IEA
International Epidemiological Association

International Journal of Epidemiology, 2023, 1–12
https://doi.org/10.1093/ije/dyad032
Original article

OXFORD

Original article

A comparison of quasi-experimental methods with data before and after an intervention: an introduction for epidemiologists and a simulation study

Roch A Nianogo ,[1,2]* Tarik Benmarhnia[3] and Stephen O'Neill[4]

[1]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles (UCLA), Los Angeles, CA, USA, [2]California Center for Population Research, UCLA, Los Angeles, CA, USA, [3]Scripps Institution of Oceanography, UC San Diego, La Jolla, CA, USA and [4]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK

*Corresponding author. Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles (UCLA), 650 Charles E Young Dr. South, Los Angeles, CA 90095, USA. E-mail: niaroch@ucla.edu

Received 14 January 2022; Editorial decision 7 February 2023; Accepted 7 March 2023

Abstract

Background: As the interest in and use of quasi-experimental methods to evaluate impacts of health policies have dramatically increased in the epidemiological literature, we set out this study to (i) systematically compare several quasi-experimental methods that use data before and after an intervention and contrast their performance within a simulation framework while providing a brief overview of the methods; and (ii) discuss challenges that could arise from using these methods as well as directions for future research in the context of epidemiological applications.
Methods: We considered single-group designs [pre-post and interrupted time series (ITS)] and multiple-group designs [controlled interrupted time series/difference-in-differences, synthetic control methods (SCMs): traditional SCMs and generalized SCMs]. We assessed performance based on bias and root mean squared error.
Results: We identified settings in which each method failed to provide unbiased estimates. We found that, among the methods investigated, when data for multiple time points and for multiple control groups are available (multiple-group designs), data-adaptive methods such as the generalized SCM were generally less biased than other methods evaluated in our study. In addition, when all of the included units have been exposed to treatment (single-group designs) and data for a sufficiently long pre-intervention period are available, then the ITS performs very well, provided the underlying model is correctly specified.
Conclusions: When using a quasi-experimental method using data before and after an intervention, epidemiologists should strive to use, whenever feasible, data-adaptive methods that nest alternative identifying assumptions including relaxing the parallel trend assumption (e.g. generalized SCMs).

Key words: Quasi-experiments, difference-in-differences, synthetic control method, interrupted time series, simulation

Key Messages

- Interest in and use of quasi-experimental methods to evaluate impacts of health policies have dramatically increased in the epidemiological literature.
- When data for multiple time points and for multiple control units are available (multiple-group designs), data-adaptive methods such the generalized synthetic control method can account for rich forms of unobserved confounding.
- When all of the included units have been exposed to treatment and there exist data for a sufficiently long pre-intervention period (single-group designs) then the interrupted time series design performs very well, provided the underlying model is correctly specified.
- Adjustment for any observed unit-time-varying confounders is critical for reducing bias and improving precision (provided they are not themselves influenced by the intervention).

Introduction

Policymakers and programme evaluators are often interested in assessing the impact of an intervention (e.g. health policy or programme) on some outcome of interest (e.g. incidence of a disease) in one or several units (e.g. communities, states, countries, etc.).[1] Such evaluation is a core task in many social sciences including epidemiology and economics. The potential outcomes (or counterfactual) framework is particularly useful for considering the assumptions under which causal effects of interventions can be identified.[2] Within the potential outcomes framework, causal effects are defined as a contrast of two outcome values: one observed (factual) and one unobserved (counterfactual). Given the fundamental problem of causal inference,[3] that the counterfactual outcome is impossible to observe, it is important to identify a valid means to infer such counterfactual outcomes, and hence causal effects.

Several strategies have been proposed to infer the counterfactual outcomes needed to estimate intervention impacts. Randomization, notably through randomized–controlled trials (RCTs), has been proposed as a key identification strategy to infer causal effects when the policy/intervention can be manipulated, given their ability to, on average, eliminate systematic observed and unobserved differences between the group(s). In many cases, however, randomization is not feasible due to costs or ethical objections or simply because the intervention of interest has already been implemented (ex-post evaluation).[4,5] In such cases, researchers often rely on observational data, and consider a range of alternative ‘quasi-experimental’ designs to evaluate causal estimates given the available data. However, threats to internal validity by confounding due to observed or unobserved omitted variables can be present if these designs are not carefully applied or their identification assumptions are not being satisfied. As there exist several classical and recent quasi-experimental designs, it is important to understand their relative performance as well as instances in which they will fail to yield unbiased causal estimates.

