Matching and Regression to the Mean in Difference-in-Differences Analysis

Jamie R. Daw and Laura A. Hatfield

Objective. To demonstrate regression to the mean bias introduced by matching on preperiod variables in difference-in-differences studies.

Data Sources. Simulated data.

Study Design. We performed a Monte Carlo simulation to estimate the effect of a placebo intervention on simulated longitudinal data for units in treatment and control groups using unmatched and matched difference-in-differences analyses. We varied the preperiod level and trend differences between the treatment and control groups, and the serial correlation of the matching variables. We assessed estimator bias as the mean absolute deviation of estimated program effects from the true value of zero.

Principal Findings. When preperiod outcome level is correlated with treatment assignment, an unmatched analysis is unbiased, but matching units on preperiod outcome levels produces biased estimates. The bias increases with greater preperiod level differences and weaker serial correlation in the outcome. This problem extends to matching on preperiod level of a time-varying covariate. When treatment assignment is correlated with preperiod trend only, the unmatched analysis is biased, and matching units on preperiod level or trend does not introduce additional bias.

Conclusions. Researchers should be aware of the threat of regression to the mean when constructing matched samples for difference-in-differences. We provide guidance on when to incorporate matching in this study design.

Key Words. Observational research, matching, difference-in-differences

Difference-in-differences is one of the most popular approaches for measuring the effect of health policies and programs (Ryan, Burgess, and Dimick 2014). A difference-in-differences analysis can be employed any time one measures an outcome in units that have been exposed to a policy or program (the treatment group) and those that have not (the control group) both before and after the intervention is implemented. Here, “group” means the level at which treatments are applied (e.g., states, clinics) and “unit” means the level at which measurements are made (e.g., counties within states, patients within clinics). The effect of the treatment is calculated as the difference in the outcome between

the treatment and control groups after the intervention minus the difference before the intervention. One concludes that the intervention affected the outcome if the difference between the two groups changes from the preperiod to the postperiod.

As the popularity of difference-in-differences has risen, so has the application of matching methods to this study design. The objective of matching is to reduce potential confounding by improving the comparability of units in the treatment and control groups. In the context of difference-in-differences, researchers identify a subset of potential confounders and match units from the treatment and control group on measures of these variables prior to the intervention. The effect of the intervention is then estimated using this matched sample.

Difference-in-differences studies with matched samples have been used to evaluate the impact of a variety of health policies and programs, including high-deductible health plans (Wharam et al. 2007; Waters et al. 2011), team-based and coordinated care programs (Scanlon et al. 2008; Xing, Goehring, and Mancuso 2015; Dale et al. 2016), multipayer medical homes (Rosenthal et al. 2015), telehealth programs (Baker et al. 2011), home-visiting programs (Mattke et al. 2015), hospital closures (Joynt et al. 2015), workplace wellness programs (Caloyeras et al. 2014), and quality reporting (Hollenbeak et al. 2008; Osborne et al. 2015). In this study, we demonstrate how matching in difference-in-differences can introduce a well-known statistical phenomenon, regression to the mean, resulting in biased estimates of intervention effects. We also offer practical guidance to researchers on how to select matching variables to minimize this important threat to validity.

Basics of Difference-in-Differences

We focus on the widely used “microlevel” difference-in-differences model in which the treatment is assigned to a group (e.g., health plans adopt benefit design changes) and observations are available for units within groups (e.g., health spending of enrollees within health plans) before and after an intervention (Ryan, Burgess, and Dimick 2014).

Address correspondence to Jamie R. Daw, Ph.D., Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115; e-mail: jdaw@fas.harvard.edu. Laura A. Hatfield, Ph.D., is with the Department of Health Care Policy, Harvard Medical School, Boston, MA.

A difference-in-differences study is usually analyzed with a regression model such as

$$Y_{ijt} = \beta_0 + \beta_1 \text{treatment}_j + \beta_2 \text{post}_t + \delta(\text{treatment}_j * \text{post}_t) + \varepsilon_{ijt}, \quad (1)$$

where i indexes units, j indexes groups, t indexes time, treatment is an indicator for whether a group was treated, post is an indicator for whether a measurement was taken in the posttreatment period, and ε is the random error term. We consider the simplest case, in which there are only two groups: one treatment group and one control group. The difference-in-difference estimate of the average treatment effect on the treated is $\hat{\delta} = (\bar{Y}_{\text{treatment,post}} - \bar{Y}_{\text{treatment,pre}}) - (\bar{Y}_{\text{control,post}} - \bar{Y}_{\text{control,pre}})$.

