The Evolving Toolkit of Difference-in-Differences: What Comes Next?

Pedro H. C. Sant'Anna
Emory University

Southern Economic Association
November 22, 2026

What Comes Next?



I should have checked this before choosing the title.

“The really challenging thing is asking good questions.”

Guido Imbens

Nobel Prize Museum quote



EMORY UNIVERSITY


DDDiD


did
DDDiD

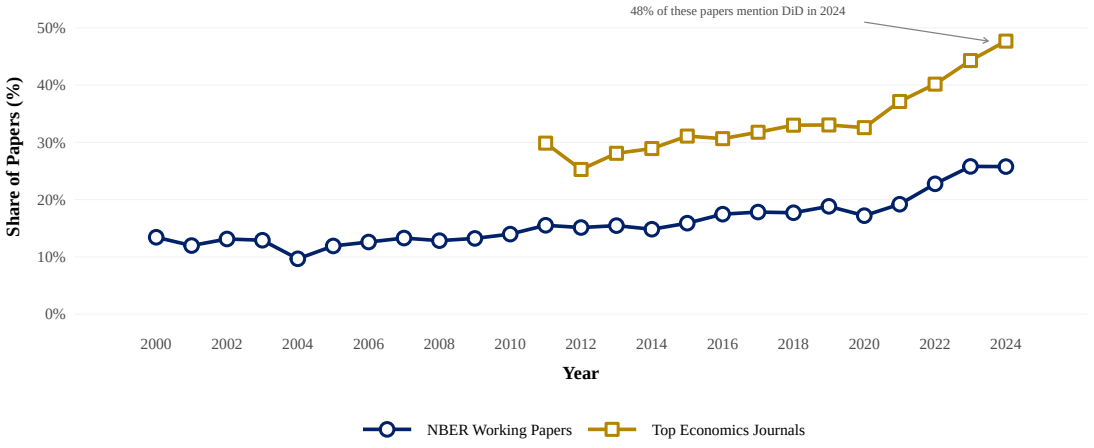
DiD or DDDiD





The Rise of Difference−in−Differences in Economics

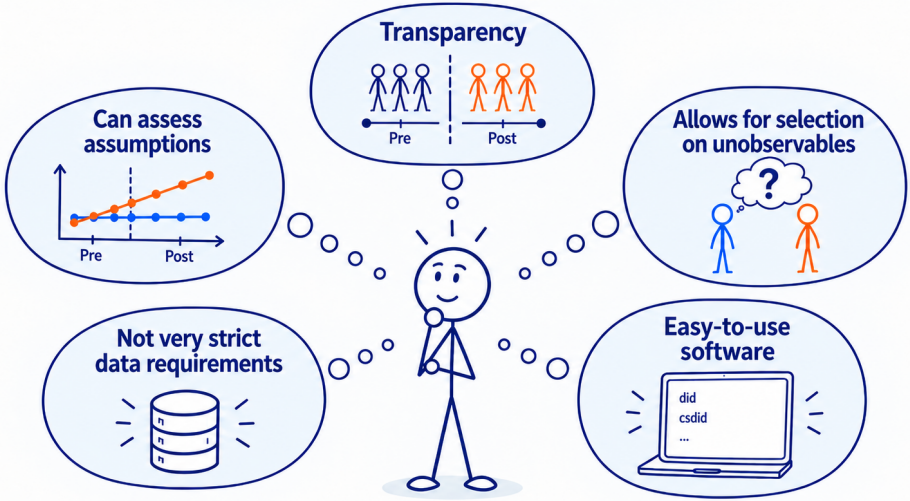
Top Journals: AER, QJE, JPE, AEJ:Applied, AEJ:Policy



Search terms: "Difference−in−Differences" | "Difference−in−Difference" | "DiD" | "DD" | "D−i−D" | "Difference in Differences" | "Difference in Difference" | "Diff in Diff" | "Diff−in−Diff"

Keyword mentions in the five named journals. Data from Goldsmith-Pinkham (2026)'s literature database.

Why is DiD so popular?



EMORY UNIVERSITY

The static two-way fixed-effects regression

Canonical DiD

$$Y_{it} = \underbrace{\alpha_i}_{\text{unit FE}} + \underbrace{\lambda_t}_{\text{time FE}} + \underbrace{\beta D_{it}}_{\text{Treatment}} + \varepsilon_{it}$$

One treated city, one comparison city

Treated city: $10 \to 14$ Change $+4$

Comparison city: $8 \to 9$ Change $+1$

$$\widehat{\beta} = 4 - 1 = 3$$

Assume both cities would have changed by the same amount without the policy.

- With staggered adoption and heterogeneous effects, TWFE can give some effects **negative weights**. Then β may miss the policy-relevant average (Goodman-Bacon, 2021).

Meanwhile

The DiD revolution

A growing toolkit makes the comparisons explicit and lets effects vary across groups and over time.

Callaway and Sant'Anna (2021) · Sun and Abraham (2021)
Borusyak, Jaravel and Spiess (2024) · Wooldridge (2025)
de Chaisemartin and D'Haultfœuille (2020), among others.

How can we be “future-ready”?

- **What is the starting point of our analysis?**

1. A regression specification of the type

$$Y_{i,t} = \alpha_i + \lambda_t + \beta D_{i,t} + X'_{i,t}\gamma + \varepsilon_{i,t}$$

alluding that β captures the treatment effect of interest under some intuitive conditions.

2. A conscious choice of what we want to estimate (target parameters), paired with well-defined identification conditions.