Some recent papers describing the use of these methods in epidemiologic or health services settings have focused on a single quasi-experimental method at a time such as interrupted time series (ITS),[6–8] difference-in-differences (DID)[9] or synthetic control method (SCM),[10–13] whereas few have compared their performance.[14,15] However, there is no available resource for epidemiologists where these methods are systematically compared with a specific focus on situations in which each method fails at identifying unbiased estimates. Furthermore, extensions of the traditional SCM, such as the generalized SCM, have recently been introduced but have not yet been widely adopted by epidemiologists, despite their ability to provide estimates with minimal bias in many settings.

In this paper, we aim to (i) systematically compare several quasi-experimental methods that use data before and after an intervention and contrast their performance within a simulation framework while providing a brief overview of the methods; and (ii) discuss challenges that could arise from using these methods as well as directions for future research in the context of epidemiological applications.

Methods

Here, we consider six quasi-experimental methods that can be grouped into (i) single-group designs (i.e. pre-post and ITS) and (ii) multiple-group designs [i.e. controlled interrupted time series (CITS) and DID, SCM: traditional SCM and generalized SCM].[1,16–19] Single-group designs are

Table 1 Relationship between different quasi-experimental study designs

	Two time points	Multiple time points
Single-group designs (one treated group)	Pre-post	Interrupted time series
Multiple-group designs (two groups, one treated and one control group)	Controlled pre-post and 2 × 2 DID or simple DID	Controlled interrupted time series/DID, traditional SCM and generalized SCM

DID, difference-in-difference; SCM, synthetic control method. A group can contain one or several units.

designs in which all of the included units have been exposed to treatment whereas multiple-group designs include both treated and untreated units. As seen in Table 1, we can further categorize these designs based on whether they require observing two or more time points, with at least one period before and after the intervention. Regardless, all of these quasi-experimental methods fundamentally aim at estimating the average treatment effect on the treated (ATT) but differ in terms of their data requirements and their identification assumptions, which need to be carefully examined before interpreting obtained estimates as 'causal' effects of interest. We briefly review some of these identification assumptions below. We describe each model and provide estimating equations for each approach in Table 2 to facilitate comparisons across methods. Additional details about each of the six quasi-experimental methods can be found in Box 1 and Box 2 and further methodological considerations described in this paper can be found in the Supplementary material (available as Supplementary data at *IJE* online).

Overview of identification assumptions

Parallel trends assumption

The parallel trends assumption, which underlies the DID estimation, states that outcomes for the treated and control unit would have followed parallel paths in the absence of the intervention, conditional on included variables.[1] This assumption is violated if, for instance, time-varying confounders are present and their coefficients differ by group. Whereas the plausibility of this assumption is often assessed visually, it can also be examined empirically considering multiple periods before the implementation of the intervention using a linear trend model as follows: $E(Y|\ A=a,\ Time=time,\ T=0,\ C=c)=\alpha_0+\alpha_A a+\alpha_{Time}Time+\alpha_{A-Time}\ (a*Time)+\alpha_C c$. Here we restrict the sample to the pre-intervention period ($T=0$), then the P-value for the test of $\alpha_{A-Time}=0$ indicates the level of support for the parallel trend assumption. However, concerns with such testing have been raised.[20] A test based on linear trends may also fail to detect nonlinear violations, so graphical comparison of the outcome trend (conditional on covariates) should always be examined in addition to such statistical tests.

The parallel trends assumption is made in both DID and CITS designs (albeit what they condition on differs), whereas the pre-post and ITS methods do not involve comparisons with other units and instead assume that differencing or temporal modelling are sufficient to control for unobserved confounding, respectively. The identifying assumption underlying SCM is instead that, in the absence of treatment, the expected outcomes of the units in each post-intervention period equals the weighted sum of the expected outcomes for the control units for that period. Unlike the DID/CITS designs, the generalized SCM is robust to some violations of the parallel trend assumption as it allows more complex patterns of unobserved confounding by allowing an interactive fixed effects structure.

Common shock assumption

The common shock assumption requires that any idiosyncratic shocks in the post-intervention are on average similar for both groups. Such an assumption would be violated if for instance an unmeasured confounder affects treated and untreated units differently in the post-intervention period.[1] In such cases, one cannot disentangle the impact of this shock/event/other policy from that of the intervention of interest. The common shock assumption is usually untestable empirically but more knowledge about the context in which the intervention was implemented can help determine whether there were other events or policies that occurred around the time of the intervention of interest (see ref. 21 for further discussion). This assumption is made in all of the designs explored in this study. In particular, since single-group designs do not include untreated units, a similar assumption for single-group designs, termed the absence of 'history bias', is therefore made. Such bias can occur when the effects of time-varying confounding from unexpected events or co-interventions (i.e. time shock) mix with the effects of the intervention of interest.[13]

