bias, evaluating and noting risk of model selection-related distortions.

ii. If the test fails to reject the null, reconsider the design, comparison group, and/or model specification, evaluating and noting risk of model selection-related distortions.

6.2 Threshold choice

As noted above, we can avoid committing to a single value of δ by using the relationship between hypothesis tests and confidence intervals to determine the range of impacts that we can rule out. The 95% confidence interval bounds on $\hat{\beta} - \hat{\beta}^{(e)}$ represent the magnitudes that can be ruled out at a 2.5% error rate. For example, we could interpret a 95% confidence interval on $\hat{\beta} - \hat{\beta}^{(e)}$ of $(-1, 4)$ as follows: "We can rule out (at the 2.5% level) violations that would reduce the treatment effect by more than 1 or increase it by more than 4." This is the procedure recommended by Hartman and Hidalgo in the related setting of wanting to provide evidence to support the appropriateness of a regression discontinuity design.[20]

However, even if using a confidence interval, researchers will benefit from context to help guide their evaluation of this range. Other researchers may prefer to test a specific threshold δ, which introduces the challenge of selecting this value. We can use the power of the overall study to help determine a threshold. For instance, we might say that we wish to rule out differential trends that would change our treatment effect by some fraction of the effect size our study is powered to detect. In our simulation study, we found that when using a threshold equal to the treatment effect for which the reduced model had 80% power (i.e., lines with Threshold=100 in Table 2), non-inferiority tests had moderate-to-high power when trends were truly parallel. Even in the presence of a violation equal to half the treatment effect for which we were powered, both non-inferiority and equivalence tests with a linear trend expanded model retained reasonable power. However, the power of non-inferiority and especially of equivalence tests was reduced when the expanded model was an event study specification.

Another strategy to inform our threshold selection is to examine related effect estimates from the literature, as we did in our applied analysis in Section 5. We can take advantage of the greater power of non-inferiority tests compared to equivalence formulations if we know the plausible direction and magnitude of the treatment effect. Suppose we have an expected effect of δ^* (e.g., based on previous studies) and we are most concerned with differential trends that would lead us to erroneously conclude that the intervention had a true effect. Then a non-inferiority test of $H_0 : \beta - \beta^{(e)} \geq \delta^*$ if $\delta^* > 0$ (or $H_0 : \beta - \beta^{(e)} \leq \delta^*$ if $\delta^* < 0$) can rule out differential trends that

would lead us to estimate a treatment effect of δ^* in the reduced model when there is truly none (in the expanded model).

6.3 Limitations

Our proposed approach has several limitations. Although we show good statistical properties for reduced model estimates following a test, this only enables researchers to bound bias under specific assumptions, not guarantee unbiased effect estimates. Our approach also requires researchers to impose parametric functional form restrictions on trend differences. Furthermore, when conditioning on test results, there is a risk of distortions beyond what we explored. For example, highly autocorrelated data, observed in some contexts[48], could drive instances of significant test-related bias in reduced model specifications. This may merit further study and exploration in specific applications. Conducting multiple tests (e.g., on event study coefficients) could also inflate false discovery rates, for which researchers could apply standard corrections. In addition, although we give bias results for the impact of pre-test conditioning, we have not provided results for variance outside of special ("no covariance condition") cases, noting only that any distortions from using unconditional variance estimates would be conservative. We leave this problem to future work. Last, for equivalence tests, our Wald-based approach is conservative;[17] future work could extend recent innovations that improve equivalence test power to the DiD context.[42]

Despite these, we believe this non-inferiority formulation is practical for a wide range of clinical and policy applications. It uses the familiar tools of regression and testing while avoiding the pitfalls of conventional parallel trends testing by providing more transparent bounds on the likely impacts of non-parallel trends on DiD estimates.

Conflicts of interest

The authors declare no potential conflict of interests.

Supporting information

Appendices, Tables, and Figures referenced in Sections 2-5 are available with this paper.

References

[1] Eli Ben-Michael, Avi Feller, and Elizabeth A. Stuart. A Trial Emulation Approach for Policy Evaluations with Group-Level Longitudinal Data. *Epidemiology*, 32(4):533–540, July 2021. ISSN 1531-5487. doi: 10.1097/EDE.0000000000001369.

[2] Michael Lechner. The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics*, 4(3):165–224, 2010. ISSN 1551-3076, 1551-3084. doi: 10.1561/0800000014. URL http://www.nowpublishers.com/article/Details/ECO-014.

[3] Audrey Renson, Michael G. Hudgens, Alexander P. Keil, Paul N. Zivich, and Allison E. Aiello. Identifying and Estimating Effects of Sustained Interventions Under Parallel Trends Assumptions. *Biometrics*, (Early View)(n/a), 2023. ISSN 1541-0420. doi: 10.1111/biom.13862.

[4] Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*, 235(2):2218–2244, August 2023. ISSN 0304-4076. doi: 10.1016/j.jeconom.2023.03.008. URL https://www.sciencedirect.com/science/article/pii/S0304407623001318.

[5] Eric Tchetgen Tchetgen, Chan Park, and David Richardson. Universal Difference-in-Differences for Causal Inference in Epidemiology, February 2023. arXiv:2302.00840.

[6] Andrew M Ryan, Evangelos Kontopantelis, Ariel Linden, and James F Burgess. Now Trending: Coping with Non-Parallel Trends in Difference-in-Differences Analysis. *Statistical Methods in Medical Research*, 28(12):3697–3711, December 2019. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280218814570. URL http://journals.sagepub.com/doi/10.1177/0962280218814570.

[7] Amol S. Navathe, Joshua M. Liao, Sarah E. Dykstra, Erkuan Wang, Zoe M. Lyon, Yash Shah, Joseph Martinez, Dylan S. Small, Rachel M. Werner, Claire Dinh, Xinshuo Ma, and Ezekiel J. Emanuel. Association of Hospital Participation in a Medicare Bundled Payment Program with Volume and Case Mix of Lower Extremity Joint Replacement Episodes. *JAMA*, 320(9):901, September 2018. ISSN 0098-7484. doi: 10.1001/jama.2018.12345.

[8] Justin B. Dimick and Andrew M. Ryan. Methods for Evaluating Changes in Health Care Policy: The Difference-in-Differences Approach. *JAMA*, 312(22):2401, December 2014. ISSN 0098-7484. doi: 10.1001/jama.2014.16153. URL `http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2014.16153`.

[9] Jeremy M. Kahn, Billie S. Davis, Jonathan G. Yabes, Chung-Chou H. Chang, David H. Chong, Tina Batra Hershey, Grant R. Martsolf, and Derek C. Angus. Association Between State-Mandated Protocolized Sepsis Care and In-Hospital Mortality Among Adults with Sepsis. *JAMA*, 322(3):240–250, July 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.9021.

[10] Rahi Abouk, Rosalie Liccardo Pacula, and David Powell. Association Between State Laws Facilitating Pharmacy Distribution of Naloxone and Risk of Fatal Overdose. *JAMA Internal Medicine*, 179(6):805–811, June 2019. ISSN 2168-6106. doi: 10.1001/jamainternmed.2019.0272.

[11] Simon Freyaldenhoven, Christian Hansen, and Jesse M. Shapiro. Pre-Event Trends in the Panel Event-Study Design. *American Economic Review*, 109(9):3307–3338, September 2019. ISSN 0002-8282. doi: 10.1257/aer.20180609. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20180609`.

[12] Ariella Kahn-Lang and Kevin Lang. The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications. *Journal of Business & Economic Statistics*, 38(3):613–620, July 2020. ISSN 0735-0015, 1537-2707. doi: 10.1080/07350015.2018.1546591. URL `https://www.tandfonline.com/doi/full/10.1080/07350015.2018.1546591`.

[13] Jonathan Roth. Pretest with Caution: Event-Study Estimates After Testing for Parallel Trends. *American Economic Review: Insights*, 4(3):305–322, September 2022. doi: 10.1257/aeri.20210236. URL `https://www.aeaweb.org/articles?id=10.1257/aeri.20210236&&from=f`.

[14] Ashesh Rambachan and Jonathan Roth. A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, page rdad018, feb 2023. ISSN 0034-6527. doi: 10.1093/restud/rdad018. URL `https://doi.org/10.1093/restud/rdad018`.

[15] Jamie R. Daw and Laura A. Hatfield. Matching and Regression to the Mean in Difference-in-Differences Analysis. *Health Services Research*, 53(6):4138–4156, 2018. ISSN 1475-6773. doi: 10.1111/1475-6773.12993. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.12993`.

[16] W. C. Blackwelder. "Proving the Null Hypothesis" in Clinical Trials. *Controlled Clinical Trials*, 3(4):345–353, December 1982. ISSN 0197-2456. doi: 10.1016/0197-2456(82)90024-1.

[17] Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, Boca Raton, 2nd edition, June 2010. ISBN 978-1-4398-0818-4.

[18] Holger Dette and Martin Schumann. Testing for Equivalence of Pre-Trends in Difference-in-Differences Estimation. *Journal of Business & Economic Statistics*, 0(0):1–13, 2024. ISSN 0735-0015. doi: 10.1080/07350015.2024.2308121. URL `https://doi.org/10.1080/07350015.2024.2308121`.

[19] Erin Hartman and F. Daniel Hidalgo. An Equivalence Approach to Balance and Placebo Tests. *American Journal of Political Science*, 62(4):1000–1013, October 2018. ISSN 0092-5853, 1540-5907. doi: 10.1111/ajps.12387. URL `https://onlinelibrary.wiley.com/doi/10.1111/ajps.12387`.

[20] Erin Hartman. Equivalence Testing for Regression Discontinuity Designs. *Political Analysis*, 29(4):505–521, oct 2021. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2020.43. URL `https://www.cambridge.org/core/journals/political-analysis/article/equivalence-testing-for-regression-discontinuity-designs/43F77CDC6337A63AE0A5E6DC3EE01A41`.

[21] Ricardo Mora and Iliana Reggio. Alternative Diff-in-Diffs Estimators with Several Pre-treatment Periods. *Econometric Reviews*, 38(5):465–486, May 2019. ISSN 0747-4938. doi: 10.1080/07474938.2017.1348683. URL `https://doi.org/10.1080/07474938.2017.1348683`.

[22] Alisa Tazhitdinova and Gonzalo Vazquez-Bare. Difference-in-Differences with Unequal Baseline Treatment Status. Available at NBER, March 2023. URL `https://www.nber.org/papers/w31063`.

[23] Anton Strezhnev. Group-Specific Linear Trends and the Triple-Differences in Time Design. Preprint, September 2024. URL `https://osf.io/dg5ps_v1`.

[24] Ting Ye, Luke Keele, Raiden Hasegawa, and Dylan S. Small. A Negative Correlation Strategy for Bracketing in Difference-in-Differences. Preprint, available at arXiv:2006.02423, June 2022. URL `http://arxiv.org/abs/2006.02423`. arXiv:2006.02423 [econ, stat].

[25] Tyler J. VanderWeele and Peng Ding. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4):268, August 2017. ISSN 0003-4819. doi: 10.7326/M16-2607. URL `http://annals.org/article.aspx?doi=10.7326/M16-2607`.

[26] Clifford C. Clogg, Eva Petkova, and Adamantios Haritou. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, 100(5):1261–1293, March 1995. ISSN 0002-9602, 1537-5390. doi: 10.1086/230638. URL `https://www.journals.uchicago.edu/doi/10.1086/230638`.

[27] W. Liu, F. Bretz, A. J. Hayter, and H. P. Wynn. Assessing Nonsuperiority, Noninferiority, or Equivalence When Comparing Two Regression Models Over a Restricted Covariate Region. *Biometrics*, 65(4):1279–1287, 2009. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2008.01192.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2008.01192.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2008.01192.x.

[28] Justin Wolfers. Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results. *American Economic Review*, 96(5):1802–1820, December 2006. ISSN 0002-8282. doi: 10.1257/aer.96.5.1802. URL `https://www.aeaweb.org/articles?id=10.1257/aer.96.5.1802`.

[29] Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225(2):200–230, December 2021. ISSN 0304-4076. doi: 10.1016/j.jeconom.2020.12.001. URL `http://www.sciencedirect.com/science/article/pii/S0304407620303948`.

[30] Andrew Goodman-Bacon. Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics*, 225(2):254–277, December 2021. ISSN 0304-4076. doi: 10.1016/j.jeconom.2021.03.014. URL `https://www.sciencedirect.com/science/article/pii/S0304407621001445`.

[31] Liyang Sun and Sarah Abraham. Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, 225(2):175–199, December 2021. ISSN 0304-4076.

[32] Susan Athey and Guido W. Imbens. Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption. *Journal of Econometrics*, 226(1):62–79, January 2022. ISSN 0304-4076. doi: 10.1016/j.jeconom.2020.10.012. URL `https://www.sciencedirect.com/science/article/pii/S0304407621000488`.

[33] Clément de Chaisemartin and Xavier D'Haultfœuille. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996, September 2020. ISSN 0002-8282. doi: 10.1257/aer.20181169. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20181169`.

[34] Kosuke Imai and In Song Kim. On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, 29(3):405–415, July 2021. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2020.33. URL `https://www.cambridge.org/core/product/identifier/S1047198720000339/type/journal_article`. Publisher: Cambridge University Press (CUP).

[35] Beth Ann Griffin, Megan S. Schuler, Elizabeth A. Stuart, Stephen Patrick, Elizabeth McNeer, Rosanna Smart, David Powell, Bradley D. Stein, Terry L. Schell, and Rosalie Liccardo Pacula. Moving Beyond the Classic Difference-in-Differences Model: A Simulation Study Comparing Statistical Methods for Estimating Effectiveness of State-Level Policies. *BMC Medical Research Methodology*, 21(1):279, December 2021. ISSN 1471-2288. doi: 10.1186/s12874-021-01471-y. URL `https://doi.org/10.1186/s12874-021-01471-y`.

[36] Pedro H. C. Sant'Anna and Qi Xu. Difference-in-Differences with Compositional Changes. arXiv:2304.13925, April 2023. arXiv:2304.13925.

[37] Bret Zeldow and Laura A. Hatfield. Confounding and Regression Adjustment in Difference-in-Differences Studies. *Health Services Research*, 56(5):932–941, October 2021. ISSN 1475-6773. doi: 10.1111/1475-6773.13666.

[38] Seokyung Hahn. Understanding Noninferiority Trials. *Korean Journal of Pediatrics*, 55(11):

403–407, November 2012. ISSN 1738-1061. doi: 10.3345/kjp.2012.55.11.403. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3510268/`.

[39] J. A. Hausman. Specification Tests in Econometrics. *Econometrica*, 46(6):1251–1271, 1978.

[40] Martha J. Bailey and Andrew Goodman-Bacon. The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans. *American Economic Review*, 105(3):1067–1104, mar 2015. ISSN 0002-8282. doi: 10.1257/aer.20120070. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20120070`.

[41] Pedro H C Sant'Anna and Jun B Zhao. Doubly Robust Difference-in-Differences Estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

[42] Holger Dette, Kathrin Möllenhoff, Stanislav Volgushev, and Frank Bretz. Equivalence of Regression Curves. *Journal of the American Statistical Association*, 113(522):711–729, April 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1281813. URL `https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1281813`.

[43] Robin A. Cohen and Michael E. Martinez. Health Insurance Coverage: Early Release of Estimates From the National Health Interview Survey, 2011. Technical report, Centers for Disease Control and Prevention, Atlanta, GA, 2012. Available at: http://www.cdc.gov/nchs/data/nhis/earlyrelease/Insur201206.pdf.