- Almost all **problems** highlighted by the recent DiD lit can be credited to using the “reverse-engineering” approach 1:
Extrapolate from a specification formalized to one setup to other more complex ones.

- Almost all recent DiD **solutions** follow the “forward-engineering” approach 2:
Use estimators that target pre-specified parameters that respect the ident. assumptions

Forward-Engineering Approach to DiD

Journal of Economic Literature 2026, 64(2), 498–557
https://doi.org/10.1257/jel.20251650

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

Andrew Baker, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna*

Difference-in-differences (DiD) is arguably the most popular quasi-experimental research design. Its canonical form, with two groups and two periods, is well understood. However, empirical practices can be ad hoc when researchers go beyond that simple case. This article provides an organizing framework for discussing different types of DiD designs and their associated DiD estimators. It discusses covariates, weights, handling multiple periods, and staggered treatments. The organizational framework, however, applies to other extensions of DiD methods as well. (*JEL* C23, H75, I12, I38)

Baker, Callaway, Cunningham, Goodman-Bacon and Sant'Anna (2026)

Forward engineering in eight steps

Design the study

1. **Define target parameters.**
   What causal effect answers your question?
2. **State (formally) the identification assumptions.**
   Defend parallel trends in context.
3. **Justify the estimation method.**
   Choose a method that targets the effect under those assumptions.
4. **Discuss sources of uncertainty.**
   Say what is random. Match inference to the design.

Estimate and learn

5. **Estimate.**
   Apply the chosen method to your data.
6. **Conduct sensitivity analysis.**
   How much would plausible violations change the answer?
7. **Conduct heterogeneity analysis.**
   Do effects vary by group, time, or exposure?
8. **Keep learning.**
   Revisit the design or build better tools when needed.

Baker et al. (2026), Section 6.

Making a Difference

A large-scale meta-analysis of DiD estimates



Joint work with Victor Rangel, building and expanding on his master's thesis.

What happens when we revisit published DiD studies?

Reanalysis in political science and economics

Yiqing Xu and coauthors

49

Political science studies reanalyzed

Six estimators and checks on parallel trends.

Chiu, Lan, Liu and Xu (2026)

Generalized synthetic control (Xu, 2017).

Peter Hull

116

Economics papers reanalyzed

1,430 staggered-adoption specifications.

Hull (2026)

We screened 85,571 abstract records

Joint work with Victor Rangel spanning 88 economics journals

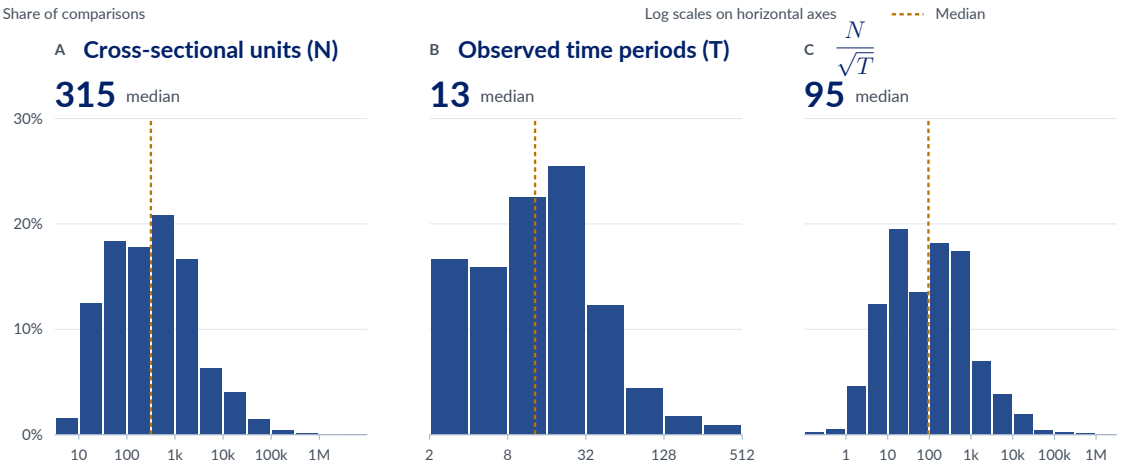
Journal	Abstract records	Full-text papers	DiD papers	Staggered DiD	Replicated papers	With binary treatment
American Economic Review	4,852	622	263	34	50	47
Econometrica	1,670	75	26	3	6	6
Journal of Political Economy	1,528	193	76	9	15	14
Quarterly Journal of Economics	1,252	256	122	11	29	27
Review of Economic Studies	1,732	71	38	3	8	8
Review of Econ. and Statistics	2,369	320	176	34	49	48
AEJ: Applied Economics	783	391	200	36	49	49
AEJ: Economic Policy	863	450	247	33	83	80
Other screening records	70,522	1,482	1,029	130	212	188
Total	**85,571**	**3,860**	**2,177**	**293**	**501**	**467**

Staggered counts use all full-text papers. Binary counts use the replicated papers.

Literature scan by Rangel and Sant'Anna (2026).