Other assumptions

Other assumptions include: no interference or SUTVA (stable unit treatment value assumption), conditional exchangeability, conditional exogeneity, consistency, no model mis-specifications, no measurement error, no selection bias and that, once exposed, units remain exposed in all

Table 2 Equations and data used for considered design in simulation study

Designs	Equations	Data included in simulation[a]
Single-group designs (all units are treated)		
Pre-post	Equation (1) $E(Y\| \ T=t,\ C=c)=\beta_0+\beta_T T+\beta_C C$	• One unit (one treated unit = California) • Two time periods (1 year before policy = 39 and 1 year after = 45)
ITS	Equation (2) $E(Y\| \ T=t,\ Time=time,\ C=c)$ $=\beta_0+\beta_T T+\beta_{Time}Time$ $+\beta_{T,Time}(T*Time)+\beta_C C$	• One unit (one treated unit = California) • Several time periods (39 years pre-policy and 10 years post-policy)
Multiple-group designs (control group available)		
Controlled pre-post	Equation (3) $E(Y\| \ A=a,\ T=t,\ C=c)=\beta_0+\beta_A A$ $+\beta_T T+\beta_{AT}(A*T)+\beta_C C$	• Two units (one treated unit = California and one untreated unit = Georgia) • Two time periods (1 year before policy = 39 and 1 year after = 45)
2 × 2 DID (simple DID)	Equation (3)	• Two units (one treated unit = California and one untreated unit = Georgia) • Several time periods (39 years pre-policy and 10 years post-policy)
DID/CITS (without temporal dynamics)	Equation (3)	• Several units (one treated unit = California and 49 untreated states) • Several time periods (39 years pre-policy and 10 years post-policy)
DID/CITS (with temporal dynamics)	Equation (4) $E(Y\|A=a,\ T=t,\ Time=time,\ C=c)=$ $\beta_0+\beta_A A+\beta_T T+\beta_{Time}Time+\beta_{T,Time}(T*Time)$ $+\beta_{A,Time}(A*Time)+\beta_{A,T}(A*T)$ $+\beta_{A,T,Time}(A*T*Time)+\beta_C C$	
Traditional SCM	Equation (5) $ATT=\hat{\tau}_{1t}=(Y_{1t}-\hat{Y}^0_{1t})=\left(Y_{1t}-\sum_j w_j\hat{Y}_{jt}\right),$ $\forall t>T_0, j\ \in$ Controls Variables included in estimation of weights: • average unit-specific pre-intervention outcomes • average unit-specific pre-intervention unit-time-varying confounders • average unit-specific pre-intervention unit-varying confounders	
Generalized SCM	Equation (6) $Y_{it}=C_{it}\beta+\lambda_t\mu_i+\tau_{it}AT+\varepsilon_{it}$	

T is an indicator for time period ($T=0$, before the policy was implemented, $T=1$, after the policy has been implemented), A is an indicator for the treatment variable ($A=0$ if the unit did not receive the policy, $A=1$ if the unit received the policy), $A*T$ is an interaction between the treatment indicator and the time indicator and C is a set of unit-time-varying covariates that affects (and not affected by) the outcome and which represents the set of covariates sufficient for confounding control.

[a] The policy occurs in Year 40.

ITS, interrupted time series; CITS, controlled interrupted time series; DID, difference-in-difference; SCM, synthetic control method.

periods with exposure starting at the same time point for all units. Recent advances in the DID literature have relaxed this latter assumption.[22,23] Furthermore, the traditional SCM further assumes that the treated units lie within the 'convex hull' of the control units to avoid bias.[18]

The presence of 'invalid controls' [i.e. controls generated by a different data-generating process (DGP) to that of the treated unit(s)] could lead to bias, since outcomes for these controls are unlikely to be informative about the true counterfactual for the treated unit over time. This represents a challenge to any method that relies on control units when constructing counterfactual outcomes (explicitly as in the case of SCM or implicitly as in the case of DID). Moreover, parametric approaches require that the

Box 1 Single-group design overview

Pre-post design. This design (sometimes referred to as before–after) can be used when two outcome measurements are available (one before and one after the intervention). In this design, all of the included units have been exposed to treatment. This design contrasts outcomes before and after the intervention:

$$E(Y|\ T=t,\ C=c)=\beta_0+\beta_T T+\beta_C C \qquad \text{Equation (1)}$$

Interrupted time series (ITS) design. This design can be used when multiple outcome measurements are available (several before and several after). Like the simple pre-post method, the ITS adjusts by design for time-invariant confounding, although ITS can also adjust for unit-time-varying confounding (those varying by unit *and* time) by explicit temporal modelling:

$$E(Y|\ T=t,\ Time=time,\ C=c)=\beta_0+\beta_T T+\beta_{Time}Time+\beta_{T,Time}(T*Time)+\beta_C C \qquad \text{Equation (2)}$$