[44] Yaa Akosa Antwi, Asako S. Moriya, and Kosali 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(4):1–28, November 2013. ISSN 1945-7731. doi: 10.1257/pol.5.4.1. URL `https://www.aeaweb.org/articles?id=10.1257/pol.5.4.1`.

[45] Silvia Barbaresco, Charles J. Courtemanche, and Yanling Qi. Impacts of the Affordable Care Act Dependent Coverage Provision on Health-Related Outcomes of Young Adults. *Journal of Health Economics*, 40:54–68, March 2015. ISSN 0167-6296. doi: 10.1016/j.jhealeco.2014.12.004. URL `http://www.sciencedirect.com/science/article/pii/S0167629614001519`.

[46] Benjamin D. Sommers, Thomas Buchmueller, Sandra L. Decker, Colleen Carey, and Richard Kronick. The Affordable Care Act Has Led to Significant Gains in Health Insurance and Access

to Care for Young Adults. *Health Affairs*, 32(1):165–174, January 2013. ISSN 0278-2715. URL `https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2012.0552`.

[47] Joel C. Cantor, Alan C. Monheit, Derek DeLia, and Kristen Lloyd. Early Impact of the Affordable Care Act on Health Insurance Coverage of Young Adults. *Health Services Research*, 47(5):1773–1790, 2012. ISSN 1475-6773. doi: 10.1111/j.1475-6773.2012.01458.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6773.2012.01458.x`.

[48] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275, February 2004. ISSN 0033-5533. doi: 10.1162/003355304772839588. URL `https://academic.oup.com/qje/article/119/1/249/1876068`.

Table 1: Possible expanded model specifications.

Model	Parameterization
Reduced	$y_{it} = \sum_{k=T_1}^{T} \beta_k G_i \mathbb{I}(t = k) + \alpha_i + \gamma_t + \epsilon_{it}$
Expanded	$y_{it} = \sum_{k=T_1}^{T} \beta_k^{(e)} G_i \mathbb{I}(t = k) + \alpha_i^{(e)} + \gamma_t^{(e)} + \epsilon_{it}^{(e)} + \square$
Linear time trends	
treatment group	$\square = \theta G_i t$
unit	$\square = \theta_i t$
covariate group[1]	$\square = \theta_{\ell(i)} t$
Differential time fixed effects	
event study[2]	$\square = \sum_{k=1}^{T_1-2} \theta_k G_i \mathbb{I}(t = k)$
covariate group[1]	$\square = \sum_{k=1}^{T_1-2} \theta_{\ell(i)k} \mathbb{I}(t = k)$

[1] $\ell(i)$ indicates the covariate group to which unit i belongs.
[2] An event study has differential time fixed effects by treatment status.

Table 2: Simulation results.

Error	Model	Violation[†]	Threshold[†]	Power				Bias (% of TX80)				
				NI	EQ	R	E	R	Test (R) — Pass	E	Test (E) — Pass	Test (E) — Fail
Independent, heteroskedastic	Linear	0	100	88	75	80	52	0	0	0	8	-55
			20	12	0	80	52	0	0	0	56	-8
		50	100	39	38	99	53	50	0	0	33	-21
			20	$0^\ddagger$	$0^\ddagger$	99	53	50	-	0	-	0
Independent, heteroskedastic	Event study	0	100	40	0	80	25	0	0	0	64	-43
			20	8	0	80	25	0	0	0	124	-11
		50	100	21	0	99	30	50	0	12	91	-25
			20	$2^\ddagger$	$0^\ddagger$	99	30	50	0	12	157	-4
AR(1) $\rho = 0.2$	Linear	0	100	87	74	80	53	0	0	0	8	-53
			20	12	0	80	53	0	-2	0	55	-8
		50	100	38	37	99	53	50	-1	0	33	-20
			20	$0^\ddagger$	$0^\ddagger$	99	53	50	-	0	-	0
AR(1) $\rho = 0.2$	Event study	0	100	50	5	80	32	0	-3	0	43	-43
			20	9	0	80	32	0	-6	0	99	-9
		50	100	26	4	99	39	50	-5	12	67	-23
			20	$2^\ddagger$	$0^\ddagger$	99	39	50	-9	12	128	-3

Simulations vary (n = 100,000 per scenario): (1) residual error structure (heteroskedastic across clusters and independent or AR(1) with ρ = 0.2); (2) the expanded model (linear trend difference or event study); (3) the true violation magnitude; (4) the threshold for a non-inferiority test.

Power indicates power for: (1) a non-inferiority test (NI); (2) an equivalence test (EQ); (3) the reduced model (R); or (4) the expanded model (E).

Bias indicates percentage bias in: (1) the reduced model treatment effect estimator (R); (2) incremental bias conditional on passing a non-inferiority test (Test (R) — Pass); (3) the expanded model treatment effect estimator (E); (4) incremental bias conditional on passing (Test (E) — Pass) or failing (Test (E) — Fail) a non-inferiority test.

[†] Violation magnitude and threshold are given as a percentage of the treatment effect for which the reduced model has 80% power under no violation (TX80).

[‡] Violation exceeds threshold; power indicates Type I error (controlled at < 5%).

Table 3: Effects of the ACA dependent coverage provision on insurance coverage.

Outcome	Model	Effect (95% CI)	Diff (95% CI)	Rule out |2.1|?	Rule out 2.1*?	Rule out |5.3|?
Any	Original	2.9 (1.3, 4.4)				
	+ trend	3.4 (0.9, 5.9)	-0.6 (-2.8, 1.6)	No	Yes[†]	Yes[†]
Dependent	Original	7.0 (5.6, 8.3)				
	+ trend	3.6 (1.2, 6.1)	3.3 (0.9, 5.8)	No	No	No
Employer	Original	-3.2 (-4.3, -2.1)				
	+ trend	-1.5 (-4.0, 1.1)	-1.7 (-4.1, 0.7)	No	No	Yes
Individual	Original	-0.8 (-1.2, -0.4)				
	+ trend	-0.2 (-1.4, 1.0)	-0.6 (-1.6, 0.4)	Yes[†]	Yes[†]	Yes
Government	Original	-0.4 (-1.6, 0.7)				
	+ trend	1.4 (-0.7, 3.6)	-1.9 (-3.8, 0.0)	No	No	Yes

Treatment effects were estimated by fitting the models in Eqs. (13) and (14), each representing the differential change, on the percentage point scale, averaged over the post-implementation period. Diff = difference in treatment effect in reduced versus expanded model; CI = confidence interval

[†] Indicates rule out in the main specification without the time*unemployment interaction, but not in the model with the interaction.

Table 4: Empirical simulation results.

		No violation		Minor violation (1.1%)	
		Normal	Bootstrap	Normal	Bootstrap
Non-inferiority	Rule out |5|	95	100	88	97
test power	Rule out 2.1	47	60	21	26
	Rule out 1	20	23	5	5
Reduced model					
	ATT	3.4	3.4	4.5	4.5
	Bias (%)	0	0	31	31
Incremental	— Rule out |5|	0	0	0	0
test-induced	— Rule out 2.1	0	-2	1	-4
bias (%)	— Rule out 1	1	-3	1	-6
Expanded model					
	ATT	3.4	3.4	3.4	3.4
	Bias (%)	0	0	0	0
Incremental	— Rule out |5|	0	0	8	2
test-induced	— Rule out 2.1	33	18	55	36
bias (%)	— Rule out 1	55	38	80	58
	— Cannot rule out |5|	1	1	-61	-64
	— Cannot rule out 2.1	-30	-28	-15	-12
	— Cannot rule out 1	-15	-12	-5	-3

This table reports results from empirical simulations (n = 15,000 per scenario) from normal and bootstrap-based data-generating processes. The top section describes non-inferiority test power over different thresholds. The bottom two sections display the ATT in the reduced and expanded models, percentage bias, and incremental percentage bias conditional on passing (— Rule out) or failing (— Cannot rule out) non-inferiority tests at different thresholds.

[†] Violation exceeds threshold; thus, power column represents Type I error (controlled at $\alpha = 0.05$).

Supplement for "Nothing to See Here? A non-inferiority approach to parallel trends"

Alyssa Bilinski, PhD and Laura A. Hatfield, PhD

Table S1: Literature review of *JAMA* and *JAMA IM* difference-in-differences (DiD) studies. We summarize all DiD studies published in the *Journal of the American Medical Association* (*JAMA*) and *JAMA Internal Medicine* (*IM*) from 2018 through 2022. Studies were identified by searching the journals' archives for the term difference(s)(-)in(-)difference(s). They are described in terms of whether they were research letters (RL), whether they mentioned the parallel trends (PT) assumption, what tests were used to assess parallel trends (a linear slope test, a joint-F test for pre-intervention deviations, or other), and whether outcome plots or event study (E-S) plots were shown.

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
(1) *Association Between COVID-19 Lockdown Measures and Emergency Department Visits for Violence-Related Injuries in Cardiff, Wales*						
10.1001/jama.2020.25511						
Yes	No	No	No	No	Yes	No
(2) *Association Between Hospital Voluntary Participation, Mandatory Participation, or Nonparticipation in Bundled Payments and Medicare Episodic Spending for Hip and Knee Replacements*						
10.1001/jama.2021.10046						
"There were nondivergent trends in episodic spending across hospital groups during the period before starting the bundled payment program."						
Yes	Yes	No	Yes	No	No	No
(3) *Association Between State-Mandated Protocolized Sepsis Care and In-hospital Mortality Among Adults With Sepsis*						
10.1001/jama.2019.9021						
"We directly examined for this possibility by fitting a model containing a treatment indicator, a continuous time variable, the interaction of these 2 variables, and all patient- and hospital-level covariates, restricted to the preregulation period...We considered parallel trends as being present if the interaction term from this model was not significant. In cases in which there were parallel trends, we simplified the comparative interrupted time series model to a difference-in-differences model by excluding the term for the interaction of the treatment indicator with the continuous time variable."						
No	Yes	Yes	No	No	Yes	No
(4) *Association Between the Experimental Kickoff Rule and Concussion Rates in Ivy League Football*						
10.1001/jama.2018.14165						
Yes	No	No	No	No	No	No
(5) *Association Between the Implementation of a Population-Based Primary Care Payment System and Achievement on Quality Measures in Hawaii*						
10.1001/jama.2019.8113						
"We estimated the risk-standardized probability of achieving the primary outcome, which indicated no significant difference between the groups in trends before intervention (eFigures 2-4 in the Supplement)."						
No	Yes	No	No	No	Yes	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(6) *Association of a Beverage Tax on Sugar-Sweetened and Artificially Sweetened Beverages With Changes in Beverage Prices and Sales at Chain Retailers in a Large Urban Setting*

10.1001/jama.2019.4249

"Analyses focused on the years 2016 and 2017 because the parallel trends assumption (ie, that the preintervention trend in the outcome is similar for the treatment and control locations) held for beverage volume sales in Philadelphia compared with Baltimore during 2016 but did not hold from January 1, 2014, to December 31, 2015, based on generalized estimating equations using a continuous time variable, the locations, and the interaction between the 2 (eAppendix 1 A.4.a in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(7) *Association of Coronary Artery Bypass Grafting vs Percutaneous Coronary Intervention With Memory Decline in Older Adults Undergoing Coronary Revascularization*

10.1001/jama.2021.5150

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	No	No	No	No	No	No

(8) *Association of Hospital Participation in a Medicare Bundled Payment Program With Volume and Case Mix of Lower Extremity Joint Replacement Episodes*

10.1001/jama.2018.12345

"The assumption of parallel trends under our difference-in-differences method was tested using a generalized linear regression of volume on a BPCI market indicator, time variable, and the interaction, using a Wald test that did not indicate divergent secular trends during the pre-BPCI period ($P = .92$) (eTable 1 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(9) *Association of Hospital Participation in Bundled Payments for Care Improvement Advanced With Medicare Spending and Hospital Incentive Payments*

10.1001/jama.2022.18529

"Visual inspection (eFigure 2 in the Supplement) and statistical tests (eTable 3 in the Supplement) revealed small, nonsignificant differences in preintervention episode spending for episodes at BPCI-A vs comparison hospitals. Nevertheless, we adjusted for these differences in preintervention trends to produce a conservative estimate of the association of BPCI-A participation with changes in episode spending (eTable 4 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	Yes

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(10) *Association of Initiation of Basal Insulin Analogs vs Neutral Protamine Hagedorn Insulin With Hypoglycemia-Related Emergency Department Visits or Hospital Admissions and With Glycemic Control in Patients With Type 2 Diabetes*
10.1001/jama.2018.7993
"This model was based on the counterfactual assumption that if patients who initiated insulin analogs had instead initiated NPH insulin, their changes in hemoglobin A1c level would be similar to the changes observed in the NPH insulin reference group, who were frequency matched based on the propensity score quintile."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	No

(11) *Association of Medicaid Expansion With 1-Year Mortality Among Patients With End-Stage Renal Disease*
10.1001/jama.2018.16504
"Mortality rates were similar in expansion and nonexpansion states prior to 2014 and then diverged, with declines in mortality rates beginning in the first 6 months of 2014." "Pre-2014 trends in outcomes for expansion and nonexpansion states did not differ significantly (eTable 4 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(12) *Association of Participation in the Oncology Care Model With Medicare Payments, Utilization, Care Delivery, and Quality Outcomes*
10.1001/jama.2021.17642
"For each claims-based measure, we tested the null hypothesis that OCM and comparison episodes had parallel trends during the 18-month baseline period (eTable 5 in Supplement 1). Difference-in-differences results are not reported for outcome measures for which we rejected the parallel trends assumption ($\alpha = .05$)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	No	No

(13) *Association of Real-time Continuous Glucose Monitoring With Glycemic Control and Acute Metabolic Events Among Patients With Insulin-Treated Diabetes*
10.1001/jama.2021.6530
"The difference-in-differences method assumes outcomes in the exposed group—had they not been exposed to the intervention (ie, counterfactual case)—would be qualitatively similar to the observed outcomes in the unexposed (reference) group" "No violations of model assumptions (ie, experimental treatment assignment, parallel trends, common shock, and no spillover) were detected."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	No

(14) *Association of Remote vs In-Person Benefit Delivery With WIC Participation During the COVID-19 Pandemic*
10.1001/jama.2021.14356
"There was no statistical evidence of differing trends in WIC participation across these states prior to the pandemic ($\beta = .0002; P = .91$)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	Yes	Yes	No	No	No	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(15) *Association of Skilled Nursing Facility Participation in a Bundled Payment Model With Institutional Spending for Joint Replacement Surgery*

10.1001/jama.2020.19181

"The validity of the difference-in-differences approach was assessed by testing the parallel trends assumption, comparing the slope over time in the pre-BPCI period for BPCI participants and nonparticipants. None of the outcomes had violations of this assumption (eTable 1 in the Supplement)."

No	Yes	Yes	No	No	No	No

(16) *Association of State Medicaid Expansion Status With Low Birth Weight and Preterm Birth*

10.1001/jama.2019.3678

"This Figure was used to visualize trends across the study period as well as whether trends were parallel prior to the expansion date, which was formally tested using linear indicator variables in the DID and DDD models. The DDD comparison of relative disparities in rates of low birth weight between black and white infants as well as 2 DID comparisons among outcomes in black infants (preterm birth and low birth weight) failed the parallel trends test; however, all other significant DID and DDD comparisons in the primary analyses passed this test (Table 3 and Table 4)."

