HSR Health Services Research

DOI: 10.1111/1475-6773.12270
METHODS CORNER

Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences

Andrew M. Ryan, James F. Burgess Jr., and Justin B. Dimick

Objective. To evaluate the effects of specification choices on the accuracy of estimates in difference-in-differences (DID) models.

Data Sources. Process-of-care quality data from Hospital Compare between 2003 and 2009.

Study Design. We performed a Monte Carlo simulation experiment to estimate the effect of an imaginary policy on quality. The experiment was performed for three different scenarios in which the probability of treatment was (1) unrelated to pre-intervention performance; (2) positively correlated with pre-intervention levels of performance; and (3) positively correlated with pre-intervention trends in performance. We estimated alternative DID models that varied with respect to the choice of data intervals, the comparison group, and the method of obtaining inference. We assessed estimator bias as the mean absolute deviation between estimated program effects and their true value. We evaluated the accuracy of inferences through statistical power and rates of false rejection of the null hypothesis.

Principal Findings. Performance of alternative specifications varied dramatically when the probability of treatment was correlated with pre-intervention levels or trends. In these cases, propensity score matching resulted in much more accurate point estimates. The use of permutation tests resulted in lower false rejection rates for the highly biased estimators, but the use of clustered standard errors resulted in slightly lower false rejection rates for the matching estimators.

Conclusions. When treatment and comparison groups differed on pre-intervention levels or trends, our results supported specifications for DID models that include matching for more accurate point estimates and models using clustered standard errors or permutation tests for better inference. Based on our findings, we propose a checklist for DID analysis.

Key Words. Hospitals, econometrics, health economics, quality of care, health policy

Health care delivery in the United States is changing at a dramatic pace. Millions of uninsured citizens, unsustainable cost growth, and uneven quality of care have prompted numerous policy responses at the state and national level.

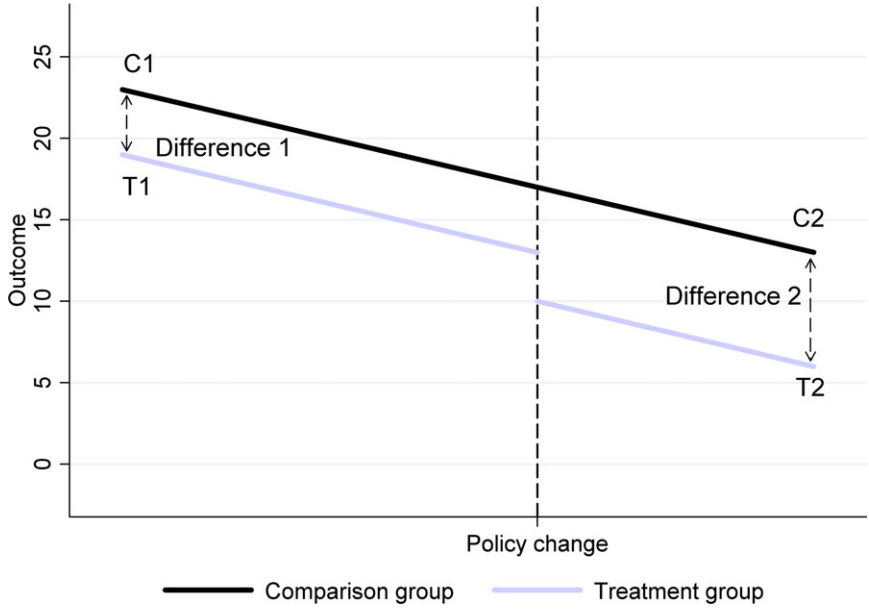
The 2010 Patient Protection and Affordable Care Act includes landmark expansions to health insurance and the introduction of delivery system reforms to reduce costs and improve quality. Evaluating the effectiveness of these efforts is essential in understanding what works—and what doesn't—so policy can be redesigned accordingly. Randomized control trials, the gold standard for understanding causal relationships, are impractical or impossible in many circumstances, and are rarely used to evaluate public policies or large-scale delivery system interventions. To maximize learning from this rapidly changing environment, observational studies are needed.

There are many challenges, however, in drawing valid conclusions from observational study designs. Difference-in-differences methods have emerged as a powerful tool to address confounding in observational studies and evaluate the impact of health care policies. Although the implementation of difference-in-differences models can be complex, the idea is simple. First, identify an intervention, an outcome of interest, and two groups—one that was exposed to the intervention (the treatment group) and one that was not (the comparison group). Then, take the difference in the outcome between the treatment and comparison groups at baseline (difference 1) and again after the intervention (difference 2) has occurred. The policy effect is estimated as the difference-in-differences (difference 2—difference 1).

We can conclude that the intervention had an impact if the outcome changed more for the treatment group than the comparison group. If the differences are the same between the two groups, then there was no effect of the intervention (see Figure 1). For example, a recent study used difference-in-differences methods to evaluate the impact of Medicare's bariatric surgery coverage decision, which limited Medicare reimbursement to "Centers of Excellence." The study compared rates of surgical complications in Medicare patients (whose care was subject to the policy) and commercially insured patients (whose care was not subject to the intervention) before and after the policy was initiated in 2006. Prior studies, using a simple "pre-post" study design found a beneficial impact of the Centers of Excellence program, but had failed to account for secular trends toward improved outcomes (Nguyen et al. 2010; Flum et al. 2011). Using difference-in-differences, we found simi-

Address correspondence to Andrew Ryan, Ph.D., University of Michigan School of Public Health, 1415 Washington Heights, SPH II RM. M3124, Ann Arbor, MI 48109; e-mail: amryan@umich.edu. James F. Burgess Jr., Ph.D., is with the Veterans Affairs Boston Health Care System, US Department of Veteran Affairs, Boston University School of Public Health, Boston, MA. Justin B. Dimick, M.D., M.P.H., is with the Department of Surgery, School of Medicine University of Michigan, Center for Healthcare Outcomes and Policy, Ann Arbor, MI.

Figure 1: Conceptual Depiction of Difference-in-Differences Analysis

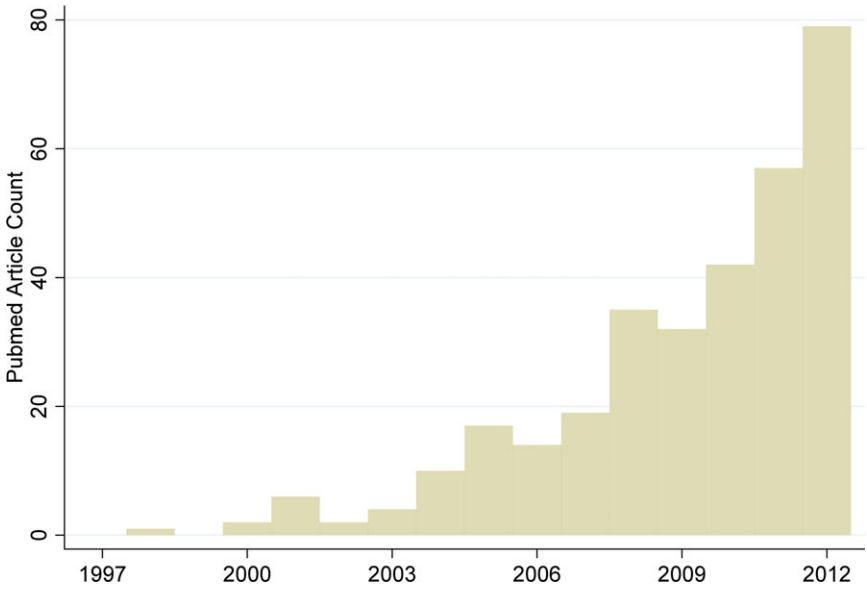


Note: Difference-in-differences estimate is equal to: (T2-C2)—(T1-C1).

lar declines in complications for both the Medicare and commercially insured patients, suggesting that the program did not lead to a reduction in complications (Dimick et al. 2013).