T is an indicator for the time period (*T* = 0 before the policy was implemented, *T* = 1 after the policy has been implementation), *A* is an indicator for the treatment variable (*A* = 0 if the unit did not receive the policy, *A* = 1 otherwise) and *C* is a set of unit-time-varying covariates sufficient for confounding control.

models are correctly specified. For instance, in the case of the ITS design, failing to correctly model temporal changes could lead to bias.[6] Such modelling can be particularly challenging when data are available for a limited duration. We explore the implications of each of these assumptions for the methods described above through a simulation study.

Simulation study

Overview of the DGP

Fifty units (representing the 50 states in the USA) were simulated over a hypothetical 50-year period with the intervention occurring in Year 40 (i.e. 10 periods were post-intervention). One state (i.e. California) received the hypothetical intervention whereas the rest of the states did not. We simulated the data according to the DGP depicted in Supplementary Figure S1 (available as Supplementary data at *IJE* online). Two potential outcomes (Y_0 and Y_1), and hence the observed outcome (Y), were simulated as a function of unit-varying, time-varying and unit-time-varying measured and unmeasured covariates drawn from a multivariate normal distribution as described below:

$$\begin{aligned}Y^0_{it} &= \beta_0+\beta_1 u_{i1}+\beta_2\lambda_{1t}+\beta_3 f(\lambda_{1t})+\beta_4 u_{i2}\lambda_{2t}+\beta_5 U_{it}\\ &\quad+\beta_6 u_{i3}\lambda_{3t}+\beta_7 x_i+\beta_8 x_t+\beta_9 x_{it}+\epsilon\\ Y^1_{it} &= \beta_0+\beta_1 u_{i1}+\beta_2\lambda_{1t}+\beta_3 f(\lambda_{1t})+\beta_4 u_{i2}\lambda_{2t}+\beta_5 U_{it}\\ &\quad+\beta_6 u_{i3}\lambda_{3t}+\beta_7 x_i+\beta_8 x_t+\beta_9 x_{it}+\epsilon+\tau\cdot TreatedPost_{it}\end{aligned}$$

$$Y_{it}=Treated_i*\left(Y^1_{it}\right)+(1-Treated_i)*Y^0_{it}$$

where *TreatedPost* is an interaction between the time indicator (Post = 1 in periods after the intervention, Post = 0 in periods before the intervention) and treatment intervention indicator ($Treated_i = 1$ if unit i received the intervention and 0 if it did not). x_i,x_t and x_{it} represent unit-varying, time-varying and unit-time-varying observable covariates, whereas u_{i1}, λ_{1t} and $u_{i2}\lambda_{2t}$ represent unit-varying, time-varying and unit-time varying unobservable (or unmeasured) covariates. Likewise, $u_{i3}\lambda_{3t}$ represent unit-time-varying unobservable covariates present in some control units. In our simulations, some controls were simulated in such a way that the treated unit would always lie inside the convex hull of the controls (i.e. have higher and lower outcome and covariate values than at least one control unit). This is especially important to consider as the Abadie's SCM requires the control weights to be restricted to between 0 and 1 (i.e. does not allow extrapolation). We assume that the ATT in period *t* is defined as $E[Y^1_{it}-Y^0_{it}\ |Treated=1]=\tau$, i.e. it does not vary with time such that the average treatment in the total population (ATE) = ATT. All parameters had a unit value (i.e. $\beta's=1$) with the exception of $\beta_2=20$, $\beta_4=0.15$, $\beta_5=100$, $\beta_6=0.001$ and $\tau=100$.

Simulated scenarios

We considered 12 scenarios defined by combinations of: (i) whether or not the parallel trend assumption was violated, (ii) whether or not the common shock assumption was violated, (iii) whether or not there were some invalid controls in the donor pool and (iv) whether there were non-linear outcome trends. The features of the DGP were implemented as follows:

- The parallel trend assumption: this assumption is violated when the trends in outcome (i.e. slope of the time unit on

Box 2 Multiple-group design overview

Controlled pre-post/2 × 2 difference-in-difference (DID). This design (also known as controlled before–after or pre-post with a non-equivalent controlled group) can be used when there is one treated group and one control group, and one period before and after the intervention. This design is also sometimes referred to as a (two-by-two or simple) DID design, 2 × 2 DID, and extends the pre-post design, contrasting the change in the outcome between the pre- and post-intervention periods for the treated units with that in the untreated units This design controls for both observed and unobserved time-invariant confounders:

$$E(Y|\ A=a,\ T=t,\ C=c)\ =\ \beta_0\ +\ \beta_A A\ +\ \beta_T T\ +\ \beta_{\mathrm{AT}}(A*T)\ +\ \beta_{\mathrm{C}} C \qquad \text{Equation (3)}$$