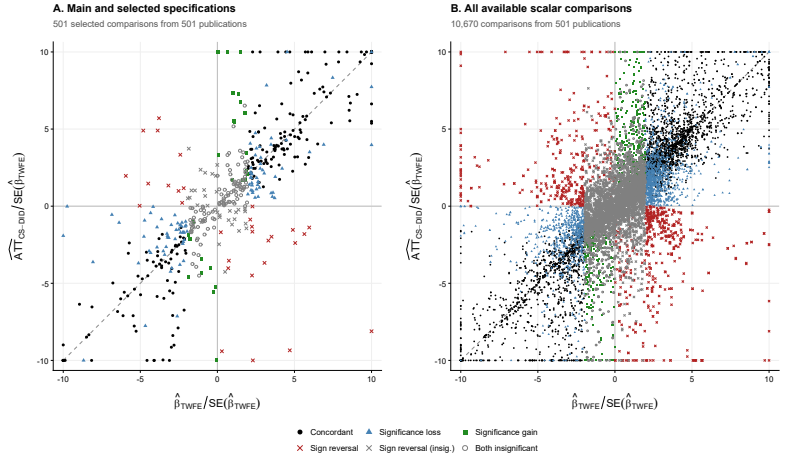
How large are the panels in the DiD reanalysis?

8,078 panel comparisons with verified dimensions across 404 publications



What changes when we re-estimate DiD?

Original and Callaway–Sant'Anna estimates from 501 papers



17%

of 10,670 estimate pairs switch signs

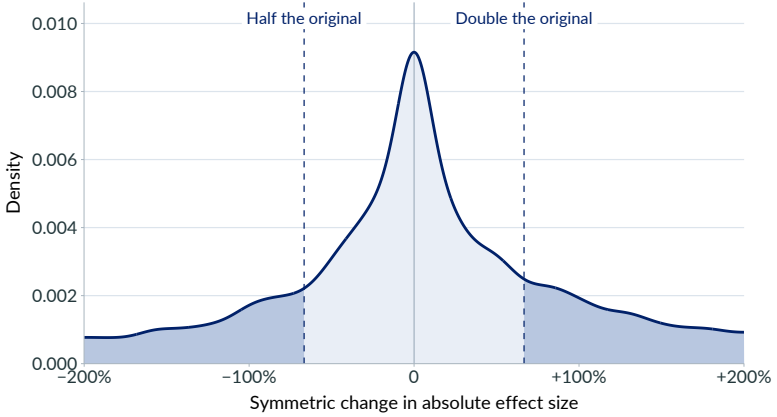
For 107 of the 1,820 reversals, both estimates' individual 95% intervals exclude zero.

Original estimate on x, CS-DID on y. The diagonal shows agreement.

Reanalysis by Rangel and Sant'Anna (2026).

Effect sizes often move by a factor of two

Changes in absolute magnitude across 10,670 estimate pairs



37%

at least double in magnitude or fall by half or more

3,952 of 10,670 pairs

Same sign?

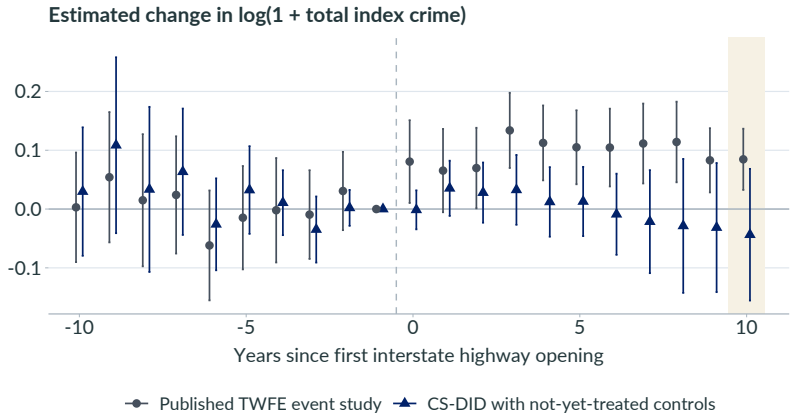
31% still double or fall by half.

Reanalysis by Rangel and Sant'Anna (2026).

EMORY UNIVERSITY

The long-run crime estimate changes sign

U.S. counties gained access to the interstate highway network at different times.



Bars show pointwise 95% confidence intervals.

Year ten

Published TWFE

+0.085

SE 0.027. CI excludes zero.

CS-DID reanalysis

−0.044

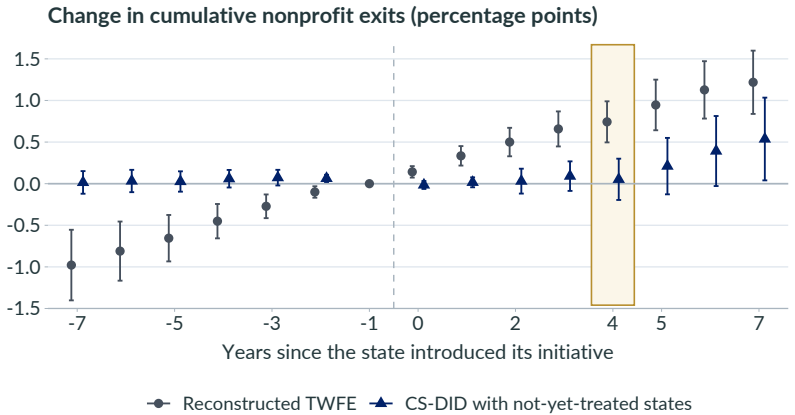
SE 0.057. CI includes zero.

The baseline and county-year sample also change.

EMORY UNIVERSITY

The TWFE exit path is already rising before adoption

States began promoting partnerships with faith-based nonprofits at different times.



Bars show pointwise 95% confidence intervals.

Year four

Reconstructed TWFE

+0.74 pp

SE 0.12. CI excludes zero.