The increasing popularity of difference-in-differences in health policy and medicine (Figure 2) is a result of both the credibility of these designs and their ease of implementation and estimation (Angrist and Pischke 2010). Recent difference-in-differences studies have evaluated the impact of health insurance expansions (Cunningham, Hadley, and Reschovsky 2002; Zhu et al. 2010; Long and Stockley 2011; Blum et al. 2012; Cantor et al. 2012; Graves and Gruber 2012; Dhingra et al. 2013; Joynt et al. 2013), high deductible health plans (Wharam et al. 2012; Kozhimannil et al. 2013), changes in payment policy and providers' financial incentives (Shen and Zuckerman 2005; Dusheiko et al. 2006; Mitchell 2008; Scanlon et al. 2008; Song et al. 2011; Jha et al. 2012; Sutton et al. 2012; Kantarevic and Kralj 2013; Werner, Konetzka, and Polsky 2013), malpractice reform (Kessler, Sage, and Becker 2005), behavioral health parity laws (Goldman et al. 2006; Azrin et al. 2007), resident work hour reform (Volpp et al. 2007), public quality reporting (Dra-

Figure 2: Health Policy and Medicine Articles Using Difference-in-Differences Analysis, 1997–2012



Source: PubMed database. Search term "difference-in-differences" or "difference-in-difference."

nove et al. 2003; Chen and Meinecke 2012; Joynt et al. 2012; Ryan, Nallamothu, and Dimick 2012b), changes in clinical practice (Baxter, Ray, and Fireman 2010; Zivin et al. 2010; Leonhardt et al. 2011; Suehs et al. 2014), smoking laws (Anger, Kvasnicka, and Siedler 2011; Nguyen 2013), electronic medical record implementation (Jones et al. 2010; McCullough, Christianson, and Leerapan 2013), and conflict of interest policies (Epstein et al. 2013).

Assumptions and Estimation for Difference-In-Differences Analysis

Point estimates of policy effects using difference-in-differences can be generated by simply calculating the difference in means for a given outcome between treatment and comparison groups, before and after the intervention was initiated. However, regression models make it possible to test whether difference-in-differences estimates are statistically significant. Regression models

also allow researchers to develop more advanced specifications for difference-in-differences analysis, which may improve the accuracy of point estimates and statistical inference.

Because difference-in-difference analysis is implemented using regression analysis, these methods are subject to standard statistical assumptions (e.g., Gauss-Markov assumptions for linear regression) (Wooldridge 2009). The additional key assumptions for difference-in-difference analysis are the "common shocks" and "parallel trends" assumptions (Angrist and Pischke 2008). The common shocks assumption holds that other phenomena occurring at the same time or after the start of treatment will equally affect the treatment and comparison groups. The parallel trends assumptions says that, although treatment and comparison groups may have different levels of the outcome prior to the start of treatment, their trends in pretreatment outcomes should be the same (Figure 1). The parallel trends assumption implies that, absent treatment, outcomes for the treatment and comparison groups are expected to change at the same rate. Thus, any difference in the differences in outcomes between groups can be attributed to the policy, rather than to differential pre-existing trends in outcomes. With multiple pre-intervention periods, the parallel trends assumption is often examined by statistically testing whether linear pre-intervention trends are statistically different between the treatment and comparison groups (Ryan 2009).

As an example, consider an evaluation of a policy that provided technical assistance to improve mortality rates among U.S. hospitals for which 30 percent or more of their discharges were from Medicaid. No hospitals were eligible to receive the intervention in the pre-intervention period, some hospitals were eligible to receive treatment in the postintervention, and data exist for all hospitals in both periods. There are two general cases in which difference-in-differences analysis could be applied to this evaluation question. First is the "group-level" difference-in-differences specification in which data exist at the level at which treatment occurs (e.g., hospital-level mortality rates). For hospital j at time t, we would estimate:

$$Y_{jt} = b_0 + b_1 \text{post}_t + \delta(\text{treatment}_j \cdot \text{post}_t) + u_j + e_{jt} \quad (1)$$

where *treatment* is a dummy variable indicating whether a hospital received the treatment in period 2, *post* is a dummy variable for whether an observation occurred in the postintervention period, u is a vector of hospital fixed effects, and e is the idiosyncratic error term. Because the equation includes hospital fixed effects, and because the treatment is time invariant, we do not include a

main effect for *treatment*. Estimating this equation with linear regression yields the following interpretation for $\hat{\delta}$:

$$\hat{\delta} = (\overline{\text{Mortality}}_{treat,post} - \overline{\text{Mortality}}_{treat,pre}) - (\overline{\text{Mortality}}_{comp,post} - \overline{\text{Mortality}}_{comp,pre}) \quad (2)$$

where "treat" and "comp" denote the treatment and comparison groups. As shown by equation (2), $\hat{\delta}$ provides the difference-in-differences estimate of the policy effect. DID estimates are typically considered to be average treatment effects on the treated, rather than average treatment effects. This is because DID estimates are generally thought of as applying to a particular group that was treated (rather than to a population that could have been treated).

Second is the "micro-level" difference-in-differences in which data exist at a lower level nested within the treatment unit (e.g., patient-level mortality within hospitals). For patient i in hospital j at time t, we estimate:

$$Y_{ijt} = b_0 + b_1 \text{post}_t + b_2 X_{ijt} + \delta(\text{treatment}_j \cdot \text{post}_t) + u_j + e_{ijt} \quad (3)$$

where X_{ijt} is a set of patient-level severity adjusters and all other terms are defined as before. The interpretation of $\hat{\delta}$ is the same as in equation (2). Note that, by testing for changes in mortality within hospitals over time, this specification accounts for any compositional differences resulting from where patients receive care over time (i.e., lower mortality hospitals receiving a greater share of patients in the post period). A variation on the micro-level specification is to replace hospital fixed effects with a set of hospital characteristics and include a main effect for *treatment*.

The group-level and micro-level DID models give rise to somewhat different issues. In group-level DID models, variation over time in the composition of groups can confound estimates of the program effect. This is why the group-level outcomes are often adjusted prior to estimation (e.g., risk-adjusted mortality). One advantage of the micro-level specification is the potential to control for individual heterogeneity with the vector X. This may reduce the variance estimate of $\hat{\delta}$ and can also account for time-varying differences in patient severity across the treatment and comparison groups. Also, while each hospital receives equal weight in the group-level specification (equation 1), hospitals in which more patients are treated would receive more weight for the micro-level specification (equation 2). Both the microlevel and group-level DID specifications are subject to autocorrelation and clustering of errors, creating challenges for statistical inference (Bertrand, Duflo, and Mullainathan 2002).