Confounding in Difference-in-Differences

One advantage of difference-in-differences, relative to cross-sectional designs, is that it does not require treatment and control groups to have similar baseline means, often referred to as preperiod “levels,” in the outcome or other covariates. This is because the design measures the effect of an intervention as the *relative change* in the outcomes between units in the treatment and control groups over time. As a result, the definition of confounding differs from cross-sectional studies. A confounder of a difference-in-differences study is any variable related to both treatment assignment and the *change in the outcome over time* (i.e., the trend). Contrast this to a confounder in a cross-sectional study, which is any variable related to both treatment assignment and the *level of the outcome at a point in time*. Of course, a variable may be related to both the level and trend in the outcome and therefore a confounder in both senses. The point is that variables related only to treatment assignment and outcome level (not trend) do not bias difference-in-differences studies. They are not confounders and therefore not a useful target of matching that is intended to reduce bias due to observable confounders.

Like other observational designs, difference-in-differences still requires a strong assumption to produce unbiased causal estimates of the treatment effect: The change from pre- to postperiod in the control group is a valid counterfactual for the change that would have occurred in the treatment group in the absence of the intervention. This is often stated as two assumptions, “common shocks” and “parallel trends,” referring, respectively, to the assumption that events during the study period affect the treatment and control groups equally and the assumption that the two groups would have equal trends in the postperiod if not for the intervention (Angrist and Pischke 2009). This is

also equivalent to assuming no unobserved confounding, where we emphasize that confounding in difference-in-differences relates only to variables correlated with treatment assignment and outcome trends. Direct assessment of these assumptions, which we refer to collectively as the “counterfactual assumption,” would require observing an alternative reality (i.e., the counterfactual change in the treatment group outcomes in the absence of intervention) and is therefore impossible.

Matching in Difference-in-Differences

Even though differences between treatment and control units at baseline are not a threat to validity per se, researchers often match units from the treatment and control groups on preperiod measures of the outcome or other variables. Doing so attempts to correct for confounding bias by balancing on variables that are different in the treatment and control group. Matching typically uses preperiod measurements of three kinds of variables: covariates, outcome levels, and outcome trends.

In this study, we do not consider the possibly beneficial application of matching on covariates that differ between the groups and are correlated with future outcome *trends* (i.e., matching on confounders in the difference-in-difference sense). Instead, we consider matching on covariates that are correlated with *levels* of the outcome (or are the outcome level itself), which is the source of the regression to the mean bias we discuss here.

We also consider matching on preperiod outcome trends. Difference-in-differences does not strictly require the treatment and control groups to have similar trends in the outcome prior to the intervention. However, divergence in preperiod trends is usually seen as a strong indication that the counterfactual assumption is violated. Thus, researchers may match on preperiod outcome trends when they suspect that preperiod trends are correlated with the change in the outcome from the preperiod to postperiod. The hope is that balance on preperiod trends strengthens the plausibility of the counterfactual assumption. In this study, we show that matching on preperiod trends may have limited benefit.

Regression to the Mean and Bias in Matched Difference-in-Differences Analysis

The statistical phenomenon of regression to the mean (aka the “regression fallacy”) occurs when a group is selected for extreme values of one variable and then another variable is measured for that group. In the longitudinal setting

relevant to difference-in-difference studies, it occurs between repeated measures of the same variable; for example, if we measure weight for a sample of individuals today and select a subset of individuals with higher-than-average weight, those individuals will have a mean weight that is closer to average upon subsequent measurement.

Two factors determine the magnitude of regression to the mean effects (Barnett, van der Pols, and Dobson 2005). First, regression effects increase as units are selected further away from their group mean. Second, regression effects decrease with increased correlation between the sample selection variable and the other measured variable. In longitudinal settings, this means that regression effects decrease with increased correlation between measures over time (i.e., the serial correlation). Selecting units that have extreme values of a variable that is unstable over time will produce large regression to the mean effects. Variables that do not vary over time, such as sex or region, are not subject to longitudinal regression to the mean effects.