CS-DID reanalysis

+0.05 pp

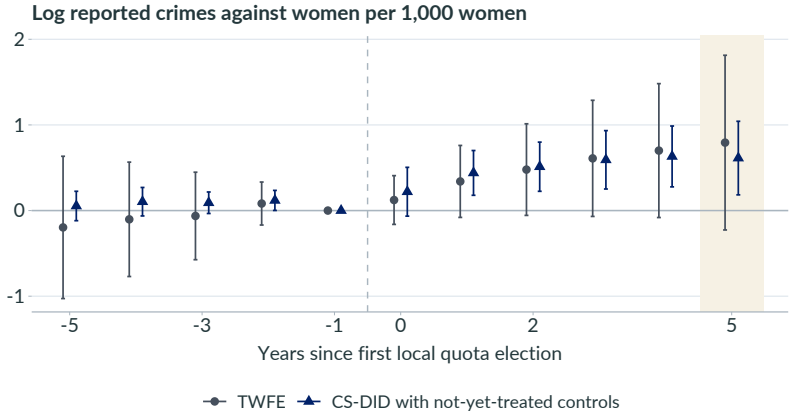
SE 0.13. CI includes zero.

The published study uses a stacked design.

EMORY UNIVERSITY

Similar point estimates, very different precision

Indian states began reserving local council seats for women at different times.



Bars show pointwise 95% confidence intervals.

Year five

TWFE +0.79

SE 0.48. CI includes zero.

CS-DID +0.61

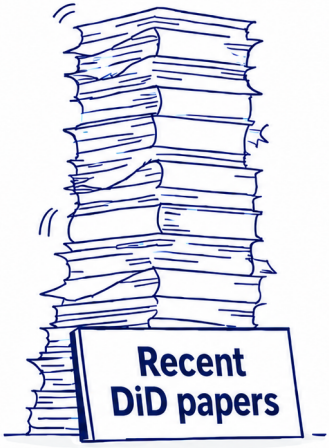
SE 0.22. CI excludes zero.

Same 357 state-years, with no covariates.

The original study links the increase to greater reporting.

EMORY UNIVERSITY

DiD has evolved so much in recent years...





But so many basic questions remain



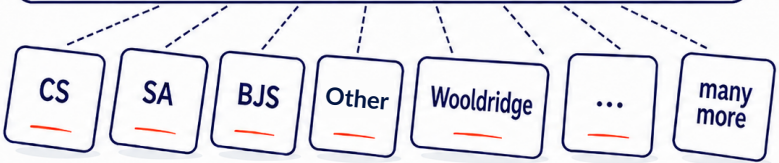
DiD is often nonparametrically overidentified



This paper leverages NP overidentification to better understand DiD

What does this mean?

Under (conditional) PTs, many different DiD estimators can identify the same effect

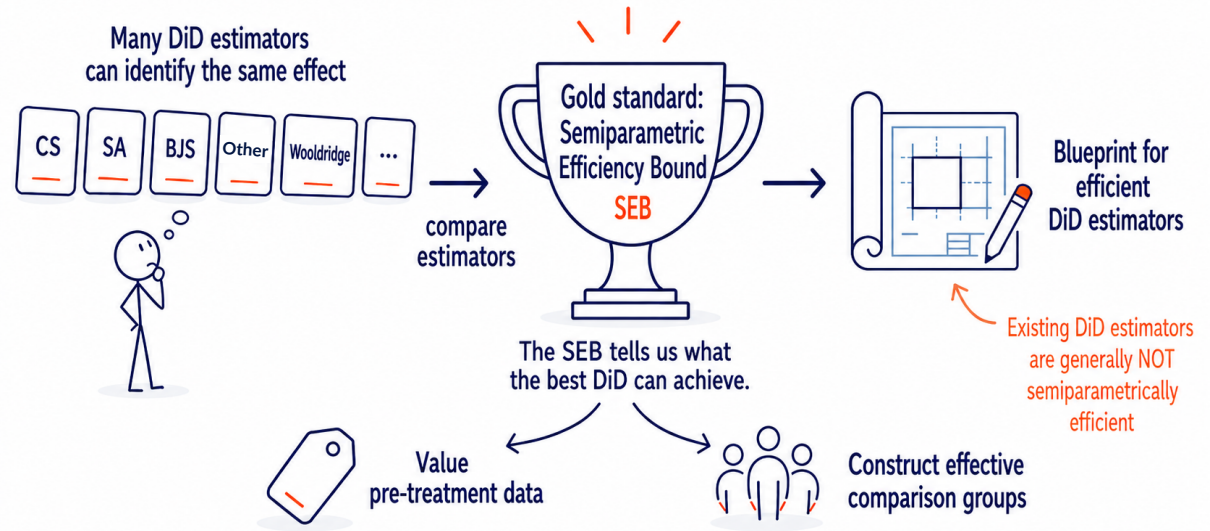


The familiar procedures are only some of the possibilities



How should we choose between the many options?

We need a gold standard for comparison



For block and staggered adoption; with and without covariates; and for unit or clustered sampling.

Overidentification also gives us diagnostics





1 **Restriction-aligned pre-treatment ES**
Do the comparisons agree before treatment?

2 **Target and system diagnostics**
What changes with added history?
H_{RU}: target movement
J / incremental: system restrictions

3 **Precision–sensitivity frontier**
What is the trade-off?
Andrews, Chen and Tecchio (2025)

Overidentification helps us assess the coherence of our design.

