Health Services and Outcomes Research Methodology (2019) 19:127–144
https://doi.org/10.1007/s10742-018-0189-0

Difference-in-differences and matching on outcomes: a tale of two unobservables

Stephan Lindner[1] · K. John McConnell[1]

Received: 24 October 2017 / Revised: 14 September 2018 / Accepted: 27 September 2018 /
Published online: 3 October 2018

Abstract

Difference-in-differences combined with matching on pre-treatment outcomes is a popular method for addressing non-parallel trends between a treatment and control group. However, previous simulations suggest that this approach does not always eliminate or reduce bias, and it is not clear when and why. Using Medicaid claims data from Oregon, we systematically vary the distribution of two key unobservables—fixed effects and the random error term—to examine how they affect bias of matching on pre-treatment outcomes levels or trends combined with difference-in-differences. We find that in most scenarios, bias increases with the standard deviation of the error term because a higher standard deviation makes short-term fluctuations in outcomes more likely, and matching cannot easily distinguish between these short-term fluctuations and more structural outcome trends. The fixed effect distribution may also create bias, but only when matching on pre-treatment outcome levels. A parallel-trend test on the matched sample does not reliably distinguish between successful and unsuccessful matching. Researchers using matching on pre-treatment outcomes to adjust for non-parallel trends should report estimates from both unadjusted and propensity-score matching adjusted difference-in-differences, compare results for matching on outcome levels and trends and examine outcome changes around intervention begin to assess remaining bias.

Keywords Difference-in-differences · Matching · Simulation

1 Introduction

Difference-in-differences is a widely used method in health care research to evaluate initiatives such as Medicaid expansions (Sommers et al. 2014, 2012), payment reforms (Song et al. 2011) or the creation of Accountable Care Organizations (McConnell et al.

✉ Stephan Lindner
lindners@ohsu.edu

K. John McConnell
mcconnjo@ohsu.edu

[1] Department of Emergency Medicine, Center for Health System Effectiveness, Oregon Health & Science University, 3030 SW Moody Ave, Portland, OR 97201, USA

2017; McWilliams et al. 2016). It requires that the expected outcomes of the treatment and control group in the absence of the intervention exhibited parallel trends. When this assumption fails, difference-in-differences results in a biased estimate because it confounds changes unrelated to the intervention with the effect of the intervention.

One popular method for reducing bias when outcome trends are not parallel is to first match treatment and control observations on pre-treatment outcomes before applying difference-in-differences on the matched sample (Ryan et al. 2015). This approach is intuitively appealing: matching on outcomes corrects for non-parallel trends between the two groups while subsequent difference-in-differences eliminates any remaining outcome level differences. However, simulations suggest that this approach does not always eliminate or even reduce bias (Chabé-Ferret 2014; O'Neill et al. 2016; Ryan et al. 2015).

The goal of this paper is to identify sources of bias for difference-in-differences and matching on outcomes, and thereby to gain a better understanding under which conditions this approach improves or does not improve the estimate of the treatment effect. Using a standard framework that allows for parallel and non-parallel trends, we first discuss the purpose of matching on outcomes in the context of difference-in-differences. We then use Medicaid claims data from Oregon to simulate a number of scenarios by systematically varying one parameter, holding all other parameters constant. By doing so, we are able to isolate two types of biases connected to two unobservables—the distribution of fixed effects affecting outcome levels (i.e., time-invariant differences among observations) and the standard deviation of the random error term—and to assess how they affect matching on outcomes combined with difference-in-differences.

2 Purpose of matching on outcomes when combined with difference-in-differences

We use the potential outcome framework developed by Rubin (1974) throughout this study. Suppose there are $i = 1, \dots, N$ units (e.g., primary care physicians or hospitals) and T time periods, were $t = 1, \dots, (t^* - 1)$ periods are before and $t = t^*, \dots, T$ periods are after intervention begin. The potential outcome for unit i in period t is denoted by y_{it}^0 in the absence of treatment and y_{it}^1 in the presence of treatment.

Let D_i be an indicator equal to one if unit i is in the treatment group, i.e., is exposed to the intervention for $t \geq t^*$. Throughout this study, we make the standard difference-in-differences assumptions that (1) the outcome of one unit is unaffected by treatment assignment of another unit (stable unit treatment value assumption, or SUTVA), (2) covariates X_i are not influenced by the treatment (exogeneity assumption) and that (3) the treatment effect has no effect on the pre-treatment population (Lechner 2010). We specify the model for the potential outcome in the absence of treatment based on Abadie et al. (2010) as follows:

$$y_{it}^0 = X_i\beta + \gamma_t + \eta_i + \lambda_t \cdot \mu_i + \varepsilon_{it}, \tag{1}$$

where y_{it}^0 is the potential outcome absent treatment, X_i is time-invariant a vector of covariates, γ_t is a time shock common to all, η_i represents a fixed effect, μ_i represents a second fixed effect that is multiplied with a time shock λ_t, and ε_{it} is the error term, which is independent of the other regressors. In what follows, we refer to η_i as the time-invariant fixed effect or *level effect* because it affects outcome levels and to μ_i as the time-varying fixed-effect or *trend effect* because it affects (together with λ_t) outcome trends.

We denote the treatment effect by α and assume that it is additive, time-constant and homogeneous for all treatment units. The potential outcome if every unit was treated can then be written as follows:

$$y^1_{it} = X_i\beta + \gamma_t + \eta_i + \lambda_t \cdot \mu_i + \alpha \cdot I(t \geq t^*) + \varepsilon_{it}\,, \tag{2}$$

where $I(t \geq t^*)$ is an indicator equal to one after the intervention starts. The observed outcomes can then written as follows:

$$y_{it} = D_i \cdot y^1_{it} + (1 - D_i) \cdot y^0_{it}\,, \tag{3}$$

where the fact that one, and only one, of the potential outcomes is indeed observed for each unit i follows from SUTVA (Lechner 2010).

The key assumption for the difference-in-differences is that after controlling for observable confounders, changes in expected non-treatment outcomes between periods t and t' are mean independent of treatment assignment (Abadie 2005; Angrist and Pischke 2008; Lechner 2010; O'Neill et al. 2016):

$$A1\ :\ E[(y^0_{it} - y^0_{it'})|D_i = 1, X_i] = E[(y^0_{it} - y^0_{it'})|D_i = 0, X_i], \quad \text{all } t \neq t'\,. \tag{4}$$

In the context of the model shown by Eqs. (1) and (2), assumption (A1) requires that either $E(\lambda_t) = E(\lambda_{t'})$, all t, t' (i.e., no trend in outcomes) or, if $E(\lambda_t) \neq E(\lambda_{t'})$, that $E(\mu_i|D_i = 1) = E(\mu_i|D_i = 0)$ (balanced trend effects). In the remainder, we assume that $E(\lambda_t) \neq E(\lambda_{t'})$, all t, t' such that imbalanced trend effects imply violation of the parallel trends assumption.