The vulnerability of matched samples to regression to the mean has been known since the work of McNemar in the 1940s, yet it receives little contemporary attention in guidance on matching methods or in discussions of the validity of published evaluations using matched samples (McNemar 1940; Althausen and Rubin 1971). The simple idea is that matching is a sample selection technique; it selects units that are extreme relative to their respective group means to achieve balance in the matched sample. For example, when the treatment group mean is larger than the control group mean, matching will select control units that are higher than average (relative to all control units) and treatment units that are lower than average (relative to all treatment units). If the variables on which units are matched vary over time, matched units will “regress back” toward the means of the groups from which they were selected. More precisely, this phenomenon might be called “regression to the means” as matched units from treatment and control groups regress back to their respective means over time. For simplicity, we use the idiomatic expression “regression to the mean.”

Longitudinal studies such as difference-in-differences may reach biased conclusions in the presence of regression to the mean effects. Recall that a confounder of a difference-in-differences study is any variable related to both treatment assignment and the change in the outcome over time (i.e., the trend). The process of matching *introduces* such a confounder—the indicator of whether a unit is selected into the sample by matching—that is (inadvertently) correlated with the change in the outcome over time because of regression to the mean. As an illustration, consider a voluntary program that increases cost-

sharing for enrollees in employer-sponsored health plans. Suppose that a firm that adopts the program does so in response to higher mean enrollee health spending compared to a firm that does not adopt the program. If we match enrollees on baseline health spending, we will select control enrollees with higher-than-average spending (relative to the control firm mean) and treatment enrollees with lower-than-average spending (relative to the treatment firm mean) to achieve a balanced sample. Even if the program has no true effect, on subsequent measurement, average spending will decrease among the matched control enrollees and increase among matched treatment enrollees purely because of regression toward their respective group means. The divergent trends between the two matched groups violate the critical counterfactual assumption of difference-in-differences. This could lead to the false conclusion that the program increased health spending. This threat to validity is a particular concern because it results from a researcher applying matching to what otherwise would be an unbiased analysis.

The magnitude of the bias introduced in a matched difference-in-differences analysis will vary with the magnitude of the regression to the mean, which depends on (1) how extreme the measures of the matching variable are in the matched sample relative to those of the unmatched sample (e.g., the magnitude of the preperiod differences between the treatment and control groups), and (2) the correlation between the matching variable and the postperiod outcome (e.g., the serial correlation between measures of the outcome).

METHODS

We use a Monte Carlo simulation to demonstrate the bias resulting from regression to the mean in matched difference-in-differences analysis. We generate data under different causal scenarios, create both matched and unmatched samples, and analyze them with standard difference-in-differences regression models to estimate the effect of a null intervention, for which the true treatment effect is zero.

Data Generation

We generate data under four causal scenarios, detailed in Appendix SA2 and summarized here. Three of the scenarios are unconfounded (for difference-in-differences) because treatment assignment is (1) completely random (i.e.,

unrelated to group mean levels or trends), (2) correlated only with group mean outcome level, or (3) correlated with group mean level of a time-varying covariate that is in turn correlated only with the group mean outcome level. The fourth scenario is confounded because treatment assignment is correlated with group mean outcome trend, thus violating the counterfactual assumption. In each scenario, there are no additional unobserved variables that determine treatment assignment.

For each scenario, we generate data for 1,000 units in the treatment group and 1,000 units in the control group. For each unit, we generate eight observations of a normally distributed outcome, centered around the respective group mean levels and trends. Because we generate under a null intervention, the group mean levels and trends of the outcome are constant over the pre- and postperiods. The unit measurements are assumed to be equally spaced and could be conceptualized, for example, as eight quarterly measurements over 2 years. The repeated measures follow an autoregressive covariance structure of order 1 [AR(1)] with constant variance and a single correlation parameter (see Appendix Figure A1). The AR(1) structure implies that a unit's measurements from one time period to the next are positively correlated and that this correlation decays exponentially for measurements that are further apart in time.