These tools do not prove parallel trends; they help us understand agreement and sensitivity.

Efficient Difference-in-Differences and Event Study Estimators

Xiaohong Chen, Pedro H. C. Sant'Anna, and Haitian Xie

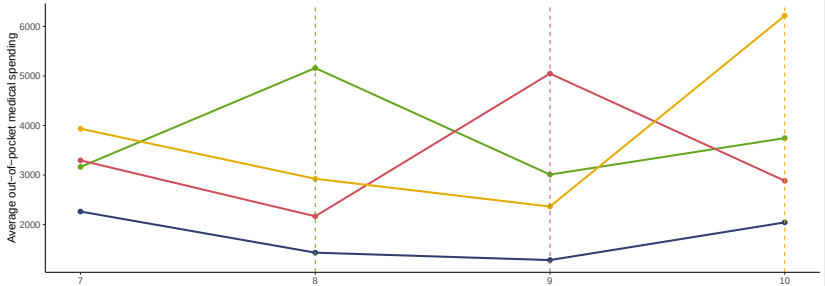
Cowles Foundation Discussion Paper 2470 (Chen, Sant'Anna and Xie, 2025)

Let's look at one application

Spending spikes when each group is admitted

Out-of-pocket spending for 656 U.S. adults hospitalized at ages 50–59

Average spending between interviews (dollars)



Admission wave **Wave 8** **Wave 9** **Wave 10** **Wave 11**

When were they admitted?

Dashes mark admissions in waves 8–10. Wave 11 admissions come later.

Who can be a control?

Patients who will be admitted later.

Original study: Dobkin, Finkelstein, Kluender and Notowidigdo (2018). Reanalysis of 656 adults: Chen et al. (2025).

EMORY UNIVERSITY

All four methods put the average increase near $1,500

Estimated change in out-of-pocket spending (2005 dollars). Standard errors in parentheses.

	Admitted in wave 8			Wave 9		Wave 10	Interviews since admission			
Estimation method	8	9	10	9	10	10	At adm.	Next	2 later	Avg.
Efficient DiD	**3,071** (805)	**1,115** (635)	**1,041** (812)	**3,067** (690)	**95** (636)	**2,911** (889)	**3,025** (486)	**695** (470)	**1,041** (812)	**1,587** (519)
Callaway–Sant'Anna Late controls only	2,826 (1,035)	825 (909)	800 (1,008)	3,031 (702)	107 (651)	3,092 (995)	2,960 (539)	530 (585)	800 (1,008)	**1,430** (647)
Callaway–Sant'Anna All not-yet-admitted	3,029 (913)	1,248 (861)	800 (1,008)	3,324 (959)	107 (651)	3,092 (995)	3,134 (536)	779 (570)	800 (1,008)	**1,571** (566)
Imputation (BJS–G–W)	3,029 (913)	1,285 (765)	1,021 (849)	3,239 (859)	77 (727)	2,758 (955)	3,017 (554)	788 (586)	1,021 (849)	**1,609** (580)

The last column gives equal weight to the first three interviews from admission onward.

Original study: Dobkin et al. (2018). Reanalysis: Chen et al. (2025).

EMORY UNIVERSITY

How much more data buys the same precision?

Variance relative to Efficient DiD. Read 1.55 as about 55% more data for the same precision.

	Admitted in wave 8			Wave 9		Wave 10	Interviews since admission			
Estimation method	8	9	10	9	10	10	At adm.	Next	2 later	Avg.
Efficient DiD	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Callaway–Sant'Anna Late controls only	1.65	**2.05**	1.54	1.04	1.05	1.25	1.23	1.55	1.54	**1.55**
Callaway–Sant'Anna All not-yet-admitted	1.29	1.84	1.54	1.93	1.05	1.25	1.22	1.47	1.54	**1.19**
Imputation (BJS–G–W)	1.29	1.45	1.09	1.55	1.31	1.15	1.30	1.56	1.09	**1.25**

656 adults. Late controls alone would need about 1,020 for the same precision.

Original study: Dobkin et al. (2018). Reanalysis: Chen et al. (2025).

EMORY UNIVERSITY

Triple Differences

Treatment requires both criteria

Policy in force

AND

Eligible



Better Understanding Triple Differences Estimators

Marcelo Ortiz-Villavicencio
Pedro H. C. Sant'Anna

Public working paper **arXiv:2505.09942**



What if wage trends differ across states?

Triple differences uses an additional comparison within each state.

Gruber's setting

Illinois, New Jersey and New York required employer health plans to cover childbirth. Did employers pass the cost on in wages?

How wages might change without the mandate

	Married women 20–40	Comparison workers	Change in wage gap
Mandate states	**+6**	+4	**+2**
Other states	**+3**	+1	**+2**

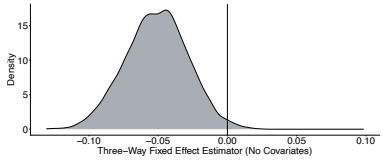
Illustrative numbers. The design needs parallel trends in the wage gap.

Original study: Gruber (1994). Triple-differences design: Ortiz-Villavicencio and Sant'Anna (2026).

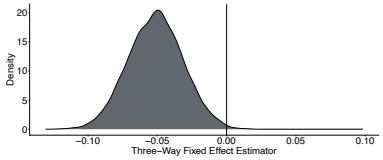
EMORY UNIVERSITY

Adding worker controls can still leave bias

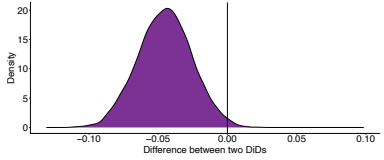
5,000 maternity-mandate simulations. The vertical line marks zero estimation error.



