HSR Health Services Research

DOI: 10.1111/1475-6773.13015
DEBATE-COMMENTARY

Editorial

Well-Balanced or too Matchy–Matchy? The Controversy over Matching in Difference-in-Differences

The United States' health care system is experiencing remarkable change. The Patient Protection and Affordable Care Act—and its subsequent revisions—reshaped insurance for millions of Americans (Courtemanche et al. 2017). The Affordable Care Act also introduced myriad payment reforms to improve quality and value (Zuckerman et al. 2016). Within commercial insurance, provider consolidation (Post, Buchmueller, and Ryan 2018) and experiments with reference pricing (Robinson and Brown 2013) are having potentially profound effects on health care prices.

These remarkable changes demand evaluation. Difference-in-differences has emerged as the key strategy to evaluate changes in health care. These methods are intuitive, simple, and relatively easy to implement (Angrist and Pischke 2010). This, coupled with researchers' increasing access to rich datasets and enhanced computing power, has led to an explosion in the use of difference-in-differences as a tool to evaluate programs and policy.

This is a welcome development. Yet despite its widespread adoption, major issues related to specification of difference-in-differences have gone unresolved. Chief among them is the choice of comparison group. Researchers require a comparison group, not exposed to the target intervention, to serve as the counterfactual for the treated group. One option for the comparison group is all untreated units. Yet the treated units and comparison units may be meaningfully different, either with respect to the study outcomes or covariates related to the outcomes. Outcomes for the treated and comparison groups may also not be "parallel" prior to the start of the intervention. Parallel trends are often considered to be the key assumption underlying the validity of difference-in-differences.

In the face of these challenges, researchers may opt to select a subset of untreated units as a more appropriate comparison group. For instance, a study of Medicaid expansion prior to the Affordable Care Act identified comparison states as those that neighbored expansion states and were most similar with respect to population and demographic characteristics (Sommers, Katherine, and Epstein 2012). Alternatively, researchers may choose to explicitly match treated units to one or more comparison group on the basis of preintervention levels in outcomes (O'Neill et al. 2016; Ryan et al. 2017), trends in outcomes, or covariates thought to be relevant to outcomes (Werner et al. 2011; Figueroa et al. 2016). The point of choosing a subset from the universe of untreated units is the expectation that this subset may serve as a more appropriate counterfactual. In other words, without the program or policy, the future outcomes of this subset may be expected to change at the same rate as the treated group.

This brings us to this study. Daw and Hatfield perform a Monte Carlo simulation study to assess the impact of matching on treatment effects in the context of difference-in-differences (Daw and Hatfield 2018). They evaluate differences in bias across estimators that are unmatched, matched on covariates that are related to the outcome, matched on preintervention levels of outcomes, and matched on preintervention trends in outcomes. They also explore bias under different scenarios related to the correlation between treatment assignment and levels of preintervention outcomes, trends in preintervention outcomes, and levels of covariates.

Daw and Hatfield find that matching tends to increase bias, rather than decrease it. The bias introduced by matching is particularly severe when matching is based on preintervention levels and a key assumption of difference-in-differences is met (no correlation between treatment and outcome trend). In these cases, bias increases with the difference in preintervention levels. This bias is caused by mean reversion. Because the treatment and comparison groups are drawn from different distributions, any observed overlap in outcomes between the treatment and comparison groups is a result of noise. Matching on this noise will not purge bias but rather will lead to mean reversion as the control group returns to its natural mean in the postintervention period.

This is an extremely important, timely, and practical investigation. It highlights a specification issue in difference-in-differences that has received limited attention in the literature (Chay and McEwan 2005). By identifying

---

Address correspondence to Andrew M. Ryan, Ph.D., Department of Health Management and Policy, University of Michigan School of Public Health, 1415 Washington Heights, SPH II, Room M3124, Ann Arbor, MI 48109; e-mail: amryan@umich.edu.

the circumstances under which matching is most likely to lead to bias, it provides a practical guide to researchers.

Yet Daw and Hatfield identify a researcher error that is relatively narrow: a case where researchers try to solve a problem that does not exist. In their study, the bias introduced by matching occurs when assumptions of difference-in-differences hold and researchers could derive an unbiased estimate using standard methods. By matching treatment and control groups when there is no need to do so, a research "self-own" is committed.

In practice, researchers often pursue matching when they have reason to believe that assumptions in difference-in-differences do not hold. Specifically, when preintervention trends are not parallel between treatment and comparison groups, researchers worry that treatment assignment will be correlated with future outcome trends. It is under these circumstances researchers may attempt to use a subset of the comparison group (through matching or qualitative selection of units) that could be a more appropriate counterfactual. When Daw and Hatfield evaluate the scenario where treatment assignment is correlated with preintervention trends (i.e., when a key assumption of difference-in-differences is violated), they find that, while there is considerable bias for all the estimators, matching on preintervention trends reduces bias (particularly when the outcome is highly serially correlated).