For the third causal scenario, we also generate eight observations of a normally distributed covariate for each unit, centered around the respective means of the treatment and control groups. These repeated covariate measures also follow an AR(1) covariance structure and have an additional correlation parameter that controls the strength of the relationship with the unit outcome. Note that this covariate is not a confounder because it is only correlated with outcome level, not with changes in the outcome over time.

Within each causal scenario, we also vary factors that affect the magnitude of regression to the mean effects: serial correlation across observations of both the outcome and the covariate, strength of correlation between the outcome and covariate, and preperiod level and trend difference between the treatment and control groups. Table 1 summarizes the causal data-generating scenarios.

Analysis

Of the eight observations for each unit, we assume four occur in the preperiod and four in the postperiod. The unmatched samples are simply samples of all the simulated units. We generate matched samples by matching on (1) the

Table 1: Summary of Simulation Scenarios

Group-Level Treatment Assignment	Mean Preperiod Difference Between Treatment and Control Groups				Results Figure
	Outcome Level	Outcome Trend	Covariate Level	Confounded?	
Random			n/a	No	1a 1b
Correlated with outcome level	No difference	No difference	n/a	No	1c
	1a 1b
Correlated with covariate level	1 or 2 SD	No difference		No	1c 2a

Correlated with outcome trend	Depends on X,Y correlation (weak, moderate, or strong)	No difference	1 SD	Yes	2b
	n/a		3a 3b
		0.01, 0.05, or 0.1 SD			3c

Notes: The serial correlation of the matching variable varies across simulation iterations within each scenario. SD refers to standard deviation of the covariate or outcome.

preperiod level of the outcome, (2) the preperiod level of a covariate correlated with the outcome (for causal scenario 3), or (3) the preperiod trend in the outcome, calculated using linear regression estimates of the preperiod slope for each unit. We use one-to-one, nearest-neighbor matching (without replacement) with a caliper of 0.2 SD of the matching variable.

For each unmatched and matched sample, we estimate the effect of the intervention using the “microlevel” difference-in-differences regression estimator (equation 1) with standard errors clustered by unit. We summarize bias using the mean absolute deviation between the estimated effect and the true value of zero. We average the absolute deviation over the 1,000 simulation replicates and scale the result in terms of standard deviations of the outcome variable. Table 1 indicates the figures that display the corresponding results.

Limitations

As our objective is to demonstrate a well-known statistical phenomenon (regression to the mean) in the setting of a popular study design (difference-in-differences with matched samples), we apply relatively simple scenarios, which do not reflect all the potential causal scenarios in which difference-in-differences analysis could be applied. However, the underlying mechanism of treatment assignment in real data is rarely known, and thus, these simple cases are often assumed to hold. The magnitude of the bias estimated in our study is specific to data-generating processes that resemble our simulation scenarios. For instance, we simulate only under null intervention effects because we are interested in bias. In addition, while the AR(1) correlation structure is reasonable, other correlation structures are possible and this may have implications for the magnitude of the results. Similarly, we focus exclusively on normally distributed variables and OLS regression, which is the standard modeling approach used for difference-in-differences analysis. While regression to the mean effects do not depend on the specific distribution of the outcomes and covariates, the magnitude of the results may vary for non-normal variables and other treatment estimators.

RESULTS

Treatment Randomly Assigned or Correlated with Preperiod Level Only

When treatment is correlated with preperiod levels of the outcome, matching units on baseline outcome measures can introduce bias in an otherwise

Figure 1: Bias of Matching Strategies for Group-Level Treatment *Randomly Assigned* or Correlated with Preperiod *Level Only* [Color figure can be viewed at wileyonlinelibrary.com]

Notes. The serial correlation of the outcome refers to the autoregressive parameter of the AR(1) correlation structure. Bias is measured as the absolute deviation of the treatment estimate from zero in standard deviations of the outcome. SD is standard deviations of the outcome. When mean preperiod level difference is 0 SD, group-level treatment is randomly assigned. When mean preperiod level difference is 1 or 2 SD, group-level treatment is correlated with preperiod level.