DID/controlled interrupted time series (CITS). This design can be used when there is at least one treated group and one control group and several pre-intervention periods. This design extends the 2 × 2 DID model by including a set of indicators for each period to account for period-specific shocks common to both groups, which can be implemented as a two-way fixed effect [Equation (3)]. Table 3 describes how the DID/CITS estimator can be obtained via a contrast of two differences. Alternatively, one could explicitly model temporal dynamics, extending the ITS analysis (as such it is sometimes referred to as a CITS or comparative interrupted time series) design [Equation (4) is a two-way fixed effect in which temporal dynamics are explicitly modelled]:

$$\begin{aligned} E(Y|A=a,\ T=t,\ Time=time,\ C=c) = {} & \beta_0 + \beta_A A + \beta_T T + \beta_{Time} Time + \beta_{T,Time}(T*Time) \\ & + \beta_{A,Time}(A*Time) + \beta_{A,T}(A*T) \qquad \text{Equation (4)} \\ & + \beta_{A,T,Time}(A*T*Time) + \beta_C C \end{aligned}$$

Traditional synthetic control method (SCM). As in the DID/CITS design, this design can be used when there is at least one treated group and one control group and several pre-intervention periods. In this design, a synthetic control is formed by finding the vector of weights *W* that minimizes the imbalance between the treated unit and a weighted average of the control units across a set of variables, subject to the weights in *W* being positive and summing to 1. The weights for each unit should ensure: $\sum_{j\,\in\,Control} w_j Y_{jt} \approx Y_{1t},\ \forall t\ \leq\ T_0$

$$ATT = \hat{\tau}_{1t} =\ (Y_{1t} - \hat{Y}^0_{1t}) = \left(Y_{1t} - \sum\nolimits_j w_j \hat{Y}_{jt}\right),\ \forall t > T_0, j\ \in\ \text{Controls} \qquad \text{Equation (5)}$$

Generalized SCM (GSCM). The design extends the two-way fixed effects models [Equation (3), i.e. the DID model] to allow interactive fixed effects (IFE), where time-invariant unobserved confounders (factor loadings, $\boldsymbol{\mu}_i$) have time-varying effects (factors, $\boldsymbol{\lambda}_t$) on outcomes. The factors, $\boldsymbol{\lambda}_t$, are assumed to be the same for the treated and control units. By allowing a more complex factor structure for unobserved confounders than the two-way fixed effects considered earlier, this model relaxes the parallel trends assumption. An IFE model is estimated for the control units only, for the entire sample period, providing estimates of the parameters $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}_t)$ for the control units:

$$Y_{it} = C_{it}\beta + \lambda_t \mu_i + \tau_{it} AT + \varepsilon_{it} \qquad \text{Equation (6)}$$

A group can contain one or several units. *T* is an indicator for the time period (*T* = 0 before the policy was implemented, *T* = 1 after the policy has been implementation), *A* is an indicator for the treatment variable (*A* = 0 if the unit did not receive the policy, *A* = 1 otherwise) and *C* is a set of unit-time-varying covariates sufficient for confounding control.

outcome) are different in the treated and untreated groups. Here, β_4 acts as a switch to turn 'on' $\beta_4 \neq 0$ for non-parallel trends or 'off' $\beta_4 = 0$ for parallel trends. The parallel trend assumption is violated when time-specific common shocks λ_{2t} have unit-varying impacts u_{i2} on the outcome, i.e. when $u_{i2}\lambda_{2t}$ is different from 0 (and $\beta_4 \neq 0$). When $\beta_4 = 0$, the unobserved component in the DGP reduces to $\beta_1 u_{i1} + \beta_2 \lambda_{1t}$ in the absence of non-linear trend, so two-way fixed effects, i.e. time and unit (or 'post' and 'treated') fixed effects, are sufficient to control for unobserved confounders, i.e. parallel trends holds.

- The common shock assumption: when a unit-time-varying unmeasured confounder affects treated and untreated units differently at the time of the policy intervention or after reception of the policy, the common shock assumption is violated.