No	Yes	Yes	No	No	Yes	No

(17) *Association of the Affordable Care Act Dependent Coverage Provision With Prenatal Care Use and Birth Outcomes*

10.1001/jama.2018.0030

"We detected small but significant differential linear trends prior to 2010 for early prenatal care and neonatal intensive care unit (NICU) admission overall and among unmarried women only (eTable 2 in the Supplement). Estimating our primary difference-in-differences regression as if the policy took effect in July 2009 (post hoc placebo testing), we identified a similar pattern, with significant results for early prenatal care and NICU admission overall and among unmarried women only (eTable 3 in the Supplement)."

No	Yes	Yes	No	No	Yes	No

(18) *Association of the Healthy, Hunger-Free Kids Act With Dietary Quality Among Children in the US National School Lunch Program*

10.1001/jama.2020.9517

No	No	No	No	No	No	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(19) *Hospital Quality Improvement Interventions, Statewide Policy Initiatives, and Rates of Cesarean Delivery for Nulliparous, Term, Singleton, Vertex Births in California*
10.1001/jama.2021.3816

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	No	No	No	No	Yes	No

(20) *Implementation of a Health Plan Program for Switching From Analogue to Human Insulin and Glycemic Control Among Medicare Beneficiaries With Type 2 Diabetes*
10.1001/jama.2018.21364

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	No	No	No	No	Yes	No

(21) *Medical Debt in the US, 2009-2020*
10.1001/jama.2021.8694
[Placebo outcome test:] "To assess whether the association between Medicaid expansion and medical debt reflected confounding factors (such as differential economic trends), we conducted the analyses separately using nonmedical debt as the outcome."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	No	No	No	No	Yes	No

(22) *Pass-Through of a Tax on Sugar-Sweetened Beverages at the Philadelphia International Airport*
10.1001/jama.2017.16903

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	No	No	No	No	No	No

(23) *Prescription Drug Monitoring Program Mandates and Opioids Dispensed Following Emergency Department Encounters for Patients With Sickle Cell Disease or Cancer With Bone Metastasis*
10.1001/jama.2021.10161
[In main text:] "Parallel trends assumptions before mandate implementation were met (Supplement)." [In supplement:] "For a staggered design, the state-of-the-art approach to testing the parallel trend assumption is to conduct an event study analysis." "All estimated differences were not statistically different from 0, supporting parallel trends in the two outcomes leading up to implementation of any mandate." "Here, results of the event study analysis suggest parallel trends in all outcome-sample combinations except one....Moreover, our finding that comprehensive mandates were associated with a reduction in MMEs dispensed to patients with SCD represented a reversal of the temporal trend seen over 7-18 months before implementation, rather than a continuation of a pre-existing trend. We thus do not consider this temporal deviation from the parallel trend assumption a threat to the validity of our findings."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	Yes	No	No	Yes	No	Yes

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(24) *Seasonal Influenza Activity During the SARS-CoV-2 Outbreak in Japan*
10.1001/jama.2020.6173

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	No	No	No	No	No	No

(25) *Trends in US Ambulatory Care Patterns During the COVID-19 Pandemic, 2019-2021*
10.1001/jama.2021.24294
"Second, 2019 utilization rates were assumed to be reasonable counterfactual control rates for 2020 had the pandemic not occurred. We concluded that this assumption was plausible after visually comparing trends between 2018 and 2019 (Figure 1 and eFigure 7 in the Supplement) and tested this assumption using a placebo test for parallel trends (eTable 3 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	Yes	No

(26) *Association Between Automotive Assembly Plant Closures and Opioid Overdose Mortality in the United States*
10.1001/jamainternmed.2019.5686
[In main text:] "Prior to plant closures, baseline opioid overdose mortality rates in exposed counties were lower than those in unexposed counties, with no evidence of differential trends in the primary outcomes." [In supplement:] "The event study specifications, which also provide a more transparent test of violations of the parallel trends assumption required for causal inference, avoid this problem by indexing the reference point to time since the event (as opposed to calendar time) and by allowing associations to vary over time."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	Yes

(27) *Association Between Medicaid Expansion and Rates of Opioid-Related Hospital Use*
10.1001/jamainternmed.2020.0473
"However, we estimated event study regressions comparing changes between expansion and nonexpansion states to assess parallel trends in the prepolicy period. eFigure2 in the Supplement provides statistical data suggesting that the expansion and nonexpansion states mostly changed similarly in preexpansion years, which increases our confidence that the states would have continued to trend similarly were it not for the expansion."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	Yes

(28) *Association Between State Laws Facilitating Pharmacy Distribution of Naloxone and Risk of Fatal Overdose*
10.1001/jamainternmed.2019.0272
"We tested the parallel trends assumption, which is necessary for obtaining unbiased estimates using the difference-in-differences framework, through event study analyses, which is recommended when evaluating health policies." "Small and statistically non-significant estimates before adoption suggest that the parallel trends assumption was satisfied."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	Yes

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(29) *Association between US state medical cannabis laws and opioid prescribing in the Medicare Part D population*

10.1001/jamainternmed.2018.0266

"We tested our data for parallel trends in prescribing between "never-MCL" states and pre-MCL years for states that implement the policy during our study period; we cannot reject the null hypothesis of parallel trends, which supports the use of our models (see online eAppendix eTable 8 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	No	No

(30) *Association of Cigarette Sales With Comprehensive Menthol Flavor Ban in Massachusetts*

10.1001/jamainternmed.2021.7333

"There were nondivergent trends in state-level sales of menthol and nonflavored cigarette packs per 1000 people in Massachusetts and comparison states during the period before Massachusetts's comprehensive flavor ban."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	Yes	No	No	No	No	No

(31) *Association of Coded Severity With Readmission Reduction After the Hospital Readmissions Reduction Program*

10.1001/jamainternmed.2017.6148

"Trends in rates of readmission were parallel between control and exposed hospitals before implementation of the HRRP (Figure, B and C)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	Yes	No	No	No	Yes	No

(32) *Association of County-Level Prescriptions for Hydroxychloroquine and Ivermectin With County-Level Political Voting Patterns in the 2020 US Presidential Election*

10.1001/jamainternmed.2022.0200

[Placebo outcome test:] "We assessed countylevel rates of new prescriptions for hydroxychloroquine and ivermectin (ie, patients with no fills for the medication in the previous 6 months) per 100 000 enrollees and 2 control medications, methotrexate sodium and albendazole (which have similar clinical applications as hydroxychloroquine or ivermectin, respectively, but are not proposed as COVID-19 treatments)....There were no substantive changes in overall prescribing volume for methotrexate or albendazole."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	Yes	No	No	Yes	Yes	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(33) *Association of Disability Compensation With Mortality and Hospitalizations Among Vietnam-Era Veterans With Diabetes*

10.1001/jamainternmed.2022.2159

"we tested the significance of an interaction between BOG status and quarter (indicating that trends in outcomes were not different for BOG and NOG in the six quarters prior to the policy change). We did not examine parallel pre-policy trends for disability compensation payments because these data were available on an annual basis, leaving only two pre-policy measurements."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(34) *Association of Medicaid Expansion With Quality in Safety-Net Hospitals*

10.1001/jamainternmed.2020.9142

"Formal tests of preexpansion trends did not reach statistical significance (eTable 5 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(35) *Association of Medical and Adult-Use Marijuana Laws With Opioid Prescribing for Medicaid Enrollees*

10.1001/jamainternmed.2018.1007

[In main text:] "we performed "parallel-trend assumption" tests by statistically and graphically comparing the prepolicy trends between medical marijuana states, adult-use marijuana states, and the comparison states." "Moreover, the "parallel-trend assumption" tests and falsification tests lent weight to the validity of the methods (Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(36) *Association of Physician Management Companies and Private Equity Investment With Commercial Health Care Prices Paid to Anesthesia Practitioners*

10.1001/jamainternmed.2022.0004

"The coefficients in the precontract period did not differ significantly between PMC and non-PMC facilities."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	Yes

(37) *Association of Scheduled vs Emergency-Only Dialysis With Health Outcomes and Costs in Undocumented Immigrants With End-stage Renal Disease*

10.1001/jamainternmed.2018.5866

"Our DiD approach accounts for between-group differences, assuming that patients in the scheduled dialysis group would have had utilization and cost trends parallel to those of patients in the emergency-only group had they not received coverage."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	Yes	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(38) *Association of state opioid prescription duration limits with changes in opioid prescribing for Medicare beneficiaries*

10.1001/jamainternmed.2021.4281

"Before the start of duration limits in 2016, days of opioid prescribed were parallel in exposed states and control states."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
Yes	Yes	No	No	No	Yes	No

(39) *Association of Surprise-Billing Legislation with Prices Paid to In-Network and Out-of-Network Anesthesiologists in California, Florida, and New York: An Economic Analysis*

10.1001/jamainternmed.2021.4564

[Found pre-trends and then:] "Because anticipatory effects would invalidate the difference-in-differences assumption of no preexisting trends, eTable 5 and eFigures 2 and 3 in the Supplement present results of the analyses for California and Florida using the quarter during which the law was introduced to the state legislature."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	No	Yes

(40) *Association of Team-Based Primary Care With Health Care Utilization and Costs Among Chronically Ill Patients*

10.1001/jamainternmed.2018.5118

"We visually and formally checked the assumption of parallel trends in the pre-period for valid difference-in-difference inference." "In eFigure 2 in the Supplement, we present figures and empirical tests of the parallel trends assumption and find that parallel trends exist in all variables except for total cost of care and outpatient visits in the more than 2 comorbidity sample. Parallel trends were not met in any outcome in the less than 2 comorbidity subsample."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

(41) *Association of the Comprehensive End-Stage Renal Disease Care Model With Medicare Payments and Quality of Care for Beneficiaries With End-Stage Renal Disease*

10.1001/jamainternmed.2020.0562

"A core assumption of the difference-in-differences design is that the intervention and comparison populations have parallel trends for a given outcome during the baseline period. Parallel trend tests were conducted for all outcomes at the 5% level. All outcomes, except catheter use for wave 1 facilities, passed parallel trend tests (eTable 2 in the Supplement). Although statistical trend tests of wave 1 catheter use did not pass, visual inspection of the relative trends to the comparison group appeared parallel. In addition, the coefficient of the difference in trends at baseline, although significant, equalled 0.00046."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	No	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot

(42) *Changes in Health Care Use and Outcomes After Turnover in Primary Care*

10.1001/jamainternmed.2020.6288

[In main text:] "Modeled pretrends of exposed and unexposed beneficiaries' health care use and health outcomes did not systematically differ prior to PCP exit (eMethods 2 in the Supplement)." [In the supplement:] "Treated and control patients see their assigned PCPs at the same rate as illustrated by curves moving in parallel. Both curves slope downward due to mean reversion and patients dying over time."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	No	Yes	Yes

(43) *Changes in Health Care Use Associated With the Introduction of Hospital Global Budgets in Maryland*

10.1001/jamainternmed.2017.7455

"This approach, which is standard in difference-in-difference studies, relies on the assumption of parallel preintervention trends in Maryland and the control group. However, it could produce biased estimates if preintervention trends differ. Therefore, in a second analysis, we assumed that differences between Maryland and the control group would have continued to change at the preintervention rate in the absence of Maryland's program."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	No	Yes	Yes	No

(44) *Changes in Hospital Income, Use, and Quality Associated With Private Equity Acquisition*

10.1001/jamainternmed.2020.3552

"A joint F test to assess differences in preacquisition trends for hospital income and use measures as well as process quality measures did not show a significant difference for any of the 11 measures (eTable 14 in the Supplement)."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	No	Yes	No	Yes	No

(45) *Clinical Outcomes After Intensifying Antihypertensive Medication Regimens Among Older Adults at Hospital Discharge*

10.1001/jamainternmed.2019.3007

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	No	No	No	No	No	No

(46) *Evaluation of Economic and Clinical Outcomes Under Centers for Medicare & Medicaid Services Mandatory Bundled Payments for Joint Replacements*

10.1001/jamainternmed.2019.0480

[In the main text:] "Tests of preintervention spending trends between treatment and control showed that differences were not statistically significant (eTable 3 in the Supplement)." [In the supplement:] "This table reports model estimates of differential changes in pre-intervention spending at the episode level associated with random assignment of the index hospital into the CJR model. The model was adjusted for age, sex, CMS Hierarchical Condition Category (CMS-HCC) risk score, indicator for hip fracture status, hospital fixed effect, a vector of quarter indicators, and interactions between each quarter and CJR status."

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
No	Yes	Yes	No	No	Yes	No

RL	Mention PT	Slope	Joint F	Other	Plot	E-S plot
(47) *Health Care Utilization and Cost Outcomes of a Comprehensive Dementia Care Program for Medicare Beneficiaries*						
10.1001/jamainternmed.2018.5579						
No No		No	No	No	No	No
(48) *Hospital Responses to Incentives in Episode-Based Payment for Joint Surgery: A Controlled Population-Based Study*						
10.1001/jamainternmed.2021.1897						
[In main text:] "In a test of the assumptions in a difference-in-differences model, we did not observe differential trends in our outcomes in the period before CJR implementation (eTable 2 in the Supplement); nor did we see differential shifts in LEJR volumes (eMethods 4 and eTable 3 in the Supplement)." [In the supplement:] "As shown in eTable 2, we found that LEJR spending and risk were increasing differentially in MSAs selected for the CJR program in the pre-period, but statistically we were unable to reject the null hypothesis of no differential pre-period trend for any of our outcomes."						
No Yes		Yes	No	No	Yes	No
(49) *Medicare Accountable Care Organization Enrollment and Appropriateness of Cancer Screening*						
10.1001/jamainternmed.2017.8087						
No No		No	No	No	No	No
(50) *Rates of Advanced Imaging by Practice Peers After Malpractice Injury Reports in Florida, 2009-2013*						
10.1001/jamainternmed.2019.0163						
Yes No		No	No	No	No	Yes
(51) *Utilization of Long-Acting Reversible Contraceptives in the United States After vs Before the 2016 US Presidential Election*						
10.1001/jamainternmed.2018.7111						
Yes No		No	No	No	Yes	No

1 Appendix A. Difference between reduced model and expanded model with a linear slope

Recall from the main text that we assume a balanced panel with the constrained model specification:

$$y_{it} = \sum_{k=T_1}^{T} \beta_k G_i \mathbb{I}(t = k) + \alpha_i + \gamma_t + \epsilon_{it}, \tag{S1}$$

where y_{it} is the outcome for unit i at time t, γ_t is a time fixed effect, and α_i is a unit fixed effect. The β_k represent differential changes in the treated group relative to the comparison group in each post-intervention period. If the model is correctly specified, the average of these, $\beta = \frac{1}{T - T_1 + 1} \sum_{k=T_1}^{T} \beta_k$, is the ATT.

When we add a slope difference θ, we obtain the expanded model specification,

$$y_{it} = \sum_{k=T_1}^{T} \beta_k^{(e)} G_i \mathbb{I}(t = k) + \alpha_i^{(e)} + \gamma_t^{(e)} + \theta G_i t + \epsilon_{it}^{(e)}, \tag{S2}$$

in which (e) distinguishes model components from those of the previous model.