unbiased analysis. Figure 1 shows the results for scenarios where treatment is randomly assigned or correlated with preperiod outcome level only. The figure displays the estimator bias for three matching strategies across different strengths of serial correlation in the outcome and different baseline level differences between the treatment and control groups (ranging from zero to two standard deviations). The estimates from the unmatched samples are unbiased across all strengths of serial correlation and for all mean preperiod level differences (see Figure 1a). This is what we expect: Preperiod differences in level do not bias difference-in-differences analyses. However, when we apply difference-in-differences analysis to samples matched on baseline level, bias is introduced. As shown in Figure 1B, the form of the bias is a classic manifestation of regression to the mean: Regression effects increase with decreasing serial correlation in the outcome and increasing baseline level differences between the treatment and control groups. In the most extreme case, where the serial correlation in the outcome is zero, the two groups regress entirely back to their original baseline differences, resulting in large, spurious treatment effects. As shown in Figure 1C, when treatment is not correlated with trends in the outcome, matching on preperiod trend does not introduce bias. As in the case where

treatment is assigned randomly, this is because matching on trend when there is no difference in group-level trend does not result in a selection of units that are extreme relative to their group means.

Matching to reduce baseline differences in a time-varying covariate that is correlated with the outcome can produce the same bias problems. Figure 2b shows the bias of matching on the preperiod level of a covariate that is correlated with the outcome. The bias behaves similar to matching on the outcome itself, increasing with decreasing serial correlation in the covariate, as well as with increasing preperiod difference between the two groups. However, compared to Figure 1b, the magnitude of the bias is proportional to the correlation between the covariate and the outcome. In other words, the stronger the relationship between the covariate and the outcome, and the greater the mean difference between the two groups at baseline, the greater the bias. We find that matching on preperiod covariates does not introduce bias when (1) the covariates are not correlated with both outcome and treatment, and (2) the covariates are fixed or very highly correlated over time. As before, the unmatched analysis is unbiased, see Figure 2a.

Figure 2: Bias of Matching Strategies for Group-Level Treatment Correlated with Preperiod *Level Only* [Color figure can be viewed at wileyonlinelibrary.com]

Notes. The serial correlation of the outcome and covariate refers to the autoregressive parameter of the AR(1) correlation structure. The mean preperiod level difference in the covariate between the treatment and control groups is 1 standard deviation across all iterations. Bias is measured as the absolute deviation of the treatment estimate from zero in standard deviations of the outcome. SD is standard deviations of the outcome.

Figure 3: Bias of Matching Strategies for Group-Level Treatment Correlated with Preperiod *Trend Only* [Color figure can be viewed at wileyonlinelibrary.com]

Notes. The serial correlation of the outcome refers to the autoregressive parameter of the AR(1) correlation structure. Bias is measured as the absolute deviation of the treatment estimate from zero in standard deviations of the outcome. SD is standard deviations of the outcome.

Treatment Correlated with Preperiod Trends

When the counterfactual assumption is violated, we find that matching is insufficient to overcome the bias. Figure 3 shows the results for unmatched and matched samples for scenarios where treatment assignment is correlated with preperiod trend only. In this scenario, the causal assumption is violated because the preperiod differences in trend persist in the postperiod, making the control group's change over time an invalid counterfactual for the treatment group's change. As expected, this violation results in biased estimates when using the unmatched samples. As shown in Figure 3a, the bias is proportional to the differences in trends between the treatment and control groups. Matching on preperiod level does not introduce additional bias (see Figure 3b) because the samples were generated with no correlation between treatment and preperiod outcome trend, so matching does not select extreme units and there is no regression to the mean. However, as shown in Figure 3c, matching on trend neither introduces additional bias nor fully corrects for the violation. We find a small decrease in bias in the presence of very high serial correlation in the outcome. When serial correlation is low, preperiod trend is a poor predictor of the pre- and postperiod difference; however, when serial correlation is high, preperiod trends are more stable and

more predictive of future trends, resulting in a modest reduction in bias in samples matched on trend (see Appendix Figure A2).