Table 3 CITS/DID estimands—a contrast of two differences

	Before policy (pre) $T=0$	After policy (post) $T=1$	Before and after difference Δ_T
Control group $A=0$	$\hat{Y}_{00}=E(Y\mid A=0, T=0, C=c)$ $=\beta_0+\beta_{\mathrm{C}}C$	$\hat{Y}_{01}=E(Y\mid A=0, T=1, C=c)$ $=\beta_0+\beta_{\mathrm{T}}+\beta_{\mathrm{C}}C$	$\hat{Y}_{01}-\hat{Y}_{00}=\beta_{\mathrm{T}}$
Treatment group $A=1$	$\hat{Y}_{10}=E(Y\mid A=1, T=0, C=c)$ $=\beta_0+\beta_{\mathrm{A}}+\beta_{\mathrm{C}}C$	$\hat{Y}_{11}=E(Y\mid A=1, T=1, C=c)$ $=\beta_0+\beta_{\mathrm{A}}+\beta_{\mathrm{T}}+\beta_{\mathrm{AT}}+$ $\beta_{\mathrm{C}}C$	$\hat{Y}_{11}-\hat{Y}_{10}=\beta_{\mathrm{T}}+\beta_{\mathrm{AT}}$
Control and treatment difference (Δ_A)	$\hat{Y}_{10}-\hat{Y}_{00}=\beta_{\mathrm{A}}$	$\hat{Y}_{11}-\hat{Y}_{01}=\beta_{\mathrm{A}}+\beta_{\mathrm{AT}}$	$\Delta_A\Delta_T Y=(\hat{Y}_{11}-\hat{Y}_{10})-(\hat{Y}_{01}-\hat{Y}_{00})$ $=(\hat{Y}_{11}-\hat{Y}_{01})-(\hat{Y}_{10}-\hat{Y}_{00})$ $=\beta_{\mathrm{AT}}$

$\hat{Y}_{A=a,T=t}$ is the expected value of the outcome Y in group A at time T.
The contrast are based on Equation (3). CITS, controlled interrupted time series; DID, difference-in-difference.

This is the case since one cannot separate the impacts of this confounder from those of the intervention. Here, β_5 acts as a switch to turn 'on' ($\beta_5 \neq 0$) for differential shocks or 'off' ($\beta_5 = 0$) for common shock. The U_{it} represents another intervention/event/shock that occurred simultaneously as the intervention of interest.

- Invalid controls: when invalid controls are switched 'on' in scenarios, 25 of the controls are generated using a different DGP than the rest of the donor pool. In other words, invalid controls have different time-varying observables patterns compared with the rest of the observations. Here, β_6 acts as a switch to turn 'on' ($\beta_6 \neq 0$) for introduction of invalid controls or 'off' ($\beta_6 = 0$) for no invalid controls present.
- Non-linear trends: we further consider a scenario in which outcomes follow non-linear trends over time and the analyst mistakenly models temporal dynamics as if they were linear trends. Here, β_3 acts as a switch to turn 'on' ($\beta_3 = f(\lambda_{1t})$) for $non-linear$ trends or 'off' ($\beta_3 = 0$) for linear trends.

Implementation of methods

For each scenario, we conducted eight different analyses including: single-group designs: (i) pre-post (ii) ITS; and multiple-group designs: (iii) controlled pre-post (iv) 2 × 2 DID (v) DID/CITS (without temporal dynamics), (vi) DID/CITS (with temporal dynamics), (vii) traditional SCM and (viii) generalized SCM. For each analytical strategy, we included only the relevant data as outlined in Table 2 (Column 3). We consider a single treated unit to allow more comparability across methods such as SCM, which were originally designed for use with a single treated unit.

Bias and root mean squared error

We evaluated the level of bias and the root mean squared error (RMSE) across 500 simulations for each scenario. To do so, we evaluated these metrics across our 12 scenarios defined by the parallel trend assumption (parallel trend holds vs not) and common shock assumption (common shock holds vs not) across three settings (linear trends + valid controls, linear trends + invalid controls, non-linear trends + valid controls). Bias and RMSE were calculated across 500 simulation runs to account for Monte Carlo simulation error. Each time, we recorded the point estimate. We then calculated the bias and RMSE using the formulae below:

$$Bias = ATT_{est} - ATT_{true}$$

$$RMSE = \sqrt{\frac{\sum (ATT_{est_i} - ATT_{true})^2}{n}}$$

All simulations and analyses were conducted in the R software package.[24] Sample codes to generate the simulated data sets and run the analyses are available at the following link: https://github.com/nianogo/quasiexperimentalmethods.

Results

We present the single-group designs followed by the multiple-group designs. There were 12 scenarios defined by the parallel trend assumption (parallel trend holds vs not) and common shock assumption (common shock holds vs not) across three settings: Row 1, no invalid controls + linear trends; Row 2, some invalid controls + linear trends; Row 3, no invalid controls + non-linear trends.