Relatedly, in their simulation, Daw and Hatfield consider only the case where the levels and trends for the treatment group are drawn from one distribution, while the levels and trends for the comparison group are drawn from a different distribution. In other words, there is no overlap in true distributions between the treatment and comparison groups. There is only overlap between the observed distributions. In practice, there is likely to be overlap in true outcome distributions between treatment and comparison groups (e.g., a high-performing academic medical center in treated state compared to a high-performing academic medical center in neighboring comparison state). Greater overlap between the true distributions of levels and trends would likely decrease the problem associated with mean reversion. In such circumstances, matching has the potential to be beneficial if covariates, preintervention levels, and preintervention trends are correlated with future outcomes.

The performance of matching estimators in the context of difference-in-differences has been evaluated in other simulation work. Our research team performed a simulation analysis that started with real data on clinical process performance from acute care hospitals in the United States. We assigned an intervention to hospitals under different scenarios concerning the correlation between treatment, preintervention levels of outcomes, and preintervention

trends in outcomes. When treatment was correlated with preintervention trends, our matching estimator—which matched on lagged levels for each of the three preintervention periods—had substantially lower bias than the other estimators examined, including standard difference-in-differences (Ryan, Burgess, and Dimick 2015). In another study, O'Neil and colleagues specified a simulation model to evaluate the performance of several estimators (including the synthetic control method, a lagged dependent variable regression approach, and matching on lagged outcomes) in the presence of parallel and nonparallel trends in outcomes (O'Neill et al. 2016). They found that, when trends in outcomes were parallel, the standard difference-in-differences estimator was the most accurate. However, the matching estimator was reasonably accurate, particularly when a large number of preintervention periods were available for matching. The authors also found that, when outcomes trends were not parallel, the standard difference-in-differences estimator had the greatest bias. Together, this work suggests that matching estimators can outperform standard difference-in-differences under certain circumstances.

It is challenging to compare results across simulations. Yet one reason why the matching estimator may have performed better in the studies by our research team and O'Neil and colleagues is because these simulation models were not constructed to include no overlap in true levels and trends between the treatment and comparison groups.

Difference-in-differences is an essential tool to understand our changing health care system. Daw and Hatfield highlight a case when statistical matching can undermine difference-in-differences. Future research should continue to develop, implement, and examine the properties of other new tools, such as generalized synthetic control methods (Xu 2017), that can generate unbiased estimates while relaxing the standard assumptions of difference-in-differences analysis.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The author is supported by the University of Michigan School of Public Health.

Disclosure: None.

Disclaimer: None.

Andrew M. Ryan

REFERENCES

Angrist, J. D., and J.-S. Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Chay, K. Y., and P. J. McEwan. 2005. "The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools." *American Economic Review* 95 (4): 1237–58.

Courtemanche, C., J. Marton, B. Ukert, A. Yelowitz, and D. Zapata. 2017. "Early Impacts of the Affordable Care Act on Health Insurance Coverage in Medicaid Expansion and Non-Expansion States." *Journal of Policy Analysis and Management* 36 (1): 178–210.

Daw, J. R., and L. A. Hatfield. 2018. "Matching and Regression to the Mean in Difference-in-Differences Analysis." *Health Services Research* 53 (6 Pt 1): 4138–56. https://doi.org/10.1111/1475-6773.12993

Figueroa, J. F., Y. Tsugawa, J. Zheng, E. J. Orav, and A. K. Jha. 2016. "Association between the Value-Based Purchasing Pay for Performance Program and Patient Mortality in US Hospitals: Observational Study." *British Medical Journal* 353: i2214.

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

Post, B., T. Buchmueller, and A. M. Ryan. 2018. "Vertical Integration of Hospitals and Physicians: Economic Theory and Empirical Evidence on Spending and Quality." *Medical Care Research and Review* 75 (4): 399–433. https://doi.org/10.1177/1077558717727834

Robinson, J. C., and T. T. Brown. 2013. "Increases in Consumer Cost Sharing Redirect Patient Volumes and Reduce Hospital Prices for Orthopedic Surgery." *Health Affairs* 32 (8): 1392–7.

Ryan, A. M., J. F. Burgess, and J. B. Dimick. 2015. "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences." *Health Services Research Methods* 50 (4): 1211–35.

Ryan, A. M., S. Krinsky, K. A. Maurer, and J. B. Dimick. 2017. "Changes in Hospital Quality Associated with Hospital Value-Based Purchasing." *New England Journal of Medicine* 376 (24): 2358–66.

Sommers, B., B. Katherine, and A. Epstein. 2012. "Mortality and Access to Care Among Adults after State Medicaid Expansions." *New England Journal of Medicine* 367 (11): 1025–34.

Werner, R. M., J. T. Kolstad, E. A. Stuart, and D. Polsky. 2011. "The Effect of Pay-for-Performance in Hospitals: Lessons for Quality Improvement." *Health Affairs (Millwood)* 30 (4): 690–8.

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

Zuckerman, R. B., S. H. Sheingold, E. J. Orav, J. Ruhter, and A. M. Epstein. 2016. "Readmissions, Observation, and the Hospital Readmissions Reduction Program." *New England Journal of Medicine* 374 (16): 1543–51.