JAMA Guide to Statistics and Methods

Estimating Intervention Effects With Difference-in-Differences

Alyssa Bilinski, PhD; Shuo Feng, PhD; Laura A. Hatfield, PhD

Governments, health systems, and payers often adjust approaches to health care delivery and payment with the objective of improving outcomes or reducing costs. To isolate and estimate the effects of these interventions, researchers use strategies developed for nonrandomized studies. A common approach focuses on difference-in-differences, contrasting the change in the outcome of interest from before the intervention to after, between settings experiencing and not experiencing the intervention.[1]

Editor's Note

CME at jamacmelookup.com

The difference-in-differences strategy is often used when an intervention or treatment is applied at a group level—for example, when a hospital or state implements a new policy that is not implemented in other hospitals or states—and when outcomes in both the treated and untreated groups are available before and after the intervention in either longitudinal or repeated cross-sectional data.

In *JAMA Internal Medicine*, Huckfeldt and colleagues[2] reported a study of intensive lifestyle interventions for patients with type 2 diabetes on long-term labor market outcomes. Although the intervention was evaluated in a randomized clinical trial, a nonrandom subset of participants agreed to long-term follow-up and was included in this subsequent study. Huckfeldt et al[2] used a difference-in-differences analysis to account for some baseline differences between treated and untreated groups.

Why Is a Difference-in-Differences Strategy Used?

A difference-in-differences identification strategy is used to answer the causal question, "What was the effect of the intervention in the treated group or groups, compared with what would have happened without the intervention?" The quantity to be estimated—the target estimand—is typically the average treatment effect on those who were treated ("the average treatment effect on the treated," abbreviated ATT).[1,3] This may differ from the average treatment effect that would be observed in untreated groups or the overall population.

A difference-in-differences analysis accounts for several sources of bias that may arise with simpler designs. A pre-post analysis comparing outcomes before and after an intervention in the treated group alone can be biased by unrelated trends (eg, hospital market consolidation, economic trends). An analysis comparing outcomes in treated and untreated groups at a single point after the intervention may be biased by baseline differences between groups (eg, preintervention differences in hospital quality or state resources). By comparing the changes before the intervention to after the intervention between treated and untreated groups, ie, calculating the difference-in-differences, investigators account for both baseline differences between groups that are stable over time and any time trends shared across groups.

For a difference-in-differences estimate to have a causal interpretation, several assumptions are required.[1,3] These include that first, in the absence of the intervention, outcomes would have evolved on average the same way in treated and untreated groups of individuals over the period from before to after the intervention ("parallel trends"). Second, the treatment status of each group does not affect the treatment status or outcomes of the others ("no spillovers" or "no interference"). Third, treated groups do not change their behavior before the intervention takes effect ("no anticipation"). When these assumptions hold, the average change in outcomes in the untreated group can stand in for what would have happened in the treated group without the intervention; therefore, the difference-in-differences estimator is unbiased for the average treatment effect on the treated.

How Is a Difference-in-Differences Analysis Performed and Evaluated?

In its simplest form with 2 groups and 2 time periods, the treatment effect estimate from a difference-in-differences analysis could be calculated in 2 ways. One could first calculate the change from before to after treatment in each group and then, second, subtract the pretreatment to posttreatment change in the untreated group from that in the treated group. Alternatively, one could first calculate the differences between the treated and untreated groups before and after the treatment and then, second, subtract the pretreatment difference from the posttreatment difference. These 2 calculations are mathematically equivalent.

However, in practice, difference-in-differences analyses are typically implemented with regression techniques, with the choice of the regression model depending on the statistical and causal assumptions. While models with fixed effects for time and treatment group and a treatment group by post period interaction can be used in simple cases, recent work has developed alternative regression specifications for more complex situations, eg, when an intervention is adopted by different groups at different times, when incorporating covariates, or when treatments are continuous rather than binary.[1,3,4] Researchers should calculate standard errors to reflect clustering at the level of treatment (eg, for each state when evaluating state policies).[1,3] Common clustered standard error estimators require both many treated and many untreated clusters to avoid inappropriately narrow confidence intervals. If there are too few clusters, researchers should consider alternatives like the wild cluster bootstrap.[5]

To evaluate the appropriateness of drawing causal conclusions from difference-in-differences analyses, researchers should consider the underlying assumptions, especially parallel trends. Because these assumptions cannot be directly tested, researchers generally draw on both context and preintervention data to support their plausibility.[1,3] Ideal untreated comparison groups have outcomes driven by the same underlying trends and are subject to the same effects of external events as treated groups. For example, researchers might choose untreated comparison groups from the same health care market as the treated groups. Time-varying factors that affect the treated and untreated groups differently could introduce bias. Researchers may also provide a rationale for why treated and untreated groups would likely have followed parallel trajectories absent intervention.

Further evidence for the plausibility of the parallel trends assumption is often sought by studying preintervention trends in the

outcomes of the treated and untreated groups. Although parallel preintervention trends do not guarantee that parallel trends would have continued absent intervention, these can provide some reassurance that untreated outcomes in both groups followed similar paths in recent history. Traditionally, researchers evaluated the parallel trends assumption by testing a hypothesis of no difference in trends prior to an intervention.[6] However, these tests are often underpowered and likely to miss meaningful violations of parallel trends.[1,3] Instead, authors can quantify the magnitude of violations of the assumption that can reasonably be ruled out, eg, with noninferiority tests or confidence intervals, or conduct sensitivity analyses of how possible violations would impact treatment effect estimates.[1,3]

A widely used approach for examining preintervention trends is an event study. Event studies consist of a series of separate difference-in-differences analyses, keeping the definition of the pretreatment (or reference) period constant—typically the time just before the intervention of interest—but varying the period being compared across both pretreatment and posttreatment points.[1,3] Analyses comparing 2 periods that occur before the real treatment (leads) should yield results near zero if the parallel trends assumption holds. Leads can also be examined for potential anticipation effects (eg, after a policy was announced but before implementation).