When assumption (A1) is satisfied, difference-in-differences identifies an unbiased estimate of the average treatment effect on the treated (ATT), $E[y^1_{it} - y^0_{it}|D_i = 1]$ (Lechner 2010). However, when $E(\mu_i|D_i = 1) \neq E(\mu_i|D_i = 0)$, difference-in-differences leads to a bias.

The purpose of matching on outcomes when combined with difference-in-differences is to correct for non-parallel trends by balancing trend effects.[1] Put differently, matching is successful if $E(\mu_i|D_i = 1; y^0_{it,t<t^*}) = E(\mu_i|D_i = 0; y^0_{it,t<t^*})$, i.e., if trend effects are balanced after matching on outcomes prior to treatment begin $y^0_{it,t<t^*}$.

One approach is to match on pre-intervention outcome *levels* (Ryan et al. 2015). Such matching is successful if, conditional on pre-intervention outcome levels and covariates, potential outcomes are independent of treatment status (O'Neill et al. 2016):

$$A2\ :\ E[y^0_{it}|D_i = 1, X_i, y^0_{it,t<t^*}] = E[y^0_{it}|D_i = 0, X_i, y^0_{it,t<t^*}]\,, \quad \text{all } t. \tag{5}$$

This assumption is satisfied if matching balances both level and trend effects. In the context of difference-in-differences, assumption (A2) requires expected pre-intervention outcome levels of the treatment and control group to be identical after matching, but assumption (A1) only requires pre-intervention outcome trends to be identical (Lechner 2010).[2]

[1] Other methods to address non-parallel trends include the synthetic control methods (Abadie et al. 2010) and lagged dependent variable regression. See O'Neill et al. (2016) for a comparison of these methods.

[2] Used as separate approaches, difference-in-differences versus matching on covariates have non-nested assumptions (Angrist and Pischke 2008; Imbens and Wooldridge 2009). Difference-in-differences requires parallel trends but allows for level effect imbalance between the treatment and control group. Matching requires all confounders to be balanced between the two groups but does not require parallel trends. When matching is applied to pre-treatment outcome levels, successful matching implies that all potential outcome trends are perfectly aligned. In this case, assumption (A1) is nested in (A2). Conversely, if level effects

A variation of the matching approach that matches on outcome *differences* or trends may therefore appear to be more suitable for difference-in-differences. This approach differences out level effects and therefore ignores level differences in outcomes when matching the two groups. Matching on outcome trends is successful if, conditional on covariates and pre-intervention outcomes, differences in potential outcomes between two time periods are independent of treatment assignment:

$$\begin{aligned} A3\ :\ & E[(y^0_{it} - y^0_{it'})|D_i = 1, X_i, (y^0_{it,t<t^*} - y^0_{i(t-1),t<t^*})] \\ & = E[(y^0_{it} - y^0_{it'})|D_i = 1, X_i, (y^0_{it,t<t^*} - y^0_{i(t-1),t<t^*})]\,, \quad \text{all } t \neq t'\,. \end{aligned} \tag{6}$$

Returning to Eq. (1), matching on outcome levels or trends may not be fully successful because similar levels or trends in outcomes do not necessarily imply balanced trend effects between the treatment and control group. Specifically, observations with the same outcome may exhibit different trend effects as well as different level effects and short-term fluctuations due to the error term, but matching cannot distinguish between these types of unobservables. In the next section, we examine how differences in the distribution of the level effect and the standard deviation of the error term affect performance of matching on outcome levels or trends combined with difference-in-differences by simulating scenarios that differ in the distribution of these two unobservables.

3 Data, simulation and model estimation

We use Medicaid claims for the year 2011. Claims are aggregated to the primary care physician (PCP) level using national provider identifier (NPI) codes. We exclude claims with missing PCP information and PCPs with less than 100 claims per year. When aggregating to the PCP level, we create as outcome variable the log of the number of PCP visits. In addition, the following variables are used in regressions as control variables: percent of claims from male patients, average age of patients; average Chronic Illness and Disability Payment System (CDPS) risk score [a score to predict costs specifically designed for the Medicaid population developed by the University of California, San Diego (Kronick et al. 2000)], and fraction of patients from an urban area (based on their zip code). The aggregated data has 1716 PCPs.

As a first step in simulating the sample, we randomly draw 1000 practices and assign one half of them to the treatment group and the other healf of them to the control group. We then use parameter values of the data-generating process shown by Eqs. (1) and (3) to simulate nine more time periods, for a total of 10 time periods. Intervention begins with the sixth period.

We vary parameter values for the trend effect, level effect and standard deviation of the error term to create 10 scenarios (see Table 1 for an overview of scenarios and key parameters). For all scenarios, we assume $\alpha = 0.125$ for the intervention effect.[3] We also set the

Footnote 2 (continued)

remain imbalanced after matching, assumption (A2) is violated but assumption (A1) may still hold, as long as trend effects are balanced.

[3] Simulations with alternative values for α do not change our results. Intuitively, matching methods only use pre-intervention observations, which are assumed to be unaffected by the intervention effect.

Table 1 Simulation parameters by data-generating scenario

Scenario	Group	Fixed effects		
		Trend (mean)	Level (mean)	Error term (standard deviation)
1	Control:	0.0	0.0	0.1
	Treatment:	0.0	0.6	0.1
2	Control:	0.0	0.0	0.0
	Treatment:	0.1	0.6	0.0
3	Control:	0.0	0.0	0.1
	Treatment:	0.1	0.6	0.1
4	Control:	0.0	0.0	0.2
	Treatment:	0.1	0.6	0.2
5	Control:	0.0	0.0	0.0
	Treatment:	0.1	0.0	0.0
6	Control:	0.0	0.0	0.1
	Treatment:	0.1	0.0	0.1
7	Control:	0.0	0.0	0.2
	Treatment:	0.1	0.0	0.2
8	Control:	0.0	0.0	0.0
	Treatment:	0.1	− 0.6	0.0
9	Control:	0.0	0.0	0.1
	Treatment:	0.1	− 0.6	0.1
10	Control:	0.0	0.0	0.2
	Treatment:	0.1	− 0.6	0.2

Trend effects and level effects are drawn using a normal distribution with mean values as specified above and a standard deviation of 0.1. The expected value of the error term is zero in all scenarios

expected value of the level and trend effect of the control group to zero for all scenarios [i.e., $E(\eta_i|D_i = 0) = 0$ and $E(\mu_i|D_i = 0) = 0$, respectively]. These parameter values imply that the expected outcome of the control group as an intercept of zero and slope of zero, i.e., does not change over time. We further assume that both level and trend effects are distributed normally with a standard deviation of 0.1 in all scenarios. Simulations are repeated 1000 times for each scenario. For each iteration, we draw a new sample of practices and repeat the random assignment into treatment and control group.

