ANNUAL REVIEWS

Annual Review of Public Health

Designing Difference in Difference Studies: Best Practices for Public Health Policy Research

Coady Wing,[1] Kosali Simon,[2] and Ricardo A. Bello-Gomez[1]

[1]School of Public and Environmental Affairs, Indiana University, Bloomington, Indiana 47405, USA; email: cwing@indiana.edu, rabellog@indiana.edu

[2]School of Public and Environmental Affairs, Indiana University, Bloomington, Indiana 47405, USA, and National Bureau of Economic Research; email: simonkos@indiana.edu

Annu. Rev. Public Health 2018. 39:453–69

First published as a Review in Advance on January 12, 2018

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-040617-013507

Keywords

causal inference, difference in difference, policy analysis, quasi-experiments, research design

Abstract

The difference in difference (DID) design is a quasi-experimental research design that researchers often use to study causal relationships in public health settings where randomized controlled trials (RCTs) are infeasible or unethical. However, causal inference poses many challenges in DID designs. In this article, we review key features of DID designs with an emphasis on public health policy research. Contemporary researchers should take an active approach to the design of DID studies, seeking to construct comparison groups, sensitivity analyses, and robustness checks that help validate the method's assumptions. We explain the key assumptions of the design and discuss analytic tactics, supplementary analysis, and approaches to statistical inference that are often important in applied research. The DID design is not a perfect substitute for randomized experiments, but it often represents a feasible way to learn about casual relationships. We conclude by noting that combining elements from multiple quasi-experimental techniques may be important in the next wave of innovations to the DID approach.

INTRODUCTION

Causal inference is a key challenge in public health policy research intended to assess past policies and help decide future priorities. The causal effects of policies and programs related to vaccines, vehicle safety, toxic substances, pollution, legal and illegal drugs, and health behaviors are difficult to measure. But scientific research and sound policy analysis demand information about causal relationships. The standard advice is to implement a randomized controlled trial (RCT) to avoid confounding and isolate treatment effects. But large-scale RCTs are rare in practice. Without an RCT, researchers often seek answers from natural experiments, including regression discontinuity designs, instrumental variables, covariate matching, and synthetic control strategies (for recent methods reviews, see 9, 10, 13, 41). In this article, we focus on the design of quasi-experimental studies that compare the outcomes of groups exposed to different policies and environmental factors at different times. Most people describe the approach as a difference in difference (DID) design, but it is sometimes called a comparative interrupted time series design or a nonequivalent control group pretest design (6, 55, 92, 99, 105).

Regardless of nomenclature, the DID design is well established in public health research (45). It has been around since the middle of the nineteenth century, when John Snow published the results of his DID study showing that cholera is transmitted through the water supply rather than air (97). Since Snow's study, researchers have developed tools and tactics that can strengthen the credibility of DID studies. Our goal in this article is to review principles and tools that researchers can use to design and implement a high-quality DID study. Throughout the article, we point to theoretical work and empirical examples that help clarify important techniques or challenges that are common in health research. Observing a variety of applied examples that implement these techniques is a very useful complement to describing the DID challenges in abstract.

THE DIFFERENCE IN DIFFERENCE DESIGN AND WORKHORSE STATISTICAL MODELS

Potential Outcomes Notation

Throughout the article, we use $g = 1 \ldots G$ to index cross-sectional units and $t = 1 \ldots T$ to index time periods. In DID studies, g often refers to geographical areas such as states, counties [e.g., when studying the historical rollout of a food stamp program (61)], or census tracts, although it could also refer to distinct groups such as those separated by age [as used in studies of Medicare Part D (e.g., 3, 65, 101) or the young adult mandate of the Affordable Care Act (e.g., 91)]. Most of the time, t represents years, quarters, or months. In most applications, researchers are concerned with outcomes in two alternative treatment regimes: the treatment condition and the control condition. To make the idea concrete, let $D_{gt} = 1$ if unit g is exposed to treatment in period t, and $D_{gt} = 0$ if unit g is exposed to the control condition in period t. In public health applications, the set of treatments might consist, for example, of two alternative approaches to the regulation of syringe exchange programs that are adopted in different states in different years (23).

Research on the causal effects of the treatment condition revolves around the outcomes that would prevail in each unit and time period under the alternative levels of treatment. One way to make this idea more tangible is to define potential outcomes that describe the same unit under different (hypothetical) treatment situations. To that end, let $Y(1)_{gt}$ represent an outcome of interest for unit g in period t under a hypothetical scenario in which the treatment was active in g at t; $Y(0)_{gt}$ is the outcome of the same unit and time under the alternative scenario in which the control condition was active in g at t. The treatment effect for this specific unit and time period is

$\Delta_{gt} = Y(1)_{gt} - Y(0)_{gt}$, which is simply the difference in the value of the outcome variable for the same unit across the two hypothetical situations. The notation suggests this would be easily done, but applied researchers cannot observe the identical unit under two different scenarios as one could through a lab experiment; in practice, each unit is exposed to only one treatment condition in a specific time period, and we observe the corresponding outcome. Specifically, for a given unit and time, we observe $Y_{gt} = Y(0)_{gt} + [Y(1)_{gt} - Y(0)_{gt}]D_{gt}$.

The notation so far describes the counterfactual inference problem that arises in every causal inference study. In a typical study, researchers have access to data on Y_{gt} and D_{gt}, and they aim to combine the data with research design assumptions to learn about the average value of $Y(1)_{gt} - Y(0)_{gt}$ in a study population. The DID design is a quasi-experimental alternative to the well-understood and straightforward RCT design, seen for example in the health insurance context in the RAND Health Insurance Experiment in the 1970s and more recently in the Oregon Health Insurance Experiment (12, 67; see 74 for new techniques in external validity).

RCT and DID share some characteristics: Both involve a well-defined study population and set of treatment conditions, where it is easy to distinguish between a treatment group and a control group and between pretreatment and post-treatment time periods. The most important distinction is that treatment conditions are randomly assigned across units in an RCT but not in a DID design. Under random assignment, treatment exposure is statistically independent of any (measured or unmeasured) factor that might also affect outcomes. In a DID design, researchers cannot rely on random assignment to avoid bias from unmeasured confounders and instead impose assumptions that restrict the scope of the possible confounders. Specifically, DID designs assume that confounders varying across the groups are time invariant, and time-varying confounders are group invariant. Researchers refer to these twin claims as a common trend assumption. In the next two sections, we describe the DID design further and explain how the key assumptions of the design lead to a statistical modeling framework in which treatment effects are easy to estimate. We start with the simple two-group two-period DID model and then examine a more general design that allows for multiple groups and time periods.

