Well-balanced or too matchy-matchy? The controversy over matching in difference-in-differences analysis

Andrew M. Ryan

This article is protected by copyright. All rights reserved
Acknowledgements: None

Disclosures: The author has no conflicts of interest to disclose.

Disclaimers: None

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
These remarkable changes demand evaluation. Difference-in-differences methods have 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 their widespread adoption, major issues related to the specification of these 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 a key condition 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 pre-intervention levels in outcomes (O’Neill et al. 2016; Ryan et al. 2017), trends in outcomes, or covariates thought to be relevant to outcomes (Figueroa et al. 2016; Werner et al. 2011). 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 the current 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 analysis (Daw and Hatfield, in press). They evaluate differences in bias across estimators that are unmatched, matched on pre-intervention levels of outcomes, and matched on pre-intervention trends in outcomes. They also
explore bias under different scenarios related to the correlation between treatment assignment and levels of pre-intervention outcomes, trends in pre-intervention 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 pre-intervention levels and a key assumption of difference-in-difference analysis is met (no correlation between treatment and outcome trend). In these cases, bias increases with the difference in pre-intervention 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 post-intervention 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 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 doesn’t exist. In their study, the bias introduced by matching occurs when assumptions of difference-in-differences hold and researchers could derive an unbiased estimate by 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 don’t hold. Specifically, when pre-intervention 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 pre-intervention 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 pre-intervention 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, pre-intervention levels, and pre-intervention trends are correlated with future outcomes.

This 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, pre-intervention levels of outcomes, and pre-intervention trends in outcomes. When treatment was correlated with pre-intervention trends our matching estimator – which matched on lagged levels for each of the three pre-intervention period – 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 non-parallel trends in outcomes (O’Neill et al. 2016). They found that, when trends in outcomes were parallel, the standard difference-in-difference estimator was the most accurate. However, the matching estimator was reasonably accurate, particularly when a large number of pre-intervention periods were available for matching. The authors also found that, when outcomes trends were not parallel, the standard difference-in-difference 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 analysis 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 generated unbiased estimates while relaxing the standard assumptions of difference-in-differences analysis.

References


This article is protected by copyright. All rights reserved