Proposition 1 (Reduced vs. expanded model estimators (linear trend difference)). *The difference between ordinary least squares (OLS) ATT estimators corresponding to model specifications in Eqs. (S1) and (S2) is a linear transformation of the differential trends parameter estimate $\hat{\theta}$ from Eq. (S2):*

$$\hat{\beta} - \hat{\beta}^{(e)} = \left(\frac{1}{T - T_1 + 1} \sum_{t=T_1}^{T} t - \frac{1}{T_1 - 1} \sum_{t=1}^{T_1 - 1} t \right) \hat{\theta} = \frac{T}{2} \hat{\theta}. \tag{S3}$$

Proof. The OLS specification from Eq. (S1) implies sample moment conditions:

$$\frac{1}{n_1} \sum_{i:G_i=1} y_{ik} = \hat{\bar{\alpha}}_1 + \hat{\gamma}_k + \hat{\beta}_k, \qquad\qquad k \in T_1, \ldots, T$$

$$\frac{1}{n_0} \sum_{i:G_i=0} y_{ik} = \hat{\bar{\alpha}}_0 + \hat{\gamma}_k, \qquad\qquad k \in T_1, \ldots, T$$

$$\frac{1}{n_1(T_1 - 1)} \sum_{i:G_i=1,\, t:t<T_1} y_{it} = \hat{\bar{\alpha}}_1 + \hat{\bar{\gamma}}_{t<T_1}$$

$$\frac{1}{n_0(T_1 - 1)} \sum_{i:G_i=0,\, t:t<T_1} y_{it} = \hat{\bar{\alpha}}_0 + \hat{\bar{\gamma}}_{t<T_1},$$

where $\hat{\bar{\alpha}}_g$ indicates the average unit fixed effect over units with $G_i = g$ and $\hat{\bar{\gamma}}_{t<T_1}$ indicates the average time fixed effect over $\{1, \ldots, T_1 - 1\}$. Rearranging these, we obtain:

$$\hat{\beta}_k = \frac{1}{n_1} \sum_{i:G_i=1} y_{ik} - \frac{1}{n_0} \sum_{i:G_i=0} y_{ik} - \frac{1}{T_1 - 1} \left(\frac{1}{n_1} \sum_{i:G_i=1,\, t:t<T_1} y_{it} - \frac{1}{n_0} \sum_{i:G_i=0,\, t:t<T_1} y_{it} \right) \tag{S4}$$

By contrast, the OLS specification from Eq. (S2) implies sample moment conditions:

$$\frac{1}{n_1} \sum_{i:G_i=1} y_{ik} = \hat{\tilde{\alpha}}_1^{(e)} + \hat{\gamma}_k^{(e)} + \hat{\beta}_k^{(e)} + \hat{\theta}k, \qquad\qquad k \in T_1, \ldots, T$$

$$\frac{1}{n_0} \sum_{i:G_i=0} y_{ik} = \hat{\tilde{\alpha}}_0^{(e)} + \hat{\gamma}_k^{(e)}, \qquad\qquad k \in T_1, \ldots, T$$

$$\frac{1}{n_1(T_1-1)} \sum_{i:G_i=1,\, t:t<T_1} y_{it} = \hat{\tilde{\alpha}}_1^{(e)} + \hat{\tilde{\gamma}}_{t<T_1}^{(e)} + \frac{1}{T_1-1} \sum_{t=1}^{T_1-1} \hat{\theta}t$$

$$\frac{1}{n_0(T_1-1)} \sum_{i:G_i=0,\, t:t<T_1} y_{it} = \hat{\tilde{\alpha}}_0^{(e)} + \hat{\tilde{\gamma}}_{t<T_1}^{(e)}$$

Rearranging these, we obtain:

$$\hat{\beta}_k^{(e)} + \hat{\theta}k - \frac{1}{T_1-1} \sum_{t=1}^{T_1-1} \hat{\theta}t = \frac{1}{n_1} \sum_{i:G_i=1} y_{ik} - \frac{1}{n_0} \sum_{i:G_i=0} y_{ik} - \frac{1}{T_1-1} \left(\frac{1}{n_1} \sum_{i:G_i=1,\, t:t<T_1} y_{it} - \frac{1}{n_0} \sum_{i:G_i=0,\, t:t<T_1} y_{it} \right)$$

$$(S5)$$

Noting that the right sides of Eqs. (S4) and (S5) are equal, we have:

$$\hat{\beta}_k^{(e)} = \hat{\beta}_k - \left(k - \frac{1}{T_1-1} \sum_{t=1}^{T_1-1} t \right) \hat{\theta}$$

Averaging over post-intervention periods from T_1 to T leaves:

$$\hat{\beta} - \hat{\beta}^{(e)} = \left(\frac{1}{T-T_1+1} \sum_{t=T_1}^{T} t - \frac{1}{T_1-1} \sum_{t=1}^{T_1-1} t \right) \hat{\theta}$$

$$= \left(\frac{T_1+T}{2} - \frac{1+T_1-1}{2} \right) \hat{\theta} = \frac{T}{2} \hat{\theta}$$

$$\square$$

If the expanded model is correctly specified but the reduced model is not, reduced model treatment effect bias is larger in magnitude when there is a longer study period or when the slope difference between groups is larger.

2   Appendix B. Test statistics

Recall that in our generalized framework, we have a reduced model with p parameters, $\beta_1, \ldots, \beta_p$, and an expanded model with an additional q parameters, $\theta_1, \ldots, \theta_q$:

$$\textbf{Reduced: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{S6}$$

$$\textbf{Expanded: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{(e)} + \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}^{(e)} \tag{S7}$$

In this setup, if the reduced model is correctly specified, the ATT of interest is an average of a subset $\mathcal{K}$ of the parameters in $\boldsymbol{\beta}$ where $|\mathcal{K}| = K$: $\beta = \frac{1}{K} \sum_{k \in \mathcal{K}} \beta_k$. In the expanded model, the corresponding quantity is: $\beta^{(e)} = \frac{1}{K} \sum_{k \in \mathcal{K}} \beta_k^{(e)}$. We assume that $q \geq 1$ (i.e., the expanded model adds at least one parameter) and that $\begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}$ has full column rank.

Lemma 1 (Form of $\left(\left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1} \right)_{1:p,1:p}$). *Denote the model matrix of Eq. (S7) as* $\mathbf{X}^{(e)}$.
We can write the upper left $p \times p$ *terms of* $\left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1}$:

$$\left(\left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1} \right)_{1:p,1:p} = \left(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \right)^{-1}$$
$$= \left(\mathbf{X}'\mathbf{M}_Z \mathbf{X} \right)^{-1},$$

where $\mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ *denotes the "annihilator matrix" associated with a projection onto the orthogonal complement of the span of* $\mathbf{Z}$, *the added covariates in the expanded model.*

Proof. Following prior work,[1] we can write the expanded model matrix as a block matrix including the reduced model matrix and added variables in the expanded model: $\mathbf{X}^{(e)} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}$. Then,

$$\left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}^{-1}$$

Per standard approaches to inverting block matrices[2] and noting that because $\begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}$ has full column rank, $\mathbf{D} = \mathbf{Z}'\mathbf{Z}$ and $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} = \mathbf{X}'\mathbf{M}_Z\mathbf{X}$ are invertible:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}$$

Applying this to $\left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1}$ as written above, the upper left entry is $p \times p$:

$$\left(\left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1} \right)_{1:p,1:p} = \left(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \right)^{-1}$$
$$= \left(\mathbf{X}'\mathbf{M}_Z \mathbf{X} \right)^{-1}$$

$\square$

Lemma 2 (OLS estimator for $\hat{\boldsymbol{\beta}}^{(e)}$). *The OLS estimator* $\hat{\boldsymbol{\beta}}^{(e)}$, *the coefficients in the expanded model associated with variables shared in both reduced and expanded models, can be written:*

$$\hat{\boldsymbol{\beta}}^{(e)} = \left(\mathbf{X}'\mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{M}_Z \mathbf{y}$$

Proof. By construction, the OLS estimator corresponding to Eq. (S7) is:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{(e)} \\ \hat{\boldsymbol{\theta}} \end{bmatrix} = \left(\mathbf{X}^{(e)\prime} \mathbf{X}^{(e)} \right)^{-1} \mathbf{X}^{(e)\prime} \mathbf{y}$$

$$= \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} \mathbf{y} \qquad \text{as defined in Lemma 1}$$

$$\hat{\boldsymbol{\beta}}^{(e)} = \left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{y} - \left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{Z} \left(\mathbf{Z}' \mathbf{Z} \right)^{-1} \mathbf{Z}' \mathbf{y}$$

$$= \left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_Z \mathbf{y}$$

$\square$

Lemma 3 (Covariance between coefficients in reduced and expanded models). *Assume reduced and expanded models per Eq. (S6) and Eq. (S7), and the expanded model is correctly specified, with $\epsilon_{it}^{(e)} \overset{i.i.d.}{\sim} N\left(0, \sigma_{(e)}^2\right)$. Then, $Cov\left(\hat{\beta}_k, \hat{\beta}_j^{(e)}\right) = Cov\left(\hat{\beta}_k, \hat{\beta}_j\right)$ for all $k, j \in \{1, \ldots, p\}$.*

Proof. Our reduced model estimator is:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{y}$$

Per Lemma 2, the expanded model for coefficients $\{1, \ldots, p\}$ can be written:

$$\hat{\boldsymbol{\beta}}^{(e)} = \left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_Z \mathbf{y}$$

Then, we can find the covariance matrix $Cov\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^{(e)}\right)$, such that $Cov\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^{(e)}\right)_{k,j} = Cov\left(\hat{\beta}_k, \hat{\beta}_j^{(e)}\right)$:

$$\begin{aligned} Cov\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^{(e)}\right) &= Cov\left(\left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{y}, \left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_Z \mathbf{y} \right) \\ &= \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' Var(\mathbf{y}) \left(\left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_Z \right)' \\ &= \sigma_{(e)}^2 \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_Z \mathbf{X} \left(\mathbf{X}' \mathbf{M}_Z \mathbf{X} \right)^{-1} \\ &= \sigma_{(e)}^2 \left(\mathbf{X}' \mathbf{X} \right)^{-1} \\ &= Var(\hat{\boldsymbol{\beta}}) \end{aligned}$$

Because $Cov\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^{(e)}\right) = Var(\hat{\boldsymbol{\beta}})$ (where the latter is the variance-covariance matrix such that $Var(\hat{\boldsymbol{\beta}})_{k,j} = Cov\left(\hat{\beta}_k, \hat{\beta}_j\right)$), we know $Cov\left(\hat{\beta}_k, \hat{\beta}_j^{(e)}\right) = Cov\left(\hat{\beta}_k, \hat{\beta}_j\right)$ for all $k, j \in \{1, \ldots, p\}$.

$\square$

Proposition 2 (Reduced vs. expanded model estimators (Gaussian errors)). *Assume reduced and expanded models as in Eq. (S6) and Eq. (S7) and that the expanded model is correctly specified, with $\boldsymbol{\epsilon}^{(e)} \sim N(0, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is an $nT \times nT$ matrix. Let $\mathbf{V} = \left(\mathbf{X}' \mathbf{X} \right)^{-1}$*

15

and denote $\mathbf{V}^{(e)}$ analogously for the expanded model. The difference between OLS estimators $\hat{\beta}_k$ and $\hat{\beta}_k^{(e)}$ is:

$$\hat{\beta}_k - \hat{\beta}_k^{(e)} \sim N\left(\beta_k - \beta_k^{(e)}, \boldsymbol{\Sigma}_{k,k} + \boldsymbol{\Sigma}_{k,k}^{(e)} - 2\boldsymbol{\Sigma}_{k,k}^*\right),$$

where $\boldsymbol{\Sigma} = \mathbf{V}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\mathbf{V}$, $\boldsymbol{\Sigma}^{(e)} = \mathbf{V}^{(e)}\mathbf{X}^{(e)'}\boldsymbol{\Omega}\mathbf{X}^{(e)}\mathbf{V}^{(e)}$, $\boldsymbol{\Sigma}^* = \mathbf{V}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}^{(e)}\mathbf{V}^{(e)}$, and $\mathbf{A}_{k,k}$ indicates the entry in the kth row and kth column of the matrix $\mathbf{A}$.

Proof. Coefficients in our misspecified reduced model are estimated:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

The OLS sampling variance of the misspecified $\hat{\boldsymbol{\beta}}$ coefficient estimator is:

$$\boldsymbol{\Sigma} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Similarly, for the expanded model:

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}^{(e)} \\ \hat{\boldsymbol{\theta}}\end{bmatrix} = \left(\mathbf{X}^{(e)'}\mathbf{X}^{(e)}\right)^{-1}\mathbf{X}^{(e)'}\mathbf{y}$$

$$\boldsymbol{\Sigma}^{(e)} = \left(\mathbf{X}^{(e)'}\mathbf{X}^{(e)}\right)^{-1}\mathbf{X}^{(e)'}\boldsymbol{\Omega}\mathbf{X}^{(e)}\left(\mathbf{X}^{(e)'}\mathbf{X}^{(e)}\right)^{-1}$$

We can take the covariance of $\hat{\boldsymbol{\beta}}$ and $\begin{bmatrix}\hat{\boldsymbol{\beta}}^{(e)'} & \hat{\boldsymbol{\theta}}'\end{bmatrix}'$:

$$\boldsymbol{\Sigma}^* = Cov\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}, \left(\mathbf{X}^{(e)'}\mathbf{X}^{(e)}\right)^{-1}\mathbf{X}^{(e)'}\mathbf{y}\right)$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}^{(e)}\left(\mathbf{X}^{(e)'}\mathbf{X}^{(e)}\right)^{-1}$$

Extracting entries corresponding to a shared covariate k and applying standard covariance properties $(Var(a-b) = Var(a) + Var(b) - 2Cov(a,b))$ completes the proof. □

Proposition 3 (Reduced vs. expanded model estimators (*i.i.d.* errors))**.** *Assume reduced and expanded models as in Eq. (S6) and Eq. (S7) and that the expanded model is correctly specified, with $\epsilon_{it}^{(e)} \overset{i.i.d.}{\sim} N\left(0, \sigma_{(e)}^2\right)$. Recall that $\beta = \frac{1}{K}\sum_{k\in\mathcal{K}}\beta_k$ and $\beta^{(e)} = \frac{1}{K}\sum_{k\in\mathcal{K}}\beta_k^{(e)}$ are the parameters of interest. The difference between the corresponding OLS ATT estimators is:*

$$\hat{\beta} - \hat{\beta}^{(e)} \sim N\left(\beta - \beta^{(e)}, \sigma_{\hat{\beta}^{(e)}}^2 - \sigma_{\hat{\beta}}^2\right), \tag{S8}$$

where $\sigma_{\hat{\beta}^{(e)}}^2$ is the variance of $\hat{\beta}^{(e)}$ (corresponding to the expanded model), and $\sigma_{\hat{\beta}}^2$ is the variance of $\hat{\beta}$ (corresponding to the potentially misspecified reduced model but defined using common error variance, $\sigma_{(e)}^2$).

Proof. The true variance of $\hat{\beta}$ in the reduced model assuming *i.i.d.* errors is:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'Var(\mathbf{y})\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1}$$
$$= \sigma_{(e)}^2\,(\mathbf{X}'\mathbf{X})^{-1},$$

where $\sigma_{(e)}^2$ corresponds to the expanded model because we assume it to be correctly specified. We denote its kth diagonal element as $\sigma_{\hat{\beta}_k}^2$.[1] Then:

$$Var\left(\hat{\beta}_k - \hat{\beta}_k^{(e)}\right) = Var\left(\hat{\beta}_k\right) + Var\left(\hat{\beta}_k^{(e)}\right) - 2Cov\left(\hat{\beta}_k, \hat{\beta}_k^{(e)}\right)$$
$$= Var\left(\hat{\beta}_k^{(e)}\right) - Var\left(\hat{\beta}_k\right) \qquad\qquad \text{by Lemma 3}$$
$$= \sigma_{\hat{\beta}_k^{(e)}}^2 - \sigma_{\hat{\beta}_k}^2 \qquad\qquad \text{adopting notation above}$$

The extension to $\hat{\beta}$ follows by averaging over $\hat{\beta}_k$. See Clogg et al. (1995) for further discussion.[1] $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

2.1 Walk-through

We apply the testing procedure outlined in Section 2.4.1, evaluating the difference between a linear combination of some $\hat{\beta}_k$ in the reduced model and corresponding $\hat{\beta}_k^{(e)}$ in the expanded model while accounting for clustering, heteroskedastic errors, and survey weights per standard practice.[3] (Note that in this subsection $\hat{\boldsymbol{\beta}}^{(e)}$ is assumed to be of length $p + q$, distinct from its usage in prior sections, to mirror standard regression output.)