Two Groups in Two Periods

The simplest form of the DID design is a special case in which there are only two groups ($g = 1, 2$) observed in two time periods ($t = 1, 2$); this situation is often represented by a 2×2 box. In the first period, both groups are exposed to the control condition. In the second period, the treatment rolls out in group 2 but not in group 1. Let $T_g = 1[g = 2]$ be a dummy variable identifying observations on group 2. T_g has no time subscript because group membership is time invariant. $P_t = 1[t = 2]$ indicates observations from period 2, and P_t has no group subscript because the time period does not vary across the groups. In the simple DID, the treatment variable is the product of these two dummy variables: $D_{gt} = T_g \times P_t$. It is easy to see the connection between the description of the design and the notation. For example, $D_{gt} = 0$ for both groups in the first period because $P_t = 0$; and $D_{gt} = 1$ only for group 2 in period 2 because that is the only way that both T_g and P_t are equal to 1.

In the two-group two-period DID design, the common trend assumption amounts to a simple statistical model of the treated and untreated potential outcomes. Under the simple DID, the untreated potential outcome is $Y(0)_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \epsilon_{gt}$. In the absence of treatment, the average outcome in group 1 is β_0 in period 1 and $\beta_0 + \beta_2$ in period 2. Likewise, the average untreated outcome in group 2 is equal to $\beta_0 + \beta_1$ in period 1 and $\beta_0 + \beta_1 + \beta_2$ in period 2. Under the common trend assumption, the coefficient on T_g captures the time-invariant difference in outcomes between the two groups. Implicitly, the group coefficient captures the combined effects

of all unmeasured covariates that differ systematically between the two groups and that do not change over the course of the study period. In a similar manner, the coefficient on P_t captures the combined effects of any unmeasured covariates that change between the two periods but affect outcomes the same way in both groups. In practice, researchers call β_1 the group effect and β_2 the time trend.

The model for the treated potential outcome is the untreated outcome plus a treatment effect, which is usually restricted to be constant across observations: $Y(1)_{gt} = Y(0)_{gt} + \beta_3$. The two potential outcome specifications combine with the treatment indicator to produce realized outcomes according to the general formula $Y_{gt} = Y(0)_{gt} + D_{gt}[Y(1)_{gt} - Y(0)_{gt}]$. Replacing the potential outcomes with the model specification gives $Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \epsilon_{gt} + D_{gt}[Y(0)_{gt} + \beta_3 - Y(0)_{gt}]$. In the two-group two-period setting, $D_{gt} = T_g \times P_t$, which means that after canceling the $Y(0)_{gt}$ terms we can rewrite the observed outcome equation in terms of the group and time period indicators to obtain the standard DID estimating equation:

$$Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \times P_t) + \epsilon_{gt}.$$

The model is easy to estimate with data on outcomes, group membership, and time periods. The coefficient on the interaction term is an estimate of the treatment effect under the common trend assumption.

Multiple Groups and Time Periods

The two-group two-period DID design is intuitive, but it does not accommodate the complexity encountered in applications, which often involve treatment exposures in multiple groups and multiple time periods. An example of this is the state adoption of medical marijuana laws, which remains an active area of state policy. Research in this area includes a study by Harper et al. (57), who reexamined earlier research that did not include state fixed effects, and one by Anderson et al. (4), who incorporated more DID techniques. Luckily, the main features of the DID design also apply in a broader set of conditions. When $G \geq 2$ groups and $T \geq 2$ periods, $D_{gt} = 1$ if the treatment is active in group g and period t; otherwise, $D_{gt} = 0$. As in the two-group two-period case, the core assumption in the generalized DID is that any unmeasured determinants of the outcomes are either time invariant or group invariant.

The generalized design is easy to analyze using a two-way fixed effects regression model to describe the potential outcomes. The model for the untreated outcome is $Y(0)_{gt} = a_g + b_t + \epsilon_{gt}$. In the model, a_g represents the combined effects of the time-invariant characteristics of group g, and b_t represents the combined effects of the time-varying but group-invariant factors.[1] The average untreated outcome for group 3 in period 5 is given by $a_3 + b_5$. Likewise, the average untreated outcome for group 4 in period 5 is $a_4 + b_5$. The two groups have different levels in every period, but any changes over time within a group come from the group-invariant trend terms described by b_t. Researchers call a_g a group-fixed effect and b_t a time-fixed effect. The time-fixed effects trace out the common time trend. A key point is that the group effects and time trends stem from underlying differences in unmeasured covariates across groups and time periods. The DID design

[1] It may be more revealing to think of $a_g = x_g \alpha$, where x_g is a vector of time-invariant covariates associated with group g and α is a coefficient vector. Likewise, we can think of $b_t = z_t \gamma$, where z_t is a vector of time-varying but group-invariant covariates, and γ is a coefficient vector. In practice, x_g and z_t are unmeasured, and we do not attempt to estimate each of the covariate specific coefficients. Instead, we estimate or eliminate the combined effects of all covariates using fixed effects differencing techniques.

is meant to control for these unmeasured confounders even though the underlying variables are not measured explicitly.

Like the two-group two-period design, the generalized DID also specifies that the treated outcome is a shifted version of the untreated outcome so that $Y(1)_{gt} = Y(0)_{gt} + \delta$. Combining the equations shows that the observed outcome is $Y_{gt} = Y(0)_{gt} + D_{gt}[Y(1)_{gt} - Y(0)_{gt}]$. Substitute the fixed effects structure for the potential outcomes to obtain $Y_{gt} = a_g + b_t + \epsilon_{gt} + D_{gt}[Y(0)_{gt} + \delta - Y(0)_{gt}]$ and cancel the remaining $Y(0)_{gt}$ terms to find the generalized DID estimating equation:

$$Y_{gt} = a_g + b_t + \delta D_{gt} + \epsilon_{gt}.$$

The two-way fixed effects parameterization stems from the same common trend assumption involved in the two-group two-period DID, but it accommodates considerably more variation in the details of the research design. In practice, researchers estimate the treatment effect parameter, δ, using fixed effects regression models; they simply regress the observed outcome on the treatment variable and a full set of group- and time-fixed effects. For an example, see the main specification in Bitler & Carpenter (21).

The Common Trends Assumption

Both the simple and generalized DID designs rely on the assumption that the important unmeasured variables are either time-invariant group attributes or time-varying factors that are group invariant. Together, these restrictions imply that the time series of outcomes in each group should differ by a fixed amount in every period and should exhibit a common set of period-specific changes. Loosely speaking, a graph of the time series should look like a set of parallel lines. For an example, see the graphs in Kaestner et al. (64) of the treatment group and synthetic control group trends among low-educated adults prior to state Medicaid expansion or figures in other Medicaid expansion studies (e.g., 96). Note that parallel lines do not have to be linear: Time-fixed effects allow for flexible time trends that move up or down across from period to period, as they do, for example, in the study of Sommers et al. (100), who examine state Medicaid expansions using as a control group low-income adults in states that did not expand Medicaid.

