Article:
O'Neill, Stephen, Kreif, Noemi, Sutton, Matt et al. (1 more author) (2020) A comparison of methods for health policy evaluation with controlled pre-post designs. Health services research. ISSN 1475-6773

https://doi.org/10.1111/1475-6773.13274

DOI: 10.1111/1475-6773.13274

RESEARCH ARTICLE

HSR Health Services Research

A comparison of methods for health policy evaluation with controlled pre-post designs

Stephen O'Neill PhD[1,2] | Noemi Kreif PhD[2,3] | Matt Sutton PhD[4,5] | Richard Grieve PhD[2]

[1]J.E. Cairnes School of Business and Economics, National University of Ireland Galway, Galway, Ireland

[2]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK

[3]Centre for Health Economics, University of York, York, UK

[4]Health Organisation, Policy and Economics, School of Health Sciences, University of Manchester, Manchester, UK

[5]Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, Victoria, Australia

Correspondence
Stephen O'Neill, PhD, JE Cairnes School of Business and Economics, National University of Ireland Galway, Galway, Ireland.
Email: stephen.oneill@nuigalway.ie

Funding information
Medical Research Council, Grant/Award Number: MR/L012332/1; National Institute for Health Research, Grant/Award Number: SRF-2013-06-016

Abstract

Objective: To compare interactive fixed effects (IFE) and generalized synthetic control (GSC) methods to methods prevalent in health policy evaluation and re-evaluate the impact of the hip fracture best practice tariffs introduced for hospitals in England in 2010.

Data Sources: Simulations and Hospital Episode Statistics.

Study Design: Best practice tariffs aimed to incentivize providers to deliver care in line with guidelines. Under the scheme, 62 providers received an additional payment for each hip fracture admission, while 49 providers did not. We estimate the impact using difference-in-differences (DiD), synthetic control (SC), IFE, and GSC methods. We contrast the estimation methods' performance in a Monte Carlo simulation study.

Principal Findings: Unlike DiD, SC, and IFE methods, the GSC method provided reliable estimates across a range of simulation scenarios and was preferred for this case study. The introduction of best practice tariffs led to a 5.9 (confidence interval: 2.0 to 9.9) percentage point increase in the proportion of patients having surgery within 48 hours and a statistically insignificant 0.6 (confidence interval: −1.4 to 0.4) percentage point reduction in 30-day mortality.

Conclusions: The GSC approach is an attractive method for health policy evaluation. We cannot be confident that best practice tariffs were effective.

KEYWORDS

difference-in-differences, interactive fixed effects, pay-for-performance, policy evaluation, synthetic control

1 | INTRODUCTION

Health policy evaluations commonly use data before and after a policy change and assume that, without the intervention, the expected outcomes for the treated and control groups would have followed parallel trends. This assumption underpins the standard difference-in-differences (DiD) estimator and implies that any differences between the comparator groups due to unobserved confounders are time-constant. However, the "parallel trends" assumption is often implausible, particularly in a health policy setting. When the parallel trends assumption is violated, DiD approaches provide biased estimates of the effect of the health policy.[1,2] DiD has been widely applied to policy evaluations within health economics[4-10] and health services research.[11-15] As recently illustrated in re-evaluating

a pay-for-performance (P4P) scheme,[3] a study's policy conclusions can rest on the approach taken to causal inference.[3]

The synthetic control (SC) method[16,17] has been viewed as an attractive alternative to DiD as it avoids the parallel trends assumption. In essence, the SC method constructs a comparator for the intervention group, the synthetic control, as a weighted average of the available control units. Each unit is weighted to ensure that the mean outcomes of the synthetic control track those of the treated unit(s) prior to the intervention.[3,17-24] However, despite its wide use, critics have shown that the SC approach may provide biased estimates in settings when few pre-intervention periods are available[2,25]; treatment assignment is correlated with time-varying unobserved confounders,[26] or where the outcomes of the treated units cannot be obtained by weighting the control units' outcomes by values between 0 and 1 (ie, the treated units are not within the "convex hull"), leading to poor overlap. [17,25,27] Statistical inference is also somewhat problematic under the SC approach.[28] Concerns about the DiD and SC approaches have encouraged recent methodological advances.[29-35] However, these methods have not been considered in the health policy evaluation domain, which is characterized by particular challenges, notably the (im)plausibility of the parallel trends assumption, the possibility of heterogeneous treatment effects, and that there may be few pretreatment periods. Here, we consider two of these approaches: (a) interactive fixed effect (IFE) models, and (b) the generalized synthetic control (GSC) method, both are novel to this context.

IFE models are flexible regression approaches that allow for multiple time-constant unobserved covariates, each of which may have effects that vary across time[36-39] relaxing the parallel trends assumption.[40] IFE models nest the fixed effects models routinely used within DiD estimation, but may produce biased estimates when policy effects are modified by unobserved covariates, that is effects are heterogeneous.[41] For instance, hospital quality, which is generally unobserved, may moderate the effect that a new health policy has on outcomes.

The GSC method[41] seeks to overcome this limitation by combining insights from the SC literature with the efficiency gains of IFE models. The GSC approach allows a separate (counterfactual) potential outcome to be estimated for each treated unit, allowing heterogeneous treatment effects to be consistently estimated. It has been argued that the GSC method maintains the approximately unbiasedness property of the SC estimator but offers improved efficiency. Despite these desirable features, the GSC method has not been considered in a published health policy evaluation.[a]

We contrast the IFE and GSC methods with DiD and SC methods in a case study and in Monte Carlo simulations. We revisit an evaluation of a pay-for-performance scheme, best practice tariffs (BPT) for hip fractures, introduced for hospitals in the English NHS.[2,42] The incidence of hip fractures in the UK is rising annually and is currently estimated at 10.2 per 10 000 per year.[43] The cost to the hospital services of hip fracture are substantial, and have been estimated to be £1,131 million in the year of the fracture.[44] Thus the impact of policies such as BPT are of interest to policymakers. Our simulation study extends the precedent comparison of Xu,[41] by considering settings relevant to the HSR context, namely few (<10) pretreatment periods, highly imbalanced numbers of treatment vs control units, and serial correlation. While all methods were susceptible to shocks that impacted treated and control units differently in the post-treatment period, the simulations show that the IFE approach otherwise avoids bias when treatment effects are homogenous but provides biased estimates under heterogeneity. By contrast, the GSC method reports efficient estimates with low bias in the presence of nonparallel trends, heterogeneous effects, and relatively few pretreatment periods.

What this study adds

- Health policy evaluations with pre-post designs are challenging as the parallel trends assumption underlying difference-in-differences estimation often does not hold for all outcomes.
- This was the case for the evaluation of the best practice tariffs (BPT) for hip fractures, a pay-for-performance scheme, introduced for hospitals in the English NHS.
- Alternative estimation methods have yielded contrasting estimates of the impacts of this BPT.
- In our simulations, the generalized synthetic control approach outperformed more commonly used methods (difference-in-differences and synthetic control methods) and hence was the preferred approach for the case study.
- It suggests that the BPT for hip fractures increased the proportion of patients who had surgery within 48 hours of admission, but did not statistically significantly reduce 30-day mortality.