1. Fit the reduced model using weighted linear regression with vector of sampling weights $\mathbf{w}$. Extract the parameter vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ (of length p), design matrix $\mathbf{X}$, and $\mathbf{V} = (\mathbf{X}'diag(\mathbf{w})\mathbf{X})^{-1}$.

2. Fit the expanded model. Extract the parameter vector $\hat{\boldsymbol{\beta}}^{(e)}$ (which is of length $p + q$), residual vector $\hat{\mathbf{u}}^{(e)}$, design matrix $\mathbf{X}^{(e)}$, and $\mathbf{V}^{(e)}$.

3. Compute within-model cluster robust variance-covariance matrices under the expanded model,[1] assuming a total of n clusters:

$$\hat{\boldsymbol{\Sigma}} = \frac{n}{n-1}\frac{nT-1}{nT-p-q}\mathbf{V}\left[\sum_g \mathbf{X}_g'\left(\mathbf{w}_g \circ \hat{\mathbf{u}}_g^{(e)}\right)\left(\mathbf{w}_g \circ \hat{\mathbf{u}}_g^{(e)}\right)'\mathbf{X}_g\right]\mathbf{V}$$

$$\hat{\boldsymbol{\Sigma}}^{(e)} = \frac{n}{n-1}\frac{nT-1}{nT-p-q}\mathbf{V}^{(e)}\left[\sum_g \mathbf{X}_g^{(e)'}\left(\mathbf{w}_g \circ \hat{\mathbf{u}}_g^{(e)}\right)\left(\mathbf{w}_g \circ \hat{\mathbf{u}}_g^{(e)}\right)'\mathbf{X}_g^{(e)}\right]\mathbf{V}^{(e)}$$

---

[1] In contrast, the expectation of the OLS estimator for the variance of $\hat{\boldsymbol{\beta}}$, corresponding to the reduced model in Eq. (S1), is $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$, where σ^2 corresponds to the expectation of the reduced model residual variance estimator. Assuming the expanded model is correctly specified (but reduced model may not be), this is biased by a factor of $\frac{\sigma^2}{\sigma_{(e)}^2}$.

where ∘ indicates element-wise multiplication and $\mathbf{X}_g$, $\mathbf{X}_g^{(e)}$, $\hat{\mathbf{u}}_g^{(e)}$, and $\mathbf{w}_g$ are group-specific subsets of their respective elements.

Note that $\hat{\mathbf{u}}^{(e)}$ must be used in $\hat{\mathbf{\Sigma}}$ because we assume the expanded model is correctly specified, but the reduced model may not be.

4. Compute the covariance between estimators:

$$\hat{\mathbf{\Sigma}}^* = \frac{n}{n-1}\mathbf{V}\left[\sum_g \mathbf{X}_g'\left(\mathbf{w}_g \circ \hat{\mathbf{u}}_g^{(e)}\right)\left(\mathbf{w}_g \circ \hat{\mathbf{u}}_g^{(e)}\right)'\mathbf{X}_g^{(e)}\right]\mathbf{V}^{(e)}$$

This matrix is $p \times (p+q)$; the top left $p \times p$-submatrix contains covariances between corresponding parameters of the two models.

5. Let $\hat{\kappa} = \mathbf{a}'\hat{\boldsymbol{\beta}}$ and $\hat{\kappa}^{(e)} = \mathbf{a}^{(e)'}\hat{\boldsymbol{\beta}}^{(e)}$ denote the linear combinations of parameters from each model that we want to test.

6. Compute the variance of the difference between these two linear combinations,

$$Var\left(\hat{\kappa} - \hat{\kappa}^{(e)}\right) = \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} - 2\mathbf{a}'\hat{\mathbf{\Sigma}}^*\mathbf{a}^{(e)} + \mathbf{a}^{(e)'}\hat{\mathbf{\Sigma}}^{(e)}\mathbf{a}^{(e)}$$

We can then use this variance in a Wald test to evaluate hypotheses of interest (e.g., $H_0 : \beta - \beta^{(e)} \geq \delta$).

3   Appendix C. Event studies

Recall the event study specification of an expanded model from Table 1:

$$\textbf{Expanded:}\ y_{it} = \sum_{k=T_1}^{T}\beta_k^{(e)}G_i\mathbb{I}(t=k) + \alpha_i^{(e)} + \gamma_t^{(e)} + \sum_{\ell=1}^{T_1-2}\theta_\ell G_i\mathbb{I}(t=\ell) + \epsilon_{it}^{(e)} \qquad \text{(S9)}$$

Note that $t = T_1 - 1$ is the omitted reference period, i.e., the corresponding pre-period coefficient is normalized to 0.

Supplement Proposition 1 (Event study coefficients as the difference between ATT estimators corresponding to reduced and expanded models). *Assume a reduced model as in Eq. (S1) and expanded model as in Eq. (S9). Following standard practice, let $\mathcal{K} = \{T_1, \ldots, T\}$ (i.e., the ATT of interest is the average effect over all post-intervention periods, with $\hat{\beta} = \frac{1}{T-T_1+1}\sum_{k=T_1}^{T}\hat{\beta}_k$ and likewise for $\hat{\beta}^{(e)}$). Then, when estimated with OLS, $\hat{\beta} - \hat{\beta}^{(e)} = -\frac{1}{T_1-1}\sum_{\ell=1}^{T_1-2}\hat{\theta}_\ell$.*

Proof. Applying the logic outlined in Appendix A, $\hat{\beta}_k$ from Eq. (S1) is:

$$\hat{\beta}_k = \frac{1}{n_1}\sum_{i:G_i=1}y_{ik} - \frac{1}{n_0}\sum_{i:G_i=0}y_{ik} - \frac{1}{T_1-1}\left(\frac{1}{n_1}\sum_{i:G_i=1,\,t:t<T_1}y_{it} - \frac{1}{n_0}\sum_{i:G_i=0,\,t:t<T_1}y_{it}\right) \qquad \text{(S10)}$$

and similarly,

$$\hat{\beta}_k^{(e)} - \frac{1}{T_1-1}\sum_{\ell=1}^{T_1-2}\hat{\theta}_\ell = \frac{1}{n_1}\sum_{i:G_i=1} y_{ik} - \frac{1}{n_0}\sum_{i:G_i=0} y_{ik} - \frac{1}{T_1-1}\left(\frac{1}{n_1}\sum_{i:G_i=1,\,t:t<T_1} y_{it} - \frac{1}{n_0}\sum_{i:G_i=0,\,t:t<T_1} y_{it}\right)$$

(S11)

Noting that the right sides of Eqs. (S10) and (S11) are equal, we have:

$$\hat{\beta}_k = \hat{\beta}_k^{(e)} - \frac{1}{T_1-1}\sum_{\ell=1}^{T_1-2}\hat{\theta}_\ell$$

As $\hat{\beta}_k - \hat{\beta}_k^{(e)}$ does not depend on k, the average (or any particular one) can be written:

$$\hat{\beta} - \hat{\beta}^{(e)} = -\frac{1}{T_1-1}\sum_{\ell=1}^{T_1-2}\hat{\theta}_\ell$$

$\square$

Corollary 1 (Alternative event study models). *Assume that we modify our event study to specify the expanded model:*

$$\textbf{Expanded: } y_{it} = \sum_{k=T_1}^{T}\beta_k^{(e)}G_i\mathbb{I}(t=k) + \alpha_i^{(e)} + \gamma_t^{(e)} + \sum_{\ell\in\mathcal{P}}\theta_\ell G_i\mathbb{I}(t=\ell) + \epsilon_{it}^{(e)},$$

where $\mathcal{P}$ denotes a set of pre-intervention periods. Then $\hat{\beta} - \hat{\beta}^{(e)} = -\frac{1}{T_1-1}\sum_{\ell\in\mathcal{P}}\hat{\theta}_\ell$. In this alternative, we estimate θ_ℓ only for $\ell \in \mathcal{P}$ and let the complement serve as the pooled reference period in the expanded model, rather than the last pre-treatment period as in the traditional event study.

Proof. The proof follows by the same logic as above in Supplement Proposition 1. $\square$

4   Appendix D. No covariance condition

We next introduce a correct model specification:

$$\textbf{Correct:}\quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{(w)} + \mathbf{Z}\boldsymbol{\theta}^{(w)} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^{(w)},\tag{S12}$$

which may add additional terms to the expanded model in Eq. (S7).

Supplement Proposition 2 (Misspecification). *Assume that the correct model specification follows Eq. (S12) and has $\epsilon_{it}^{(w)} \overset{i.i.d.}{\sim} N\left(0,\sigma_{(w)}^2\right)$. Further assume that reduced and expanded models are specified as in Eq. (S6) and Eq. (S7), with $\hat{\beta}$ and $\hat{\beta}^{(e)}$ denoting corresponding OLS ATT estimators. Then, $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$.*

Proof. Following the logic in Lemma 3:

$$Cov\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^{(e)}\right) = Cov\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{y}\right)$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathbf{y})\left((\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\right)'$$

$$= \sigma^2_{(w)}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$

$$= \sigma^2_{(w)}\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

where $Var(\mathbf{y}) = \sigma^2_{(w)}\mathbf{I}$, not $\sigma^2_{(e)}\mathbf{I}$. Similarly, we have:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'Var(\mathbf{y})\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

$$= \sigma^2_{(w)}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Because $Cov\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^{(e)}\right) = Var(\hat{\boldsymbol{\beta}})$, $Cov\left(\hat{\beta},\hat{\beta}-\hat{\beta}^{(e)}\right) = 0$. The scalar result follows by taking linear combinations of parameter estimates. $\square$

Supplement Proposition 3 (Difference between $Cov\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^{(e)}\right)$ and $Var\left(\hat{\boldsymbol{\beta}}\right)$). *Assume that the correct model specification follows Eq. (S12), with $\boldsymbol{\epsilon}^{(w)} \sim N(0,\boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is an $nT \times nT$ matrix. Further assume that reduced and expanded models are specified as in Eq. (S6) and Eq. (S7), with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{(e)}$ denoting corresponding OLS estimators. Then,*

$$\boldsymbol{\Sigma}^*_{1:p,1:p} - \boldsymbol{\Sigma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{M}_X\mathbf{M}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1},$$

where $\boldsymbol{\Sigma}^*_{1:p,1:p} = Cov\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^{(e)}\right)$, the covariance matrix between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{(e)}$; $\boldsymbol{\Sigma} = Var\left(\hat{\boldsymbol{\beta}}\right)$, the variance-covariance matrix associated with $\hat{\boldsymbol{\beta}}$.

Proof. We can write $\boldsymbol{\Sigma}^*_{1:p,1:p}$ as defined in Proposition 2:

$$\boldsymbol{\Sigma}^*_{1:p,1:p} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{M}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}(\mathbf{I}-\mathbf{M}_Z)\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1},$$

where the second line comes from adding and subtracting $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$.

Noting that $\mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ and applying the Woodbury matrix identity formula, $(\boldsymbol{A}-\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}-\boldsymbol{B})^{-1}$ to $(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}$, we can rewrite the first term:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} +$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$

$$= \boldsymbol{\Sigma} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$

Substituting back in, collecting terms, and setting $\mathbf{M}_X = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, the projection onto the orthogonal complement of the span of $\mathbf{X}$, we obtain:

$$\boldsymbol{\Sigma}^*_{1:p,1:p} = \boldsymbol{\Sigma} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\left(\mathbf{I} - \mathbf{M}_X\right)\left(\mathbf{I} - \mathbf{M}_Z\right)\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1} -$$
$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\left(\mathbf{I} - \mathbf{M}_Z\right)\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$
$$= \boldsymbol{\Sigma} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{M}_X(\mathbf{I} - \mathbf{M}_Z)\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$
$$\boldsymbol{\Sigma}^*_{1:p,1:p} - \boldsymbol{\Sigma} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{M}_X\mathbf{M}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1} \text{ because } \mathbf{M}_X\mathbf{X} = 0,$$

which completes the proof.

$\square$

4.1 Special cases

From Supplement Proposition 3, the magnitude of test-induced bias with non-*i.i.d.* error structures depends on both $\boldsymbol{\Omega}$ and $\tilde{\mathbf{X}} = \mathbf{M}_X\mathbf{M}_Z\mathbf{X}$. Let $\boldsymbol{\Sigma}^B = (\boldsymbol{\Sigma}^*_{1:p,1:p} - \boldsymbol{\Sigma})$. For the content that follows in this subsection, assume that $\mathbf{X}$ is structured according to Eq. (S1):

$$\mathbf{X} = \begin{bmatrix} \mathbf{trt}_{T_1} & \dots & \mathbf{trt}_T & \mathbf{u}_1 & \dots & \mathbf{u}_n & \mathbf{t}_2 & \dots & \mathbf{t}_T \end{bmatrix},$$

where $\mathbf{trt}_k$ corresponds to the dummy variable vector indicating $G_i = 1$ and $t = k$, $\mathbf{u}_i$ to the dummy variable vector indicating whether an entry corresponds to unit i and $\mathbf{t}_z$ a dummy variable vector indicating whether an entry corresponds to time z. We refer to its columns (and the corresponding rows of $\mathbf{X}'$) with the numbers $\{1, \dots, K, u_1, \dots, u_n, t_2, \dots, t_T\}$ and let $\mathbf{X}_i$ be the subset of rows of the matrix $\mathbf{X}$ corresponding to unit i. As with other results in this paper, we assume a balanced panel in this subsection per Section 2.1 of the main text.

Supplement Proposition 4 (Independent errors with unit-specific variance)**.** *Assume that the correct model specification follows Eq. (S12) and that errors are independent with unit-specific variance (i.e., $\epsilon_{it} \sim N(0, \sigma_i^2)$ and all ϵ_{it} independent, such that $\boldsymbol{\Omega}$ is diagonal). Further assume that reduced and expanded models are specified as in Eq. (S1) and Eq. (S7) respectively (with $\hat{\beta}$ and $\hat{\beta}^{(e)}$ denoting corresponding OLS ATT estimators) and that the expanded model is specified such that $\tilde{\mathbf{X}}_i = \tilde{\mathbf{X}}_j$ for i, j with shared treatment status.[2] Then, the entries of $\boldsymbol{\Sigma}^B$ corresponding to treatment effects β_k are 0, i.e., $\boldsymbol{\Sigma}^B_{1:K,1:K} = 0$, and thus $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$.*

Proof. Let $\boldsymbol{\Sigma}^B = (\boldsymbol{\Sigma}^*_{1:p,1:p} - \boldsymbol{\Sigma})$ and $\tilde{\mathbf{X}} = \mathbf{M}_X\mathbf{M}_Z\mathbf{X}$. Because errors are independent and heteroskedastic only across groups, we can write per Supplement Proposition 3:

$$\boldsymbol{\Sigma}^B = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\tilde{\mathbf{X}}\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$
$$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^n \sigma_i^2 \mathbf{X}_i'\tilde{\mathbf{X}}_i\right)\left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$

---

[2]The condition that $\tilde{\mathbf{X}}_i = \tilde{\mathbf{X}}_j$ for i, j with shared treatment status applies to both expanded models with a linear slope coefficient Eq. (S2) and event studies Eq. (S9) as well as for other common tests (e.g., an expanded model with unit-specific slopes) under the balanced panel setup introduced in Section 2.1 of the main text. Intuitively, it requires that the expanded model test the same trend difference across all treated units.

where $\mathbf{X}_i$ is the subset of rows of the matrix $\mathbf{X}$ corresponding to unit i.