For the first scenario, we set $E(\mu_i|D_1 = 1) = 0$ such that trend effects are balanced between the treatment and control group and potential outcomes for the two groups exhibit parallel trends in expectation. We also set $E(\eta_i|D_i = 1) = 0.6$ such that the treatment group has a higher intercept than the control group. We set the standard deviation of the error term to 0.1.

For the other nine scenarios, we set the expected value of the trend effect for the treatment group to 0.1 [i.e., $E(\mu_i|D_i = 1) = 0.1 > E(\mu_i|D_i = 0) = 0$] and the expected value of the time shocks interacted with the trend effect to $E(\lambda_t) = t$. These parameter values imply imbalanced trend effects between the treatment and control group and an expected potential outcome of the treatment group that increases linearly with a slope of 0.1.[4] We then specify three values for the expected level effect of the treatment group and three values for the standard deviation (SD) of the error term. By combining these values, we obtain the nine scenarios.

Specifically, the three expected values for the level effect of the treatment group are as follows: $E(\eta_i|D_i = 1) = 0.6$ (positive level effects difference between treatment and control observations), $E(\eta_i|D_i = 1) = 0$ (zero level effects difference) and $E(\eta_i|D_i = 1) = -0.6$ (negative level effects difference). Outcome levels and trends between the treatment group are closer together as one moves from scenarios with a positive level effects differences to scenarios with a negative level effects difference.

Regarding the error term, we set its standard deviation to 0, 0.1, and 0.2, respectively. The standard deviation of the error term determines how much short-term, random fluctuations due to the error term affect outcome trends. When set to zero, there are no such fluctuations and outcome trends are fully determined by trend effects (as well as the common time shock γ_t). By contrast, short-term fluctuations in the error term may influence outcome trends when the standard deviation is non-zero. For instance, if the standard deviation of the error term is 0.1, then an observation in the control group has a roughly 2.5% chance of receiving two consecutive positive error shocks with a magnitude of 0.1 or higher; the chance is about 10% when the standard deviation of the error term is 0.2.

We estimate four different models on the simulated data. All of them use the same difference-in-differences model but differ in how they specify their sample. Specifically, the first model (subsequently abbreviated by `DD`) uses the full sample whereas the second to fourth model (`DD + PSM 1`, `DD + PSM 2` and `DD + PSM 3`) specify a subsample by using different strategies to match on outcome levels or trends.

We use propensity score matching for all three matching models. The propensity score is defined as the probability of being in the treatment group as a function of observed covariates (Austin 2011; Stuart 2010). It has the advantage of condensing the potentially large number of observed covariates into one measure used for matching. We use logistic regressions to estimate propensity scores (Stuart et al. 2014). As in Ryan et al. (2015), we use nearest neighborhood matching with replacement, a caliper of 0.1 of the standard deviation of the propensity score and enforcement of common support.

The three matching models differ in terms of the pre-intervention outcomes used for the propensity score matching. Model `DD + PSM 1` uses only the last outcome before intervention begin for matching, as in Chabé-Ferret (2014). In this case, the logistic regression may be written as follows:

$$P(D_i = 1|y_{i,(t^*-1)}) = \Lambda(\phi + \pi y_{i,(t^*-1)})\,, \tag{7}$$

where Λ is the logistic cumulative distribution function, ϕ is the intercept and π is the coefficient for $y_{i,(t^*-1)}$, which is the last outcome before intervention begin for unit i. As defined above, D_i is an indicator equal to one if unit i is in the treatment group.[5]

[4] Assuming that the expected value of the time shock λ_t changes by a constant term implies a specific version of the Abadie et al. (2010) model with a constant slope for the expected outcomes. We make this restriction to simplify our simulations.

[5] For this model, matching is based on just one covariate, and therefore, propensity score matching does not have the advantage of condensing a large number of variables into one measure. We still use propensity score matching for this model so that all three matching models follow the same matching approach.

Model `DD + PSM 2` uses all pre-intervention outcome levels by estimating the following logistic regression:

$$P(D_i = 1|y_{it,t<t^*}) = \Lambda\left(\phi + \sum_{j=1}^{t^*-1} \pi_j y_{i,(t^*-j)}\right). \tag{8}$$

The last model (`DD + PSM 3`) uses outcome differences instead of levels by estimating the following logistic regression:

$$P(D_i = 1|(y_{it,t<t^*} - y_{i(t-1),t<t^*})) = \Lambda\left(\phi + \sum_{j=1}^{t^*-2} \pi_j \left(y_{i,(tx^*-j)} - y_{i,(t^*-j-1)}\right)\right). \tag{9}$$

The difference-in-differences regression is identical for all four models:

$$y_{it} = \beta_0 + \alpha D_i \cdot I(t \geq t^*) + \beta_1 D_i + \beta_2 I(t \geq t^*) + X_i' \gamma + \zeta_{it}, \tag{10}$$

where y_{it} is the log of the number of PCP visits and X_i includes fraction male, average age, average CDPS risk score and fraction urban. The errors component is denoted by ζ_{it} and includes all unobservable terms from Eq. (1). Standard errors are clustered at the practice level. We use weights from the matching step for the three matching models.

We also perform parallel-trend tests for the simulated sample both before and after the matching procedure. The parallel-trend test is specified as follows:

$$y_{it} = \delta_0 + \pi_i + \rho_0 \cdot t + \rho_1 \cdot t \cdot D_i + \gamma X_i + \psi_{it}, \tag{11}$$

where δ_0 is the intercept, ρ_0 is a coefficient for the slope for the comparison group, ρ_1 is a coefficient for the difference in slopes between the comparison and treatment group, and ψ_{it} is the error term. Standard errors are clustered at the practice level. The test uses only observations from the pre-intervention period. The null hypothesis of parallel trends corresponds to $\rho_1 = 0$. All analysis is performed using R version 3.4.1.