2 | MOTIVATING EXAMPLE: EVALUATION OF A BEST PRACTICE TARIFFS SCHEME (BPT)

Hospital pay-for-performance (P4P) schemes link a portion of provider income to achieving predefined quality targets. These schemes intend to encourage providers to engage in "desirable" behaviors. However, P4P schemes may shift resources toward rewarded vs unrewarded dimensions of care quality, and so have negative spill-over effects.[45] A number of studies have concluded that hospital pay-for-performance schemes have not had the desired impact.[14,46-51] The international evidence on P4P has been criticized for failing to provide reliable estimates of these schemes' relative effectiveness.[52-54]

The particular P4P scheme considered here, the BPT for hip fractures, was introduced for participating English NHS hospitals from April 2010,[2,42] who were paid a fixed sum, set at £445 in the 2010/11 financial year,[55] for each hip fracture admission if certain conditions

representing "best practice" were met.[b] The BPT payments represented a considerable share of the total payment to providers for hip fracture care, 14% in 2011/12,[55] so one might anticipate that providers would respond to these altered incentives to provide best practice care.

A published survey and qualitative interviews suggested that BPT participation was influenced by factors unobserved by researchers[42c], such as the resources required for this scheme, the quality of facilities available, and the expected benefits from participation. These may have had time-varying effects on the outcomes. Hence, a priori, it was unclear whether the parallel trends assumption held for each outcome. For one outcome, the proportion of patients who had surgery within 48 hours, the parallel trends assumption appeared plausible (Figure 1), and tests suggested this assumption could not be rejected (P = .9255).[d] However, for the primary outcome, mortality within 30 days, the parallel trends assumption appeared less plausible (Figure 2) and the null hypothesis of parallel trends was rejected (P = .039).

Previous analyses, using DiD and SC methods, found that conclusions regarding the effects of the BPT differed by method.[2] Estimates based on DiD reported that the introduction of BPTs led to a statistically significant reduction in mortality, whereas the SC method failed to reject the null of no effect across all outcomes and indicated a smaller impact on mortality compared to DiD. However, the authors raised concerns regarding the efficiency of the SC estimates, motivating this re-analysis using alternative methods.

We re-analyze the data used in a previously published study,[2] consisting of hospital admissions from 62 hospital trusts that reported receiving at least some BPT payments (treated group) and 49 trusts that reported receiving no payments under the scheme (control group). Panel data were available for twelve quarters before, and four after, the scheme's introduction. All analyses were conducted at the level of the hospital-quarter.

The outcomes considered are the proportion of patients receiving surgery within 48 hours of an emergency admission and the proportion of patients that die within 30 days of admission. We adjust for baseline covariates according to age group, gender, and source of admission.

3 | METHODS

Suppose there are $i = 1,\ldots,n$ units, and T time periods, where $t = 1,\ldots t'$ are pretreatment, and $t' + 1,\ldots,T$ are post-treatment. The potential outcomes[56] for unit i in period t in the presence and absence of treatment are denoted by Y_{it}^1 and Y_{it}^0, respectively. Let D_{it} be an indicator equal to one if unit i is treated (exposed to the policy) in period t and zero otherwise. The observed outcome can be written as:

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$$

We assume the following factor model for the potential outcome in the absence of treatment:

$$Y_{it}^0 = X_{it}'\beta + \left(\lambda_{1t}\mu_{i1} \ldots + \lambda_{Rt}\mu_{iR}\right) + \varepsilon_{it}$$

where $\boldsymbol{X}_{it}$ is a $(1 \times k)$ vector of observed time-varying covariates, β is the $(k \times 1)$ vector of their coefficients, assumed to be the same for both groups, μ_{ir} $(r = 1, \ldots, R)$ represents an unobserved time-invariant variable with λ_{rt} capturing the effect of that unobserved variable in period t, and ε_{it} represents exogenous, unobserved idiosyncratic shocks. Allowing for an additive treatment effect that may differ by individual

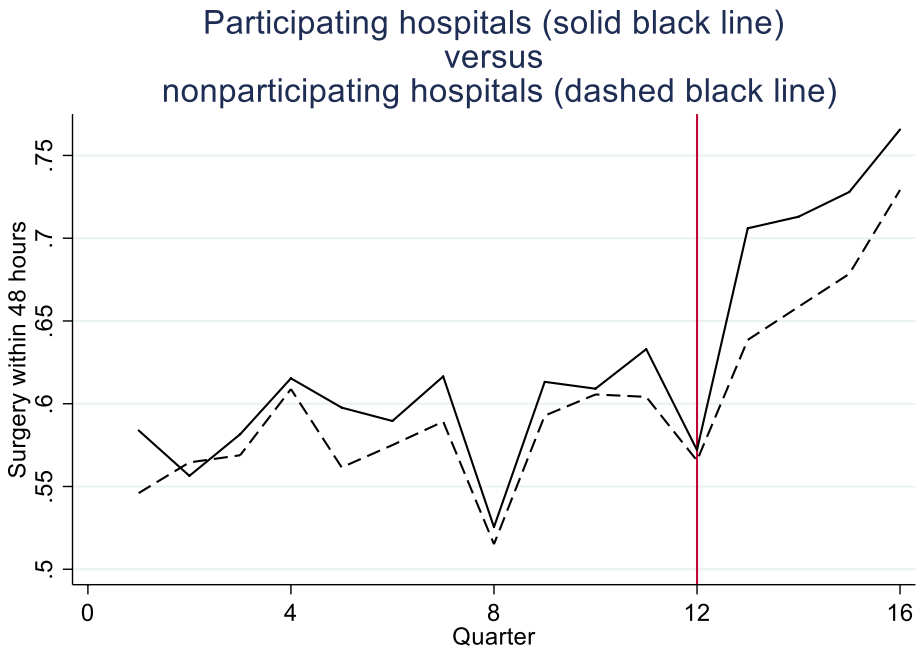


FIGURE 1 Proportions of hip fracture patients receiving surgery within 48 h of emergency admission in participating vs nonparticipating hospitals [Color figure can be viewed at wileyonlinelibrary.com]

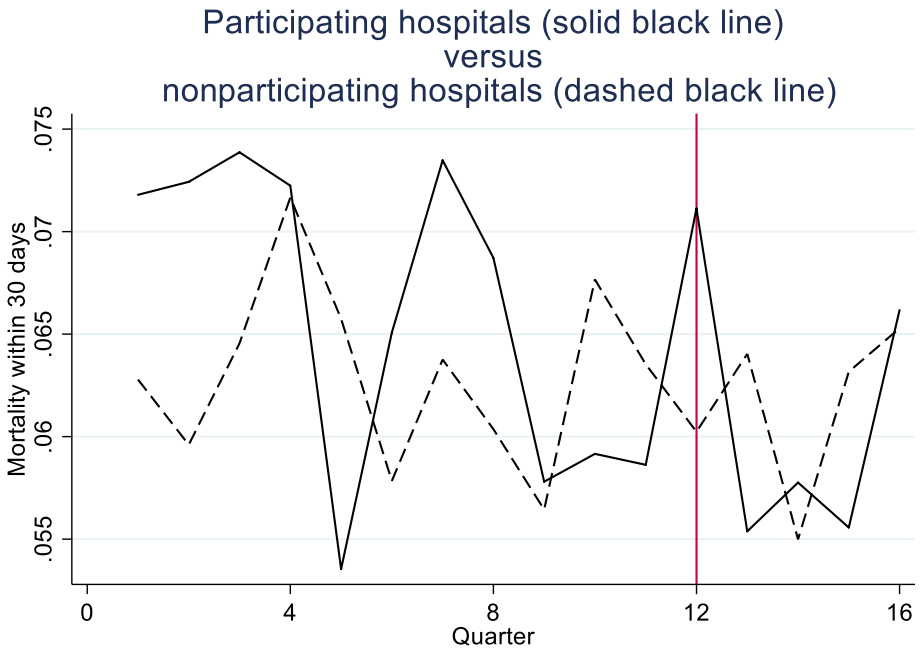


FIGURE 2 Proportions of hip fracture patients dying within 30 d of emergency admission in participating vs nonparticipating hospitals Color figure can be viewed at wileyonlinelibrary.com]