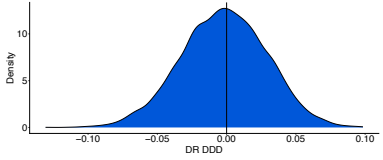
(a) 3WFE without covariates



(b) 3WFE with covariates



(c) Difference between two Doubly Robust DiDs



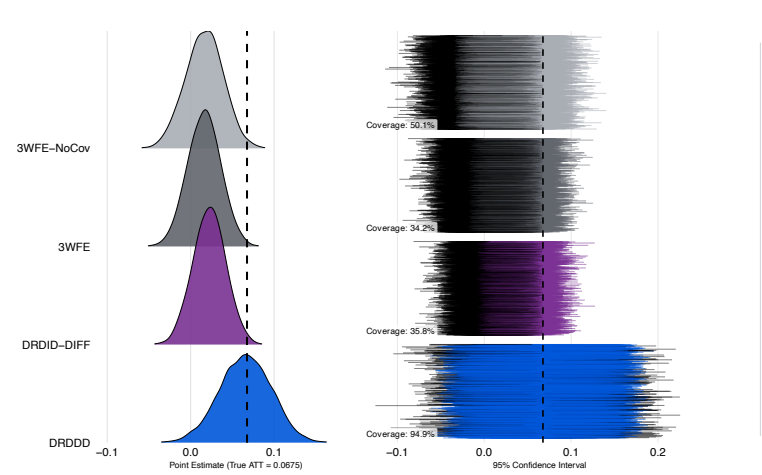
(d) Doubly Robust DDD

Compare workers at the same characteristics, then average over the treated workers.

Ortiz-Villavicencio and Sant'Anna (2026), Figure 2.

How often do 95% intervals cover the true effect?

In the maternity-mandate simulation, black intervals miss the true effect on wages.



34.2%

Three-way fixed effects with worker controls

94.9%

Doubly robust triple differences

Ortiz-Villavicencio and Sant'Anna (2026), Figure OA-1.

Did access to GM seed raise national crop yields?

Eligibility now varies across crops, and countries open access in different years.

Compare yield changes within each country

Eligible crops	Other crops
Maize, cotton, soybean and rapeseed	Including rice and wheat

Then compare that crop gap across countries

Countries that open access to GM seed	Countries without access in the same years

Later adopters and never adopters can have different crop mixes and yield trends.

Original study: Hansen and Wingender (2023). Triple-differences design: Ortiz-Villavicencio and Sant'Anna (2026).

EMORY UNIVERSITY

Pooling untreated cohorts can bias the estimate

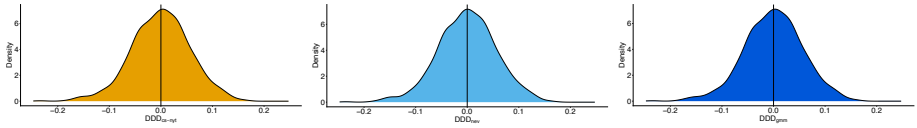
GM crop simulation. Bottom row: comparison cohorts differ in crop mix and yield trends.

(a) DDD using pooled not-yet-treated units

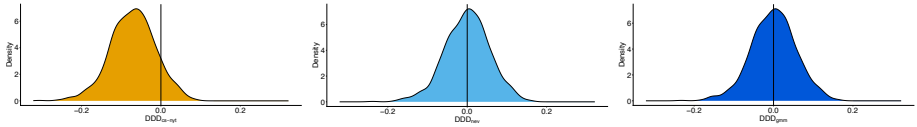
(b) DDD using never adopters only

(c) DDD GMM combining not-yet-treated cohorts

No systematic differences in comparison-cohort composition or untreated trends



Comparison cohorts differ in composition and untreated trends



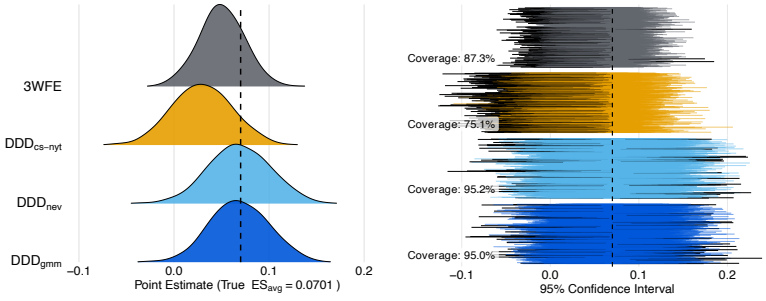
Zero means no estimation error. Compare eligible and other crops within each cohort before combining.

Ortiz-Villavicencio and Sant'Anna (2026), Figure 3.

EMORY UNIVERSITY

Separate cohort comparisons keep coverage near 95%

Crop-yield stress test. Share of 95% intervals that include the true average post-adoption effect.



75.1%
Pooled untreated cohorts

95.0%
Separate cohort comparisons

In the benchmark simulation, all four methods cover the true effect about 93% of the time.

Ortiz-Villavicencio and Sant'Anna (2026), Figure OA-6, panel (b).

What If Treatment Is Continuous?



- What are the causal parameters of interest?
- Is there anything different from the binary case?
- How do I estimate them?