DISCUSSION

Longitudinal study designs using matched samples have long been known to be vulnerable to the problem of regression to the mean. Our results emphasize that health services researchers should be aware of this threat when constructing matched samples for difference-in-differences analysis. When treatment group assignment is correlated with outcome level in the preperiod, we find that matching units on preperiod outcome level can introduce bias to an otherwise unbiased analysis. The magnitude of the bias increases with (1) greater preperiod level differences between the treatment and control groups and (2) lower serial correlation of the outcome. This problem extends to samples matched on preperiod level of a time-varying covariate when the covariate is correlated both with treatment and the outcome. The bias is minimal when preperiod level differences are small or serial correlation is very high, and is absent when there are no preperiod level differences or the matching variable is constant.

These results imply a challenging paradox. The greater the preperiod level differences between the treatment and control groups in the outcome or a covariate that is correlated with the outcome, the more inclined researchers may be to match on preperiod level. However, the greater the preperiod level difference, the larger the bias that can result from matching due to regression to the mean. Thus, researchers should not attempt to “match-away” level differences in time-varying variables, as doing so can introduce bias. Indeed, this practice is particularly unnecessary given that preperiod level differences per se are not a violation of the causal assumption of difference-in-differences.

Researchers often want to exercise prudence and include all potential confounders in their matching model. As a result, matching on baseline levels, and increasingly baseline trends, has become a mainstream practice. Common preperiod matching variables include time-varying individual-level variables, such as number of primary care visits, and time-varying practice- and regional-level variables such as rates of emergency department visits or risk-adjusted mortality. In many studies, matching models include preperiod measures of the study outcome itself, for example, matching on baseline levels of health care spending in a study of the effect of a policy on spending. To give

context to the strengths of serial correlation shown in our results, Appendix Table A3 shows the year-to-year correlation for a selection of variables measured on people and hospital service areas.

Our results caution against the popular “kitchen sink” approach to including preperiod variables in matching models. We would argue that the inclusion of matching variables ought to be considered carefully given the potential for regression-to-the-mean bias. Our results also show when matching units on baseline measures does not introduce a risk of regression to the mean: when there is good preperiod balance between the two groups, strong or perfect serial correlation in the matching variable, or when the association between the matching variable and the outcome is weak.

Guidance for Selecting Matching Variables in Difference-in-Differences Analysis

Based on previous research and the results of our simulations, we present a flowchart for selecting matching variables for difference-in-differences analysis (Figure 4). We assume a process whereby researchers consider matching approaches in response to inspections for preperiod differences such as those

Figure 4: Selecting Matching Variables for Difference-in-Differences Analysis [Color figure can be viewed at wileyonlinelibrary.com]

that have been recommended in other general checklists for this design (e.g., those proposed by Ryan, Burgess, and Dimick 2014).

Preperiod Differences in Outcome Level

While a baseline difference in outcome level between the treatment and control group is not itself a threat to validity of a difference-in-differences analysis, in practice, it may raise concerns that the control group is not a valid counterfactual. In response to finding a preperiod difference in the outcome level, researchers should consider whether preperiod level is likely to be correlated with the pre–post change, possibly with empirical tests using the control group data. If there is no or a weak relationship between preperiod level and the pre–post change (and thus no threat to validity), researchers should proceed without matching on preperiod level. As we demonstrate, in this scenario, matching units on preperiod level will unnecessarily introduce bias when the preperiod level difference is large and the serial correlation is low.

Preperiod Differences in Covariate Level

Again, while preperiod covariate differences between treatment and control groups are not a threat to validity for difference-in-differences, they may undermine confidence in the control group as a valid counterfactual. As for differences in preperiod levels of the outcome, this concern is justified if preperiod levels of the covariate are predictors of the pre–post change in the outcome. In the face of preperiod covariate differences, researchers should draw on scientific understanding and empirical evidence for the relationship between preperiod covariate levels and pre–post outcome changes. A covariate that differs between the two groups at baseline and is correlated with pre–post outcome changes is an appropriate matching variable if it is stable over time (e.g., sex, race, or region). Empirical estimates of the control group's repeated measures correlation matrix may be informative. If the covariate is fixed or highly correlated over time, researchers may proceed with matching on preperiod levels of the covariate without risk of regression to the mean. If it is weakly or moderately correlated over time, researchers should avoid matching on the variable or consider stabilizing transformations. For example, one could transform a continuous variable with moderate serial correlation (such as a hospital quality score) into a categorical variable (such as low, medium, and high).