Supplementary Figures S2–S4 (available as Supplementary data at *IJE* online) show how the data for a quasi-experimental study can be presented and organized, and how this can aid the researcher in designing an appropriate estimation strategy. Figures 1 and 2 present the bias in effect estimates and the RMSE, respectively, across all the analytical strategies. Below we limit our discussion to bias, since the narrative for the RMSE is similar.

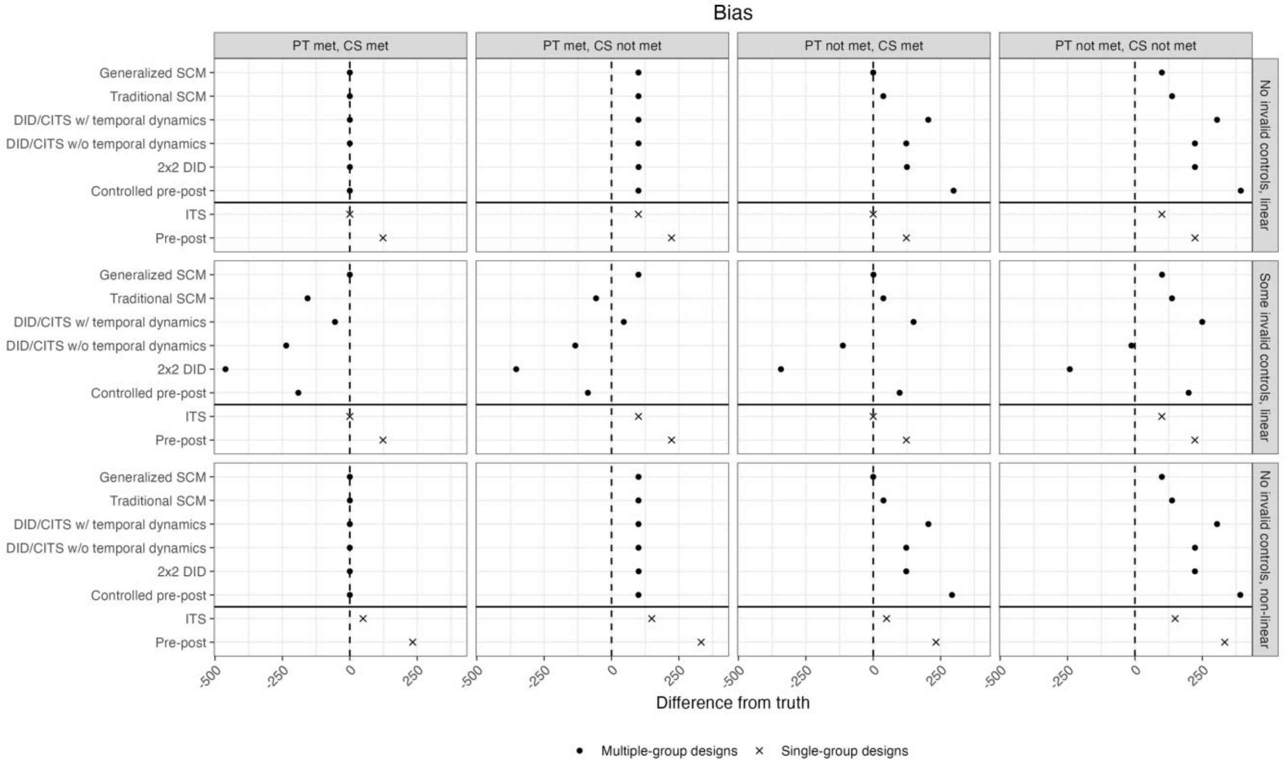

Figure 1 Bias resulting from the various analyses across scenarios. SCM, synthetic control method; DID, difference-in-difference; CITS, controlled interrupted time series; ITS, interrupted time series. For controlled pre-post, there were not sufficient degrees of freedom to control for covariates; for all other methods we controlled for x_{it}, which represents a unit-time-varying covariate. 'Invalid controls' refers to the inclusion of some control units that are generated by a different factor model than the rest of the data. DID/CITS without temporal dynamics estimates the effect of intervention using Equation (3) whereas DID/CITS with temporal dynamics uses Equation (4) to do so. The data have been simulated 500 times and the results are averaged across these replications

By analytical strategy

Single-group designs

Pre-post design. The pre-post design as described in the Supplementary material (available as Supplementary data at *IJE* online) included data from one treated unit and two time periods (one before and the other after the policy). As expected, the estimate from the pre-post design was biased in all scenarios considered even when both the parallel trend assumption and common shock assumption were met. This is unsurprising since the DGP includes time-specific unobserved confounders.