In applied work, the most difficult task is evaluating the credibility of the common trends assumption. Later in the article, we discuss statistical tests and graphical analyses that researchers can use to empirically probe the credibility of the assumption. Researchers, however, must also think carefully about the conceptual reasons for which the common trends assumption might be valid in some settings and not in others. It may be helpful to interpret the common trends assumption as a byproduct of a set of underlying variables that differ across states and change over time. Consider the case of vaccine policy (a topic studied, for example, in 102). Instead of asking the abstract question of whether vaccination rates in two states are apt to follow a common time trend absent the policy, we could ask what sorts of (unmeasured) factors likely explain variation in vaccination rates across states and over time, such as parental attitudes. Next, we would ask whether those factors are likely covered by the DID design: Are they time-invariant group attributes or group-invariant time-varying factors? Naming the unmeasured variables that the fixed effects structure is intended to capture is a good way to assess the quality of a DID design, because it is easier to construct and evaluate arguments for and against specific variables than for abstract trends that arise from unknown origins.

Being specific about unmeasured variables often points the way to stronger research designs as well. Perhaps it makes sense to exclude certain groups from the analysis if they seem likely to differ from the others with respect to the important unmeasured variable. A version of this argument is

used in forming synthetic control groups, where groups that differ in past characteristics compared to the treatment groups are excluded or given less weight when forming the control group for a single difference (as is done in 1, which forms a synthetic California from a weighted average of potential control states that do not have tobacco control programs; for longer reviews of synthetic control methods, see 10, 47, 77). The common trends assumption may hold in a restricted sample of groups or time periods even if it does not hold across all groups and times. This line of thinking is the starting point for combined research designs in which researchers use propensity score matching in a first step and then estimate treatment effects using DID methods on the matched sample [as was done, for example, in studying health effects of employment transitions in Germany (50); for use of DID and synthetic control methods together, see also 54].

Strict Exogeneity

The DID design aims to difference out unmeasured confounders using techniques that eliminate biases from group- or time-invariant factors. For the differencing technique at the core of the method to work, the timing of treatment exposures in the DID design must be statistically independent of the potential outcome distributions, conditional on the group- and time-fixed effects. This aspect of the design is harder to understand. Econometrics textbooks use the term "strict exogeneity" to describe it, pointing out that it is stronger than "contemporaneous exogeneity," which is the foundational assumption in studies based on propensity score matching and cross-sectional regression adjustment.

To better understand the distinction, suppose that a_g and b_t are functions of vectors of the underlying covariates x_g and z_t. A researcher who collects data on each covariate might estimate the causal effect of D_{gt} on Y_{gt} under the conditional mean independence assumption that $E[Y(j)_{gt}|D_{gt},\ a_g,\ b_t] = E[Y(j)_{gt}|a_g, b_t]$ for $j = 0, 1$. To put this idea into practice, the researcher might form matched pairs of treated and control observations and estimate the treatment effect using the mean difference in outcomes in the matched sample, as, for example, Obermeyer et al. (83) did in studying Medicare's hospice benefit. The situation is different in the DID design. To remove confounding using differencing, the entire sequence of past and future treatment exposures must be independent of unmeasured determinants of the outcome variables. Formally, strict exogeneity requires that $E[Y(j)_{gt}|a_g, b_t,\ D_{g1}, \ldots D_{gT}] = E[Y(j)_{gt}|a_g, b_t]$ for $j = 0, 1$.

The idea is that—after conditioning on the group and period effects—treatment exposures that occur at $t + 1$ are not anticipated by outcomes measured in an earlier period such as t. The restriction could fail in practice for many reasons. Perhaps states change their regulations in response to changes in the outcome variable of interest (19), or perhaps companies change their behavior in anticipation of a regulation that seems likely to occur in the near future. Such behavioral patterns almost certainly occur in the real world, and they represent important threats to the validity of DID designs. One way that researchers investigate such effects is to include the policy variables on the left-hand side and show that the factors that most concern us do not predict the passage of the law. Some studies use these specifications to show that political variables are influential, and to the extent that they can be considered exogenous, they could be used as instruments for the policy (71).

SENSITIVITY ANALYSIS AND ROBUSTNESS CHECKS OF THE COMMON TRENDS ASSUMPTION

Modern applications of the DID design devote much attention to sensitivity analysis and robustness checks designed to probe the main assumptions that support the internal validity of

the research design. Although the specific details involved vary with the context and data limitations of individual studies, this section provides a short summary of the analytical techniques researchers use to shed light on the validity of the common trends assumption and threats to the strict exogeneity condition.

Graphical Evidence

In the simple two-group two-period DID, the common trend assumption is not testable. In settings with multiple pretreatment time periods, however, researchers can partially validate the common trends assumption. For example, researchers often plot the mean outcomes by group and time period and then ask whether the lines appear to be approximately parallel (e.g., see 8, figure 1, for an example related to the young adult mandate of the ACA, where the visual plot serves as a precursor to a statistical test of the parallel trends assumption). When the annual means are precisely estimated and year-to-year volatility is relatively low, it is easy to spot deviations from the common trends assumption in a long time series.

Visual evidence may be less compelling when the data are noisy or the time series is short. In such cases, it may be difficult to distinguish between statistical noise and genuine deviations from the common trends. A graph also helps convey the strength of the policy shock, as measured for example by the impact of a health insurance policy on coverage rates. This is important because studies often go on to examine the impact of a policy on downstream effects (such as health care use or health status). The interpretability of graphical evidence is related to the broader issue of statistical power in DID designs. The statistical power of DID designs often requires more analysis than the standard power analysis for simple mean differences and linear regression coefficients considered in standard textbooks, and it is important to consider the size of effects that such studies can reliably detect (see 26, p. 46; 70, 80).

Group-Specific Linear Trends

Another strategy for evaluating the common trend assumption in studies with more than two time periods is to fit an augmented DID regression that allows for group-specific linear trends [as done, for example, by Hansen et al. (56) in studying state cigarette taxes]. In practice, this amounts to a regression of the outcome on the treatment variable, group and period effects, and each group effect interacted with the linear time index: $Y_{gt} = a_g + b_t + \beta_g(a_g \times t) + D_{gt}\delta + \epsilon_{gt}$. The common trends model is nested in the group-specific trend model. An F-test of the compound null in which all the coefficients of the group-specific linear trends are jointly zero is a test of the common trends model. Rejecting the null hypothesis implies that common trends is not a valid assumption. In practice, most researchers interpret the group-specific linear trends model more casually by comparing the treatment effect estimates in the restricted and unrestricted models. If the treatment effect is not sensitive to the alternative specification, most researchers consider the core results more credible.

Balancing Tests for Changes in Composition

In RCTs and matching studies, researchers often present evidence that the distribution of covariates is very similar in the treatment and control groups (59, 63). The basic goal in this case is to show that the two groups were comparable prior to treatment exposure. In a DID study, the groups are usually nonequivalent prior to treatment exposure, so that a simple covariate balancing table is not very informative about the validity of the research design; however, readers tend to

be more reassured when covariates are similar. What matters for DID validity is that differences between the two groups are stable over time and that the changes in treatment exposure are not associated with changes in the distribution of covariates.