Postintervention event study estimates—in which different periods after the true intervention are compared with the preintervention period—show time trends in treatment effects and can also lend credence to a causal relationship. For instance, estimated effects may appear immediately after the intervention introduction or grow slowly. These patterns should be evaluated for plausibility, given policy implementation and hypothesized mechanisms.

What Are Limitations of Difference-in-Differences Studies?

Difference-in-differences can only be used when there are untreated groups and therefore cannot evaluate interventions that affect everyone simultaneously (eg, many national policies). A causal interpretation is also complicated when multiple related policies change over a similar period. Last, difference-in-differences may be less appropriate when the outcomes are unusually variable over time (eg, due to COVID-19 in the preintervention period).

When the parallel trends assumption is implausible, researchers may choose different untreated comparison groups for which the assumption is more plausible. Alternatively, they may try to quantify how violation of the parallel trends assumption would affect the results by extrapolating preintervention trends seen in each group to the postintervention period or using other sensitivity analyses.[7,8]

How Was Difference-in-Differences Used by Huckfeldt et al?

In Huckfeldt et al,[2] approximately 60% of participants in the randomized trial agreed to data linkage necessary for long-term follow-up. Overall, in this subsample, the treated and untreated groups remained similar. The primary outcomes of employment and receipt of federal disability benefits for 3091 participants were compared for 7 years before and 13 to 15 years after participation in the randomized trial.[2] The difference-in-differences analysis model included study year fixed effects, an indicator for treatment group at baseline, the effect of treatment over time, and additional variables.[2]

How Should the Analysis Be Interpreted?

Huckfeldt et al[2] found a 2.9 (95% CI, 0.3 to 5.5) percentage-point increase in employment among individuals who received the intensive lifestyle intervention, with a smaller, statistically insignificant impact on receipt of disability benefits (−0.9 percentage points [95% CI, −2.1 to 0.3]). The authors' preintervention event study estimates, or leads, were close to zero with relatively narrow confidence intervals and lacked a trend over time, lending confidence to the idea that preintervention outcomes were evolving in parallel (eFigures 3 and 4 in the Supplement to Huckfeldt et al[2]). Sensitivity analyses demonstrated stability of the estimated treatment effects across different subgroups and model specifications (eg, different pre periods and fixed effects). The absence of an effect among college-educated participants, who were less likely to hold physically demanding jobs and thus less likely to experience employment effects from the lifestyle intervention, served as an informal placebo test supporting the proposed mechanism. Taken together, these results support the conclusion that the lifestyle intervention led to a modest increase in employment rates among the treated.

ARTICLE INFORMATION

Author Affiliations: Department of Health Services, Policy, and Practice, Brown University School of Public Health, Providence, Rhode Island (Bilinski); Department of Biostatistics, Brown University School of Public Health, Providence, Rhode Island (Bilinski, Feng); Department of Statistics and Data Science, NORC at the University of Chicago, Chicago, Illinois (Hatfield); Statistical Editor, *JAMA* (Hatfield).

Corresponding Author: Alyssa Bilinski, PhD, Brown University School of Public Health, 121 S Main St, 8th Floor, Providence, RI 02903 (alyssa_bilinski@brown.edu).

Published Online: July 16, 2026. doi:10.1001/jama.2026.12551

Conflict of Interest Disclosures: Dr Bilinski reported receiving grants from the National Institute of General Medical Sciences (NIGMS) (1R35GM155224), National Institute on Aging, Council of State and Territorial Epidemiologists (CSTE)/CDC, National Center for HIV, Viral Hepatitis, STD, and Tuberculosis Prevention, National Institute of Environmental Health Sciences, and 1Day Sooner and receiving consulting fees from Adam Goff/Renaissance Philanthropies and the Weiss Asset Management Foundation. Dr Feng reported receiving grants from NIGMS (1R35GM155224) and CSTE/CDC and participating in a summer internship with Analysis Group. No other disclosures were reported.

REFERENCES

1. Feng S, Ganguli I, Lee Y, Poe J, Ryan A, Bilinski A. Difference-in-differences for health policy and practice: a review of modern methods. *Stat Med*. 2025;44(23-24):e70247. doi:10.1002/sim.70247

2. Huckfeldt PJ, Yu JC, O'Leary PK, et al. Association of intensive lifestyle intervention for type 2 diabetes with labor market outcomes. *JAMA Intern Med*. 2023;183(10):1071-1079. doi:10.1001/jamainternmed.2023.3283

3. Roth J, Sant'Anna PHC, Bilinski A, Poe J. What's trending in difference-in-differences? *J Econom*. 2023;235(2):2218-2244. doi:10.1016/j.jeconom.2023.03.008

4. Callaway B, Goodman-Bacon A, Sant'Anna PHC. *Difference-in-Differences With a Continuous Treatment*. National Bureau of Economic Research; 2024. doi:10.3386/w32117

5. Rokicki S, Cohen J, Fink G, et al. Inference with difference-in-differences with a small number of groups. *Med Care*. 2018;56(1):97-105. doi:10.1097/MLR.0000000000000830

6. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA*. 2014;312(22):2401-2402. doi:10.1001/jama.2014.16153

7. Rambachan A, Roth J. A more credible approach to parallel trends. *Rev Econ Stud*. 2023;90(5):2555-2591. doi:10.1093/restud/rdad018

8. Ye T, Keele L, Hasegawa R, Small DS. A negative correlation strategy for bracketing in difference-in-differences. *J Am Stat Assoc*. 2024;119(547):2256-2268. doi:10.1080/01621459.2023.2252576