To assess performance of the four difference-in-differences estimators, we present the following simulation statistics: (1) mean absolute bias (i.e., the average difference between simulated estimates and the true intervention effect α); (2) coverage probability (i.e., the percent of simulations for which the 95-percent confidence interval of the difference-in-differences estimates includes the true intervention effect); (3) mean standard errors; and (4) Monte Carlo standard errors, which is a measure of the accuracy of the simulations (Koehler et al. 2009; Lee and Young 1999).[6] For the parallel-trend tests, we report the average simulated value of ρ_1 before and after matching, the rejection percentage (i.e., the percentage of simulations for which the null hypothesis of parallel trends is rejected) before and after matching, and the mean absolute bias of the corresponding difference-in-differences estimator.

[6] We estimate the Monte Carlo standard error based on Koehler et al. (2009) using the following formula: $\widehat{MCSE} = \frac{1}{R}\sqrt{\sum_{r=1}^{R}(\hat{\beta}_{2,r} - \overline{\hat{\beta}_2})^2}$, where R is the number of replications, $\hat{\beta}_{2,r}$ is the estimate of the treatment effect for the difference-in-differences model shown by Eq. (10) for replication r, and $\overline{\hat{\beta}_2} = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{2,r}$.

Table 2 Description of Medicaid sample at baseline

Variable	Mean	Standard deviation	Minimum	Maximum
Number of visits (log)	1.4	0.3	0.2	2.9
Male patients	42.3	13.9	0.0	81.0
Age	19.2	10.7	3.5	49.5
CDPS risk score	1.1	0.4	0.6	6.6
Urban residents	56.1	38.9	0.0	100.0

The table shows mean values, standard deviations and minimum and maximum values of variables used for the simulation at baseline. The unit of observation is a practice. Mean values of male and urban patients are average values of the percentage of male and urban patients in each practice, respectively. Mean values of number of visits, age and CDPS risk score are average values of respective practice averages.

CDPS chronic illness and disability payment system score

4 Results

Table 2 describes the sample at baseline (i.e., first period). The average log of visits is 1.4, which corresponds to about 4 practice visits per year. Medicaid enrollees tend to be young (reflecting the large fraction of children on Medicaid before the Medicaid expansion) and female. The average fraction of patients residing in urban areas across practices is slightly above 50% and the average CDPS risk score across patients and practices is 1.1 (range 0.6–6.6).

When trends are parallel, simple difference-in-differences performs best (Table 3): the absolute mean bias of the estimator is zero and its mean standard error is the lowest. `DD + PSM 1` has a significant negative bias, `DD + PSM 2` a small negative bias and `DD + PSM 3` a zero bias. Coverage rates correspond to these biases, with the simple difference-in-differences and `DD + PSM 3` having the highest coverage probability, followed by `DD + PSM 2` and `DD + PSM 1` (with a coverage probability of zero percent). Monte Carlo standard errors are small across all scenarios and models.

In the other nine scenarios with non-parallel trends, difference-in-differences has a substantial bias of about 0.5. In these scenarios, simple difference-in-differences confounds a steeper outcome trend of the treatment group compared to the control group with the intervention effect.[7] `DD + PSM 1` also performs relatively poorly across most scenarios, and it even has a higher bias than simple difference-in-differences in some scenarios. Models `DD + PSM 2` and `DD + PSM 3` have low or even zero bias in some scenarios but a

[7] The expected mean bias for the simple difference-in-differences is exactly 0.5 across all scenarios. To understand this result, note that $\hat{\alpha}$ in Eq. (10) can be written as follows (Angrist and Pischke 2008): $\hat{\alpha} = (\bar{y}_{Treat,t \geq t^{*}t} - \bar{y}_{Treat,t<t^{*}t}) - (\bar{y}_{Control,t \geq t^{*}t} - \bar{y}_{Control,t<t^{*}t})$, where $\hat{y}$ denotes averages. The second difference is zero in expectations for all simulations because expected values of the level and trend effects for the comparison group are set to zero. Regarding the first difference, note that treatment group observations have an expected intercept and slope of 0.6 and 0.1, respectively. The slope implies that expected average outcome values are shifted upwards by $0, 0.1, \ldots 0.4$ for the five pre-intervention periods and by $0.5, 0.6, \ldots 0.9 = 2.5 + (0, 0.1, \ldots 0.4)$ for the five post-intervention scenarios. It then follows that $(E(\hat{\alpha}) = \alpha + 0.6 + 0.5 + (0 + 0.1 + 0.2 + 0.3 + 0.4)/5) - (0.6 + (0 + 0.1 + 0.2 + 0.3 + 0.4)/5) = \alpha + 0.5$.

Table 3 Difference-in-differences simulation results

Scenario	Parallel trends?	Level effect difference	SD error	Model	Mean bias	Cover. prob.	Mean S.E.	Monte Carlo S.E.
1	Yes	Positive	0.1	DD	− 0.001	95.2	0.032	0.001
				DD + PSM 1	− 0.512	0.0	0.050	0.001
				DD + PSM 2	− 0.041	93.3	0.098	0.003
				DD + PSM 3	− 0.001	98.4	0.049	0.001
2	No	Positive	0.0	DD	0.500	0.0	0.032	0.001
				DD + PSM 1	− 0.312	13.2	0.088	0.003
				DD + PSM 2	0.206	54.2	0.114	0.004
				DD + PSM 3	0.000	100.0	0.042	0.000
3	No	Positive	0.1	DD	0.500	0.0	0.032	0.001
				DD + PSM 1	− 0.284	13.2	0.086	0.003
				DD + PSM 2	0.183	60.3	0.112	0.004
				DD + PSM 3	0.046	98.5	0.042	0.000
4	No	Positive	0.2	DD	0.501	0.0	0.033	0.001
				DD + PSM 1	− 0.188	32.4	0.081	0.003
				DD + PSM 2	0.146	70.9	0.113	0.004
				DD + PSM 3	0.141	2.8	0.043	0.001
5	No	Zero	0.0	DD	0.500	0.0	0.032	0.001
				DD + PSM 1	0.130	3.5	0.041	0.001
				DD + PSM 2	0.000	100.0	0.042	0.000
				DD + PSM 3	0.000	100.0	0.042	0.000
6	No	Zero	0.1	DD	0.500	0.0	0.032	0.001
				DD + PSM 1	0.150	0.7	0.042	0.001
				DD + PSM 2	0.044	98.6	0.042	0.001
				DD + PSM 3	0.046	97.3	0.042	0.001