A covariate that differs between the two groups at baseline but is not associated with the outcome is an instrument and should not be used as a

matching variable. Doing so can increase bias and decrease the precision of treatment estimates (Brookhart et al. 2006; Bhattacharya and Vogt 2007).

Finally, baseline covariates that are not substantially different between treatment and control but are correlated with the outcome may be good candidates for matching, as their inclusion may yield more precise treatment estimates (Stuart 2010). As we show, matching on baseline variables that do not differ between treatment and control does not introduce the threat of regression to the mean.

Preperiod Differences in Outcome Trend

There is no empirical test of the validity of the counterfactual assumption of a difference-in-differences analysis. However, researchers try to bolster confidence that the assumption holds by looking for evidence of differences in preperiod trends, which may indicate that the control group is not a valid counterfactual. However, similar preperiod trends are no guarantee that the similarities would persist in the postperiod in the absence of the intervention. More important, in our data-generating scenarios, matching on preperiod trend does not address violations of the counterfactual assumption, although matching may provide small bias reductions when the trends are highly stable. Matching on trend does not introduce additional bias (relative to unmatched samples) because the regression to the mean effect simply pulls the postperiod trends back to the original violation. However, matching on trend can lead to misleading assessments of the validity of the counterfactual assumption because plots of preperiod trends in matched samples will appear parallel (as matching forces them to be so); yet unless the trends are highly stable, this comparability will break down in the postperiod due to regression to the mean trend in each group (see Appendix Figure A3).

It is possible that other matching techniques could address violations of the counterfactual assumption. Methodologists and applied researchers are experimenting with new approaches such as cross-temporal matching algorithms (Stuart et al. 2014; Gozalo et al. 2015). The synthetic control method is yet another alternative for constructing a comparable control group based on preperiod measures (Abadie, Diamond, and Hainmueller 2010). There is a need for further simulation studies to compare the performance of techniques researchers are applying in the field—including the potential for regression to the mean bias—under varying data-generating and treatment assignment scenarios.

CONCLUSION

Matching is an increasingly popular approach for improving the comparability of treatment and control groups in difference-in-differences analysis. Limited attention has been given to how matching, as a sample selection technique, can introduce bias due to regression to the mean. Our results show that matching can have an important impact on estimated intervention effects, particularly when matching on preperiod levels of the outcome itself or on time-varying covariates with low serial correlation. We provide guidance on when to incorporate matching in difference-in-differences based on observed evidence of potential violations of the assumptions of the design.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This article was benefited from valuable feedback provided by Katherine Swartz, Benjamin Sommers, Tom Kelley, and Sherri Rose. The authors would also like to acknowledge the constructive comments provided by the editors and anonymous reviewers. Research reported in this publication was supported by the National Institute of Aging of the National Institutes of Health under award number P01AG032952. Jamie Daw was supported by a Frank Knox Memorial Fellowship and a Canadian Institutes for Health Research Doctoral Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors have no additional financial disclosures or conflict of interests.

Disclosure: None.

Disclaimer: None.

REFERENCES

- Abadie, A., A. Diamond, and J. Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.
- Althausen, R. P., and D. Rubin. 1971. "Measurement Error and Regression to the Mean in Matched Samples." *Social Forces* 50 (2): 206–14.
- Angrist, J., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