Specification Choices for Difference-in-Differences Analysis: A Simulation Study

The theoretical underpinnings of DID are well understood (Wooldridge 2002; Angrist and Pischke 2008). What is less known is how well the assumptions of DID hold up in empirical practice and what can be done for more robust estimation and inference when the assumptions don't hold. In DID models, point estimates should be correct and variance estimates should appropriately reflect the uncertainty in parameter estimates due to sampling variation. Most research related to specification in difference-in-differences has focused on the issues related to the variance estimates of program effects (Bertrand, Duflo, and Mullainathan 2002; Donald and Lang 2007). However, specification choices related to obtaining accurate point estimates in difference-in-differences models—including the choice of comparison groups, the choice of the pre-intervention time interval, and addressing violations to the "parallel trends" assumption—have received scant attention in the literature. Researchers do not have good guides to implement difference-in-differences analyses.

We address specification questions related to difference-in-differences analysis by conducting a Monte-Carlo simulation experiment. We use hospital-level data on quality of care from Hospital Compare between 2004 and 2009. For each hospital in each year, we measure quality of care using a composite measure of process-of-care quality from 37 individual measures. The composite is created by using the "opportunities model," which is calculated as the sum of successfully achieved measures divided by the sum of opportunities that practices have to achieve these measures (Landrum, Bronskill, and Normand 2000). This quality measure is expressed as a percentage. We exclude hospitals without quality data in each year. We also exclude hospitals that participated in the Premier Hospital Quality Incentive Demonstration, as these hospitals experienced greater quality improvement during part of the study period (Ryan, Blustein, and Casalino 2012a). Our final data file includes 3,192 hospitals.

We use these data to estimate the effect of an imaginary policy, initiated in 2007 and continuing through 2009. To do this, we randomly assign hospitals to treatment and comparison groups in 2007. We then assume different effects of the imaginary intervention, ranging from no effect, a "small" effect (+0.2 standard deviations, or +2.3 percentage points), or a "medium" effect (+0.5 standard deviations, or +5.8 percentage points) (Cohen 1988). We add

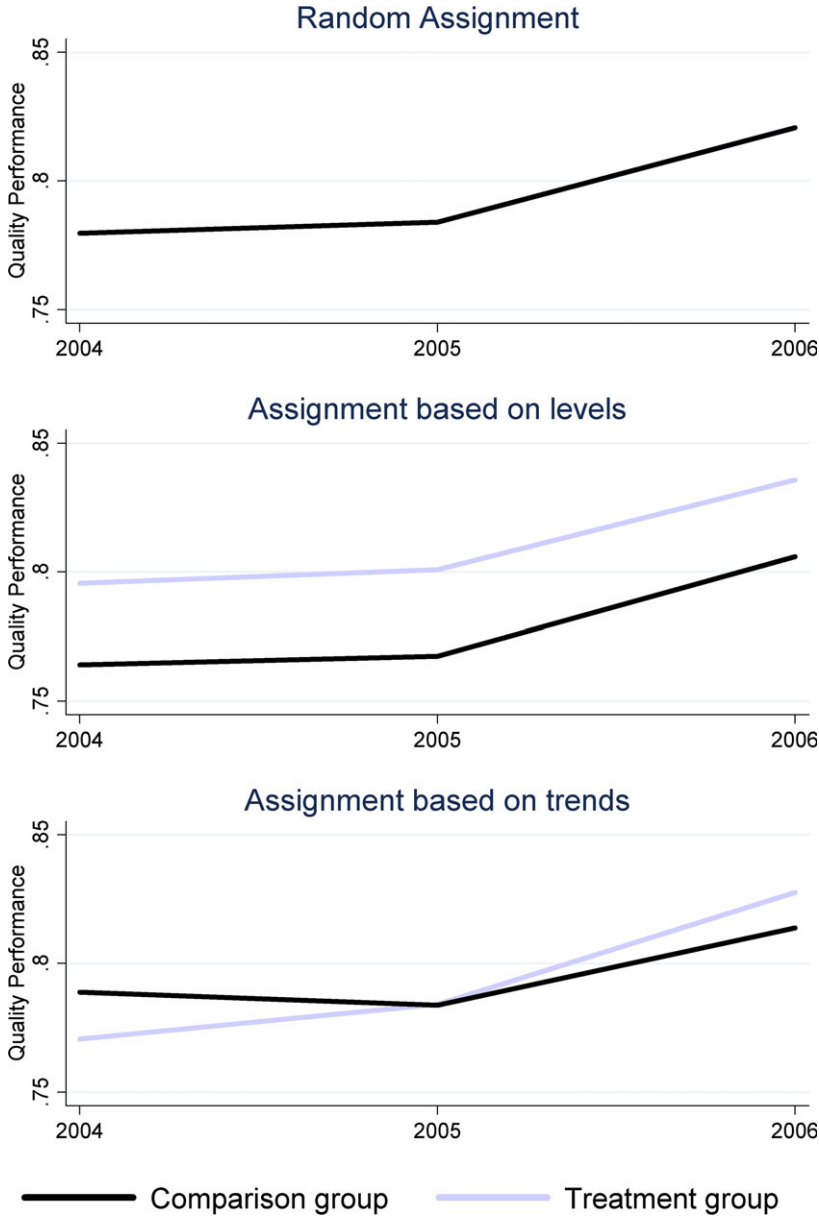
these assumed effects to the actual quality scores for those hospitals in the treatment group after the intervention has begun.[1]

We randomly assign hospitals to treatment using three scenarios (Figure 3). In the first scenario, each hospital has an equal probability of being assigned to the treatment or comparison group. Assignment is completely random and unrelated to performance in the pre-intervention period. In the second scenario, pre-intervention performance is associated positively with the probability of assignment to the program: hospitals in the bottom quartile of pre-intervention performance have a 35 percent probability of assignment to treatment increasing by 10 percentage points for each quartile up to 65 percent for hospitals in the highest quartile of pre-intervention performance. Recent research highlights that higher pre-intervention quality performance can affect the expectation for change in quality (Ryan and Blustein 2011; Ryan, Blustein, and Casalino 2012a). The third scenario is similar to the second scenario, except that the probability of assignment to treatment is associated positively with pre-intervention trends in quality, rather than levels. Hospitals in the bottom quartile (those with the weakest trends toward improved performance) have a 35 percent probability of assignment to treatment increasing by 10 percentage points for each quartile up to 65 percent for hospitals in the highest quartile (strongest trends toward improved performance).

We then estimate the impact of the imaginary policy using group-level difference-in-differences (equation 1). We perform this analysis using alternative specifications. First, we vary the data used in the analysis: (1) using hospital-level data from all observations (multiple pre and post-intervention periods); (2) collapsing data into two periods by averaging performance within a single pre and a single post-intervention period; and (3) using data from only the last pre-intervention and the first post-intervention observation.