Table 3 (continued)

Scenario	Parallel trends?	Level effect difference	SD error	Model	Mean bias	Cover. prob.	Mean S.E.	Monte Carlo S.E.
7	No	Zero	0.2	DD	0.501	0.0	0.033	0.001
				DD + PSM 1	0.198	0.0	0.043	0.001
				DD + PSM 2	0.127	6.6	0.043	0.001
				DD + PSM 3	0.144	2.6	0.043	0.001
8	No	Negative	0.0	DD	0.499	0.0	0.032	0.001
				DD + PSM 1	0.691	0.0	0.037	0.001
				DD + PSM 2	− 0.155	76.1	0.108	0.003
				DD + PSM 3	0.000	100.0	0.042	0.000
9	No	Negative	0.1	DD	0.500	0.0	0.032	0.001
				DD + PSM 1	0.682	0.0	0.038	0.001
				DD + PSM 2	0.004	94.5	0.104	0.003
				DD + PSM 3	0.046	97.5	0.042	0.001
10	No	Negative	0.2	DD	0.498	0.0	0.033	0.001
				DD + PSM 1	0.653	0.0	0.040	0.001
				DD + PSM 2	0.249	24.2	0.091	0.003
				DD + PSM 3	0.141	2.5	0.043	0.001

The table shows mean absolute bias, empirical coverage probability, mean standard error and Monte Carlo standard error of ten simulated scenarios. Standard errors are clustered at the PCP level. The difference-in-differences models are defined as follows: `DD`: Difference-in-differences without matching to adjust for diverging trends; `DD + PSM 1`: Propensity score matching using the last outcome level before intervention begin, followed by difference-in-differences; `DD + PSM 2`: Propensity score matching using all outcome levels before intervention begin, followed by difference-in-differences; `DD + PSM 3`: Propensity score matching using all outcome differences before intervention begin, followed by difference-in-differences.

moderate bias in others. A higher bias corresponds to a lower coverage rate, with model DD + PSM 2 having higher coverage rates for a similar bias compared to model DD + PSM 3 because of its higher average standard errors. In what follows, we investigate how level effect differences and differences in the standard deviation of the error term affect bias for these two models by graphically comparing the mean absolute bias as a function of these parameters.

Starting with model DD + PSM 3 (matching on outcome trends), three results emerge (Fig. 1). First, the mean absolute bias does not change with the level effect difference because outcome trends do not depend on level effects. Second, the bias is exactly zero when the standard deviation of the error term is zero. In this case, short-term fluctuations caused by the error term do not exist and matching can perfectly balance trend effects between the treatment and control group. And third, the mean absolute bias increases in the standard deviation of the error term as matching increasingly confounds short-term fluctuations due to the error term with structural changes due to trend effects.

To better understand the role of the error variance, we calculate, for each period in scenario 4, mean differences in the error component (i.e., ε_{it}) and trend effect component (i.e., $\lambda_t \cdot \mu_i$) between treatment and control group using the sample obtained after matching on outcome trends (Fig. 2). The trend effect component increases steadily, reflecting the remaining imbalance in trend effects between the treatment and control group in the matched sample that creates bias for the subsequent difference-in-differences estimator. By contrast, the error term component trends down for the pre-treatment period and then hovers around zero for the post-treatment period. This shows that observations from the treatment group with structural, persistent upward trends in their outcomes tend to be matched to observations from the control group that happen to have, due to random fluctuations, similar positive outcome trends in the pre-treatment period but not afterwards. Put differently, matching, by attempting to eliminate one source of bias (caused by imbalanced trend effects) introduces another source of bias (caused by short-term fluctuations).

Moving to model DD + PSM 2 and focusing for now on the three scenarios with a zero error variance and thus, zero bias due to short-term fluctuations, we can see that there is a positive relationship between the level effect difference and bias (Fig. 3). Matching on outcome levels results in a bias because it confounds positive or negative differences in level effects with non-parallel trends that both result in similar outcome levels. Specifically, when the level effect difference is positive, control group observations with a smaller positive trend and higher intercept tend to be matched to treatment group observations with a higher trend and lower intercept. The result is a matched sample where the outcome trend of the control group still is flatter than the outcome trend of the treatment group, creating a positive bias. Conversely, in scenarios with a negative level effect difference, control group observations with a higher positive trend and lower intercept tend to be matched to treatment group observations with a lower positive trend and higher intercept, resulting in a negative bias. Turning to the other six scenarios with a non-zero standard deviation of the error term, we can see that short-term fluctuations create an additional bias in scenarios with a zero level effect difference and, even more so, in scenarios with a negative level effect difference. The bias increases in the degree to which outcome trends between the two groups overlap because the more they do, the more likely it is to find a control group observation with a consecutively more positive error term that also has similar outcome levels than some of the treatment group observations.

Table 4 shows mean simulated estimates and rejection rates of the parallel-trend test before and after matching as well as the corresponding absolute mean bias for the nine scenarios with diverging trends and models DD + PSM 2 and DD + PSM 3. Matching

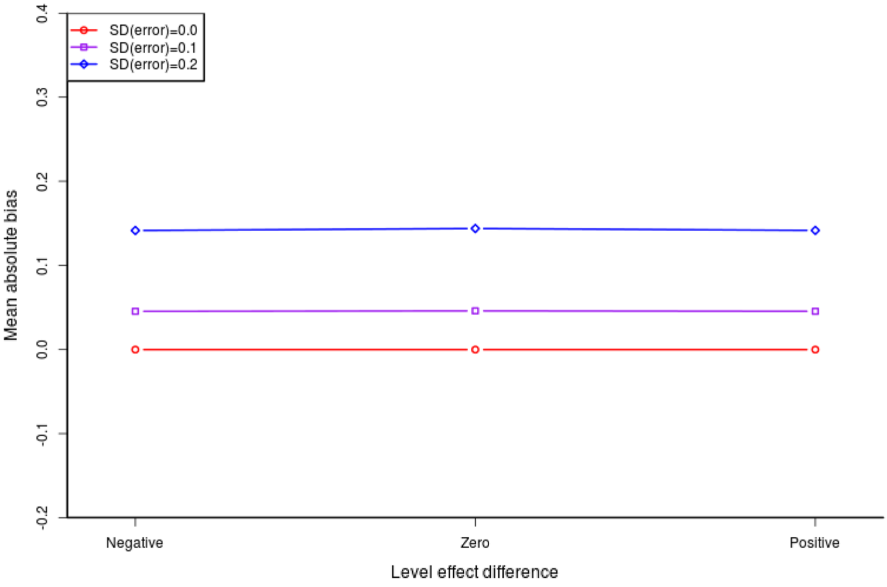

Fig. 1 Mean absolute bias as a function of level effect differences and standard deviation of the error term when matching on outcome trends (Scenarios 2–10, model `DD + PSM 3`)