One way to examine this aspect of DID validity empirically is to estimate covariate balance regressions (see, for example, 86, which uses covariate balancing to study the productivity of new surgeons). Suppose that in addition to data on Y_{gt} and D_{gt}, researchers also have access to data on a covariate C_{gt} associated with group g in period t. A simple way to test for problematic compositional changes is to replace the outcome variable with the covariate and fit the standard DID regression model: $C_{gt} = a_g + b_t + D_{gt}\delta' + \epsilon_{gt}$. Under the null hypothesis that there are no compositional changes, we expect that $\delta' = 0$. Of course, it is sensible to consider the magnitude of the change in composition rather than the pure statistical significance of the coefficient estimate. Researchers can fit the DID regression to data on a large list of available covariates to assess the relevant concept of balance across a broad range of factors.

Granger-Type Causality Tests

To examine the possibility that future treatment exposures are anticipated by current outcomes, researchers can augment the standard DID regression model to include leading values of the treatment variable. For example, researchers might fit a model with S leading values of the treatment variable:

$$Y_{gt} = a_g + b_t + D_{gt}\delta + \sum_{s=1}^{S} D_{g,t+s}\gamma_s + \epsilon_{gt}.$$

Under the strict exogeneity null, we expect that future policy changes will not be associated with current outcomes, so that $\gamma_s = 0$ for $s = 1 \ldots S$. Decisions about how many leads to examine are somewhat arbitrary and mainly have to do with the total number of periods available for analysis and the timing of the policy changes. Examples of studies that include lead tests are those by Bachhuber et al. (11) (on the relationship between medical cannabis laws and opioid overdose mortality) and Raifman et al. (88) (on the relationship between same-sex marriage laws and adolescent suicide attempts).

Time-Varying Treatment Effects

In many applications, the effect of the treatment may vary with time since exposure. Researchers can study these effects by including lagged treatment variables in the standard DID model. One common strategy is to use an event study framework examining anticipation effects and phase-in effects in a single regression such as

$$Y_{gt} = a_g + b_t + D_{gt}\delta + \sum_{s=1}^{S} D_{g,t+s}\gamma_s + \sum_{m}^{M} D_{g,t-m}\lambda_s + \epsilon_{gt}.$$

In this specification, δ captures the immediate effect of the policy, and λ_m measures any additional effects of a policy that occur m periods after adoption. If the initial effect of the policy is positive, then negative values of λ_m imply that the initial effect of the policy dissipates over time, and positive values of λ_m suggest that the policy has larger effects over time. Event study figures are included, for example, in Bellou & Bhatt (16), who study drivers' license laws; Anderson et al. (5), who study medical marijuana laws; Bitler & Carpenter (21), who study mammography mandates; Simon (93), who studies cigarette taxes; Marcus & Siedler (75), who study alcohol policy in Germany; and Paik et al. (84), who study medical malpractice. Some studies, such as the one by

Brot-Goldberg et al. (24), specifically look for anticipatory effects, in this case studying the effect of deductibles on health care prices, quantities, and spending. In general, whenever a policy includes a time gap between announcement and effective date, such behaviors are possible. In the context of a well-publicized federal policy change, Alpert (3) examines anticipatory effects before Medicare Part D implementation, exploiting the difference in behaviors observed for chronic versus acute drugs, and Kolstad & Kowalski (66) consider periods before, during, and after treatment.

Triple Differences

When the core DID assumptions are suspicious on conceptual or empirical grounds, researchers sometimes seek to strengthen the research design by adding an additional comparison group and estimating treatment effects using a difference in difference in difference (DDD) design. Suppose that the DID design is questionable because there is some time-varying confounder that changes differentially across the states that make up the study design. A time-varying confounder that is not state invariant is a problem for the DID study because it violates the common trend assumption. To address the problem with a DDD design, researchers need to find a new within-state comparison group that is not exposed to treatment but is exposed to the problematic time-varying confounder. With the two groups in hand, researchers can estimate the standard DID specification separately on the original data and on the new comparison group data. The DID estimate from the comparison group represents an estimate of the effect of the state-specific time-varying confounder that is free from any treatment effect. The DID estimate from the original data represents the combined effect of the confounder and the treatment. By subtracting one DID estimate from the other—forming a triple difference—researchers can remove the bias from the confounder and isolate the treatment effect [see Atanasov & Black (9, pp. 254–58) for a careful treatment of DDD designs].

Suppose that some states impose a tax on large hospitals but not small hospitals, and we wish to study its impact on the wages of nurses. The treatment states experience some of the same spurious shocks that affect control states, but suppose also that the tax-adopting states are mostly from geographic areas that faced a different set of regional economic booms and busts over time. The standard DID estimate might conflate the changes in the hospital tax policy with the regional economic conditions; that is, the DID model might fail to meet the common trends assumption. A DDD strategy might start by reasoning that small hospitals are subject to the same regional economic conditions as large hospitals but are not subject to the large hospital tax.

Nationwide, there are also some small-firm shocks and some large-firm shocks. Thus, either a DID that compares small and large firms within treated states or a DID that compares large firms across treatment and control states would be compromised. However, a DDD that compares changes over time in large firms in states with and without the policy, compared to the similar difference for small firms, would produce an unbiased result. In other words, the common trends assumption should hold in the DDD, whereas it would not hold in either of the two possible DID methods separately. Researchers almost always present triple difference specification results as a supplement to a main DID specification; recent examples of use in health include the studies by Chatterjee and colleagues (36) and Heim & Lin (58), both of which examine the labor market outcomes of health insurance reform. It is fairly rare to find an article presenting a parallel trends test of a DDD, but Paik and colleagues (85) offer an example of such a test and show the importance of conducting such tests.

STATISTICAL INFERENCE IN DIFFERENCE IN DIFFERENCE

So far we have focused on the assumptions and conceptual threats to the validity of DID studies. However, a substantial literature makes it clear that statistical inference is also an important

challenge in DID studies. The core message is that standard errors estimated under the assumption that errors are independent across observations are often biased downwards, which leads to over-rejection of the null hypothesis.

Moulton (79) considers statistical inference of regression coefficients on variables that do not vary within aggregate groups. His examples involve models that link micro data on labor market outcomes with aggregate geographical information. The problem is that these factors do not vary within groups (or are correlated within groups), and the groups may also have a shared error structure. Moulton uses a parametric random effects model to show that standard errors are biased downwards and that the magnitude of the bias depends positively on group size, intraclass correlations of the regression errors, and intraclass correlations of the regressors included in the model. Bertrand and colleagues (18) point out that many DID studies involve large group sizes and are apt to exhibit high levels of intraclass correlation of both errors and key independent variables. They use Monte Carlo simulations to assess the performance of several different methods of performing statistical inference in clustered data designed to mimic many DID studies. They find that many methods of inference fare poorly, especially when the number of clusters is relatively small. However, they also find that collapsing the data down to group-level cells, clustering robust standard errors, and using clustered bootstraps work relatively well.