Second, we estimate the effect of the intervention using alternative comparison groups: (1) using all non-treated hospitals as comparison hospitals and (2) using propensity score matched comparison hospitals. Differences in levels or trends prior to the start of an intervention between treatment and comparison groups may result in different expectations for changes in outcomes, and matching can alleviate this concern. Propensity score matching is performed using one-to-one matching (with replacement), calipers of 1.0 (0.09 of the pooled standard deviation), and enforcing common support. Common support excludes observations from hospitals with propensity score values that are above the maximum value or below the minimum value of that of the comparison group's propensity score distribution. Lagged levels of quality for each year prior to the start of the intervention were used for matching. The

Figure 3: Pre-Intervention Levels and Trends under Different Assignment Scenarios



Note: Data are based on average levels across 200 simulation iterations for each scenario.

matching procedure was implemented in Stata using a user-written command (Leuven and Sianesi 2003).

Third, we use alternative approaches to address differences in pre-existing trends between treatment and comparison groups (Angrist and Pischke 2008; Besley and Burgess 2004): (1) do not model trends; and (2) model differential trends using treatment and comparison group-specific pre-intervention dummies. The specification that modeled trends in outcomes took the following form:

$$Y_{jt} = b_0 + b_1 \text{year pre}_t + b_2(\text{treatment}_j \cdot \text{year pre}_t) + b_3 \text{post}_t + \delta(\text{treatment}_j \cdot \text{post}_t) + u_j + e_{jt} \quad (4)$$

where *year pre* is a vector of dummy variables for each pre-intervention period.

Three alternative approaches are used to test for statistical significance: (1) assume that errors are identically and independently distributed (i.i.d.); (2) use clustered standard errors (Rogers 1993) to account for heteroskedasticity at the hospital-level; and (3) perform permutation tests. Permutation tests are nonparametric methods recommended for exact inference in situations in which assumptions underlying other variance estimators may be violated (Ernst 2004). These tests reassign groups to treatment and control conditions and then recalculate program effects under the different permuted conditions. These alternative program effects form a "randomization distribution," which is then evaluated to obtain inference. Permutation tests have been recommended for difference-in-differences analysis (Bertrand, Duflo, and Mullainathan 2002; Abadie, Diamond, and Hainmueller 2010). In this study, permutation tests are performed using 49 random permutations of the data.

We combine these specification features, ultimately choosing six alternative difference-in-differences specifications to compare (Table 1). For each specification, our analysis yields point estimates of the magnitude of the policy effect as well as tests of the statistical significance ($\alpha \leq .05$) of the policy using the three alternative strategies for inference. Our simulation study performs this analysis for 200 simulation iterations using each of the three assignment scenarios. For each simulation iteration, we capture the rate of false rejection (i.e., type II error), the rate of rejection from "small" and "medium" program effects, and the mean absolute deviation (Pham-Gia and Hung 2001) between estimated program effects and their true value, a measure of estimator bias. We report the mean absolute deviation because it provides a clear interpretation of the magnitude of the estimator bias in units of the dependent variable.

Table 1: Simulation Results from Program Assignment Process 1: Pre-Intervention Performance is Unrelated to Program Assignment

Specification Description					*Specification Performance*			
Specification	*Data Used*	*Comparison Group*	*Model pre-Intervention Trends?**	*Inference†*	*Rejection Rate, No Effect*	*Rejection Rate, Small Effect‡*	*Rejection Rate, Medium Effect§*	*Mean Absolute Deviation*
1	2004–2009	All non-treated	N	i.i.d. se	17.0%	100%	100%	.0020
				cluster se	4.0%	100%	100%	
				permutation	0%	94.0%	100%	
2	2004–2009	All non-treated	Y	i.i.d. se	12.5%	100%	100%	.0025
				cluster se	2.5%	100%	100%	
				permutation	0%	76.0%	100%	
3	2004–2009	Matched	N	i.i.d. se	29.0%	100%	100%	.0023
				cluster se	3.0%	100%	100%	
				permutation	12.0%	100%	100%	
4	2004–2009	Matched	Y	i.i.d. se	18.0%	100%	100%	.0028
				cluster se	2.0%	100%	100%	
				permutation	9.5%	100%	100%	
5	2006–2007	All non-treated	N	i.i.d. se	5.5%	100%	100%	.0017
				cluster se	5.5%	100%	100%	
				permutation	0%	100%	100%	
6	2004/6–2007/9	All non-treated	N	i.i.d. se	3.5%	100%	100%	.0025
				cluster se	3.5%	100%	100%	
				permutation	0%	98.0%	100%	

Note. Composite process performance is the outcome. No risk adjustment was performed. The hospital-year is the unit of analysis (3,192 hospitals).
*Differential trends were modeled using group-specific pre-intervention dummy variables for the treatment and comparison groups.
†"i.i.d. se": statistical significance was determined using a two-sided test ($p < .05$) with standard errors that assumed independently and identically distributed errors; "cluster se": statistical significance was determined using a two-sided test ($p < .05$) using standard errors that were robust to hospital-level clustering; "permutation": statistical significance was determined using two-sided permutation tests ($p < .05$).
‡A small effect is defined as +0.2 standard deviations of composite quality, or + 2.3 percentage points.
§A medium effect is defined as +0.5 standard deviations, or 5.8 percentage points.

Our procedure is similar to that used by Bertrand and colleagues to evaluate the properties of variance estimates using alternative difference-in-difference specifications (Bertrand, Duflo, and Mullainathan 2002).

All analysis was performed using *Stata 12.0* (Stata Corp., College Station, TX, USA). The program used to conduct the simulation study can be found in Appendix A.

Results from the Simulation Study

Table 1 shows the results from the simulation experiment from Scenario 1, in which hospital assignment to treatment is completely random. To the left of the table, the specification description shows the data that were used, the choice of comparison group, how trends were modeled, and the method for determining statistical inference. The right of the table shows rate of rejection when there is no effect (the rate of false rejection), a small effect, and a medium effect, as well as the mean absolute deviation between the estimated point estimate and the true program effect.

For Scenario 1, alternative specifications have a relatively minor impact on rejection rates and the estimator bias. Estimator bias is similar across the specifications using different comparison groups and those modeling and not modeling pre-intervention trends in performance. Statistical power to detect small effects was similar across the specifications, although somewhat lower when using permutation tests. The one meaningful difference across specifications was that specifications that assume i.i.d. errors tend to have higher rates of false rejection than specifications using clustered standard errors and those using permutation tests for inference. These results highlight that when treatment is unrelated to the pre-intervention levels or trends, specification in DID does not substantively affect model inference.