Consider the inner term $\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i' \tilde{\mathbf{X}}_i$. Recall that $\mathbf{X}' \tilde{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i' \tilde{\mathbf{X}}_i = 0$ by construction. Both of these matrices are of dimension $p \times p$, and we structure $\mathbf{X}$ as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{trt}_{T_1} & \ldots & \mathbf{trt}_T & \mathbf{u}_1 & \ldots & \mathbf{u}_n & \mathbf{t}_2 & \ldots & \mathbf{t}_T \end{bmatrix}$$

As above, we refer to its columns (and the corresponding rows of $\mathbf{X}'$) as $\{1, \ldots, K, u_1, \ldots, u_n, t_2, \ldots, t_T\}$. We also denote entries of $\tilde{\mathbf{X}}$ as $\tilde{x}_{it,p}$, where it specifies row and p the column. We can then consider 3 types of rows of the $p \times p$ matrix $\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i' \tilde{\mathbf{X}}_i$.

1. **Rows corresponding to post-treatment indicators** $(1, \ldots, K)$.
   We first consider the first K rows, each corresponding to a post-treatment indicator row k.

   In $\sum_{i=1}^{n} \mathbf{X}_i' \tilde{\mathbf{X}}_i$ (without σ_i^2), these take the form:

   $$\left[\sum_{i \in \mathcal{N}_1} \tilde{x}_{ik,1}, \ldots, \sum_{i \in \mathcal{N}_1} \tilde{x}_{ik,p} \right] = \begin{bmatrix} 0, \ldots, 0 \end{bmatrix}$$

   Under the assumption that $\tilde{\mathbf{X}}_i = \tilde{\mathbf{X}}_j$ for i, j with shared treatment status, we can simplify:

   $$\left[n_1 \tilde{x}_{n_1^* k,1}, \ldots, n_1 \tilde{x}_{n_1^* k,p} \right] = \begin{bmatrix} 0, \ldots, 0 \end{bmatrix},$$

   where $\tilde{x}_{ik,\ell} = \tilde{x}_{n_1^* k,\ell}$ for all treated units (i.e., n_1^* indicates a representative treated unit). Extending this to $\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i' \tilde{\mathbf{X}}_i$, we have:

   $$\left[\tilde{x}_{n_1^* k,1} \sum_{i \in \mathcal{N}_1} \sigma_i^2, \ldots, \tilde{x}_{n_1^* k,p} \sum_{i \in \mathcal{N}_1} \sigma_i^2 \right] = \begin{bmatrix} 0, \ldots, 0 \end{bmatrix}$$

2. **Rows corresponding to unit fixed effects.**
   In $\sum_{i=1}^{n} \mathbf{X}_i' \tilde{\mathbf{X}}_i$, these rows take the form, when corresponding to unit fixed effect i:

   $$\left[\sum_{t=1}^{T} \tilde{x}_{it,1}, \ldots, \sum_{t=1}^{T} \tilde{x}_{it,p} \right] = \begin{bmatrix} 0, \ldots, 0 \end{bmatrix}$$

   Therefore in $\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i' \tilde{\mathbf{X}}_i$, these rows must also be 0:

   $$\left[\sigma_i^2 \sum_{t=1}^{T} \tilde{x}_{it,1}, \ldots, \sigma_i^2 \sum_{t=1}^{T} \tilde{x}_{it,p} \right] = \begin{bmatrix} 0, \ldots, 0 \end{bmatrix}$$

22

3. **Rows corresponding to time fixed effects.**

In $\sum_{i=1}^{n} \mathbf{X}_i' \tilde{\mathbf{X}}_i$, these rows take the form, when corresponding to time fixed effect t:

$$\left[\sum_{i=1}^{n} \tilde{x}_{it,1}, \ldots, \sum_{i=1}^{n} \tilde{x}_{it,p}\right] = \left[0, \ldots, 0\right]$$

In $\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i' \tilde{\mathbf{X}}_i$, these take the form:

$$\left[\sum_{i=1}^{n} \sigma_i^2 \tilde{x}_{it,1}, \ldots, \sum_{i=1}^{n} \sigma_i^2 \tilde{x}_{it,p}\right]$$

- For $t \geq T_1$: Because we assume that $\tilde{\mathbf{X}}_i = \tilde{\mathbf{X}}_j$ for i, j with shared treatment status, each column p of $\tilde{\mathbf{X}}$ has shared values $a_{t,p}$ for treated units and $b_{t,p}$ for controls. Because $\tilde{\mathbf{X}}$ is orthogonal to all reduced model regressors, orthogonality to the treatment indicator for time t forces $a_{t,p} = 0$, and orthogonality to the time fixed effect for time t then forces $b_{t,p} = 0$. Hence every row of $\tilde{\mathbf{X}}$ in post-treatment periods is the zero vector.

- For $t < T_1$: These may not be zero.

We therefore consider entries of $(\mathbf{X}'\mathbf{X})^{-1}$, where $(\mathbf{X}'\mathbf{X})_{i,j}^{-1} = adj(\mathbf{X}'\mathbf{X})/det(\mathbf{X}'\mathbf{X})$ and $adj(\mathbf{X}'\mathbf{X})_{i,j} = (-1)^{i+j} M_{ij}$, where $M_{i,j}$ is the determinant of the submatrix formed by removing the ith row and jth column. Because $M_{i,j}$ is a determinant, it is zero if columns are linearly dependent. Indeed, if $i \in \{t_2, \ldots, t_{T_1-1}\}$, $j \in \{1, \ldots, K\}$, and $\mathbf{X}$ is structured as detailed above, submatrix columns are linearly dependent. Therefore, corresponding entries of $(\mathbf{X}'\mathbf{X})_{i,j}^{-1}$ are 0.

As a result, when $(\mathbf{X}'\mathbf{X})^{-1}$ is multiplied by the second term, the rows $1, \ldots, K$ of the resultant product are 0, and thus the upper left $K \times K$ entries of $\mathbf{\Sigma}^B$ are 0, $\mathbf{\Sigma}^B_{1:K,1:K} = 0$. Applying Supplement Proposition 3, $Cov\left(\hat{\beta}, \hat{\beta}^{(e)}\right) = Var\left(\hat{\beta}\right)$, and thus $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$.
□

Supplement Proposition 5 (Errors with unit-specific variance and constant error correlation within units)**.** *Assume that the correct model specification follows Eq. (S12) and that errors have unit-specific variance, are independent across units, and have constant error correlation within units (i.e., $\epsilon_{it} \sim N(0, \sigma_i^2)$ and $Cor(\epsilon_{ij}, \epsilon_{ik}) = \rho_i$ for $j, k \in \{1, \ldots, T\}$). Further assume that reduced and expanded models are specified as in Eq. (S1) and Eq. (S7) respectively (with $\hat{\beta}$ and $\hat{\beta}^{(e)}$ denoting corresponding OLS ATT estimators) and that the expanded model is specified such that $\tilde{\mathbf{X}}_i = \tilde{\mathbf{X}}_j$ for i, j with shared treatment status. Then, the entries of $\mathbf{\Sigma}^B$ corresponding to treatment effects β_k are 0, i.e., $\mathbf{\Sigma}^B_{1:K,1:K} = 0$, and thus $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$.*

Proof. Recall that:

$$\mathbf{\Sigma}^B = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{\Omega}\tilde{\mathbf{X}} (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}$$

Because units are independent,

$$\boldsymbol{\Omega} = \operatorname{diag}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_n), \qquad \boldsymbol{\Omega}_i = \sigma_i^2 \left((1 - \rho_i) \mathbf{I}_T + \rho_i \mathbf{J}_T \right),$$

where $\mathbf{J}_T$ is the $T \times T$ all-ones matrix. Therefore, again denoting rows of $\mathbf{X}$ corresponding to unit i as $\mathbf{X}_i$, we have:

$$\mathbf{X}' \boldsymbol{\Omega} \, \tilde{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i' \, \boldsymbol{\Omega}_i \, \tilde{\mathbf{X}}_i = \sum_{i=1}^n \sigma_i^2 \left((1 - \rho_i) \mathbf{X}_i' \tilde{\mathbf{X}}_i + \rho_i \, \mathbf{X}_i' \mathbf{J}_T \, \tilde{\mathbf{X}}_i \right)$$

We note the following:

1. **The second term in the above sum $\rho_i \, \mathbf{X}_i' \mathbf{J}_T \, \tilde{\mathbf{X}}_i$ is 0 for all i.**
   Because $\tilde{\mathbf{X}} = \mathbf{M}_X \mathbf{M}_Z \mathbf{X}$ is orthogonal to the unit-fixed-effect columns of $\mathbf{X}$, $\mathbf{1}_T' \tilde{\mathbf{X}}_i = \mathbf{0}$, and therefore $\mathbf{J}_T \, \tilde{\mathbf{X}}_i = \mathbf{1}_T \mathbf{1}_T' \tilde{\mathbf{X}}_i = \mathbf{0}$.

2. **The remaining terms take a similar form to Supplement Proposition 4:**

$$\begin{aligned}
\boldsymbol{\Sigma}^B &= (\mathbf{X}'\mathbf{X})^{-1} \, \mathbf{X}' \boldsymbol{\Omega} \tilde{\mathbf{X}} \, (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \sigma_i^2 (1 - \rho_i) \mathbf{X}_i' \tilde{\mathbf{X}}_i \right) (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}
\end{aligned}$$

Applying logic from Supplement Proposition 4, we know the upper left $K \times K$ entries of $\boldsymbol{\Sigma}^B$, $\boldsymbol{\Sigma}^B_{1:K,1:K} = 0$, and as a result, applying Supplement Proposition 3, $Cov\left(\hat{\beta}, \hat{\beta}^{(e)}\right) = Var\left(\hat{\beta}\right)$, and thus $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$. $\qquad \square$

5   Appendix E. Test-induced distortions

Assumption 1 (No covariance condition). Assume that the correct model specification follows Eq. (S12), with $\boldsymbol{\epsilon}^{(w)} \sim N(0, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is an $nT \times nT$ matrix, and that the reduced and expanded models are specified as in Eq. (S6) and Eq. (S7), with $\hat{\beta}$ and $\hat{\beta}^{(e)}$ denoting corresponding OLS ATT estimators. Further assume that the combination of error structure and model specifications yields $Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right) = 0$.

Proposition 4 (Reduced model test-induced distortions under Assumption 1). *Under Assumption 1, if we conduct a non-inferiority test with a threshold δ, rejecting the null if $\hat{\beta} - \hat{\beta}^{(e)} < \delta^*$, where $\delta^* = z_\alpha \sigma_{\hat{\beta} - \hat{\beta}^{(e)}} + \delta$, then there is no distortion in the reduced model ATT induced by conditioning on the test result:*

$$\mathbb{E}\left(\hat{\beta} \,\middle|\, \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) - \mathbb{E}\left(\hat{\beta}\right) = 0$$

and likewise,

$$Var\left(\hat{\beta} \,\middle|\, \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) - Var\left(\hat{\beta}\right) = 0.$$

Proof. Under Assumption 1 and Proposition 2, $\hat{\beta}$ and $\hat{\beta} - \hat{\beta}^{(e)}$ are jointly Gaussian as linear functions of Gaussian errors, and thus, zero covariance implies independence. Therefore, $\mathbb{E}\left(\hat{\beta} \big| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) = \mathbb{E}\left(\hat{\beta}\right)$ and $Var\left(\hat{\beta} \big| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) = Var\left(\hat{\beta}\right)$. $\hfill\square$

Corollary 2 (Equivalence tests)**.** *Under Assumption 1, if we conduct an equivalence test with a threshold $|\delta|$, rejecting the null if $\hat{\beta} - \hat{\beta}^{(e)} \in (\delta_L^*, \delta_U^*) = \left(z_{1-\alpha}\sigma_{\hat{\beta}-\hat{\beta}^{(e)}} - |\delta|, z_\alpha\sigma_{\hat{\beta}-\hat{\beta}^{(e)}} + |\delta|\right)$, then there is no distortion in the reduced model ATT induced by conditioning on the test result:*

$$\mathbb{E}\left(\hat{\beta} \Big| \hat{\beta} - \hat{\beta}^{(e)} \in (\delta_L^*, \delta_U^*)\right) - \mathbb{E}\left(\hat{\beta}\right) = 0$$

and likewise,

$$Var\left(\hat{\beta} \Big| \hat{\beta} - \hat{\beta}^{(e)} \in (\delta_L^*, \delta_U^*)\right) - Var\left(\hat{\beta}\right) = 0.$$

Proof. The result follows by applying the logic in Proposition 4. $\hfill\square$

Proposition 5 (Reduced model test-induced bias)**.** *Assume setup and reduced and expanded model estimators as in Proposition 2. If we conduct a non-inferiority test with a threshold δ, rejecting the null if $\hat{\beta} - \hat{\beta}^{(e)} < \delta^*$, where $\delta^* = z_\alpha\sigma_{\hat{\beta}-\hat{\beta}^{(e)}} + \delta$, then the expectation of the reduced model ATT estimator conditional on passing the test may differ from its unconditional expectation:*

$$\mathbb{E}\left(\hat{\beta} \Big| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) - \mathbb{E}\left(\hat{\beta}\right) = -\frac{Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right)}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}} \frac{\phi\left(z_\alpha + \frac{\delta - \beta + \beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}{\Phi\left(z_\alpha + \frac{\delta - \beta + \beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)},$$

where ϕ and Φ are the probability density function and cumulative distribution function of a standard normal, respectively.