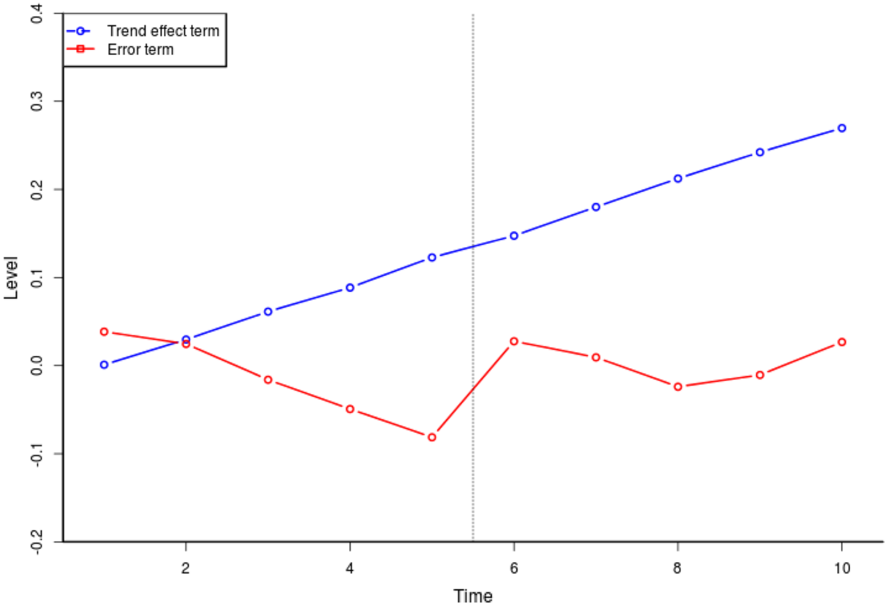

Fig. 2 Trends in error term and trend effect difference between treatment and control group (Scenario 4, model `DD + PSM 3`) *Notes:* The figure shows mean differences of the error term component and trend effect component between treatment and control group over time for the propensity-score matching model on outcome trends and scenario 4 (diverging trends, positive level effect differences, high standard deviation of error term). The error term component is the part of the outcome determined by the error term. The trend effect component is the part of the outcome determined by the trend effect. Both trends are calculated using the propensity score matched sample

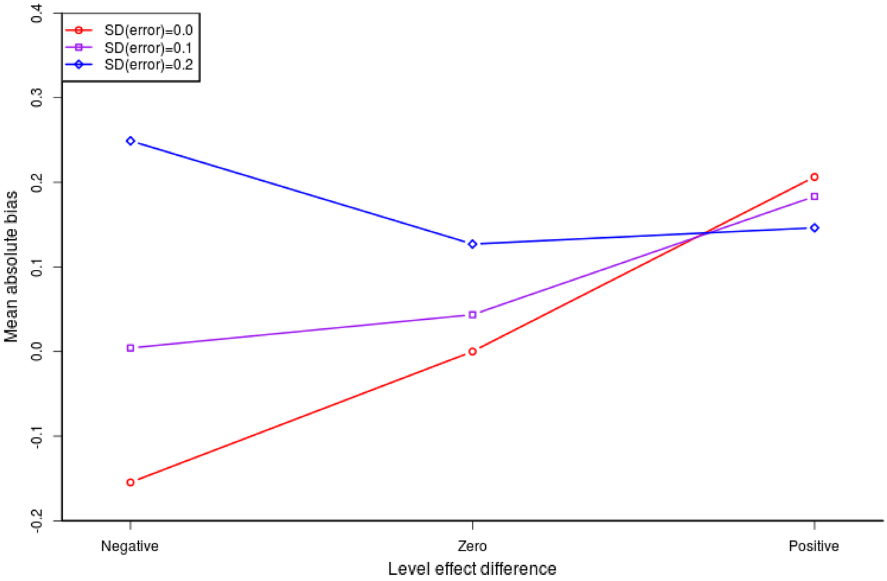

Fig. 3 Mean absolute bias as a function of level effect differences and standard deviation of the error term when matching on outcome levels (Scenarios 2–10, model `DD + PSM 2`)

on outcome levels or trends drastically reduces mean simulated coefficient values of ρ_1 as well as rejection rates. For model `DD + PSM 3` (matching on outcome trends), the rejection rate is zero for all displayed scenarios and simulations. For this model, a parallel-trend test on the matched sample is completely uninformative about the remaining difference-in-differences bias in the matched sample because an increasing imbalance in the error term component cancels out the trend effect imbalance, resulting in non-significant parallel-trend tests even when then difference-in-differences bias on the matched sample is substantial. For model `DD + PSM 2`, the parallel-trend test is informative for scenarios with positive level effects difference but not so for the scenarios with negative level effects difference.

To illustrate the relationship between parallel trend test and bias, we compare estimates of ρ_1 to mean absolute difference-in-differences biases for simulations in scenario 3 (Fig. 4). There is a clear and strong correlation between the parallel-trend test coefficient and difference-in-differences bias when matching on outcome levels (correlation: 0.91) but not when matching on outcome trends (correlation: 0.16).

5 Discussion

Simulation results suggest that there are at least two sources of biases that can arise in difference-and-differences analyses when combined with matching on outcome levels: one related to the unobserved, random error term and the other one related to the distribution of the unobserved level effect. In what follows, we describe each of these biases and then discuss implications for research using the difference-in-differences.