and period (τ_{it}), and letting $\boldsymbol{\mu}_i = [\mu_{i1},..., \mu_{iR}]$ and $\lambda_t = [\lambda_{1t},..., \lambda_{Rt}]$, the observed outcome can be written as:

$$Y_{it} = \boldsymbol{X'}_{it}\beta + \lambda_t'\boldsymbol{\mu}_i + D_{it}\tau_{it} + \varepsilon_{it} \quad (1)$$

The estimand of interest is the average treatment effect for the treated (ATT) after controlling for covariates, $E(\tau_{it}|D_{it} = 1, X_{it})$ over the post-treatment period, $t > t'$.

3.1 | Difference in Differences (DiD)

Note that if $\boldsymbol{\mu_i} = [1, \mu_i]$ and $\boldsymbol{\lambda_t} = [\lambda_t, 1]$, equation 1 would correspond to a two-way fixed effects model:

$$Y_{it} = \boldsymbol{X'}_{it}\beta + \mu_i + \lambda_t + D_{it}\tau_{it} + \varepsilon_{it} \quad (2)$$

In this case, the parallel trends assumption will hold[57,58]:

$$E\left(Y_{it}^0 - Y_{it'}^0 | D_{it} = 1, X_{it}\right) = E\left(Y_{it}^0 - Y_{it'}^0 | D_{it} = 0, X_{it}\right) \forall t > t' \text{ (A1: Parallel trends).}$$

where t' represents the final pretreatment period, and the conditional ATT can be estimated using DiD with two-way fixed effects regression.[24,59-61ef]

3.2 | Interactive fixed effects

Interactive fixed effects models rely on an alternative set of estimation approaches for the common factor structure $\boldsymbol{\lambda_t'\mu_i}$.[37] Here, we estimate the IFE model using the iterative principal component estimator.[37] This approach consists of iterating between (a) estimating $\boldsymbol{\lambda_t}$ and $\boldsymbol{\mu_i}$ using principal components while holding $\hat{\boldsymbol{\beta}}$ constant, and (b) estimating $\boldsymbol{\beta}$ by regressing $(\mathbf{Y} - \hat{\lambda}_t'\hat{\boldsymbol{\mu}}_i)$ on $\mathbf{X}$, until convergence is achieved. The number of factors to include can be chosen according to cross-validation as described in Algorithm 1 in Xu.[41] It is preferable to include too many rather than too few factors.[62]

One limitation of the IFE approach is that when treatment effects are moderated by the unobserved factors, the estimated average treatment effect may be biased, since the heterogeneity in treatment effects leads to biased estimates of the common factors and hence the implied treatment-free potential outcome.

3.3 | Synthetic control (SC) method

The synthetic control method has been shown to provide an approximately unbiased estimator of the ATT for a treated unit[17] when outcomes are determined by a linear factor model with time-invariant covariates (Z_i), such as:

$$Y_{it} = \theta_t \boldsymbol{Z}_i + \boldsymbol{\lambda_t'\mu_i} + D_{it}\tau_{it} + \varepsilon_{it} \quad (3)$$

The SC method aims to estimate the unit level causal effect τ_{it} for the treated unit, by constructing a "synthetic control," or a weighted average of the control units that has similar outcomes and observed covariates to the treated unit over the pre-intervention period:

$$\sum_{j \in Control} w_j Y_{jt} \approx Y_{1t}, \forall t \le T_0 \quad \text{and} \quad \sum_{j \in Control} w_j \boldsymbol{Z}_i \approx \boldsymbol{Z}_1, \forall t \le T_0$$

where w_j is an element of **W** representing the weight for control *j*, with $0 \le w_j \le 1$. The synthetic control is formed by finding the vector of weights **W** that minimizes $(X_1 - X_0 W)' V (X_1 - X_0 W)$ subject to the weights in *W* being positive and summing to 1, where X_1 and X_0 contain the pretreatment outcomes and covariates for the treated unit and control units, respectively,[17] and ***V*** captures the relative importance of these variables as predictors of the outcome of interest. When X_1 and X_0 include all of the pre-intervention outcomes, other covariates do not influence the weights and hence can be excluded as is done in our analysis below. If the synthetic control and treated unit have similar outcomes over an extended pre-intervention period, it is plausible that they have similar observed and unobserved predictors of the outcome.[25] Hence, the postintervention outcome for the synthetic control represents the counterfactual treatment-free potential outcome for the treated unit ($\hat{Y}^0_{1t}$). The SC method assumes conditional independence[2]/ignorability[74]:

$$Y^0_{it} \perp D_{it} \mid \left(Y^0_{ih}\right) \quad \left(\text{A2: Independence conditional on past outcomes}\right).$$

where Y^0_{ih} is a vector of potential outcomes in the *h* time periods prior to treatment.

Since the weights are restricted to be between 0 and 1, the treated unit must lie within the "convex hull" of the control units to avoid bias.[17] The treatment effect for the treated unit ($i = 1$), τ_{1t}, can be estimated by $(Y'_{1t} - \hat{Y}^0_{1t})$ for each postintervention period separately, and these can be averaged over time to obtain an ATT over the postintervention period.

The SC approach can be applied to multiple treated units by applying the method to each treated unit or, as we do here, averaging across the sample of treated units to obtain a single treated unit.[18,20]

3.4 | Generalized synthetic control (GSC) method

The GSC approach[41] assumes that treatment assignment is independent of potential outcomes conditional on the observed covariates, and *R* orthogonal, unobserved latent factors ($\boldsymbol{\lambda_t} = \lambda_{t1}, \ldots, \lambda_{tR}$) and their factor loadings ($\boldsymbol{\mu_i} = \mu_{i1}, \ldots, \mu_{iR}$)[41]:

$$\left\{Y^1_{it}, Y^0_{it}\right\} \perp D_i \mid \boldsymbol{X_{it}}, \boldsymbol{\lambda_t}, \boldsymbol{\mu_i}$$

which implies that

$$E\left(Y^1_{it}, Y^0_{it} \mid D_i = 1, \boldsymbol{X_{it}}, \boldsymbol{\lambda_t}, \boldsymbol{\mu_i}\right) = E\left(Y^1_{it}, Y^0_{it} \mid D_i = 0, \boldsymbol{X_{it}}, \boldsymbol{\lambda_t}, \boldsymbol{\mu_i}\right) \qquad \text{(A3)}$$

This will hold true if the same IFE data generating process, such as equation 1 above, underlies outcomes for the treated and the control units. The key difficulty in estimating the unobserved treatment-free potential outcome of the treated units in the post-treatment periods is estimating $\boldsymbol{\lambda_t}$ for the post-treatment period and $\boldsymbol{\mu_i}$ for each treated unit. The GSC approach tackles these difficulties as follows:[g]