- Baker, L. C., S. J. Johnson, D. Macaulay, and H. Birnbaum. 2011. "Integrated Telehealth and Care Management Program for Medicare Beneficiaries with Chronic Disease Linked To Savings." *Health Affairs* 30 (9): 1689–97.
- Barnett, A. G., J. C. van der Pols, and A. J. Dobson. 2005. "Regression to the Mean: What It Is and How to Deal with It." *International Journal of Epidemiology* 34 (1): 215–20.
- Bhattacharya, J., and W. B. Vogt. 2007. "Do Instrumental Variables Belong in Propensity Scores?," SSRN Scholarly Paper ID 1014784, Rochester, NY: Social Science Research Network [accessed on June 20, 2017]. Available at <https://papers.ssrn.com/abstract=1014784>.
- Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163 (12): 1149–56.
- Caloyeras, J. P., H. Liu, E. Exum, M. Broderick, and S. Mattke. 2014. "Managing Manifest Diseases, But Not Health Risks, Saved PepsiCo Money over Seven Years." *Health Affairs* 33 (1): 124–31.
- Dale, S. B., A. Ghosh, D. N. Peikes, T. J. Day, F. B. Yoon, E. F. Taylor, K. Swankoski, A. S. O'Malley, P. H. Conway, R. Rajkumar, M. J. Press, L. Sessums, and R. Brown. 2016. "Two-Year Costs and Quality in the Comprehensive Primary Care Initiative." *New England Journal of Medicine* 374 (24): 2345–56.
- Gozalo, P., M. Plotzke, V. Mor, S. C. Miller, and J. M. Teno. 2015. "Changes in Medicare Costs with the Growth of Hospice Care in Nursing Homes." *New England Journal of Medicine* 372 (19): 1823–31.
- Hollenbeak, C. S., C. P. Gorton, Y. P. Tabak, J. L. Jones, A. Milstein, and R. S. Johannes. 2008. "Reductions in Mortality Associated with Intensive Public Reporting of Hospital Outcomes." *American Journal of Medical Quality* 23 (4): 279–86.
- Joynt, K. E., P. Chatterjee, E. J. Orav, and A. K. Jha. 2015. "Hospital Closures Had No Measurable Impact on Local Hospitalization Rates or Mortality Rates, 2003–11." *Health Affairs* 34 (5): 765–72.
- Mattke, S., D. Han, A. Wilks, and E. Sloss. 2015. "Medicare Home Visit Program Associated with Fewer Hospital and Nursing Home Admissions, Increased Office Visits." *Health Affairs* 34 (12): 2138–46.
- McNemar, Q. 1940. "Sampling in Psychological Research." *Psychological Bulletin* 37: 331–65.
- Osborne, N. H., L. H. Nicholas, A. M. Ryan, J. R. Thumma, and J. B. Dimick. 2015. "Association of Hospital Participation in a Quality Reporting Program with Surgical Outcomes and Expenditures for Medicare Beneficiaries." *Journal of the American Medical Association* 313 (5): 496–504.
- Rosenthal, M. B., S. Alidina, M. W. Friedberg, S. J. Singer, D. Eastman, Z. Li, and E. C. Schneider. 2015. "A Difference-in-Difference Analysis of Changes in Quality, Utilization and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine* 31 (3): 289–96.
- Ryan, A. M., J. F. Burgess, and J. B. Dimick. 2014. "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences." *Health Services Research* 50: 1211–35.

- Scanlon, D. P., C. S. Hollenbeak, J. Beich, A.-M. Dyer, R. A. Gabbay, and A. Milstein. 2008. "Financial and Clinical Impact of Team-Based Treatment for Medicaid Enrollees with Diabetes in a Federally Qualified Health Center." *Diabetes Care* 31 (11): 2160–5.
- Stuart, E. A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25 (1): 1–21.
- Stuart, E. A., H. A. Huskamp, K. Duckworth, J. Simmons, Z. Song, M. E. Chernew, and C. L. Barry. 2014. "Using Propensity Scores in Difference-in-Differences Models to Estimate the Effects of a Policy Change." *Health Services and Outcomes Research Methodology* 14 (4): 166–82.
- Waters, T. M., C. F. Chang, W. T. Cecil, P. Kasteridis, and D. Mirvis. 2011. "Impact of High-Deductible Health Plans on Health Care Utilization and Costs." *Health Services Research* 46 (1p1): 155–72.
- Wharam, J. F., B. E. Landon, A. A. Galbraith, K. P. Kleinman, S. B. Soumerai, and D. Ross-Degnan. 2007. "Emergency Department Use and Subsequent Hospitalizations among Members of a High-Deductible Health Plan." *Journal of the American Medical Association* 297 (10): 1093–102.
- Xing, J., C. Goehring, and D. Mancuso. 2015. "Care Coordination Program for Washington State Medicaid Enrollees Reduced Inpatient Hospital Costs." *Health Affairs* 34 (4): 653–61.

SUPPORTING INFORMATION

Additional supporting information may be found online in the supporting information section at the end of the article:

Appendix SA1: Author Matrix.

Appendix SA2: Supplementary Appendix.