Table 4 Parallel trend test and bias

Sce nario	Level effect difference	SD error	Model	Mean value of ρ_1		Rejection rate (percent)		Mean absolute
				Before matching	After matching	Before matching	After matching	bias (diff and diff)
2	Positive	0.0	DD + PSM 2	0.10	0.04	100.0	46.4	0.206
			DD + PSM 3	0.10	0.00	100.0	0.0	0.000
3	Positive	0.1	DD + PSM 2	0.10	0.04	100.0	41.9	0.183
			DD + PSM 3	0.10	0.00	100.0	0.0	0.046
4	Positive	0.2	DD + PSM 2	0.10	0.04	100.0	28.5	0.146
			DD + PSM 3	0.10	0.00	100.0	0.0	0.141
5	Zero	0.0	DD + PSM 2	0.10	0.00	100.0	0.0	0.000
			DD + PSM 3	0.10	0.00	100.0	0.0	0.000
6	Zero	0.1	DD + PSM 2	0.10	0.00	100.0	0.0	0.044
			DD + PSM 3	0.10	0.00	100.0	0.0	0.046
7	Zero	0.2	DD + PSM 2	0.10	0.00	100.0	0.0	0.127
			DD + PSM 3	0.10	0.00	100.0	0.0	0.144
8	Negative	0.0	DD + PSM 2	0.10	− 0.03	100.0	22.9	− 0.155
			DD + PSM 3	0.10	0.00	100.0	0.0	0.000
9	Negative	0.1	DD + PSM 2	0.10	− 0.03	100.0	21.6	0.004
			DD + PSM 3	0.10	0.00	100.0	0.0	0.046
10	Negative	0.2	DD + PSM 2	0.10	− 0.02	100.0	11.5	0.249
			DD + PSM 3	0.10	0.00	100.0	0.0	0.141

The table shows average coefficient values of parallel trend tests and rejection rate (as percentage) before matching (full sample) and after matching for the nine simulation scenarios with diverging trends and matching on outcome levels (`DD + PSM 2`) as well as matching on outcome differences (`DD + PSM 3`). The average bias corresponds to the difference-in-differences estimator of the matched sample

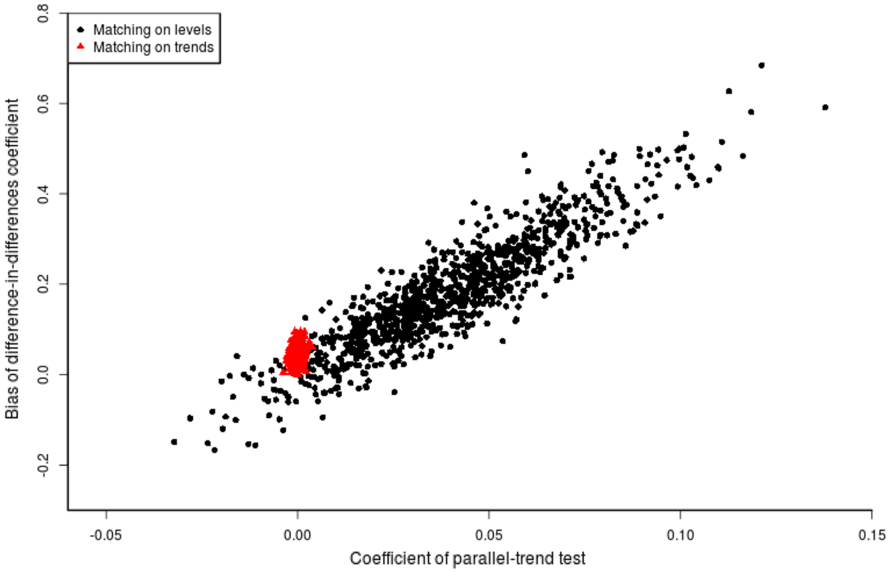

Fig. 4 Correlation between post-matching parallel-trend test and bias of subsequent difference-in-differences when matching on outcome levels versus outcome trends (Scenario 3, models `DD + PSM 2` and `DD + PSM 3`) *Notes:* Each point in the figure corresponds to one simulation for scenario 3 (non-parallel trends, positive level effect difference and a 0.1 standard deviation of the error term). Values on the horizontal axis show coefficient values of the parallel-trend test using the matched sample, where a value of zero corresponds to the null hypothesis of parallel trends. Values on the vertical axis show bias of subsequent difference-in-differences

1. Differences in short-term fluctuations versus non-parallel trends: A first type of bias occurs because matching cannot fully distinguish between temporary changes in outcomes due to the error term and permanent or structural changes in outcomes due to the trend effect. For instance, improvements in mortality rates among clinics may be due to improved processes (in which case they persist) or sheer luck (in which case they return to previous levels). We isolate this effect by varying the standard deviation of the error term, holding other parameters constant. This type of bias can affect both matching approaches, but affects matching on outcome levels only in scenarios where outcomes between the treatment and control group are similar.
2. Differences in level effects versus non-parallel trends: The second type of bias exists because matching on pre-treatment outcomes may not fully distinguish between differences in levels due to the level effects and differences in slope due to the trend effects. We isolate this second type of bias in the simulations by varying the level effect difference between treatment and control group, holding other parameters constant. Our results show that this source of bias only arises when matching on outcome levels but not when matching on outcome trends because the latter differences out level effects for matching. Using more pre-treatment outcome levels appears to ameliorate the problem when matching on outcome levels because it allows matching to better discriminate between level and trend effects.

Aside from these sources of bias, several insights can be learned from our simulations. Matching on outcome trends focuses on correcting different slopes between treatment and comparison group observations, which leads to bias when using simple difference-in-differences. It is therefore not surprising that matching on outcome trends performs relatively well when the error variance is small. Matching on outcome levels seeks to match on both levels and trends, but is identical to matching on outcome trends when there is no fixed effect difference. In these scenarios, the two approaches show similarly good performance. Our finding that both matching approaches are negatively affected by the variance of the error term relates our simulations to other work on adjusting for covariates with measurement error (Lockwood and McCaffrey 2016). Future research could explore whether solutions suggested in this literature could be applied to this difference-in-differences framework.

There are several implications for the practitioner interested in conducting a difference-in-differences analysis where the pre-intervention trends do not pass the parallel trend test. First, researchers should be aware that matching on pre-intervention outcome levels or trends may reduce but not fully eliminate the bias found in simple difference-in-differences. They should therefore report results for simple difference-in-differences along with results from matching on levels and trends to provide a more robust understanding of the range of estimates and why they differ. Second, testing for parallel trends after matching on pre-intervention outcomes is not highly informative. In addition to such a test, researchers should therefore also visually inspect outcome trends around intervention begin. For instance, a strong drop in outcomes of the control group at intervention begin might indicate a bias due to short-term fluctuations. And third, our simulations provide some guidance as to when matching on levels or trends may be preferable. Specifically, matching on outcome levels might be more suitable when the combination of intercepts and trends create a scenario with distinct, separate outcome trends. Conversely, matching on outcome trends might be preferable when outcome trends of the two groups are close together because in such a scenario, matching on outcomes levels cannot easily delineate between differences in trends and levels that contribute to similar outcomes.