First, an IFE model, $Y^0_{it} = \boldsymbol{X_{it}}\boldsymbol{\beta} + \boldsymbol{\lambda_t}\boldsymbol{\mu_i} + \varepsilon_{it}$, is estimated for the control units only, for the entire sample period, yielding estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{t}})$ for the control units. Since $\tau_{it} D_{it}$ is zero in equation 1 for the control units, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{t}})$ are consistent estimates of $(\boldsymbol{\beta}, \boldsymbol{\lambda_t})$, which are assumed to be the same for the treated and control units. If we knew $\boldsymbol{\mu_i}$ for the treated units, we could use our estimates from the control group $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{t}})$ to predict the post-treatment treatment-free potential outcome for the treated unit using:

$$\hat{Y}'_{it} = \boldsymbol{X_{it}}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\lambda}}_{\boldsymbol{t}}\boldsymbol{\mu_i} \qquad (4)$$

Since we do not know $\boldsymbol{\mu_i}$ for each treated unit, the GSC method finds the value, $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$, that minimizes the pretreatment discrepancy between the observed outcome and the predicted outcome for a given treated unit, based on [4].[41] Using the estimates for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}_{\boldsymbol{t}}$ from the control units and the resulting prediction $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$ for the treated unit, we can estimate the treatment-free potential outcome for the treated units as:

$$\hat{Y}^0_{it} = \boldsymbol{X_{it}}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\lambda}}_{\boldsymbol{t}}\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} \qquad (5)$$

The estimated treatment-free potential outcomes after the program starts can be compared to the actual outcomes for the treated units to obtain an estimated treatment effect $\hat{\tau}_{it} = (Y'_{it} - \hat{Y}^0_{it})$ for each unit in each period. Since, unlike the IFE approach, estimates of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\lambda}}_{\boldsymbol{t}}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$ do not depend on post-treatment information for the treated units, $\hat{\tau}_{it}$ is not biased by heterogeneous treatment effects.

As with the SC method, when the number of pretreatment periods is small, it becomes harder to distinguish between $\boldsymbol{\mu_i}$ and ε_{it}, which can lead to biased estimates of the treatment effect. This bias shrinks to zero as both the number of pretreatment periods and the size of the control group grow.[41] Unlike the SC method, the GSC method conveniently allows for time-varying observed covariates. The GSC approach requires data be available for *R + 1* pre-intervention periods.[h]

4 | IMPLEMENTING THE METHODS IN THE RE-ANALYSIS OF BPT FOR HIP FRACTURES

We replicated the DiD and SC estimations reported in a previously published study.[2] The DiD estimation was undertaken at the hospital-level and controlled for covariates (age, gender, source of admission), together with two-way fixed effects for time periods and hospitals. The SC method averaged the treated units to define a single treated unit, and a synthetic control was formed from the control units. In our implementation of the SC method, we included all of the pre-intervention outcomes as separate variables in the X_0 and X_1 matrices. The variable weights were determined simultaneously with the synthetic control weights[17] as implemented in the Stata package *synth*.

The IFE model was estimated using the iterative principal component estimator.[37] In our implementations of IFE and GSC, we included the time-varying covariates in the IFE model, two-way fixed effects, and up to five interactive fixed effects with the number

chosen by cross-validation, following Algorithm 1 in Xu.[41] For inference, we used a parametric bootstrap with 500 replications.

For each method, we report *p*-values using the most common approach to inference for each approach, but recognizing that there are differences across methods that limit comparability of the resultant *p*-values across methods.[i] For the SC method, we use placebo tests for inference[2,17]; for the GSC method, we use a bootstrap approach[41]; and for the DiD and IFE methods, we report *p*-values based on cluster-robust standard errors.

5 | SIMULATION STUDY

We compare the methods in a Monte Carlo Simulation study where the true ATT is known and contrast the approaches according to mean bias (%) and RMSE. Building from the case study, we create 500 datasets of 111 units, of which 62 (49) were assigned to treatment (control) as in the case study[j] and simulate data for up to 22 periods, with four of these assigned to be post-treatment. The data generating process (DGP) includes one observed covariate (X_{it}), 2-way additive fixed effects (μ_{i1} and λ_{1t}), and a further two interacted factors and an additive treatment effect:

$$Y_{it} = X_{it}\beta + \mu_{i1} + \lambda_{1t} + \lambda_{2t}\mu_{i2} + \lambda_{3t}\mu_{i3} + D_{it}\tau_{it} + \varepsilon_{it}$$

We draw X_i, μ_{i1}, μ_{i2}, and μ_{i3} from a standard multivariate normal distribution and λ_{1t} from a uniform(0,5) distribution.[k] To create a time-varying X_{it}, we then define $X_{it} = 0.5X_i + 0.5 * N(0,1)$. Here, ε_{it} is a standard normally distributed idiosyncratic error term. To introduce imbalance between the treated and control groups, the means of μ_{i1}, μ_{i2}, and μ_{i3} are set two standard deviations higher for the treated units than for the controls. In scenario A, we ensure the parallel trends assumption holds by setting $\lambda_{2t} = \lambda_{3t} = 0$, so the DGP becomes a standard two-way fixed effects model. In scenario B, we allow for monotonically increasing nonparallel trends by setting $\lambda_{2t} = 0.2 * t$ and $\lambda_{3t} = 0.1 * t$.

The performance of the SC method in scenario B may be negatively affected by our inclusion of time-varying covariates (X_{it}) since the SC weights are time-invariant, and by the imbalance in $\boldsymbol{\mu}$ leading to treated units that lie outside of the convex hull of the controls. Scenario C represents a setting without these specific challenges. Here, we use X_i in place of X_{it} so that we have time-invariant covariates, and to ensure that the average treated unit lies in the convex hull of the controls, for 25% of the control units we increase μ_{i2} and μ_{i3} by 4 standard deviations so that these unit's outcomes are likely to lie above those of the average treated unit, while the remaining 75% of controls tend to lie below. In scenario D, we include an additional postintervention shock, $\Delta\varepsilon_{it} = 2$, that only affects the treated group.

We consider scenarios (A1, B1, C1, and D1) where the treatment effect is homogenous (τ_{it}=1), and otherwise identical scenarios (A2, B2, C2 & D2) with a heterogeneous treatment effect, in which we define $\tau_{it} = (1 + (\mu_{i1} - 2))$.[l] We then apply each method to estimate the average treatment effect for the treated group as a whole over the postintervention period. We consider the methods' performance across pretreatment periods of different lengths (6, 9, 12, and 18 periods). Finally, we assess the impact of imbalance in the numbers of treated (n = 10) vs control (n = 100) units (scenario E; Appendix S1).

6 | RESULTS

6.1 | Case study results

The estimated effects of the introduction of the BPT for hip fractures according to method are reported in Table 1. For both endpoints, the IFE method reports that the magnitude of the effect of BPT is larger than for the other methods. However, since differences in unobserved covariates, such as hospital quality, are likely to modify the effects of the policy, this may reflect bias due to heterogeneous treatment effects.