Proof. Let $\mathbb{E}\left(\hat{\beta}\right) = \beta$, the unconditional expected value of the reduced model effect estimate (but not necessarily an unbiased ATT) and $\mathbb{E}\left(\hat{\beta}^{(e)}\right) = \beta^{(e)}$. The distortion in the expected value of the reduced model treatment effect estimates induced only by the selection on pre-trends can be characterized by a truncated bivariate normal distribution:[4,5]

$$\mathbb{E}\left(\hat{\beta} - \beta \Big| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) = \frac{Cov\left(\hat{\beta}, \hat{\beta} - \hat{\beta}^{(e)}\right)}{Var\left(\hat{\beta} - \hat{\beta}^{(e)}\right)} \left(\mathbb{E}\left(\hat{\beta} - \hat{\beta}^{(e)} \Big| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right) - \left(\beta - \beta^{(e)}\right)\right)$$

We substitute in $\mathbb{E}\left(\hat{\beta} - \hat{\beta}^{(e)} \Big| \hat{\beta} - \hat{\beta}^{(e)} < \delta^*\right)$, also following a truncated normal distribution:

$$\mathbb{E}\left(\hat{\beta}-\beta\Big|\hat{\beta}-\hat{\beta}^{(e)}<\delta^*\right) = \frac{Cov\left(\hat{\beta},\hat{\beta}-\hat{\beta}^{(e)}\right)}{Var\left(\hat{\beta}-\hat{\beta}^{(e)}\right)}\left[\left(\beta-\beta^{(e)}\right)-\right.$$

$$\left.\sqrt{Var\left(\hat{\beta}-\hat{\beta}^{(e)}\right)}\frac{\phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}{\Phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}-\left(\beta-\beta^{(e)}\right)\right]$$

$$=-\frac{Cov\left(\hat{\beta},\hat{\beta}-\hat{\beta}^{(e)}\right)}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\frac{\phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}{\Phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}$$

$\square$

Proposition 6 (Expanded model test-induced bias). *Assume setup and reduced and expanded model estimators as in Proposition 2. If we conduct a non-inferiority test with a threshold δ, rejecting the null if $\hat{\beta}-\hat{\beta}^{(e)}<\delta^*$, where $\delta^*=z_\alpha\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}+\delta$, then the expectation of the expanded model ATT estimator conditional on passing the test may differ from its unconditional expectation:*

$$\mathbb{E}\left(\hat{\beta}^{(e)}\Big|\hat{\beta}-\hat{\beta}^{(e)}<\delta^*\right)-\mathbb{E}\left(\hat{\beta}^{(e)}\right)=-\frac{Cov\left(\hat{\beta}^{(e)},\hat{\beta}-\hat{\beta}^{(e)}\right)}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\frac{\phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}{\Phi\left(z_\alpha+\frac{\delta-\beta+\beta^{(e)}}{\sigma_{\hat{\beta}-\hat{\beta}^{(e)}}}\right)}.$$

Proof. The result follows by the same logic as Proposition 5, applying the bivariate normal truncation formula to $\hat{\beta}^{(e)}$ conditioning on $\hat{\beta}-\hat{\beta}^{(e)}<\delta^*$. $\square$

6 Appendix F. Non-inferiority test power

Remark 1. *Under standard regularity conditions, OLS estimators are asymptotically normal:*

$$\frac{\hat{\beta}_k-\beta_k}{\sigma_{\hat{\beta}_k}}\xrightarrow{d}N(0,1),$$

where $\sigma^2_{\hat{\beta}_k}=Var\left(\hat{\beta}_k\mid\mathbf{X}\right)$ is the (k,k) element of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Consistent variance estimators satisfy $\hat{\sigma}^2_{\hat{\beta}_k}\xrightarrow{p}\sigma^2_{\hat{\beta}_k}$. In this section, we discuss power evaluated under this asymptotic normal approximation, treating $\sigma^2_{\hat{\beta}_k}$ as known.

Supplement Proposition 6 (Non-inferiority power heuristic). *Assume conditions in Proposition 2. If a one-sided Wald test has power p evaluated under an asymptotic normal approximation to reject $H_0:\beta-\beta^{(e)}\leq 0$ given $\beta-\beta^{(e)}=\beta^*-\beta^{(e)*}$ at level α, then the non-inferiority formulation will have power p to reject $H_0:\beta-\beta^{(e)}\geq\beta^*-\beta^{(e)*}$, given no violation exists.*

Proof. Suppose we want to perform a Wald test on $\beta - \beta^{(e)}$ from Equation (S7). We specify a one-sided test with $H_0 : \beta - \beta^{(e)} \leq 0$ versus the alternative $H_A : \beta - \beta^{(e)} > 0$. Using the asymptotic distribution of the test statistic under the null, the critical value should be $\Phi^{-1}(1 - \alpha)$ to control type I error at α. Then assuming $\hat{\beta} - \hat{\beta}^{(e)} \sim N\left(\beta - \beta^{(e)}, \sigma^2_{\hat{\beta} - \hat{\beta}^{(e)}}\right)$ (see Remark 1) and $\beta - \beta^{(e)} = \beta^* - \beta^{(e)*}$, we define power ($p$), the probability that the test statistic exceeds the critical value:

$$p = Pr\left(\frac{\hat{\beta} - \hat{\beta}^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} > \Phi^{-1}(1 - \alpha) \,\middle|\, \beta - \beta^{(e)} = \beta^* - \beta^{(e)*}\right)$$

We subtract $\frac{\beta - \beta^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}}$ from each side:

$$p = Pr\left(\frac{\hat{\beta} - \hat{\beta}^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} - \frac{\beta - \beta^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} > \Phi^{-1}(1 - \alpha) - \frac{\beta - \beta^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} \,\middle|\, \beta - \beta^{(e)} = \beta^* - \beta^{(e)*}\right)$$

Then, because $\hat{\beta} - \hat{\beta}^{(e)} \sim N\left(\beta^* - \beta^{(e)*}, \sigma^2_{\hat{\beta} - \hat{\beta}^{(e)}}\right)$, we can rewrite our expression as:

$$p = 1 - \Phi\left(\Phi^{-1}(1 - \alpha) - \frac{\beta^* - \beta^{(e)*}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}}\right)$$

Using $1 - \Phi(x) = \Phi(-x)$, and $-\Phi^{-1}(\alpha) = \Phi^{-1}(1 - \alpha)$ for $\alpha \in [0, 1]$, we rearrange as:

$$p = \Phi\left(\Phi^{-1}(\alpha) + \frac{\beta^* - \beta^{(e)*}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}}\right)$$

This means we can write $\beta^* - \beta^{(e)*}$, the value at which power p is achieved as:

$$\beta^* - \beta^{(e)*} = \sigma_{\hat{\beta} - \hat{\beta}^{(e)}}\left(\Phi^{-1}(p) - \Phi^{-1}(\alpha)\right)$$

That is, if $\beta - \beta^{(e)} = \beta^* - \beta^{(e)*} = \sigma_{\hat{\beta} - \hat{\beta}^{(e)}}\left(\Phi^{-1}(p) - \Phi^{-1}(\alpha)\right)$, we will have power p to reject the null that $\beta - \beta^{(e)} \leq 0$.

Now suppose we conduct a non-inferiority test. Following the reasoning above, we formulate a test with a bound based on the $\beta^* - \beta^{(e)*}$ from the main study. That is, we wish to test whether we can rule out differences (relative to the reference of 0) at least as big as the treatment effect we are powered to detect. The hypotheses are therefore $H_0 : \beta - \beta^{(e)} \geq \beta^* - \beta^{(e)*}$ versus $H_A : \beta - \beta^{(e)} < \beta^* - \beta^{(e)*}$. We write a test statistic $\frac{\hat{\beta} - \hat{\beta}^{(e)} - \left(\beta^* - \beta^{(e)*}\right)}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}}$. We reject violations when the test statistic is smaller than a critical value. As before, we determine the critical value by assuming that the null is true and controlling the Type I error rate at α, which yields a critical value of $\Phi^{-1}(\alpha)$.

We are interested in the power of the test when the parallel trends assumption is exactly met, that is, when $\beta - \beta^{(e)} = 0$. By definition, this is:

$$P\left(\frac{\hat{\beta} - \hat{\beta}^{(e)} - \left(\beta^* - \beta^{(e)*}\right)}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} < \Phi^{-1}(\alpha) \,\middle|\, \beta - \beta^{(e)} = 0\right) = P\left(\frac{\hat{\beta} - \hat{\beta}^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} < \Phi^{-1}(\alpha) + \frac{\beta^* - \beta^{(e)*}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} \,\middle|\, \beta - \beta^{(e)} = 0\right)$$

Because we assume $\beta - \beta^{(e)} = 0$, we then have $\frac{\hat{\beta} - \hat{\beta}^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} \sim N(0, 1)$. Plugging in the standard normal cumulative distribution function and our expression for $\beta^* - \beta^{(e)*}$ from above, we obtain

$$\Phi\left(\Phi^{-1}(\alpha) + \frac{\beta^* - \beta^{(e)*}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}}\right) = \Phi\left(\Phi^{-1}(\alpha) + \left(\Phi^{-1}(p) - \Phi^{-1}(\alpha)\right)\right) = p.$$

$\square$

Corollary 3 (Equivalence test power). *Assume conditions in Proposition 2. If a Wald non-inferiority test conducted as in Supplement Proposition 6 has power p to rule out H_0 : $\beta - \beta^{(e)} \geq |\beta^* - \beta^{(e)*}|$ given $\beta - \beta^{(e)} = 0$, then a Wald equivalence test has power less than or equal to p to reject $H_0 : |\beta - \beta^{(e)}| \geq |\beta^* - \beta^{(e)*}|$.*

Proof. As detailed in Supplement Proposition 6, the non-inferiority rejection event for ruling out $H_0 : \beta - \beta^{(e)} \geq |\beta^* - \beta^{(e)*}|$ is:

$$A = \left\{ \frac{\hat{\beta} - \hat{\beta}^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} < \Phi^{-1}(\alpha) + \frac{|\beta^* - \beta^{(e)*}|}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} \right\}$$

A Wald equivalence test rejects $H_0 : |\beta - \beta^{(e)}| \geq |\beta^* - \beta^{(e)*}|$ only if it rejects both one-sided null hypotheses $H_{0U} : \beta - \beta^{(e)} \geq |\beta^* - \beta^{(e)*}|$ and $H_{0L} : \beta - \beta^{(e)} \leq -|\beta^* - \beta^{(e)*}|$. In terms of the same standardized statistic, the additional rejection event is:

$$B = \left\{ \frac{\hat{\beta} - \hat{\beta}^{(e)}}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} > \Phi^{-1}(1 - \alpha) - \frac{|\beta^* - \beta^{(e)*}|}{\sigma_{\hat{\beta} - \hat{\beta}^{(e)}}} \right\}$$

Therefore the equivalence test rejection event is $A \cap B$, and since $A \cap B \subseteq A$,

$$P\left(A \cap B \mid \beta - \beta^{(e)} = 0\right) \leq P\left(A \mid \beta - \beta^{(e)} = 0\right) = p,$$

and the power of the Wald equivalence test is less than or equal to p. $\square$

Proposition 7 (Non-inferiority difference-in-differences power). *Assume setup and reduced and expanded model estimators as in Proposition 3. If a non-inferiority test has power p evaluated under an asymptotic normal approximation to rule out violations of parallel trends equal to or larger than β_k^* (i.e., to reject $H_0 : \beta_k - \beta_k^{(e)} \geq \beta_k^*$, with $\beta_k^* > 0$) in a Wald test at level α, and assuming no violation exists ($\boldsymbol{\theta} = \mathbf{0}$), then $p > p_e$, where p_e is the power to detect $\beta_k^{(e)} = \beta_k^*$ (likewise evaluated) in a one-sided Wald test at level α in an expanded model.*

Proof. We wish to compare the power to detect an effect of size $\beta_k^{(e)} = \beta_k^*$ to the power to rule out a difference $\beta_k - \beta_k^{(e)} \geq \beta_k^*$, assuming $\boldsymbol{\theta} = \mathbf{0}$. We consider the following asymptotic normal approximations (see Remark 1), noting that because $\boldsymbol{\theta} = \mathbf{0}$, we have $\beta_k^{(e)} = \beta_k$:

$$\hat{\beta}_k \sim N\left(\beta_k, \sigma_{\hat{\beta}_k}^2\right)$$

$$\hat{\beta}_k^{(e)} \sim N\left(\beta_k, \sigma_{\hat{\beta}_k^{(e)}}^2\right)$$

28

Let $\mathbf{X}_{-\mathbf{trt}_k}$ denote the matrix of the inputs in the reduced model other than $\mathbf{trt}_k = G_i \mathbb{I}(t = k)$ and $\mathbf{Z}$ indicate the additional columns in the expanded model. In what follows, σ^2 and R^2 terms denote their population analogs (i.e., probability limits of the corresponding sample quantities). Through standard regression algebra, the asymptotic variances satisfy:

$$\sigma_{\hat{\beta}_k}^2 = \frac{\sigma_\varepsilon^2}{SST_{\mathbf{trt}_k}\left(1 - R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k}}\right)}$$

$$\sigma_{\hat{\beta}_k^{(e)}}^2 = \frac{\sigma_\varepsilon^2}{SST_{\mathbf{trt}_k}\left(1 - R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k},\mathbf{Z}}\right)}$$

$$= \frac{1 - R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k}}}{1 - R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k},\mathbf{Z}}}\sigma_{\hat{\beta}_k}^2$$

$$= \kappa\sigma_{\hat{\beta}_k}^2, \quad \text{where } \kappa = \frac{1 - R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k}}}{1 - R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k},\mathbf{Z}}}$$

The second equality uses the fact that when $\boldsymbol{\theta} = \mathbf{0}$, adding $\mathbf{Z}$ to the model does not change the population residual variance. Further, because $Cov\left(\hat{\beta}_k, \hat{\beta}_k^{(e)}\right) = \sigma_{\hat{\beta}_k}^2$ per Lemma 3,[1] we have:

$$\hat{\beta}_k - \hat{\beta}_k^{(e)} \sim N\left(0, (\kappa - 1)\sigma_{\hat{\beta}_k}^2\right)$$

Note that $\kappa \geq 1$ always, because adding regressors cannot decrease the regression R^2 (i.e., $R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k},\mathbf{Z}} \geq R^2_{\mathbf{trt}_k|\mathbf{X}_{-\mathbf{trt}_k}}$). Following the logic in Supplement Proposition 6, power to detect an effect of $\beta_k^{(e)} = \beta_k^*$ at level α in a one-sided test is:

$$p_e = \Phi\left(\Phi^{-1}(\alpha) + \frac{\beta_k^*}{\sqrt{\kappa}\sigma_{\hat{\beta}_k}}\right)$$

Likewise the power to rule out a difference $\beta_k - \beta_k^{(e)} \geq \beta_k^*$ at level α is:

$$p = \Phi\left(\Phi^{-1}(\alpha) + \frac{\beta_k^*}{\sqrt{\kappa - 1}\sigma_{\hat{\beta}_k}}\right)$$

Because $\sqrt{\kappa - 1} < \sqrt{\kappa}$, we know $p > p_e$ for $\kappa > 1$. (When $\kappa = 1$, the reduced and expanded estimators coincide, and no test is needed.) $\square$

The assumption $\beta_k^* > 0$ is set because we posit that researchers aim to show that violations are small in magnitude. With a negative violation, the same logic applies with a non-inferiority null hypothesis of $H_0 : \beta_k - \beta_k^{(e)} \leq \beta_k^*$. Furthermore, tests for main effects in reduced and expanded models would typically be two-sided, further reducing their power relative to a direction-specific non-inferiority test.

Last, for completeness, note that the power to detect an effect in the reduced model at level α in a one-sided test is:

$$p_r = \Phi\left(\Phi^{-1}(\alpha) + \frac{\beta_k^*}{\sigma_{\hat{\beta}_k}}\right)$$

If $\kappa < 2$, indicating low added explanatory power for $\mathbf{trt}_k$ from $\mathbf{Z}$, then $p > p_r$.

7 Appendix G. Additional results for the ACA dependent coverage re-analysis

During most of the pre-period of this study, the US experienced rapid growth in unemployment due to a recession (see the bottom right plot of Figure S1). According to the Federal Reserve, this recession lasted from Jan 2008 to June 2009.[6]

The three age groups were affected differently by this rising unemployment. During the period of rising unemployment, both the treated group (ages 19-25) and the older comparison group (ages 27-29) lost employer coverage, while the younger comparison group (ages 16-18) lost dependent coverage (presumably as their parents became unemployed). Those in the younger comparison group were eligible for Medicaid as dependents and therefore received government insurance. Those in the treated group (even in the pre-period) often gained dependent coverage (perhaps due to a combination of voluntary coverage of people in these age groups by employers and state laws already enacted prior to the ACA). The older comparison group was more likely to become uninsured, having access to neither dependent coverage nor Medicaid.

Why didn't the original authors' evaluations of pre-intervention trends (including those reported in their Appendix Table A1) suggest violations? First, although the two comparison age cohorts had different trends, their average trends were more similar to the treated group (Figure S1). Although some visual differences may still be apparent (see Figure S2, which expands on Figure 1 of Akosa Antwi et al. by including the rest of the insurance variables), there were likely two additional factors at play. As noted in the main text, conventional tests have limited power and may have been underpowered to detect effects. This may have been exacerbated by the fact that test models were fit to pre-intervention data and tested for differential trends using the same control variables as their analytic specification, which included an interaction term between unemployment and treatment group that was highly correlated with the interaction between time and treatment group in the pre-intervention period.