Since the article by Bertrand and colleagues (18), there has been a small boom in research on alternative approaches to statistical inference in DID studies. Cameron & Miller (30) provide a helpful review of the literature. By our reading, the literature has not reached a consensus on the best way to perform inference in DID models. However, several themes have emerged. In most cases, it makes sense to aggregate the data so that outcomes are measured at the same level as the treatment variable [as is done by Bedard & Kuhn (14), who study healthy food nudging messages in a restaurant chain]. The standard cluster robust variance estimator (72) should perform well in studies based on a large number of clusters. For studies with smaller numbers of clusters [this applies to geographical variation in countries like Germany, which has 16 states (69), or Sweden, which has 4 (2)], three broad families of methods have emerged. One set of methods performs inference using cluster-level randomization distributions (38, 90). Another pursues various forms of the cluster bootstrap (28). A third approach performs finite sample corrections based on bias-reduced linearization (15, 46, 62, 87). Cameron et al. (29) provide a method for adjusting for multiway clustering; Solon et al. (98) discuss the role of sampling weights. In addition, recent work by Abadie et al. (1) revisits the rationale for cluster standard error adjustments and emphasizes that the decision to adjust for clustering should flow from the treatment assignment rule embedded in the research design and the data collection method.

POLICY VARIATION AND HETEROGENEITY

Many US health policies in the last century have been decided at the state level, reflecting principles of federalism and efforts to find locally tailored solutions (60, 82). State policy variation, however, often displays a high degree of standardization across states, making it possible to generalize from the experience of several states in one study. If each state were to adopt extremely unique legislative solutions to public health challenges, the result would be a series of one-state DID studies, which would make it difficult to develop consensus and to provide evidence to aid in future policy making. This is not to downplay the importance of single-geographic-unit studies when health policies such as indoor tobacco bans are introduced nationally, as they were in Ireland and China (51, 106), or when one US state or locality enacts a policy that is unique in its time [e.g., Massachusetts health reform, in Kolstad & Kowalski (66); Tennessee's Medicaid disenrollment, in Garthwaite et al. (49); or food policy, in Cantor et al. (31) and Cawley & Frisvold (35)]. However, researchers

are sometimes still able in such cases to compare policies across countries, as in the case of health care privatization in Latin America (27). Researchers are also able to use synthetic control methods to construct comparison groups using a weighted average of other countries, as done by Rieger and colleagues (89) in studying the effects of universal health insurance in Thailand.

One reason for this relative standardization across US state laws is the proliferation of model laws by policy organizations. For example, when states regulate access to controlled substances, they are able to consider sample legislation available through the National Alliance for Model State Drug Laws. Standardized versions of state laws for policies like the medicinal use of marijuana allow researchers to conduct studies using categorizations of states, exploiting variations in the year of adoption to implement a study with a DID design (22, 81).

Despite the forces acting toward standardization in state laws, policies do tend to differ in important ways that reflect local political marketplaces (105). Researchers often separate state laws into a reasonably small number of meaningfully different categories, but it is important to understand the degree of detail that is sacrificed in this approach, for example because of alternative classifications or sensitivity analyses that remove states that are difficult to classify. Researchers often investigate the characteristics of the policies themselves or borrow classifications from other studies or policy organizations. In the area of state small-group insurance market reforms, for example, state laws may be characterized as strong or weak depending on whether regulations apply to all or some insurance policies (94). Considering alternative classifications of state policies and testing for sensitivity to the removal of certain states with particularly ambiguous policy status are both useful additions to analyses with policy heterogeneity. However, the availability of multiple analyses using the same classification systems facilitates comparisons across studies, and providing enough details for replication is good practice.

Another way in which policy heterogeneity commonly presents itself in public health settings is a tax rate, for example in the area of regulating health behaviors (e.g., cigarette or alcohol taxes). Because each state tends to set an individualized rate, there is heterogeneity in the policy; however, the policy is linear, and its intensity can be measured continuously. Carpenter & Cook (33) advanced the study of cigarette tax effects on youth by implementing a DID model, whereas prior literature had not included state fixed effects. Non-tax-rate examples of such linear policy measures include Medicaid physician fees or minimum wage laws, the public health impacts of which have been discussed in several recent articles (25, 40, 103). Linear measures of policy variation can be placed into the DID framework directly, but researchers may also explore non-linearity in policy impacts using quadratic terms or by creating dummy variables for ranges of policy values (such as classifying tax rates as under or over certain values, or entering the values as a spline).

Even when using linear measures, researchers are faced with decisions as to whether the values should be entered in logs if the distribution of values is skewed across states, whether policy values should be measured in real or nominal terms, and whether the values should be normalized to the cost of some outside option (for example, studies of Medicaid fees often measure them relative to Medicare or private insurance fees, using a ratio as the key policy measure: e.g., 43, 44). If there are nuances in these linear forms of laws, for example if health insurance regulations only apply to large firms, or alcohol taxes apply to beer but not wine, some may use the excluded group as a within-state control (e.g., 76) or may test for unintended spillover effects onto those groups; others may prefer to simply exclude those other groups.

The use of a within-state control group is especially helpful when diagnostic tests indicate that the DID is problematic due to a violation of the common trends assumption; if a credible within-state control group can be found that trends similarly to the treatment group absent the policy, then researchers may be able to implement a DDD as well. An example that is often used to explain

DDD is the case of maternity coverage mandates and wages; Gruber (52) show that because men should not be affected by the policy, they form a convincing within-state control group. Sometimes researchers report two separate DIDs rather than explicitly estimate a DDD [e.g., Simon & Kaestner (95) estimate the effects of minimum wages for low-educated and high-educated persons using the high-educated group as a close-to-placebo group]. This way of observing effects on different groups differs from the approach taken by studies that examine policy heterogeneity (for example, researchers wishing to examine differences in the effects of cigarette taxes on smoking rates among youth versus adults would run two DIDs and report them separately, rather than run a DDD). Similarly, several health insurance studies use baseline county characteristics to examine whether the intended effects are greater in counties that are likely to benefit more from the policy (20, 48, 78). Cook & Durrance (39) take advantage of state variation in the degree to which federal alcohol taxes should be binding to construct an identification strategy.

Multidimensional policy heterogeneity can also be transformed into a linear measure, a technique that has proven popular in cases in which a formula can be created to measure the strength of the overall policy based on the fraction of people affected. Measures of Medicaid eligibility expansions in the 1980s and 1990s (42, 53) and the literature on the long-term impacts of these expansions (e.g., 37) represent a prominent example. Medicaid eligibility is determined by a formula that counts some but not other forms of income and deducts certain expenses, with different rules depending on the number and ages of children in the family. Rather than create separate variables to measure each aspect of the policy, which leads to a cumbersome interpretation of parameters, or to separate states into strong versus weak expansions, researchers collect the parameters that determine eligibility and boil down the variation into an index of stringency. Taking a nationally representative population, one could examine the percentage of the population that would be eligible by the rules in place in a certain state and year, leading to an index that increases with generosity.