Table 2 shows the results from Scenario 2, in which hospitals with higher levels of quality performance prior to the intervention were more likely to be assigned to treatment. A few key patterns emerge about the performance of alternative specifications. First, the two specifications that use a matched comparison group have much lower estimator bias: the mean absolute deviation values are .0025 and .0019 for these specifications that model and do not model trends, respectively. The bias of the other estimators is between 4 and 7 times higher than the matching estimators. Second, specifications using permutation tests have the lowest rate of false rejection for four of the six specifications. The differences in false rejection rates are more pronounced for specifications that have greater bias. However, for the matching estimators,

Table 2: Simulation Results from Program Assignment Process 2: Pre-Intervention Performance is Positively Related to Program Assignment

Specification Description					*Specification Performance*			
Specification	*Data Used*	*Comparison Group*	*Model Pre-Intervention Trends?**	*Inference†*	*Rejection Rate, No Effect*	*Rejection Rate, Small Effect‡*	*Rejection Rate, Medium Effect§*	*Mean Absolute Deviation*
1	2004–2009	All non-treated	N	i.i.d. se	100%	96%	100%	.0131
				cluster se	100%	91%	100%	
				permutation	6.0%	0%	100%	
2	2004–2009	All non-treated	Y	i.i.d. se	100%	81.5%	100%	.0131
				cluster se	98.0%	65.0%	100%	
				permutation	7.0%	0%	100%	
3	2004–2009	Matched	N	i.i.d. se	18.5%	100%	100%	.0019
				cluster se	1.5%	100%	100%	
				permutation	9.0%	100%	100%	
4	2004–2009	Matched	Y	i.i.d. se	16.5%	100%	100%	.0025
				cluster se	1.0%	100%	100%	
				permutation	5.5%	100%	100%	
5	2006–2007	All non-treated	N	i.i.d. se	95.5%	100%	100%	.0073
				cluster se	95.5%	100%	100%	
				permutation	5.0%	97.5%	100%	
6	2004/6–2007/9	All non-treated	N	i.i.d. se	81.0%	94.0%	100%	.0092
				cluster se	81.0%	94.0%	100%	
				permutation	5.5%	28.0%	100%	

Note. Composite process performance is the outcome. No risk adjustment was performed. The hospital-year is the unit of analysis (3,192 hospitals).
*Differential trends were modeled using group-specific pre-intervention dummy variables for the treatment and comparison groups.
†"i.i.d. se": statistical significance was determined using a two-sided test ($p < .05$) with standard errors that assumed independently and identically distributed errors; "cluster se": statistical significance was determined using a two-sided test ($p < .05$) using standard errors that were robust to hospital-level clustering; "permutation": statistical significance was determined using two-sided permutation tests ($p < .05$).
‡A small effect is defined as +0.2 standard deviations of composite quality, or + 2.3 percentage points.
§A medium effect is defined as +0.5 standard deviations, or 5.8 percentage points.

the specifications using clustered standard errors had somewhat lower false rejection rates (e.g., 1.5 percent vs. 9.0 percent for the specification that did not model trends). Statistical power to detect small effects was much lower when using permutation tests with the highly biased estimators, but identical when using the matching estimators.

Table 3 shows the results from Scenario 3, in which hospitals with stronger trends toward quality improvement prior to the intervention were more likely to be assigned to treatment. Similar to Scenario 2, it shows that the specifications using the matched comparison group have lower estimator bias: the mean absolute deviation values are .0023 for these specifications, which is 2–10 times lower than the other specifications. As in Scenario 2, permutation tests tend to have lower rates of false rejection in the specifications with more bias, while inference based on clustered standard errors is somewhat better for the matching estimators. Statistical power is nearly identical across the estimators.

Overall, our simulation study found that, when treatment was randomly assigned, the bias of alternative DID specifications was approximately equivalent. However, when the probability of treatment was correlated with pre-intervention levels or trends, propensity score matching resulted in much more accurate point estimates. The use of permutation tests resulted in lower false rejection rates for the highly biased estimators, but the use of clustered standard errors resulted in slightly lower false rejection rates for the matching estimators. In all specifications, standard errors that assumed that errors were identically and independently distributed resulted in elevated rates of false rejection. Despite substantial differences in the data generating process, our study found that specifications that included clustered standard errors or used permutation tests for statistical inference resulted in similar rates of false rejection that have been documented elsewhere (Bertrand, Duflo, and Mullainathan 2002).

Limitations of the Simulation Study

The inferences about specification in difference-in-differences models from our simulation study may not generalize to estimation in other datasets. For instance, differences in pre-intervention levels and trends between treatment and comparison groups may not have a similar impact on inference in other settings. Quality performance data are top-coded at 100 percent in Hospital Compare, and prior research shows strong evidence of nonlinear rates of improvement as hospitals approach the maximum score. These features of the

Table 3: Simulation Results from Program Assignment Process 3: Pre-Intervention Trends Are Positively Related to Program Assignment

Specification Description					*Specification Performance*			
Specification	*Data Used*	*Comparison Group*	*Model pre-Intervention Trends?*‡	*Inference*§	*Rejection Rate, No Effect*	*Rejection Rate, Small Effect**	*Rejection Rate, Medium Effect*†	*Mean Absolute Deviation*
1	2004–2009	All non-treated	N	i.i.d. se	96.0%	100%	100%	.0071
				cluster se	91.5%	100%	100%	
				permutation	0%	100%	100%	
2	2004–2009	All non-treated	Y	i.i.d. se	100%	100%	100%	.0240
				cluster se	100%	100%	100%	
				permutation	98.5%	100%	100%	
3	2004–2009	Matched	N	i.i.d. se	26.0%	100%	100%	.0023
				cluster se	4.0%	100%	100%	
				permutation	12.5%	100%	100%	
4	2004–2009	Matched	Y	i.i.d. se	12.0%	100%	100%	.0023
				cluster se	1.0%	100%	100%	
				permutation	3.5%	100%	100%	
5	2006–2007	All non-treated	N	i.i.d. se	70.0%	100%	100%	.0050
				cluster se	70.0%	100%	100%	
				permutation	1.5%	99.5%	100%	
6	2004/6–2007/9	All non-treated	N	i.i.d. se	100%	100%	100%	.0269
				cluster se	100%	100%	100%	
				permutation	100%	100%	100%	

*A small effect is defined as +0.2 standard deviations of composite quality, or + 2.3 percentage points.
†A medium effect is defined as +0.5 standard deviations, or 5.8 percentage points.
‡Differential trends were modeled using group-specific pre-intervention dummy variables for the treatment and comparison groups.
§i.i.d se: statistical significance was determined using a two-sided test ($p < .05$) with standard errors that assumed independently and identically distributed errors; "cluster se": statistical significance was determined using a two-sided test ($p < .05$) using standard errors that were robust to hospital-level clustering; "permutation": statistical significance was determined using two-sided permutation tests ($p < .05$).

data generating process make inference strongly susceptible to bias from differences in pre-intervention levels and trends between treatment and comparison groups. However, difference-in-differences studies in health care frequently evaluate limited dependent variables (e.g., mortality rates) that likely display similar patterns.

Our study is also limited by our choice of alternative program effects and specification features to compare. Other econometric methods, such as the synthetic control estimator (Abadie, Diamond, and Hainmueller 2010), the nonlinear "change in changes" model (Athey and Imbens 2006), and other statistical matching estimators (Diamond and Sekhon 2013), may be superior to the estimators evaluated in this study. We also did not consider the properties of microlevel difference-in-differences estimators. While the greater number of observations in microlevel can increase statistical power, these models also have especially high rates of false rejection (Bertrand, Duflo, and Mullainathan 2002). It is therefore possible that permutation tests may provide more reliable inference in these cases (Ernst 2004). Future research should consider the performance of difference-in-differences estimators using a broader range of specification features, outcomes, heterogeneous program effects, and different datasets.