The DiD, SC, and GSC methods provide similar point estimates. The *p*-values do differ somewhat across the approaches, but the interpretation of these differences must recognize that the SC approach to inference differs to the other methods. The GSC method reports that the introduction of BPT increases the proportion of patients who have surgery within 48 hours, and suggests that the scheme leads to a reduction in mortality although this difference is not statistically significant.

6.2 | Simulation results

Figure 3 presents boxplots of the simulation estimates for each scenario by method, while Table 2 reports the corresponding mean bias

TABLE 1 Best Practices Tariffs case study results: ATT on process and outcome measures according to method

	Difference-in-differences	Synthetic controls	Interactive fixed effects	Generalized synthetic controls
Surgery within 48 h	0.0403	0.0482	0.0647	0.0590
	(*P* = .196)	(*P* = .250)	(*P* = .004)	(*P* = .004)
Dead within 30 d	−0.0080	−0.0051	−0.0123	−0.0062
	(*P* = .037)	(*P* = .560)	(*P* < .001)	(*P* = .308)

Note: For difference in differences, O'Neill et al[2] report *p*-values based on cluster robust standard errors. For the synthetic control method, *p*-values are based on placebo tests as described in O'Neill et al[2]; for interactive fixed effects, we report *p*-values based on cluster robust standard errors and the generalized synthetic control approach uses a bootstrap approach as described in Xu.[41]

(%) and root mean squared error (RMSE). We begin by considering the scenarios where effects are homogenous (scenario A1, B1, C1, and D1, panel (a) of Figure 3). As expected if the parallel trends assumption holds, DiD performs best (scenario A1), although IFE and GSC perform almost as well (Table 2, Figure 3(I)). By contrast, SC performs poorly, providing biased estimates attributable to the average treated unit tending to lie outside the convex hull of controls. Where the parallel trends assumption fails (scenario B1), DiD provides biased estimates, whereas IFE and GSC report minimal bias (Table 2, Figure 3(ii)). The SC method again provides biased estimates. In scenario C1, the performance of the SC method improves markedly (Table 2, Figure 3(iii)) since here the treated units tend to lie inside the convex hull of the controls. When a shock has a differential effect for the treated vs control group in the postintervention period (scenario D1), all methods provide biased estimates (Table 2, Figure 3(iv)).

In those scenarios with heterogeneous treatment effects (scenarios A2, B2, and C2, panel (b) of Figure 3), the GSC method continues to perform well, providing estimates with low bias and low RMSE (Table 2, Figure 3(i), (ii) (iii), (iv)). DiD, IFE, and SC all report biased estimates. For DiD, the bias is due to the failure of the parallel trends assumption. For the IFE model, the heterogeneous treatment effect biases the estimated values for $\lambda_t \mu_i$, which in turn biases the treatment-free potential outcome and ultimately the ATT. For the SC method, the bias is attributable to poor overlap and is mitigated when the treated units lie in the convex hull of the controls (scenario C2). In scenario D2, all methods again report bias due to the postintervention shock.

7 | DISCUSSION

This paper critically assesses two causal inference approaches, IFE and GSC methods, new to health policy evaluation, and contrasts them with DiD estimation and the SC method. The paper extends previous papers in the health policy and political science literatures[4-15,41,74] in contrasting IFE and GSC, but also approaches often considered in the HSR literature (DID and SC). Rather than focus solely on simple scenarios,[41] the paper considers a range of settings relevant to the HSR context, including homogeneous and heterogeneous treatment effects, parallel tends and nonparallel tends, highly imbalanced numbers of treatment and control units, serial correlation, and idiosyncratic shocks. While our paper underscores the main finding from Xu's early simulation study,[41] that GSC performs better than IFE when there is treatment effect heterogeneity, it offers a wider set of insights into the relative performance of GSC vs alternative methods in settings of direct relevance to the HSR context.

Our re-evaluation of the BPT scheme exemplifies many critical issues faced in health policy evaluations. Here, there are multiple outcomes with the parallel trends assumption plausible for some but not others; the effects of the policy are anticipated to differ across hospitals; and data are only available for relatively few periods pre-intervention. An attractive feature of the IFE and GSC methods is that they allow the analyst to adopt a consistent analytical approach across all outcomes, as their factor structure allows greater flexibility in controlling for unobserved confounders. However, the IFE estimator assumes homogenous treatment effects, which is unlikely in this study. Here, the GSC method is preferred in light of its robustness to the assumption of parallel/nonparallel trends and homogeneous/heterogeneous effects. It reported that BPT led to a large[m] and statistically significant increase in the proportion of patients who had surgery within 48 hours of admission, together with a small, but not statistically significant, reduction in 30-day mortality.

The simulation study found that the GSC approach performed better than the alternatives considered across a range of challenging settings typically faced in health economic and policy evaluations that use routine data, namely nonparallel trends, heterogeneous treatment effects, and few (6) pre-intervention periods. However, when deciding which methods to apply to a particular setting, it is important to consider the underlying theory and requirements of the method. In particular, GSC and IFE approaches both require repeated observations of the same units over time (ie, panel data) and also require data for multiple pre-intervention periods (one more than the specified number of interactive fixed effects to include).

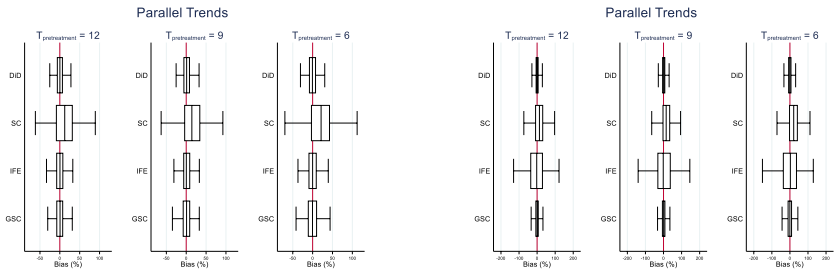
Generalized synthetic control reports relatively precise estimates across all these challenging settings. We find the method performs well even if there is limited support for particular underlying causal assumptions (eg, parallel trends). In light of this, for the case study, which has some of these features, we emphasize the policy conclusions from the GSC approach, which is that the BPT intervention increased the probability of surgery within 48 hours, but does not lead to a change in 30-day mortality. We also contribute to the growing literature that critically evaluates the SC method.[2,26] We extend O'Neill et al[2] in recognizing that the SC method can perform badly if there is poor overlap in the pretreatment outcomes between the treated and control units, specifically when treated units lie outside the convex hull of the controls[17,41n]. Conversely, we highlight that the SC method can perform well provided the treated observations do lie within the convex hull of the controls. Hence, future studies should consider carefully whether their evaluations have these features before opting for SC as an alternative for DiD estimation.