Using this variable as the policy term leads to a DID format whereby researchers can interpret how the outcome changes as generosity is increased, so that, for example, 10% more of a representative population may become eligible for the policy. This linear policy measure can then be used as the sole policy measure, although one criticism of this approach is that policy makers may want to know the effect of each actual policy lever they control (55). No matter how the policy variable is created, it can be used as an instrument for eligibility (e.g., when asking whether being eligible due to policy variation causes a reduction in private coverage) or as a reduced form (e.g., when answering how policy generosity affects the outcome).

DISCUSSION

Quasi-experimental research designs can be an effective way to learn about causal relationships that are important for public health science and public health policy. Recent innovations allow researchers to approach the design of quasi-experimental studies in much the same way that they would approach the design of a fully randomized experimental study. Quasi-experiments are apt to work best when researchers actively decide which of the possible imperfect comparison groups is likely to best satisfy the assumptions of a particular technique. A study will be most convincing when researchers have thought carefully about the substantive meaning of key assumptions for their specific study. Given that the modern technical literature is large and complex, care is needed to identify and employ the tools and techniques that are most relevant to a given study.

This article examined DID designs in detail not because DID designs are the best approach to quasi-experimental research design, but because DID designs are often feasible in public health research in large federal or decentralized countries that collect data through a wide range of surveys

and administrative databases. However, there are several cases in which methods other than DID are best for evaluating state policy: When data prior to state policy variation were not available, researchers have used age-based regression discontinuities to understand the impact of Medicare (32) or alcohol policy (34). In the United States, for instance, a wide range of regulations and environmental conditions vary across geographic areas and over time, providing many opportunities to learn about causal effects. DID designs are also applied in nongeographic units, such as in studying Medicare Part D and the ACA's young adult mandate, where groups are compared over time by age; moreover, DID designs can also be applied with neither time nor geography (e.g., access to insurance and health are the two dimensions used in 73). Honing skills at designing and implementing high-quality DID studies that can make the best of the available data is a valuable part of the public health research toolkit.

Although it is beyond the scope of our review, we anticipate that future methodological advances will often involve hybrid research designs that exploit multiple quasi-experimental design elements. For example, Wing & Cook (104) use design elements from DID and matching studies to strengthen the external validity of the regression discontinuity design [(7) and (17) also aim to expand external validity of regression discontinuity design], and Kreif et al. (68) compare the results of a synthetic control approach to those of a DID approach to evaluate the effects of hospital pay-for-performance programs. The advances in DID methods surveyed in this article, together with these future possibilities for further innovation, suggest that the DID framework will continue to be one of the workhorse models used in public health policy research.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful for helpful suggestions from Kitt Carpenter, John Cawley, and Dan Rees. We thank Ryan Bennion and Sam Kuster for research assistance.

LITERATURE CITED

1. Abadie A, Diamond A, Hainmueller J. 2010. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J. Am. Stat. Assoc.* 105(490):493–505
2. Almond D, Edlund L, Palme M. 2009. Chernobyl's subclinical legacy: prenatal exposure to radioactive fallout and school outcomes in Sweden. *Q. J. Econ.* 124(4):1729–72
3. Alpert A. 2016. The anticipatory effects of Medicare Part D on drug utilization. *J. Health Econ.* 49:28–45
4. Anderson DM, Hansen B, Rees DI. 2015. Medical marijuana laws and teen marijuana use. *Am. Law Econ. Rev.* 17(2):495–528
5. Anderson DM, Rees DI, Sabia JJ. 2014. Medical marijuana laws and suicides by gender and age. *Am. J. Public Health* 104(12):2369–76
6. Angrist JD, Pischke JS. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press
7. Angrist J, Rokkanen M. 2012. *Wanna get away? RD identification away from the cutoff*. NBER Work. Pap. 18662, Cambridge, MA
8. Antwi YA, Moriya AS, Simon K. 2013. Effects of federal policy to insure young adults: evidence from the 2010 Affordable Care Act's dependent-coverage mandate. *Am. Econ. J. Econ. Policy* 5(4):1–28
9. Atanasov V, Black B. 2016. Shock-based causal inference in corporate finance and accounting research. *Crit. Finance Rev.* 5:207–304

10. Athey A, Imbens GW. 2017. The state of applied econometrics: causality and policy evaluations. *J. Econ. Perspect.* 31(2):3–32
11. Bachhuber MA, Saloner B, Cunningham CO, Barry CL. 2014. Medical cannabis laws and opioid analgesic overdose mortality in the United States, 1999–2010. *JAMA Intern. Med.* 174(10):1668–73
12. Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, Newhouse JP, et al. 2013. The Oregon experiment—effects of Medicaid on clinical outcomes. *N. Engl. J. Med.* 368:1713–22
13. Basu S, Meghani A, Siddiqi A. 2017. Evaluating the health impact of large-scale public policy changes: classical and novel approaches. *Annu. Rev. Public Health* 38:351–70
14. Bedard K, Kuhn P. 2015. Micro-marketing healthier choices: effects of personalized ordering suggestions on restaurant purchases. *J. Health Econ.* 39:106–22
15. Bell RM, McCaffrey DF. 2002. Bias reduction in standard errors for linear regression with multi-stage samples. *Surv. Methodol.* 28(2):169–81
16. Bellou A, Bhatt R. 2013. Reducing underage alcohol and tobacco use: evidence from the introduction of vertical identification cards. *J. Health Econ.* 32(2):353–66
17. Bertanha M, Imbens GW. 2014. *External validity in fuzzy regression discontinuity designs*. NBER Work. Pap. 20773, Cambridge, MA
18. Bertrand M, Duflo E, Mullainathan S. 2004. How much should we trust differences-in-differences estimates? *Q. J. Econ.* 119(1):249–75
19. Besley T, Case A. 2000. Unnatural experiments? Estimating the incidence of endogenous policies. *Econ. J.* 110(467):672–94
20. Bleakley H. 2007. Disease and development: evidence from hookworm eradication in the American South. *Q. J. Econ.* 122(1):73–117
21. Bitler MP, Carpenter CS. 2016. Health insurance mandates, mammography, and breast cancer diagnoses. *Am. Econ. J.: Econ. Policy* 8(3):39–68
22. Bradford AC, Bradford WD. 2016. Medical marijuana laws reduce prescription medication use in Medicare Part D. *Health Aff.* 35(7):1230–36
23. Bramson H, Jarlais DCD, Arasteh K, Nugent A, Guardino V, Feelemyer J, et al. 2015. State laws, syringe exchange, and HIV among persons who inject drugs in the United States: history and effectiveness. *J. Public Health Policy* 36(2):212–30
24. Brot-Goldberg ZC, Chandra A, Handel BR, Kolstad JT. 2017. What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics. *Q. J. Econ.* 132(3):1261–318
25. Bullinger LR. 2017. The effect of minimum wages on adolescent fertility: a nationwide analysis. *Am. J. Public Health* 107(3):447–52
26. Burlig F, Preonas L, Woerman M. 2017. *Panel data and experimental design*. Work. Pap. 277a, Energy Inst. Haas, Univ. Calif., Berkeley. **https://ei.haas.berkeley.edu/research/papers/WP277Appendix.pdf**
27. Bustamante AV, Mendez CA. 2014. Health care privatization in Latin America: comparing divergent privatization approaches in Chile, Colombia, and Mexico. *J. Health Polit. Policy Law* 39(4):841–86
28. Cameron AC, Gelbach JB, Miller DL. 2008. Bootstrap-based improvements for inference with clustered errors. *Rev. Econ. Stat.* 90(3):414–27
29. Cameron AC, Gelbach JB, Miller DL. 2011. Robust inference with multiway clustering. *J. Bus. Econ. Stat.* 29(2):238–49
30. Cameron AC, Miller DL. 2015. A practitioner's guide to cluster-robust inference. *J. Hum. Resour.* 50(2):317–72
31. Cantor J, Torres A, Abrams C, Elbel B. 2015. Five years later: awareness of New York City's calorie labels declined, with no changes in calories purchased. *Health Aff.* 34(11):1893–900
32. Card D, Shore-Sheppard L. 2004. Using discontinuous eligibility rules to identify the effects of the federal Medicaid expansion on low-income children. *Rev. Econ. Stat.* 86:752–66
33. Carpenter C, Cook PJ. 2008. Cigarette taxes and youth smoking: new evidence from national, state, and local Youth Risk Behavior Surveys. *J. Health Econ.* 27(2):287–99
34. Carpenter C, Dobkin C. 2017. The minimum legal drinking age and morbidity in the United States. *Rev. Econ. Stat.* 99(1):95–104
35. Cawley J, Frisvold DE. 2017. The pass-through of taxes on sugar-sweetened beverages to retail prices: the case of Berkeley, California. *J. Policy Anal. Manag.* 36(2):303–26