Figure S3 illustrates the impact of this collinearity. It plots the pre-period slopes (θ) estimated by fitting the model with a differential linear trend to pre-period data only, varying the included control variables in $\mathbf{X}_{it}$. With the interaction between treatment group and unemployment in the model, the added differential trend parameter was nearly collinear, roughly doubling the standard errors of the estimated slope difference. This made it more likely that we would pass a conventional trends test, but fail a non-inferiority test.

For completeness, we used the original authors' model specifications (with all control variables in the regression) to replicate their results for the conventional tests of parallel pre-trends (Table S2) and estimated treatment effects (Tables S3-S4). However, we also show

results without the treatment/unemployment interaction, which has greater non-inferiority test power.

Table S2: Conventional tests of parallel trends in the pre-implementation period. These were estimated by fitting the model in Eq. (15) of the main text, using the specification and all control variables from the original analysis.[7] The estimates (and the state-clustered robust standard errors) represent monthly differential slopes on the percentage point scale. SE = standard error

Outcome	Estimate (SE)
Any	0.11 (0.09)
Dependent	0.06 (0.1)
Employer	0.05 (0.09)
Individual	-0.02 (0.04)
Government	0.01 (0.07)

Table S3: Estimated effects of the ACA dependent coverage provision on insurance coverage (replication). These were estimated by fitting the reduced model in Eq. (13) of the main text, using the original specification and all control variables from the original analysis.[7] The estimates (and their state-clustered robust standard errors) represent differential changes on the percentage point scale, averaged over the post-implementation period. SE = standard error

Outcome	Estimate (SE)
Any	3.18 (0.74)
Dependent	7.02 (0.69)
Employer	-3.12 (0.60)
Individual	-0.80 (0.23)
Government	-0.25 (0.57)

Table S4: Estimated effects of the ACA dependent coverage provision on insurance coverage. These were estimated by fitting the reduced model in Eq. (13) of the main text, using all control variables from the original analysis with month-year fixed effects but changing the specification to include saturated treatment effects (and omitting linear trend variables).[7] The estimates (and their state-clustered robust standard errors) represent differential changes on the percentage point scale, averaged over the post-implementation period. SE = standard error

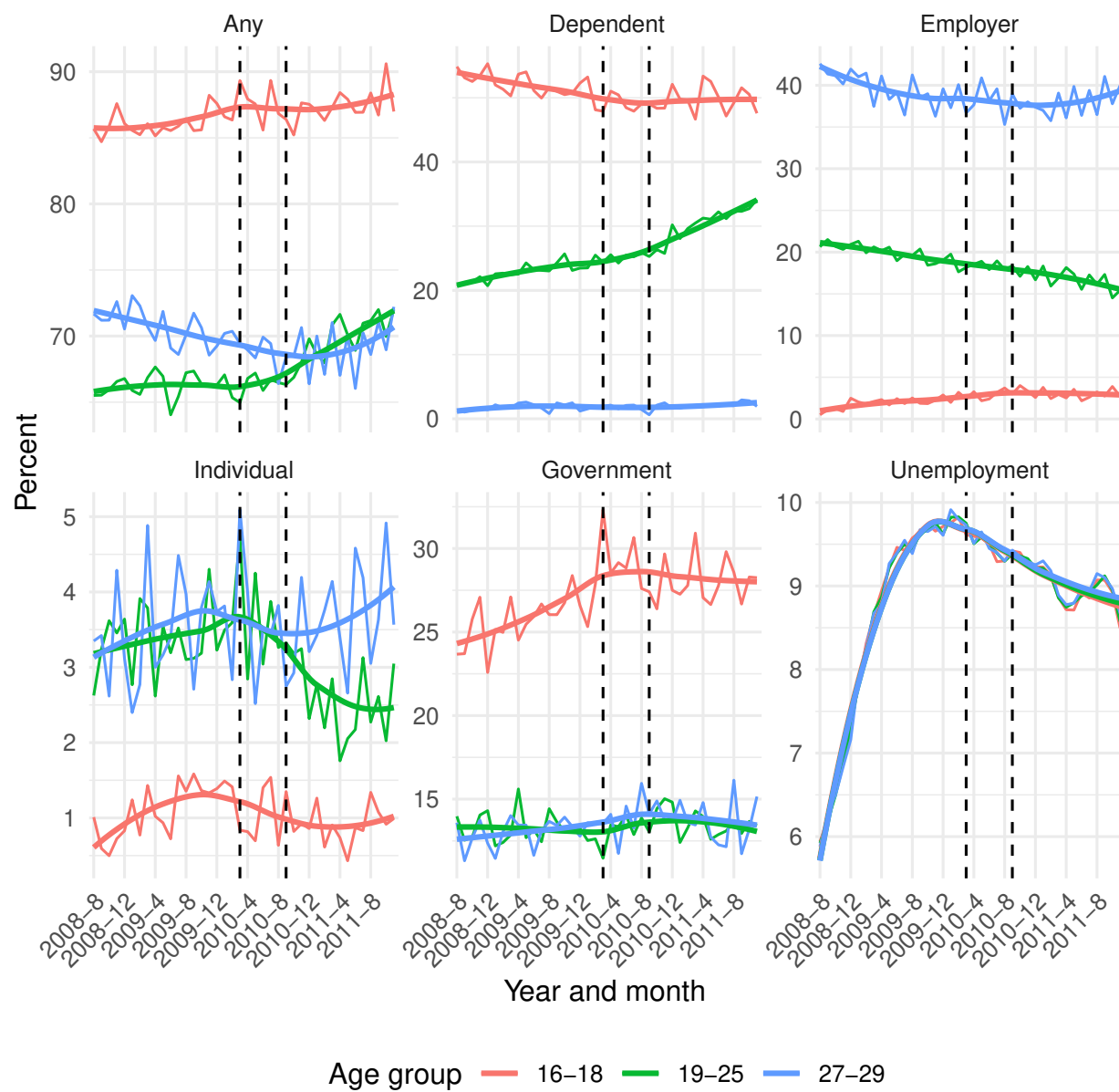| Outcome	Estimate (SE)
Any	3.16 (0.75)
Dependent	6.99 (0.71)
Employer	-3.10 (0.6)
Individual	-0.80 (0.23)
Government	-0.23 (0.57)

Figure S1: Monthly insurance coverage and unemployment by age group. Dashed vertical lines mark the beginnings of the enactment and implementation periods.
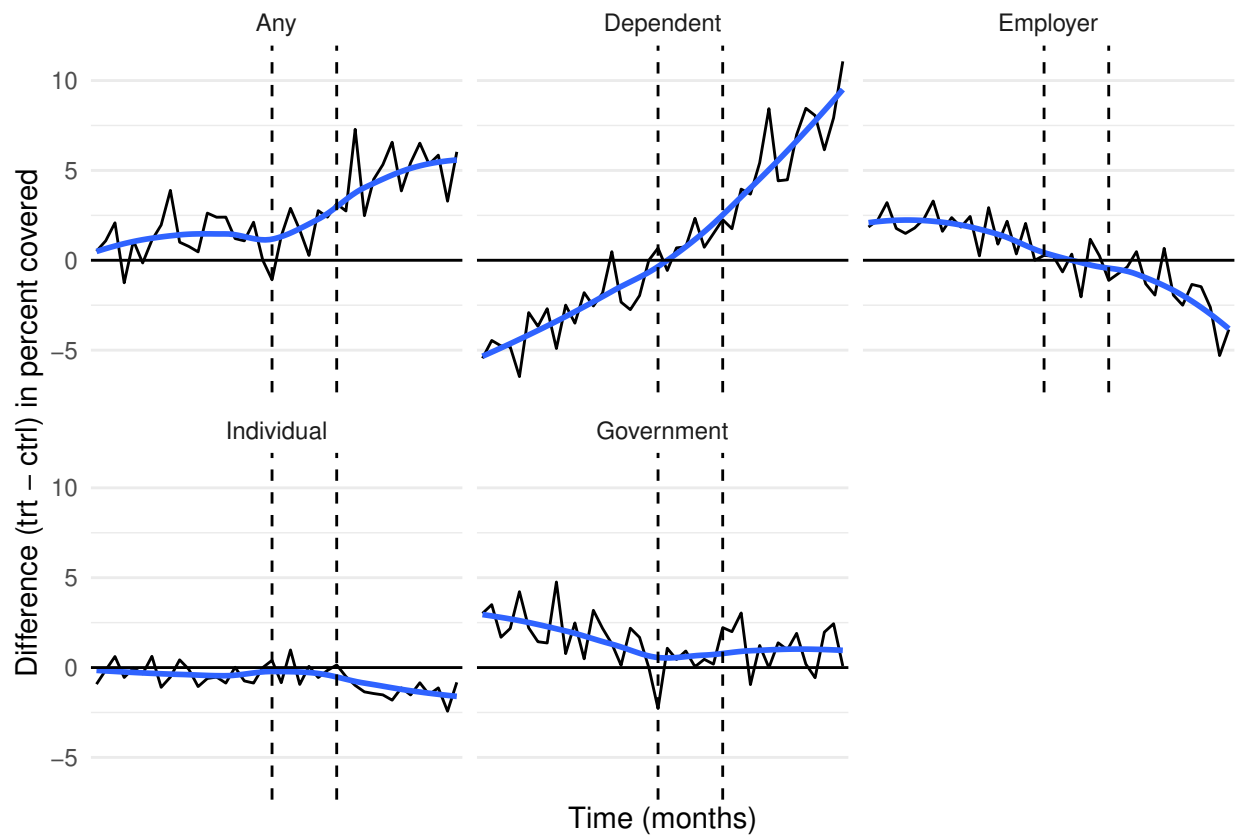
Figure S2: Differential monthly insurance coverage trends in treatment versus control. Dashed vertical lines mark the beginnings of the enactment and implementation periods.
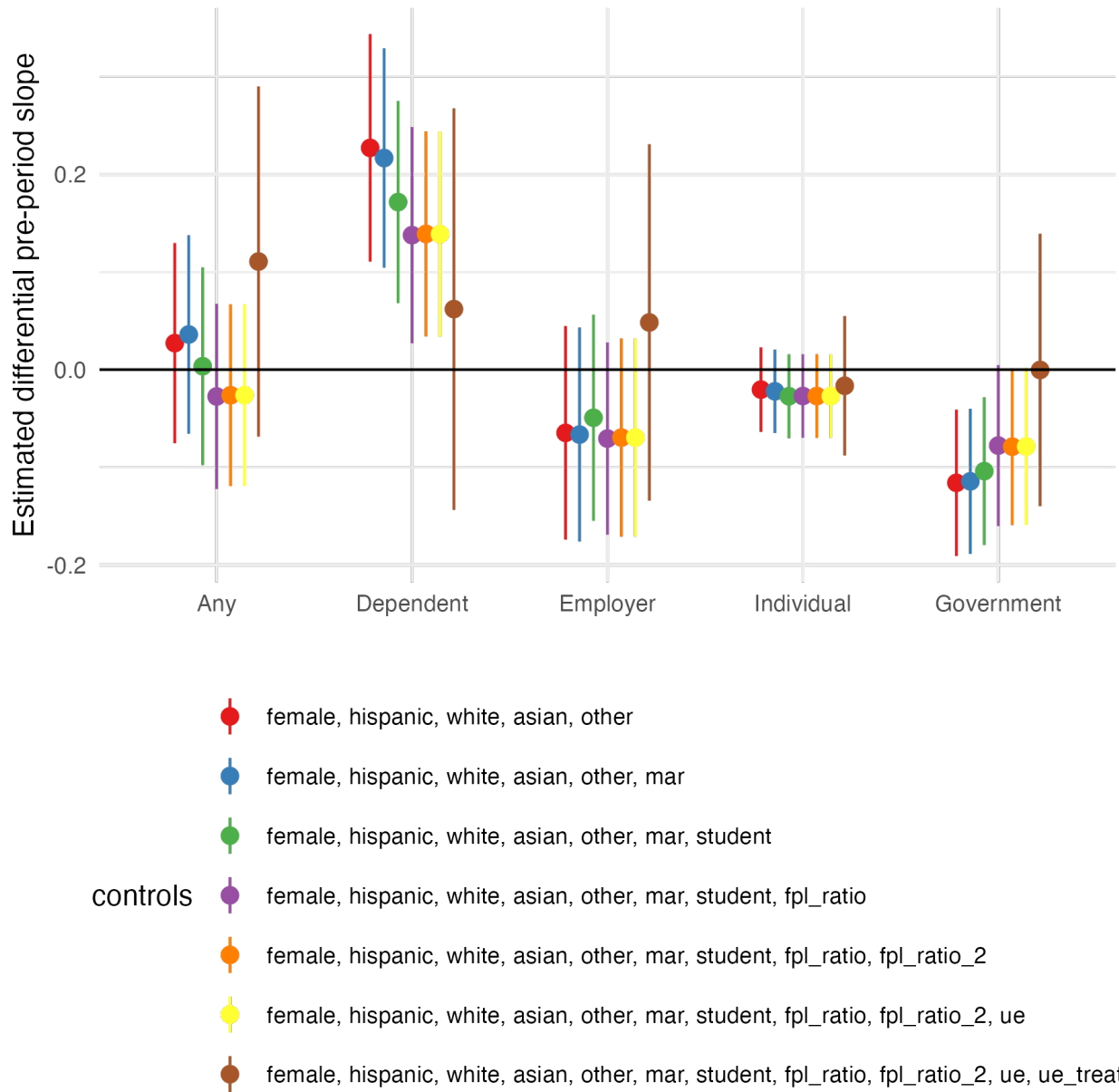
Figure S3: Estimated differential pre-period slopes from the model in Eq. (15) of the main text, using different sets of control variables (estimated only on pre-intervention data). mar = married; fpl_ratio = household income expressed as a ratio of the federal poverty limit; ue = state-month unemployment; ue_treat = state-month unemployment interacted with treatment group indicator

References

[1] Clogg CC, Petkova E, Haritou A. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology.* 1995;100(5):1261–1293. doi: 10.1086/230638

[2] Lu TT, Shiou SH. Inverses of 2×2 Block Matrices. *Computers & Mathematics with Applications.* 2002;43(1-2):119–129. doi: 10.1016/S0898-1221(01)00278-4

[3] Colin Cameron A, Miller DL. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources.* 2015;50(2):317–372. doi: 10.3368/jhr.50.2.317

[4] Lee JD, Sun DL, Sun Y, Taylor JE. Exact Post-Selection Inference, with Application to the Lasso. *The Annals of Statistics.* 2016;44(3):907–927.

[5] Roth J. Pretest with Caution: Event-Study Estimates After Testing for Parallel Trends. *American Economic Review: Insights.* 2022;4(3):305–322. doi: 10.1257/aeri.20210236

[6] Federal Reserve Bank of St. Louis . NBER Based Recession Indicators for the United States from the Period Following the Peak Through the Trough [USREC]. https://fred.stlouisfed.org/series/USREC; 2023. Accessed 7 June 2023.

[7] Akosa Antwi Y, Moriya AS, Simon K. Effects of Federal Policy to Insure Young Adults: Evidence from the 2010 Affordable Care Act's Dependent-Coverage Mandate. *American Economic Journal: Economic Policy.* 2013;5(4):1-28. doi: 10.1257/pol.5.4.1