This paper has the following limitations. First, each of the methods considered assumes that idiosyncratic shocks postintervention have the same expected effect on outcomes for the treated and control groups. Similarly, while any of these approaches can incorporate individual-level baseline information, for example, on patient case mix, by "risk adjusting" outcomes, unobserved compositional changes in the postintervention period may be wrongly attributed to the effect of the intervention. Second, to aid transparency, the Monte Carlo simulation study had a relatively simple DGP and assumed the IFE models including the one underlying the GSC method were correctly specified. A natural next step would be to contrast the IFE and GSC approaches to other relatively untested methods from the general causal inference literature.[29,31-33] Third, in empirical studies the methods would ideally be contrasted by applying

(i) Scenarios A: parallel trends and,

(a) A1: homogenous treatment effects

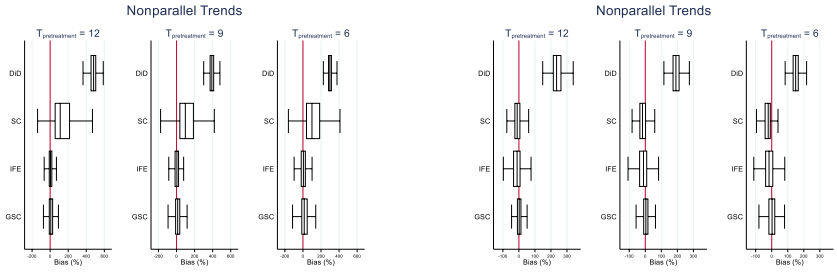
(b) A2: heterogeneous treatment effects



(ii) Scenario B: nonparallel trends

(a) B1: homogenous treatment effects

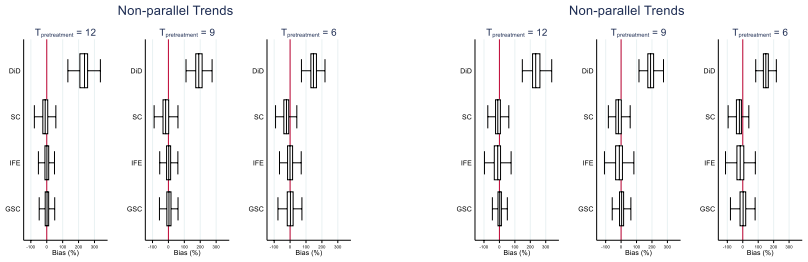
(b) B2: heterogeneous treatment effects



(iii) Scenario C: nonparallel trends, time-invariant covariates and treated units lying inside the convex hull of controls

(a) C1: homogenous treatment effects

(b) C2: heterogeneous treatment effects



(iv) Scenario D: nonparallel trends with group specific shock post-intervention

(a) D1: homogenous treatment effects

(b) D2: heterogeneous treatment effects

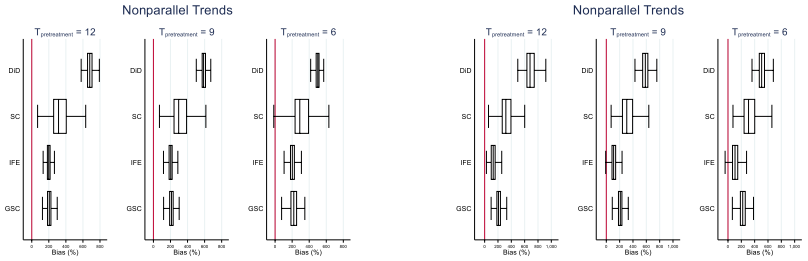


FIGURE 3 Boxplot of mean % bias in treatment effect estimates from Monte Carlo simulation. (i) Scenarios A: parallel trends, (ii) Scenario B: non-parallel trends, (iii) Scenario C: non-parallel trends, time-invariant covariates and treated units lying inside the convex hull of controls, (iv) Scenario D: non-parallel trends with group specific shock post-intervention. *Note*: 500 simulations. *T*pretreatment is the number of pre-treatment periods. Abbreviations: DiD, difference in differences; GSC, generalised synthetic control; IFE, interactive fixed effecs; SC= Synthetic control method [Color figure can be viewed at wileyonlinelibrary.com]

the same randomization inference procedure. The Conley-Taber randomization inference procedure has been recommended for this purpose, but requires the same number of observations across the treated and control groups.[63]

The findings from this paper and ongoing methods development more widely highlight two complementary areas for further research. First, a number of extensions to DiD have been proposed to increase the validity of DiD-type estimators including: allowing for unit-specific trends,[64,65] combining matching with DiD,[66] and combining instrumental variables (IV) approaches with DiD.[67] While combining IV with DiD would allow for unobserved confounding, the population this estimate relates to (compliers) may not be of policy relevance.

TABLE 2 Monte Carlo simulation study results by method and scenario

	Root mean squared error							
Scenario	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**D1**	**D2**
Parallel trends	Holds	Holds	Fails	Fails	Fails	Fails	Fails	Fails
Homogenous treatment effects	Yes	No	Yes	No	Yes	No	Yes	No
Time-invariant covariates and treated units in convex hull	No	No	No	No	Yes	Yes	No	No
Group-specific shock postintervention	No	No	No	No	No	No	Yes	Yes
18 pre-treatment periods								
Difference-in-differences	0.01	0.01	43.74	43.74	10.78	10.78	74.25	74.25
Synthetic controls	0.09	0.09	4.46	4.46	0.06	0.06	14.88	14.88
Interactive fixed effects	0.01	0.36	0.06	1.88	0.02	0.11	4.37	6.43
Generalized synthetic controls	0.01	0.01	0.07	0.07	0.02	0.02	4.28	4.28
12 pre-treatment periods								
Difference-in-differences	0.01	0.01	23.19	23.19	5.65	5.65	46.38	46.38
Synthetic controls	0.11	0.11	3.97	3.97	0.07	0.07	13.47	13.47
Interactive fixed effects	0.02	0.32	0.14	0.71	0.04	0.14	4.04	1.70
Generalized synthetic controls	0.02	0.02	0.11	0.11	0.03	0.03	4.30	4.30
9 pre-treatment periods								
Difference-in-differences	0.01	0.01	15.34	15.34	3.78	3.78	34.93	34.93
Synthetic controls	0.13	0.13	3.80	3.80	0.10	0.10	12.98	12.98
Interactive fixed effects	0.02	0.34	0.17	0.98	0.05	0.17	4.19	1.52
Generalized synthetic controls	0.02	0.02	0.18	0.18	0.05	0.05	4.58	4.58
6 pre-treatment periods								
Difference-in-differences	0.02	0.02	9.05	9.05	2.24	2.24	24.98	24.98
Synthetic controls	0.16	0.16	3.66	3.66	0.13	0.13	12.97	12.97
Interactive fixed effects	0.03	0.33	0.19	1.33	0.09	0.19	4.27	1.39
Generalized synthetic controls	0.03	0.03	0.34	0.34	0.10	0.10	5.00	5.00

Second, the limitations of the originally proposed SC method[16,17] have led to recent modifications. The augmented SC approach[71] addresses the bias due to non-exact balance on pretreatment outcomes. The imperfect SC[35] reduces the sensitivity of estimates to idiosyncratic errors by applying SC to predicted rather than actual outcomes. A number of approaches relax the overlap requirement by allowing for negative weights.[29,35,71] Extensions of the SC method using machine learning methods such as ridge regression[71] and the matrix completion approach[31] appear promising. Inference for SC type methods is an area of active research, with several authors proposing extensions to the originally proposed placebo tests.[29,32,70] Future work is required that considers the relative performance of these methods and reports the coverage of alternative inferential procedures.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This report is independent research supported by the National Institute for Health Research (Senior Research Fellowship, Dr Richard Grieve, SRF-2013-06-016) and the Medical Research Council (Early Career Fellowship in the Economics of Health, Dr Noemi Kreif MR/L012332/1). This research is part-funded by the National Institute for Health Research (NIHR) Policy Research Programme, conducted through the NIHR Policy Research Unit in Policy Innovation and Evaluation, 102/0001. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

This work has been presented at the 38th Spanish Health Economics Association (AES) Conference 2018, the 12th International Conference on Health Policy Statistics (ICHPS) 2018, the HSRUK conference 2018; the 2018 American–European Health Economics Study Group III Edition meeting, and at seminar series in the University of York, the University of Granada, and the National University of Ireland Galway. An early version of this work was presented at 6th Biennial Conference of the American Society of Health Economists in 2016 and at the 2016 UK Causal Inference meeting.