THE DIFFERENCES-IN-DIFFERENCES CHECKLIST

Based on the results of our analysis, we propose a difference-in-differences checklist, identifying the critical conditions that must be met to be able to make valid inferences using this analytic approach (Table 4). Checklists are being developed to promulgate high quality methods (e.g., the PRISMA checklist for systematic reviews [Moher et al. 2009] and the GRACE checklist for comparative effectiveness research [Dreyer et al. 2014]). In Appendix A, we provide sample code to conduct the specification tests described in the checklist.

Implementing a difference-in-differences design requires longitudinal data on groups exposed and not exposed to an intervention. Therefore, data must exist on study outcomes for groups exposed and not exposed to an intervention, both before and after the intervention was implemented (Element 1). The key strength of the difference-in-differences design is that the comparison group serves as the "counterfactual" for the treatment group; that is, the comparison group gives an estimate for the postintervention outcome of the treatment group had they not received the intervention. The ability of the

Table 4: Elements of Difference-in-Differences Checklist

Confirm That	*How to Test*	*What To Do If Violated*
1. Data exist on study outcomes for at least one observation period among groups exposed and not exposed to an intervention, both before and after the intervention was implemented	Directly observable	NA
2. Trends in outcome performance prior to an intervention are "parallel" between treatment and comparison groups	Test equivalence of linear trends between treatment and comparison groups prior to intervention by testing the significance of the interaction term between the time trend and the treatment group	If multiple comparison groups are available, match treatment and comparison units
3. Baseline outcome levels are unrelated to expectations for changes in outcomes	For both treatment and comparison groups, test whether baseline outcome is correlated with change in performance across the study period	If multiple comparison groups are available, match treatment and comparison units
4. Violations to standard statistical assumptions are appropriately addressed	Test for violations of homoscedasticity of standard errors.(Breusch and Pagan 1979; Drukker 2003)	Permutation tests or clustered standard errors will likely result in the most accurate statistical inference when using difference-in-difference analysis
5. Events or factors other than treatment, occurring at the time of treatment, do not differentially affect outcomes for treatment and comparison groups	Not directly testable	NA
6. The composition of treatment and comparison groups does not change over the course of the study	Test for difference in observed covariates between treatment and comparison rates before and after the intervention. Test for differential drop-out rates between treatment and comparison groups. (Hausman and Wise 1979)	Control for differences in observed covariates between treatment and comparison rates before and after the intervention

Continued

Table 4. *Continued*

Confirm That	*How to Test*	*What To Do If Violated*
7. Treatment does not "spill-over" from treatment group to comparison group	Test whether comparison group experiences deviation from existing trend concurrent with intervention	If multiple comparison groups are available, choose alternative comparison group that is not subject to spillovers

comparison group to provide this counterfactual requires that trends in outcome performance prior to an intervention are "parallel" between treatment and comparison groups (Element 2). If the treatment and comparison groups have parallel trends prior to the intervention, then there is a reasonable expectation that, absent the intervention, outcomes in the treatment group would change at a similar rate to outcomes in the comparison group (Figure 1). In our simulation study, when hospitals with stronger trends toward improvement were more likely to receive treatment (Scenario 3), point estimates using standard difference-in-differences estimation (equation 1) had greater bias. Matching treatment and comparison units on pre-intervention levels of performance greatly reduced this bias.

While baseline levels of the outcome do not need to be the same for treatment and comparison groups, it is crucial that baseline outcome levels are unrelated to expectations for changes in outcomes (Element 3). This element remains valid even if there are parallel trends between treatment and comparison groups in the pre-intervention period (Element 2). For instance, because many of the hospitals participating in Medicare's hospital pay-for-performance demonstration began with higher performance than comparison hospitals and achieved nearly perfect performance on the incentivized clinical process performance measures, it was challenging for researchers to identify the longer term effects of the program given these ceiling effects (Ryan, Blustein, and Casalino 2012a). Relatedly, if the treatment group is selected on the basis of previous performance (high or low) on the outcome, there is potential bias from regression to the mean (Chay, McEwan, and Urquiola 2005). We found that when hospitals with higher levels of pre-intervention performance were more likely to receive treatment (Scenario 2), point estimates using standard difference-in-differences estimation (equation 1) also had greater bias. Again, matching treatment and comparison group on pre-intervention levels greatly reduced bias.

Because difference-in-differences designs require researchers to perform longitudinal data analysis using regression techniques and it is critical that violations to standard statistical assumptions are appropriately addressed (Element 4). This includes appropriately estimating treatment effects when using nonlinear models (e.g., logistic regression) (Ai and Norton 2003) and accounting for clustering and other violations of independence when estimating standard errors (Bertrand, Duflo, and Mullainathan 2002). Our simulation study strongly supports the use of clustered standard errors and permutation tests for statistical inference in difference-in-difference analysis.

Several other elements of the checklist are not informed by our simulation study, but instead by conceptual issues related to causal inference in the context of different-in-differences studies. For instance, because difference-in-differences methods seek to identify the impact of a specific policy or intervention, events or factors other than treatment, occurring at the time of treatment, should not differentially affect outcomes for treatment and comparison groups (Element 5). Although this "common shocks" assumption is generally not testable, researchers should attempt to identify factors other than the treatment being studied that may have affected outcomes for either the treatment of comparison groups. For instance, if groups self-selected to receive treatment, it is possible that outcomes for this group would have improved even without treatment. This "expected gains bias" is a result of changes in unobserved factors over the study period. Selection into the study at the start of the program could signal that the treatment group had an increased unobserved interest in improving the study outcome that is concomitant with the start of the program. Hence, it may be the changing unobservable, rather than the program itself, that led to any effect.

Relatedly, in cases in which researchers seek to estimate the effects of programs for which treatment does not occur at the same time for all treated units, specifications should be modified to avoid confounding secular time-trends with staggered implementation. For instance, for a case with three post-intervention periods, a separate treatment effect can be estimated for each postintervention period. Here, the "treatment" group remains in the "comparison" group until the time in which they receive treatment. A linear combination of estimated effects can be used to generate an overall program effect. A variation on this specification can account for continuously "rolling" treatment commencement dates (see Ryan et al. 2013).

Difference-in-differences designs also rest on the assumption that all unobserved factors affecting outcomes between the treatment and comparison groups do not change over time. Therefore, the composition of treatment and

comparison groups should not change over the course of the study (Element 6). If these groups do change, potentially due to differential drop out between treatment and comparison groups, study inference could be biased. Researchers should statistically adjust for differences in characteristics of nested observations (e.g., patients) between treatment and comparison groups to mitigate the effects of compositional changes.

For the difference-in-differences design to be valid, the comparison group cannot be affected by the intervention. In other words, treatment does not "spill-over" from treatment group to comparison group (Element 7). For example, consider a study evaluating the impact of public quality reporting in Hospital Compare on 30-day mortality rates in Medicare beneficiaries. The treatment group consists of Medicare patients receiving care for publicly reported diagnoses in US hospitals, and the comparison group consists of Medicare patients treated for conditions that were not publicly reported (Ryan, Nallamothu, and Dimick 2012b). If public quality reporting resulted in broad changes in hospital practice that improved mortality for diagnoses and conditions that were both publicly reported and not publicly reported, then the use of this comparison group would result in bias (toward the null hypothesis).