36. Chatterjee N, Shi J, García-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17:392–406
37. Cohodes SR, Grossman DS, Kleiner SA, Lovenheim MF. 2016. The effect of child health insurance access on schooling: evidence from public insurance expansions. *J. Hum. Resour.* 51(3):727–59
38. Conley TG, Taber CR. 2011. Inference with "difference in differences" with a small number of policy changes. *Rev. Econ. Stat.* 93(1):113–25
39. Cook PJ, Durrance CP. 2013. The virtuous tax: lifesaving and crime-prevention effects of the 1991 federal alcohol-tax increase. *J. Health Econ.* 32:261–67
40. Cotti C, Tefft N. 2013. Fast food prices, obesity, and the minimum wage. *Econ. Hum. Biol.* 11(2):134–47
41. Craig P, Katikireddi SV, Leyland A, Popham F. 2017. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annu. Rev. Public Health* 38:39–56
42. Currie J, Gruber J. 1996. Health insurance eligibility, utilization of medical care, and child health. *Q. J. Econ.* 111(2):431–66
43. Currie J, Gruber J, Fischer M. 1995. Physician payments and infant mortality: evidence from Medicaid fee policy. *Am. Econ. Rev.* 85(2):106–11
44. Decker S. 2009. Changes in Medicaid physician fees and patterns of ambulatory care. *INQUIRY: J. Health Care Org. Provis. Financ.* 46(3):291–304
45. Dimick JB, Ryan AM. 2014. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA* 312(22):2401–2
46. Donald SG, Lang K. 2007. Inference with difference-in-differences and other panel data. *Rev. Econ. Stat.* 89(2):221–33
47. Doudchenko N, Imbens GW. 2016. *Balancing, regression, difference-in-differences and synthetic control methods: a synthesis*. NBER Work. Pap. 22791, Cambridge, MA
48. Finkelstein A. 2007. The aggregate effects of health insurance: evidence from the introduction of Medicare. *Q. J. Econ.* 122(1):1–37
49. Garthwaite C, Gross T, Notowidigdo MJ. 2014. Public health insurance, labor supply, and employment lock. *Q. J. Econ.* 129(2):653–96
50. Gebel M, Vossemer J. 2014. The impact of employment transitions on health in Germany: a difference-in-differences propensity score matching approach. *Soc. Sci. Med.* 108:128–36
51. Goodman P, Agnew M, McCaffrey M, Paul G, Clancy L. 2007. Effects of the Irish smoking ban on respiratory health of bar workers and air quality in Dublin pubs. *Am. J. Respir. Crit. Care Med.* 175(8):840–45
52. Gruber J. 1994. The incidence of mandated maternity benefits. *Am. Econ. Rev.* 84(3):622–41
53. Gruber J, Simon K. 2008. Crowd-out 10 years later: Have recent public insurance expansions crowded out private health insurance? *J. Health Econ.* 27(2):201–17
54. Hackmann M, Kolstadt J, Kowalski A. 2015. Adverse selection and an individual mandate: when theory meets practice. *Am. Econ. Rev.* 105(3):1030–66
55. Hamersma S, Kim M. 2013. Participation and crowd out: assessing the effects of parental Medicaid expansions. *J. Health Econ.* 32(1):160–71
56. Hansen B, Sabia JJ, Rees DI. 2017. Have cigarette taxes lost their bite? New estimates of the relationship between cigarette taxes and youth smoking. *Am. J. Health Econ.* 3(1):60–75
57. Harper S, Strumpf EC, Kaufman JS. 2012. Do medical marijuana laws increase marijuana use? Replication study and extension. *Ann. Epidemiol.* 22(3):207–12
58. Heim B, Lin L. 2016. Does health reform lead to an increase in early retirement? Evidence from Massachusetts. *ILR Rev.* 70(3):704–32
59. Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* 15(3):199–236
60. Holahan J, Weil A, Wiener JM. 2003. *Federalism and Health Policy*. Washington, DC: Urban Inst. Press
61. Hoynes H, Schanzenbach DW, Almond D. 2016. Long-run impacts of childhood access to the safety net. *Am. Econ. Rev.* 106(4):903–34
62. Imbens GW, Kolesar M. 2016. Robust standard errors in small samples: some practical advice. *Rev. Econ. Stat.* 98(4):701–12