ORCID

Stephen O'Neill https://orcid.org/0000-0002-0022-0500
Noemi Kreif https://orcid.org/0000-0001-9008-5690
Matt Sutton https://orcid.org/0000-0002-6635-2127
Richard Grieve https://orcid.org/0000-0001-8899-1301

ENDNOTES

[a]A recent working paper by Schmidt et al[69] uses the GSC method to assess whether insurance coverage of medical treatments with high out-of-pocket costs affects patients' utilization.

[b]See McDonald et al[42] and Kristensen et al [68] for further details.

[c]Here, participation status is defined according to whether the hospital trust had reported receiving any BPT payments for hip fractures in 2010/11.[42]

[d]The test for parallel trends is described in Appendix C of O'Neill et al[2]

[e]The general framework in equation 1 also nests unit-specific linear trends,[64,65] which would be obtained if we specify $\mu_i = [1, \mu_i, \mu_i]$ and $\lambda_t = [\lambda_t, t, 1]$.

[f]Under staggered adoption, and heterogeneous effects, extra care must be taken to identify the effect being estimated.[72,73]

[g]Gobillon and Magnac[30] suggest a similar approach that uses an expectation maximization approach.

[h]Sample code to estimate all of the methods is available from the authors on request.

[i]For instance, placebo tests capture whether the estimated effect for the treated group (or unit) is large relative to the effect that would have been estimated for a treatment group (or unit) chosen at random.[17] This contrasts with the more common random sampling perspective underlying standard errors for regression models.[28]

[j]Results were similar when 10 treated units and 100 control units were used instead.

[k]We allow for correlation between x_{it}, μ_{i1}, μ_{i2}, and μ_{i3} with the correlation matrix, C = (1, 0.5, 0.5, 0.3 \ 0.5, 1, 0.5, 0.3 \ 0.5, 0.5, 1, 0.5 \ 0.3, 0.3, 0.5, 1).

[l]Note that since $E(\mu_{i1} | D_{it} = 1)=2$ here, the true ATT is 1 in all scenarios.

[m]The average rate of surgery within 48 hours was 58.3%.

[n]Where the treated unit's outcomes are very different to those of the controls, the most similar control will receive a weight of 1 and be used as the counterfactual for the treated unit, even though it may be very dissimilar.

REFERENCES

1. Ryan AM, Burgess J, Dimick JB. Why we should not be indifferent to specification in difference-in-differences analysis. *Health Serv Res*. 2014;50(4):1211-1235.
2. O'Neill S, Kreif N, Grieve R, Sutton M, Sekhon JS. Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Serv Outcomes Res Method*. 2016;16(1-2):1-21.
3. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ*. 2015;25(12):1514-1528.
4. Chen Y, Zhou LA. The long-term health and economic consequences of the 1959-1961 famine in China. *J Health Econ*. 2007;26(4):659-681.
5. Schmidt L. Effects of infertility insurance mandates on fertility. *J Health Econ*. 2007;26(3):431-446.
6. Moran JR, Short PF, Hollenbeak CS. Long-term employment effects of surviving cancer. *J Health Econ*. 2011;30(3):505-514.
7. Cowan B, Schwab B. Employer-sponsored health insurance and the gender wage gap. *J Health Econ*. 2016;45:103-114.
8. Carpenter CS, McClellan CB, Rees DI. Economic conditions, illicit drug use, and substance use disorders in the United States. *J Health Econ*. 2017;52:63-73.
9. Akbulut-Yuksel M. War during childhood: the long run effects of warfare on health. *J Health Econ*. 2017;53:117-130.
10. Shai O. Is retirement good for men's health? Evidence using a change in the retirement age in Israel. *J Health Econ*. 2018;57:15-30.
11. Cantor JC, Monheit AC, DeLia D, Lloyd K. Early impact of the Affordable Care Act on health insurance coverage of young adults. *Health Serv Res*. 2012;47(5):1773-1790.
12. Werner RM, Konetzka RT, Polsky D. The effect of pay-for-performance in nursing homes: evidence from state Medicaid programs. *Health Serv Res*. 2013;48(4):1393-1414.
13. Nasseh K, Vujicic M. The impact of medicaid reform on children's dental care utilization in Connecticut, Maryland, and Texas. *Health Serv Res*. 2015;50(4):1236-1249.
14. Ryan AM, Burgess JF, Pesko MF, Borden WB, Dimick JB. The early effects of Medicare's mandatory hospital pay-for-performance program. *Health Serv Res*. 2015;50(1):81-97.
15. Benitez JA, Adams EK, Seiber EE. Did health care reform help kentucky address disparities in coverage and access to care among the poor? *Health Serv Res*. 2018;53(3):1387-1406.
16. Abadie A, Gardeazabal J. The economic costs of conflict: a case-control study for the Basque country. *Am Econ Rev*. 2003;93:112-132.
17. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's Tobacco Control Program. *J Am Stat Assoc*. 2010;105(490):493-505.
18. Nonnemaker J, Engelen M, Shive D. Are methamphetamine precursor control laws effective tools to fight the methamphetamine epidemic? *Health Econ*. 2011;20(5):519-531.
19. Acemoglu D, Johnson S, Kermani A, Kwak J, Mitton T. The value of connections in turbulent times: evidence from the United States. National Bureau of Economic Research. 2013.
20. Dube A, Zipperer B. Pooled synthetic control estimates for recurring treatment: an application to minimum wage studies, University of Massachusetts. Amherst Working Paper.2013.
21. Bauhoff S. The effect of school district nutrition policies on dietary intake and overweight: a synthetic control approach. *Econ Hum Biol*. 2014;12:45-55.
22. Callison K, Kaestner R. Do higher tobacco taxes reduce adult smoking? New evidence of the effect of recent cigarette tax increases on adult smoking. *Econ Inq*. 2014;52(1):155-172.
23. Dunn A, Shapiro AH. Physician payments under health care reform. *J Health Econ*. 2015;39:89-105.
24. Fletcher JM, Frisvold DE, Tefft N. Non-linear effects of soda taxes on consumption and weight outcomes. *Health Econ*. 2015;24(5):566-582.
25. Abadie A, Diamond A, Hainmueller J. Comparative politics and the synthetic control method. *Am J Pol Sci*. 2015;59(2):495-510.
26. Ferman B, Pinto C.Revisiting the synthetic control estimator. MPRA Paper No. 81941. 2016. https://mpra.ub.uni-muenchen.de/81941/
27. King G, Zeng L. The dangers of extreme counterfactuals. *Pol Analysis*. 2006;14:131-159.
28. Hahn J, Shi R. Synthetic control and inference. *Econometrics*. 2017;5(4):52.
29. Doudchenko N, Imbens GW. Balancing, regression, difference-in-differences and synthetic control methods: a synthesis (No. w22791). National Bureau of Economic Research; 2016.
30. Gobillon L, Magnac T. Regional policy evaluation: interactive fixed effects and synthetic controls. *Rev Econ Stat*. 2016;98(3):535-551.
31. Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K. Matrix completion methods for causal panel data models. 2017. arXiv preprint arXiv:1710.10251
32. Chernozhukov V, Wuthrich K, Zhu Y. An exact and robust conformal inference method for counterfactual and synthetic controls. 2017. arXiv preprint arXiv:1712.09089
33. Robbins MW, Saunders J, Kilmer B. A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *J Am Stat Assoc*. 2017;112(517):109-126.