CONCLUSION

Difference-in-differences methods have emerged as some of the most popular and rigorous methods to estimate the impact of medical and health policy interventions. However, limited attention has been given to how specification choices in difference-in-differences models affect inference. Results from our simulation experiment suggest that specification choices can have a major impact on the point estimates and statistical significance of estimated policy effects. A difference-in-differences checklist, based in part on our simulation study, provides a guide to practice for empirical researchers wishing to use these methods.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors would like to acknowledge Jayme Mendelsohn for research assistance. Funding for Dr. Ryan was provided by a career development award from the Agency for

Healthcare Research and Quality (1 K01 HS018546). Dr. Dimick was supported by a grant from the National Institute on Aging (R01AG039434). Dr. Dimick is a paid consultant and equity owner in ArborMetrix Inc., a company that provides software and analytic services for assessing hospital quality and efficiency.

Disclosures: None.

Disclaimers: None.

NOTE

1. Performance on the quality measures is top coded at 100, indicating perfect compliance. In our study, by adding program effects to hospitals' actual scores, we allow measure performance to exceed 100 for hospitals that are assigned to treatment.

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* 490 (105): 493–505.

Ai, C. R., and E. C. Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters* 80 (1): 123–9.

Anger, S., M. Kvasnicka, and T. Siedler. 2011. "One Last Puff? Public Smoking Bans and Smoking Behavior." *Journal of Health Economics* 30 (3): 591–601.

Angrist, J. D., and J.-S. Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

———. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Athey, S., and G. W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74 (2): 431–97.

Azrin, S. T., H. A. Huskamp, V. Azzone, H. H. Goldman, R. G. Frank, M. A. Burnam, S. L. Normand, M. S. Ridgely, A. S. Young, C. L. Barry, A. B. Busch, and G. Moran. 2007. "Impact of Full Mental Health and Substance Abuse Parity for Children in the Federal Employees Health Benefits Program." *Pediatrics* 119 (2): e452–9.

Baxter, R., G. T. Ray, and B. H. Fireman. 2010. "Effect of Influenza Vaccination on Hospitalizations in Persons Aged 50 Years and Older." *Vaccine* 28 (45): 7267–72.

Bertrand, M., E. Duflo, and S. Mullainathan. 2002. "*How Much Should We Trust Differences-in-Differences Estimates?*" National Bureau of Economic Research Working Paper Series No. 8841.

Besley, T., and R. Burgess. 2004. "Can Labor Regulation Hinder Economic Performance? Evidence from India." *Quarterly Journal of Economics* 119 (1): 91–134.

Blum, A. B., L. C. Kleinman, B. Starfield, and J. S. Ross. 2012. "Impact of State Laws That Extend Eligibility for Parents' Health Insurance Coverage to Young Adults." *Pediatrics* 129 (3): 426–32.

Breusch, T. S., and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (5): 1287.

Cantor, J. C., A. C. Monheit, D. DeLia, and K. Lloyd. 2012. "Early Impact of the Affordable Care Act on Health Insurance Coverage of Young Adults." *Health Services Research* 47 (5): 1773–90.

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

Chen, Y., and J. Meinecke. 2012. "Do Healthcare Report Cards Cause Providers to Select Patients and Raise Quality of Care?" *Health Economics* 21 (Suppl 1): 33–55.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Hillside, NJ: Lawrence Erlbaum Associates.

Cunningham, P. J., J. Hadley, and J. Reschovsky. 2002. "The Effects of SCHIP on Children's Health Insurance Coverage: Early Evidence from the Community Tracking Study." *Medical Care Research and Review: MCRR* 59 (4): 359–83.

Dhingra, S. S., M. M. Zack, T. W. Strine, B. G. Druss, and E. Simoes. 2013. "Change in Health Insurance Coverage in Massachusetts and Other New England States by Perceived Health Status: Potential Impact of Health Reform." *American Journal of Public Health* 103 (6): e107–14.

Diamond, A., and J. S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95 (3): 932–45.

Dimick, J. B., L. H. Nicholas, A. M. Ryan, J. R. Thumma, and J. D. Birkmeyer. 2013. "Bariatric Surgery Complications before vs after Implementation of a National Policy Restricting Coverage to Centers of Excellence." *Journal of the American Medical Association* 309 (8): 792–9.

Donald, S. G., and K. Lang. 2007. "Inference with Difference in Differences and Other Panel Data." *Review of Economics and Statistics* 89 (2): 221–33.

Dranove, D., D. Kessler, M. McClellan, and M. Satterthwaite. 2003. "Is More Information Better? The Effects of "Report Cards" on Health Care Providers." *Journal of Political Economy* 111 (3): 555–88.

Dreyer, N. A., P. Velentgas, K. Westrich, and R. Dubois. 2014. "The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution." *Journal of Managed Care and Specialty Pharmacy* 20 (3): 301–8.

Drukker, D. M. 2003. "Testing for Serial Correlation in Linear Panel-Data Models." *Stata Journal* 3 (2): 168–77.

Dusheiko, M., H. Gravelle, R. Jacobs, and P. Smith. 2006. "The Effect of Financial Incentives on Gatekeeping Doctors: Evidence from a Natural Experiment." *Journal of Health Economics* 25 (3): 449–78.

Epstein, A. J., S. H. Busch, A. B. Busch, D. A. Asch, and C. L. Barry. 2013. "Does Exposure to Conflict of Interest Policies in Psychiatry Residency Affect Antidepressant Prescribing?" *Medical Care* 51 (2): 199–203.

Ernst, M. D. 2004. "Permutation Methods: A Basis for Exact Inference." *Statistical Science* 19 (4): 676–85.

Flum, D. R., S. Kwon, K. MacLeod, B. Wang, R. Alfonso-Cristancho, L. P. Garrison, and S. D. Sullivan. 2011. "Bariatric Obesity Outcome Modeling Collaborative. The Use, Safety and Cost of Bariatric Surgery before and after Medicare's National Coverage Decision." *Annals of Surgery* 254 (6): 860–5.

Goldman, H. H., R. G. Frank, M. A. Burnam, H. A. Huskamp, M. S. Ridgely, S. L. Normand, A. S. Young, C. L. Barry, V. Azzone, A. B. Busch, S. T. Azrin, G. Moran, C. Lichtenstein, and M. Blasinsky. 2006. "Behavioral Health Insurance Parity for Federal Employees." *New England Journal of Medicine* 354 (13): 1378–86.

Graves, J. A., and J. Gruber. 2012. "How Did Health Care Reform in Massachusetts Impact Insurance Premiums?" *American Economic Review* 102 (3): 508–13.

Hausman, J. A., and D. A. Wise. 1979. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica* 47 (2): 455–73.

Jha, A. K., K. E. Joynt, E. J. Orav, and A. M. Epstein. 2012. "The Long-Term Effect of Premier Pay for Performance on Patient Outcomes." *New England Journal of Medicine* 366 (17): 1606–15.

Jones, S. S., J. L. Adams, E. C. Schneider, J. S. Ringel, and E. A. McGlynn. 2010. "Electronic Health Record Adoption and Quality Improvement in US Hospitals." *American Journal of Managed Care* 16 (12 Suppl HIT): SP64–71.