Difference-in-Differences with a Continuous Treatment

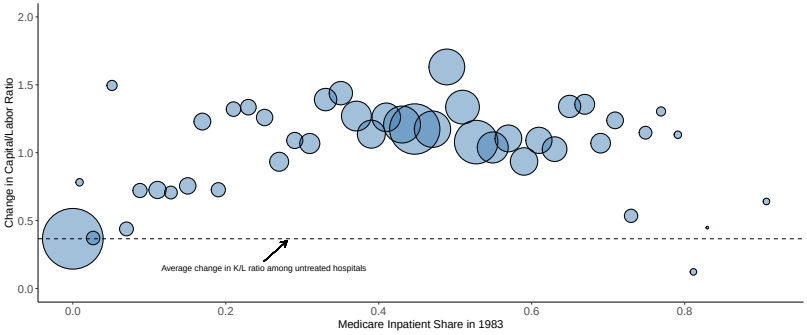
Callaway, Goodman-Bacon and Sant'Anna (2025)

American Economic Review, forthcoming

Read the paper · doi:10.1257/aer.20240137

One Medicare reform, many doses

The policy changed financial incentives more at hospitals with more Medicare patients.



1980–83 vs. 1984–86. Larger circles represent more hospitals.

What changed?

Medicare switched to fixed payments per discharge. Capital costs were still reimbursed separately.

What is the dose?

The share of inpatient days covered by Medicare before the reform.

Capital intensity is depreciation as a share of total costs.

Original study: Acemoglu and Finkelstein (2008). Dose analysis: Callaway et al. (2025), Figure 4.

EMORY UNIVERSITY

Two estimands, two questions about dose

Follow the same hospitals whose Medicare share was 40% before reform.

What did the reform do at the dose they received?

$0\% \longrightarrow 40\%$

Compare the outcome at 40% with what it would have been with zero exposure.

Effect on the treated $\text{ATT}(.40 \mid .40)$

What would a higher dose do to those same hospitals?

$40\% \longrightarrow 50\%$

Hold the hospitals fixed and compare their outcomes at 50% and 40%.

Dose increase for the same group. The local slope is the average causal response, $\text{ACRT}(.40 \mid .40)$.

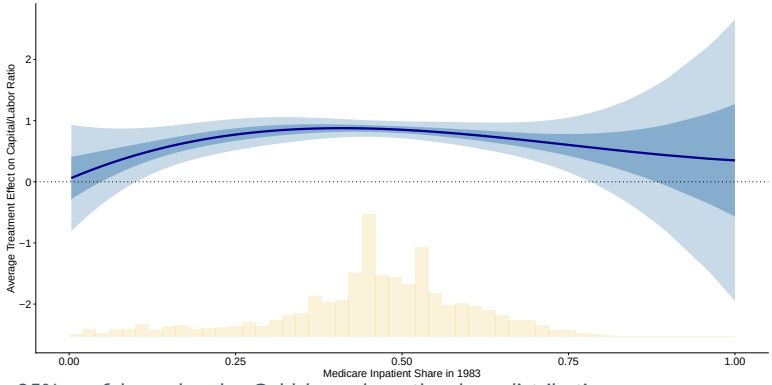
Callaway et al. (2025), Section III.

EMORY UNIVERSITY

What the Medicare curve estimates

The plotted effect compares each group at the dose it received with zero exposure.

Effect on depreciation share



95% confidence bands. Gold bars show the dose distribution.

Recover the level effect

Use zero-dose hospitals to learn how each exposed group would have changed without the reform.

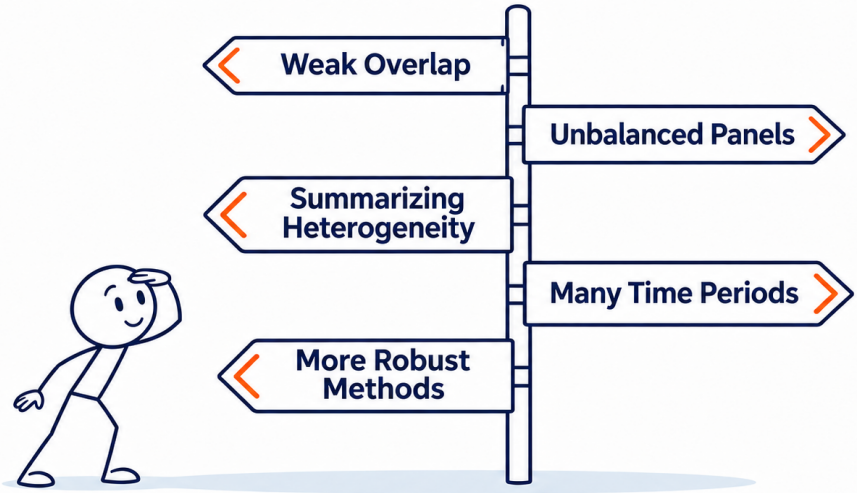
Reading the slope

Each point represents different hospitals. To infer the effect of raising the dose, we need an additional assumption about their responses.

Callaway et al. (2025)

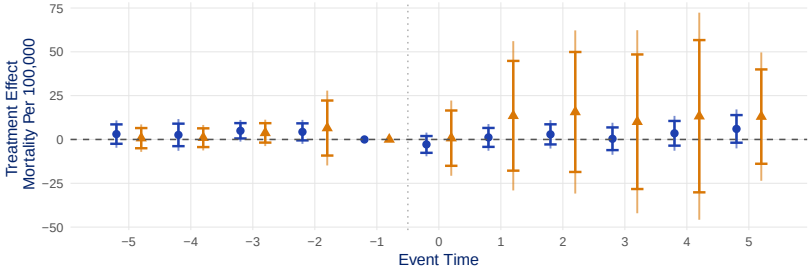
EMORY UNIVERSITY

What's Next?