34. Poulos J. Causal inference for observational time-series with encoder-decoder networks. 2017. arXiv preprint arXiv:1712.03553
35. Powell D.Imperfect synthetic controls: did the Massachusetts health care reform save lives? 2017. http://works.bepress.com/david_powell/24/
36. Pesaran H. Estimation and inference in large heterogeneous panels with a multi-factor error structure. *Econometrica*. 2006;74:967-1012.
37. Bai J. Panel data models with interactive fixed effects. *Econometrica*. 2009;77:1229-1279.
38. Gaibulloev K, Sandler T, Sul D. Dynamic panel analysis under cross-sectional dependence. *Pol Anal*. 2014;22(2):258-273.
39. Greenaway-McGrevy R, Han C, Sul D. Asymptotic distribution of factor augmented estimators for panel regression. *J Econometrics*. 2012;168:48-53.
40. Samartsidis P, Seaman SR, Presanis AM, Hickman M, De Angelis D. Review of methods for assessing the causal effect of binary interventions from aggregate time-series observational data. 2018. arXiv preprint arXiv:1804.07683.
41. Xu Y. Generalized synthetic control method: causal inference with interactive fixed effects models. *Pol Anal*. 2017;25(1):57-76.
42. McDonald R, Zaidi S, Todd S, et al. *A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs. An Evaluation Commissioned by the Department of Health*. Nottingham: University of Nottingham and University of Manchester; 2012.
43. Hawkes D, Baxter J, Bailey C, et al. Improving the care of patients with a hip fracture: a quality improvement report. *BMJ Qual Saf*. 2015;24(8):532-538.
44. Leal J, Gray AM, Prieto-Alhambra D, et al. Impact of hip fracture on hospital care costs: a population-based study. *Osteoporos Int*. 2016;27(2):549-558.
45. Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *Rand J Econ*. 2010;41(1):64-91.
46. Ryan AM, Blustein J. The effect of the MassHealth hospital pay-for-performance program on quality. *Health Serv Res*. 2011;46(3):712-728.
47. Kruse GB, Polsky D, Stuart EA, Werner RM. The impact of hospital pay-for-performance on hospital and Medicare costs. *Health Serv Res*. 2012;47(6):2118-2136.
48. Nicholas LH, Dimick JB, Do ITJ. Hospitals alter patient care effort allocations under pay-for-performance? *Health Serv Res*. 2011;46(1p1):61-81.
49. Ryan AM. Has pay-for-performance decreased access for minority patients? *Health Serv Res*. 2010;45(1):6-23.
50. Layton TJ, Ryan AM. Higher incentive payments in Medicare advantage's pay-for-performance program did not improve quality but did increase plan offerings. *Health Serv Res*. 2015;50(6):1810-1828.
51. Mandavia R, Mehta N, Schilder A, Mossialos E. Effectiveness of UK provider financial incentives on quality of care: a systematic review. *Br J Gen Pract*. 2017;67:e800-e815.
52. Lagarde M, Wright M, Nossiter J, Mays N. Challenges of payment-for-performance in health care and other public services – design, implementation and evaluation. Policy Innovation and Research Unit. 2013.
53. Ogundeji YK, Bland JM, Sheldon TA. The effectiveness of payment for performance in health care: a meta-analysis and exploration of variation in outcomes. *Health Policy*. 2016;120(10):1141-1150.
54. Mendelson A, Kondo K, Damberg C, et al. The effects of pay-for-performance programs on health, health care use, and processes of care: a systematic review. *Ann Intern Med*. 2017;166(5):341-353.
55. Khan SK, Shirley MD, Glennie C, Fearon PV, Deehan DJ. Achieving best practice tariff may not reflect improved survival after hip fracture treatment. *Clin Interv Aging*. 2014;9:2097.
56. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;1974(66):688-701.
57. Jones AM, Rice N. Econometric evaluation of health policies. In: Glied S, Smith P, eds. *The Oxford Handbook of Health Economics*. Oxford, UK: Oxford University Press; 2011:890-923.
58. Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press; 2009.
59. Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? *Q J Econ*. 2004;119(1):249-275.
60. Carpenter CS, Stehr M. The effects of mandatory seatbelt laws on seatbelt use, motor vehicle fatalities, and crash-related injuries among youths. *J Health Econ*. 2008;27:642-662.
61. Wen H, Hockenberry JM, Cummins JR. The effect of medical marijuana laws on adolescent and adult use of marijuana, alcohol, and other substances. *J Health Econ*. 2015;42:64-80.
62. Moon HR, Weidner M. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*. 2015;83(4):1543-1579.
63. MacKinnon JG, Webb MD. *Randomization Inference for Differences-in-differences with Few Treated Clusters (No. 16-11)*. Carleton University, Department of Economics. 2016.
64. Bell B, Blundell R, Van Reenen J. Getting the unemployed back to work: an evaluation of the new deal proposals. *Int Tax Public Fin*. 1999;6(3):339-360.
65. Wagstaff A, Moreno-Serra R. Europe and Central Asia's great post-communist social health insurance experiment: aggregate impacts on health sector outcomes. *J Health Econ*. 2009;28(2):322-340.
66. Abadie A. Semiparametric difference-in-differences estimators. *Rev Econ Stud*. 2005;72(1):1-19.
67. Duflo E. Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *Am Econ Rev*. 2001;91(4):795-813.
68. Kristensen SR, McDonald R, Sutton M. Should pay-for-performance schemes be locally designed? Evidence from the commissioning for quality and innovation (CQUIN) framework. *J Health Serv Res Policy*. 2013;18(2_suppl):38-49.
69. Schmidt L, Tedds L, Zaresani A. Utilization with high out-of-pocket costs: evidence from in-vitro-fertilization treatment. 2017; https://azaresani.com/wp-content/uploads/IVF2017.pdf.
70. Firpo S, Possebom V. Synthetic control method: inference, sensitivity analysis and confidence sets. *J Causal Inf*. 2018;6:1-26.
71. Ben-Michael E, Feller A, Rothstein J. The Augmented Synthetic Control Method. 2018. arXiv preprint arXiv:1811.04170.
72. Callaway B, Sant'Anna P.Difference-in-Differences with multiple time periods and an application on the minimum wage and employment. 2018. arXiv preprint arXiv:1803.09015.
73. Goodman-Bacon A. *Difference-in-Differences with Variation in Treatment Timing (No. w25018)*. National Bureau of Economic Research; 2018.
74. Ding P, Li F. A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Pol Anal*. 2019;27(4):605-615. https://doi.org/10.1017/pan.2019.25

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: O'Neill S, Kreif N, Sutton M, Grieve R. A comparison of methods for health policy evaluation with controlled pre-post designs. *Health Serv Res*. 2020;00:1–11. https://doi.org/10.1111/1475-6773.13274