63. Imbens GW, Rubin DB. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge Univ. Press
64. Kaestner R, Garret B, Chen J, Gangopadhyaya A, Fleming C. 2017. Effects of ACA Medicaid expansion on health insurance coverage and labor supply. *J. Policy Anal. Manag.* 36(3):608–42
65. Ketcham JD, Simon KI. 2008. Medicare Part D's effects on elderly patients' drug costs and utilization. *Am. J. Manag. Care* 14(11):14–22
66. Kolstad JT, Kowalski AE. 2012. The impact of health care reform on hospital and preventive care: evidence from Massachusetts. *J. Public Econ.* 96(11–12):909–29
67. Kowalski A. 2016. *Doing more when you're running LATE: applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments*. NBER Work. Pap. 22363, Cambridge, MA
68. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. 2016. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ.* 25(12):1514–28
69. Kuehnle D, Wunder C. 2016. The effects of smoking bans on self-assessed health: evidence from Germany. *Health Econ.* 26(3):321–37
70. Lane S, Hennes E, West T. 2016. "*I've got the power": how anyone can do a power analysis of any type of study using simulation*. Presented at Soc. Personal. Soc. Psychol. Annu. Conv., San Diego. **http://meeting.spsp.org/2016/sites/default/files/Lane%2C%20Hennes%2C%20West%20SPSP%20Power%20Workshop%202016.pdf**
71. Lemos S. 2005. Political variables as instruments for the minimum wage. *B. E. J. Econ. Anal. Policy* 4(1):1–33
72. Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22
73. Madrian BC. 1994. Employment-based health insurance and job mobility: Is there evidence of job-lock? *Q. J. Econ.* 109(1):27–54
74. Manning WG, Newhouse JP, Duan N, Keeler EB, Leibowitz A. 1987. Health insurance and the demand for medical care: evidence from a randomized experiment. *Am. Econ. Rev.* 77(3):251–77
75. Marcus J, Siedler T. 2015. Reducing binge drinking? The effect of a ban on late-night off-premise alcohol sales on alcohol-related hospital stays in Germany. *J. Public Econ.* 123:55–77
76. Marks MS. 2011. Minimum wages, employer-provided health insurance, and the non-discrimination law. *Ind. Relat. J. Econ. Soc.* 50(2):241–62
77. McClelland G, Gault S. 2017. *The synthetic control method as a tool to understand state policy*. Res. Rep., Urban Inst., Washington, DC
78. Miller S. 2012. The effect of insurance on emergency room visits: an analysis of the 2006 Massachusetts health reform. *J. Public Econ.* 96(11–12):893–908
79. Moulton BR. 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Rev. Econ. Stat.* 72(2):334–38
80. Muthén LK, Muthén BO. 2009. How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.: Multidiscip. J.* 9(4):599–620
81. NAMSDL (Natl. Alliance Model State Drug Laws). 2017. *Use of marijuana for medicinal purposes: map of state laws*. NAMSDL, Manchester, IA. **http://www.namsdl.org/library/CDC6A46B-F8E2-5F41-DBF6FBE02C0BC877/**
82. Nathan RP. 2005. Federalism and health policy. *Health Aff.* 24(6):1458–66
83. Obermeyer Z, Makar M, Abujaber S. 2014. Association between the Medicare hospice benefit and health care utilization and costs for patients with poor-prognosis cancer. *JAMA* 312(18):1888–96
84. Paik M, Black B, Hyman D. 2013. The receding tide of medical malpractice litigation part 2: effect of damage caps. *J. Empir. Legal Stud.* 10:639–69
85. Paik M, Black B, Hyman DA. 2016. Damage caps and the labor supply of physicians: evidence from the third reform wave. *Am. Law Econ. Rev.* 18(2):463–505
86. Pimentel SD, Kelz RR, Silber JH, Rosenbaum PR. 2015. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Am. Stat. Assoc.* 110(510):515–27

87. Pustejovsky JE, Tipton E. 2016. Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J. Bus. Econ. Stat.* In press. **https://doi.org/10.1080/07350015.2016.1247004**
88. Raifman J, Moscoe E, Austin B, McConnell M. 2017. Difference-in-differences analysis of the association between state same-sex marriage policies and adolescent suicide attempts. *JAMA Pediatr*. 171(4):350–56
89. Rieger M, Wagner N, Bedi AS. 2017. Universal health coverage at the macro level: synthetic control evidence from Thailand. *Soc. Sci. Med.* 172:46–55
90. Rosenbaum PR. 2002. Covariance adjustment in randomized experiments and observational studies. *Stat. Sci.* 17(3):286–327
91. Saloner B, Feder KA, Krawczyk N. 2017. Closing the medication-assisted treatment gap for youth with opioid use disorder. *JAMA Pediatr*. 171(8):729–31
92. Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Co.
93. Simon D. 2016. Does early life exposure to cigarette smoke permanently harm childhood welfare? Evidence from cigarette tax hikes. *Am. Econ. J. Appl. Econ.* 8(4):128–59
94. Simon KI. 2005. Adverse selection in health insurance markets? Evidence from state small-group health insurance reforms. *J. Public Econ.* 89(9–10):1865–77
95. Simon KI, Kaestner R. 2004. Do minimum wages affect non-wage job attributes? Evidence on fringe benefits. *ILR Rev*. 58(1):52–70
96. Simon K, Soni A, Cawley J. 2017. The impact of health insurance on preventive care and health behaviors: evidence from the first two years of the ACA Medicaid expansions. *J. Policy Anal. Manag.* 36(2):390–417
97. Snow J. 1855. *On the Mode of Communication of Cholera*. London: John Churchill
98. Solon G, Haider SJ, Wooldridge JM. 2015. What are we weighting for? *J. Hum. Resour.* 50(2):301–16
99. Somers MA, Zhu P, Jacob R, Bloom H. 2013. *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation*. Work. Pap. Res. Methodol., MDRC, New York. **https://www.mdrc.org/sites/default/files/validity_precision_comparative_interrupted_time_series_design.pdf**
100. Sommers BD, Gunja MZ, Finegold K, Musco T. 2015. Changes in self-reported insurance coverage, access to care, and health under the Affordable Care Act. *JAMA* 314(4):366–74
101. Sommers BD, Kronick R. 2012. The Affordable Care Act and insurance coverage for young adults. *JAMA* 307(9):913–14
102. Trogdon JG, Shafer PR, Shah PD, Calo WA. 2016. Are state laws granting pharmacists authority to vaccinate associated with HPV vaccination rates among adolescents? *Vaccine* 34(38):4514–19
103. Wehby G, Dave D, Kaestner R. 2016. *Effects of the minimum wage on infant health*. NBER Work. Pap. 22373, Cambridge, MA
104. Wing C, Cook TD. 2013. Strengthening the regression discontinuity design using additional design elements: a within-study comparison. *J. Policy Anal. Manag.* 32(4):853–77
105. Wolfe B, Scrivner S. 2005. The devil may be in the details: how the characteristics of SCHIP programs affect take-up. *J. Policy Anal. Manag.* 24(3):499–522
106. Xiao L, Jiang Y, Liu X, Li Y, Gan Q, Liu F. 2017. Smoking reduced in urban restaurants: the effect of Beijing Smoking Control Regulation. *Tob. Control* 26:e75–78