Results of our simulation may also explain why research findings have been inconsistent with respect to the utility of matching on pre-intervention outcomes. For instance, Ryan et al. (2015) appear to simulate diverging trend scenarios with small level effect differences when assignment to the treatment group is based on trends.[8] Our results suggest that matching on outcome levels performs relatively well in such a case. Chabé-Ferret (2014) uses the last pre-treatment outcome for matching, an approach that tends to perform worse than the other matching approaches because it does not exploit all pre-treatment periods.

While matching on outcome levels or trends generally decreases bias compared to simple difference-in-differences in the scenarios presented here, using just difference-in-differences may sometimes be preferable. For instance, Chabé-Ferret (2015) finds that difference-in-differences performs well when trends diverge symmetrically around the treatment date and when the same number of pre-and post-treatment periods are used. More generally, his example and this article suggest that instead of relying on one research design for all types of scenarios, researchers do best to gain a thorough understanding of

[8] Specifically, the probability of being assigned to the treatment group is associated positively with pre-intervention trends for their third scenario, which suggests that level effects were similar between the two groups.

the underlying processes determining outcome trends of their sample and then to select the research approach that fits their data best.

Acknowledgements We would like to thank participants of the CHSE brownbag for helpful comments and suggestions.

Funding Authors Lindner and McConnell received no outside funding for this study.

Compliance with ethical standards

Conflict of interest Authors Lindner and McConnell declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with animals performed by any of the authors. All procedures performed in this study that involved human subsjects were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent for this study was not required.

References

Abadie, A.: Semiparametric difference-in-differences estimators. Rev. Econ. Stud. **72**(1), 1–19 (2005). https://doi.org/10.1111/0034-6527.00321

Abadie, A., Diamond, A., Hainmueller, J.: Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J. Am. Stat. Assoc. **105**(490), 493–505 (2010). https://doi.org/10.1198/jasa.2009.ap08746

Angrist, J.D., Pischke, J.S.: Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton (2008)

Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar. Behav. Res. **46**(3), 399–424 (2011). https://doi.org/10.1080/00273171.2011.568786

Chabé-Ferret, S.: Bias of Causal Effect Estimators Using Pre-policy Outcomes. Toulouse School of Economics, Toulouse (2014)

Chabé-Ferret, S.: Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. J. Econom. **185**(1), 110–123 (2015). https://doi.org/10.1016/j.jeconom.2014.09.013

Imbens, G.W., Wooldridge, J.M.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**(1), 5–86 (2009). https://doi.org/10.1257/jel.47.1.5

Koehler, E., Brown, E., Haneuse, S.J.P.A.: On the assessment of monte carlo error in simulation-based statistical analyses. Am. Stat. **63**(2), 155–162 (2009). https://doi.org/10.1198/tast.2009.0030

Kronick, R., Gilmer, T., Dreyfus, T., Lee, L.: Improving health-based payment for Medicaid beneficiaries: CDPS. Health Care Financ. Rev. **21**(3), 29 (2000)

Lechner, M.: The estimation of causal effects by difference-in-difference methods estimation of spatial panels. Found. Trends Econom. **4**(3), 165–224 (2010). https://doi.org/10.1561/0800000014

Lee, S.M.S., Young, G.A.: The effect of monte carlo approximation on coverage error of double-bootstrap confidence intervals. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **61**(2), 353–366 (1999). https://doi.org/10.1111/1467-9868.00181

Lockwood, J.R., McCaffrey, D.F.: Matching and weighting with functions of error-prone covariates for causal inference. J. Am. Stat. Assoc. **111**(516), 1831–1839 (2016). https://doi.org/10.1080/01621459.2015.1122601

McConnell, K.J., Renfro, S., Lindrooth, R.C., Cohen, D.J., Wallace, N.T., Chernew, M.E.: Oregon's Medicaid reform and transition to global budgets were associated with reductions in expenditures. Health Aff. **36**(3), 451–459 (2017). https://doi.org/10.1377/hlthaff.2016.1298

McWilliams, J.M., Hatfield, L.A., Chernew, M.E., Landon, B.E., Schwartz, A.L.: Early performance of accountable care organizations in Medicare. N. Engl. J. Med. **374**(24), 2357–2366 (2016). https://doi.org/10.1056/nejmsa1600142

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., Sekhon, J.S.: Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv. Outcomes Res. Methodol. **16**(1–2), 1–21 (2016). https://doi.org/10.1007/s10742-016-0146-8

Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688–701 (1974). https://doi.org/10.1037/h0037350

Ryan, A.M., Burgess, J.F., Dimick, J.B.: Why we should not be indifferent to specification choices for difference-in-differences. Health Serv. Res. **50**(4), 1211–1235 (2015). https://doi.org/10.1111/1475-6773.12270

Sommers, B.D., Buchmueller, T., Decker, S.L., Carey, C., Kronick, R.: The Affordable Care Act has led to significant gains in health insurance and access to care for young adults. Health Aff. **32**(1), 165–174 (2012). https://doi.org/10.1377/hlthaff.2012.0552

Sommers, B.D., Kenney, G.M., Epstein, A.M.: New evidence on the Affordable Care Act: coverage impacts of early Medicaid expansions. Health Aff. **33**(1), 78–87 (2014). https://doi.org/10.1377/hlthaff.2013.1087

Song, Z., Safran, D.G., Landon, B.E., He, Y., Ellis, R.P., Mechanic, R.E., Day, M.P., Chernew, M.E.: Health care spending and quality in year 1 of the alternative quality contract. N. Engl. J. Med. **365**(10), 909–918 (2011). https://doi.org/10.1056/nejmsa1101416

Stuart, E.A.: Matching methods for causal inference: a review and a look forward. Stat. Sci. **25**(1), 1–21 (2010). https://doi.org/10.1214/09-sts313

Stuart, E.A., Huskamp, H.A., Duckworth, K., Simmons, J., Song, Z., Chernew, M.E., Barry, C.L.: Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Serv. Outcomes Res. Methodol. **14**(4), 166–182 (2014). https://doi.org/10.1007/s10742-014-0123-z