Joynt, K. E., D. M. Blumenthal, E. J. Orav, F. S. Resnic, and A. K. Jha. 2012. "Association of Public Reporting for Percutaneous Coronary Intervention with Utilization and Outcomes among Medicare Beneficiaries with Acute Myocardial Infarction." *Journal of the American Medical Association* 308 (14): 1460–8.

Joynt, K. E., D. Chan, E. J. Orav, and A. K. Jha. 2013. "Insurance Expansion In Massachusetts Did Not Reduce Access among Previously Insured Medicare Patients." *Health Affairs* 32 (3): 571–8.

Kantarevic, J., and B. Kralj. 2013. "Link between Pay for Performance Incentives and Physician Payment Mechanisms: Evidence from the Diabetes Management Incentive in Ontario." *Health Economics* 22 (12): 1417–39.

Kessler, D. P., W. M. Sage, and D. J. Becker. 2005. "Impact of Malpractice Reforms on the Supply of Physician Services." *Journal of the American Medical Association* 293 (21): 2618–25.

Kozhimannil, K. B., M. R. Law, C. Blauer-Peterson, F. Zhang, and J. F. Wharam. 2013. "The Impact of High-deductible Health Plans on Men and Women: An Analysis of Emergency Department Care." *Medical Care* 51 (8): 639–45.

Landrum, M. B., S. E. Bronskill, and S. T. Normand. 2000. "Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers." *Health Services Research Methodology* 1 (1): 23–47.

Leonhardt, K. K., O. Yakusheva, D. Phelan, A. Reeths, T. Hosterman, D. Bonin, and M. Costello. 2011. "Clinical Effectiveness and Cost Benefit of Universal versus

Targeted Methicillin-Resistant Staphylococcus aureus Screening Upon Admission in Hospitals." *Infection Control and Hospital Epidemiology* 32 (8): 797–803.

Leuven, E., and B. Sianesi 2003. "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing" [accessed on November 11, 2014]. Available at http://ideas.repec.org/c/boc/bocode/s432001.html

Long, S. K., and K. Stockley. 2011. "The Impacts of State Health Reform Initiatives on Adults in New York and Massachusetts." *Health Services Research* 46 (1 Pt 2): 365–87.

McCullough, J. S., J. Christianson, and B. Leerapan. 2013. "Do Electronic Medical Records Improve Diabetes Quality in Physician Practices?" *American Journal of Managed Care* 19 (2): 144–9.

Mitchell, J. M. 2008. "Do Financial Incentives Linked to Ownership of Specialty Hospitals Affect Physicians' Practice Patterns?" *Medical Care* 46 (7): 732–7.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *Journal of Clinical Epidemiology* 62 (10): 1006–12.

Nguyen, H. V. 2013. "Do Smoke-Free Car Laws Work? Evidence from a Quasi-Experiment." *Journal of Health Economics* 32 (1): 138–48.

Nguyen, N. T., S. Hohmann, J. Slone, E. Varela, B. R. Smith, and D. Hoyt. 2010. "Improved Bariatric Surgery Outcomes for Medicare Beneficiaries after Implementation of the Medicare National Coverage Determination." *Archives of Surgery* 145 (1): 72–8.

Pham-Gia, T., and T. L. Hung. 2001. "The Mean and Median Absolute Deviations." *Mathematical and Computer Modelling* 34 (7–8): 921–36.

Rogers, W. H. 1993. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 13: 19–23.

Ryan, A. M. 2009. "The Effects of the Premier Hospital Quality Incentive Demonstration on Mortality and Hospital Costs in Medicare." *Health Services Research* 44 (3): 821–42.

Ryan, A. M., and J. Blustein. 2011. "The Effect of the MassHealth Hospital Pay-for-Performance Program on Quality." *Health Services Research* 46 (3): 712–28.

Ryan, A. M., J. Blustein, and L. P. Casalino. 2012a. "Medicare's Flagship Test of Pay-For-Performance Did Not Spur More Rapid Quality Improvement among Low-Performing Hospitals." *Health Affairs (Millwood)* 31 (4): 797–805.

Ryan, A. M., B. K. Nallamothu, and J. B. Dimick. 2012b. "Medicare's Public Reporting Initiative on Hospital Quality Had Modest or No Impact on Mortality from Three Key Conditions." *Health Affairs (Millwood)* 31 (3): 585–92.

Ryan, A. M., T. Bishop, S. Shih, and L. P. Casalino. 2013. "Small Physician Practices in New York Needed Sustained Help to Realize Gains in Quality from Use of Electronic Health Records." *Health Affairs* 32 (1): 53–62.

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.

Shen, Y. C., and S. Zuckerman. 2005. "The Effect of Medicaid Payment Generosity on Access and Use among Beneficiaries." *Health Services Research* 40 (3): 723–44.

Song, Z., D. G. Safran, B. E. Landon, Y. He, R. P. Ellis, R. E. Mechanic, M. P. Day, and M. E. Chernew. 2011. "Health Care Spending and Quality in Year 1 of the Alternative Quality Contract." *New England Journal of Medicine* 365 (10): 909–18.

Suehs, B. T., A. Louder, M. Udall, J. C. Cappelleri, A. V. Joshi, and N. C. Patel. 2014. "Impact of a Pregabalin Step Therapy Policy among Medicare Advantage Beneficiaries." *Pain Practice* 14 (5): 419–26.

Sutton, M., S. Nikolova, R. Boaden, H. Lester, R. McDonald, and M. Roland. 2012. "Reduced Mortality with Hospital Pay for Performance in England." *New England Journal of Medicine* 367 (19): 1821–8.

Volpp, K. G., A. K. Rosen, P. R. Rosenbaum, P. S. Romano, O. Even-Shoshan, A. Canamucio, L. Bellini, T. Behringer, and J. H. Silber. 2007. "Mortality among Patients in VA Hospitals in the First 2 Years Following ACGME Resident Duty Hour Reform." *Journal of the American Medical Association* 298 (9): 984–92.

Werner, R. M., R. T. Konetzka, and D. Polsky. 2013. "The Effect of Pay-for-Performance in Nursing Homes: Evidence from State Medicaid Programs." *Health Services Research* 48 (4): 1393–414.

Wharam, J. F., A. J. Graves, F. Zhang, S. B. Soumerai, D. Ross-Degnan, and B. E. Landon. 2012. "Two-Year Trends in Cancer Screening among Low Socioeconomic Status Women in an HMO-Based High-Deductible Health Plan." *Journal of General Internal Medicine* 27 (9): 1112–9.

Wooldridge, J. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

———. 2009. *Introductory Econometrics: A Modern Approach.* Mason, OH: Thomson.

Zhu, J., P. Brawarsky, S. Lipsitz, H. Huskamp, and J. S. Haas. 2010. "Massachusetts Health Reform and Disparities in Coverage, Access and Health Status." *Journal of General Internal Medicine* 25 (12): 1356–62.

Zivin, K., P. N. Pfeiffer, B. R. Szymanski, M. Valenstein, E. P. Post, E. M. Miller, and J. F. McCarthy. 2010. "Initiation of Primary Care-Mental Health Integration Programs in the VA Health System: Associations with Psychiatric Diagnoses in Primary Care." *Medical Care* 48 (9): 843–51.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.
Appendix SA2: Stata Code for Analysis.