Weak Overlap
Unbalanced Panels
Summarizing Heterogeneity
Many Time Periods
More Robust Methods

When cohorts are thin, a few counties can dominate - weak overlap

Ma, Sant'Anna, Sasaki and Ura (2026): use the same Medicaid application as in Baker et al. (2026) but handle weak overlap



● Bias-corrected estimate ▲ Standard estimate

The longer intervals are the simultaneous bands. Both include zero.

Average width after expansion

86 to 18

Deaths per 100,000 adults. 95% simultaneous bands.

Correcting for weak overlap narrows the bands.

Can one comparison work under several designs?

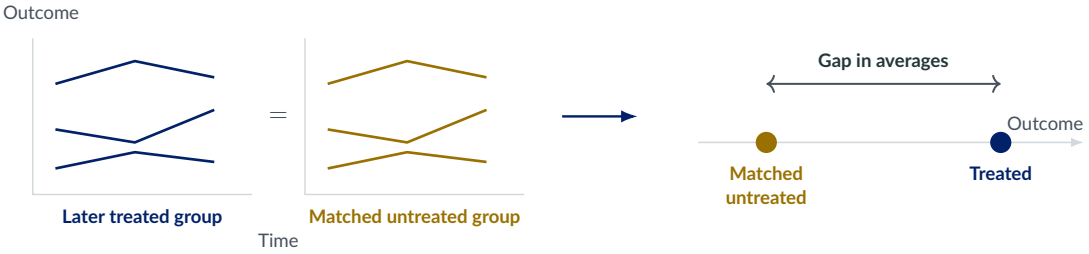
Beyond Parallel Trends · Callaway, Dyal, Sant'Anna and Tsyawo (2026), Section 3.1.

Before treatment

Match the full distribution of outcome histories.

After treatment

Compare the group averages.



The gap identifies the effect if any one set of assumptions holds.

Difference-in-differences • Change-in-changes • Adjustment for lagged outcomes

Schematic.

EMORY UNIVERSITY

What comes next?

Heterogeneity

How do effects vary with a few characteristics, after accounting for the others?

Long panels

As the number of periods grows, when should we move beyond DiD, and what should we use?

Treatment on and off

Where do current methods still fall short?

Unbalanced panels

When some units are missing in some periods, whose treatment effect are we averaging?

Tools can keep different observations and weight treatment cohorts differently.

Whatever the design, the recipe stays the same.

Forward engineering

Define the estimand.

State the identifying assumptions.

Build the analysis from there.

References i

Acemoglu, Daron and Amy Finkelstein, "Input and Technology Choices in Regulated Industries: Evidence from the Health Care Sector," *Journal of Political Economy*, 2008, *116* (5), 837–880.

Andrews, Isaiah, Jiafeng Chen, and Otavio Tecchio, "The Purpose of an Estimator Is What It Does: Misspecification, Estimands, and Over-Identification," 2025. arXiv:2508.13076v4.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna, "Difference-in-Differences Designs: A Practitioner's Guide," *Journal of Economic Literature*, 2026, *64* (2), 498–557.

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

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

References ii

_____ , **Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna**, "Difference-in-Differences with a Continuous Treatment," 2025. Forthcoming in the *American Economic Review*. Publisher record.

_____ , **Derek Dyal, Pedro H. C. Sant'Anna, and Emmanuel S. Tsyawo**, "Beyond Parallel Trends: An Identification-Strategy-Robust Approach to Causal Inference with Panel Data," 2026. November 22 draft. Public working paper.

Chen, Xiaohong, Pedro H. C. Sant'Anna, and Haitian Xie, "Efficient Difference-in-Differences and Event Study Estimators," Discussion Paper 2470, Cowles Foundation 2025.

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu, "Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study," *American Political Science Review*, 2026, *120* (1), 245–266.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

References iii

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo, "The Economic Consequences of Hospital Admissions," *American Economic Review*, 2018, *108* (2), 308–352.

Goldsmith-Pinkham, Paul, "Tracking the Credibility Revolution across Fields," Working Paper 35051, National Bureau of Economic Research 2026.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

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

Hansen, Casper Worm and Asger Mose Wingender, "National and Global Impacts of Genetically Modified Crops," *American Economic Review: Insights*, 2023, *5* (2), 224–240.

Hull, Peter, "How Different are Different Differences-in-Differences? A Large-Scale Meta-Analysis," Working Paper 35974, National Bureau of Economic Research 2026.

References iv

Ma, Yukun, Pedro H. C. Sant'Anna, Yuya Sasaki, and Takuya Ura, "Doubly Robust Estimators with Weak Overlap," 2026. November draft. Public April version.

Ortiz-Villavicencio, Marcelo and Pedro H. C. Sant'Anna, "Better Understanding Triple Differences Estimators," 2026. November 22 draft. Public preprint.

Rangel, Victor and Pedro H. C. Sant'Anna, "Difference-in-Differences Reanalysis," 2026. Unpublished screening and replication archive, November 21 release.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Wooldridge, Jeffrey M., "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," *Empirical Economics*, 2025, *69* (5), 2545–2587.

Xu, Yiqing, "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 2017, *25* (1), 57–76.