ITS. The estimate from the ITS design was biased when there were time shocks (e.g. time-varying shocks/events occurring at the same time as the policy and affecting the outcome) or when the trends were not linear but modelled as if they were linear.

Multiple-group designs

Controlled pre-post/2 × 2 DID. For the controlled pre-post design, there were insufficient degrees of freedom to control for covariates as there are only two units and two periods. Hence covariates are unadjusted for in this design, which is an important limitation. Both controlled pre-post and 2 × 2 DID yielded estimates that were biased in several circumstances: when the common shock assumption was violated, when the parallel trends assumption was violated or when there were invalid controls. Of note, the RMSE as expected was generally smaller for the 2 × 2 DID design compared with the controlled pre-post when the identification assumptions held. This is due to the fact the 2 × 2 DID design makes use of the full pre-intervention and post-intervention periods increasing efficiency.

DID/CITS with and without temporal dynamics and the traditional SCM. The estimate from the DID/CITS design with and without temporal dynamics as well as the traditional SCM yielded estimates that were biased in several circumstances: when either the common shock assumption or the parallel trends assumption was violated, or when there were invalid controls.

Generalized SCM. The estimate from the generalized SCM design was only found to be biased when the common shock assumption was violated. Of note, in our simulation,

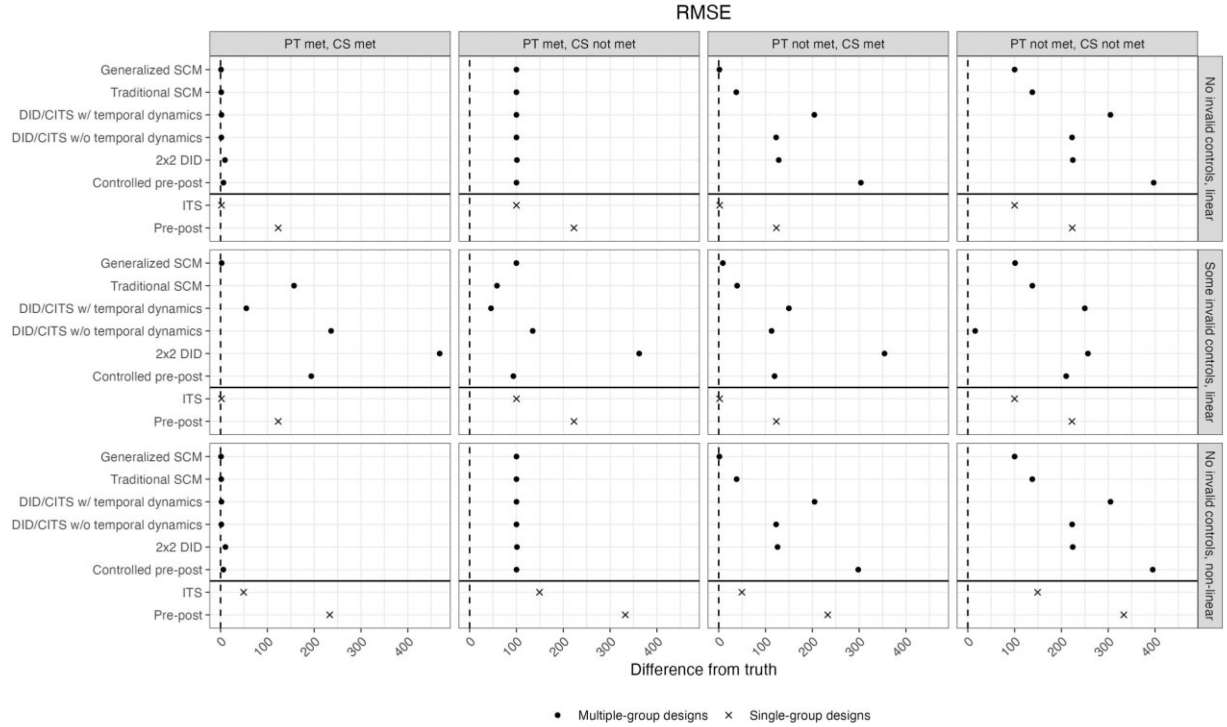

Figure 2 RMSE resulting from the various analyses across scenarios. SCM, synthetic control method; DID, difference-in-difference; CITS, controlled interrupted time series; ITS, interrupted time series; PT, parallel trend assumption; CS, common shock assumption. For controlled pre-post, there were not sufficient degrees of freedom to control for covariates; for all other methods we controlled for x_{it}, which represents a unit-time-varying covariate. 'Invalid controls' refers to the inclusion of some control units that are generated by a different factor model than the rest of the data. DID/CITS without temporal dynamics estimates the effect of intervention using Equation (3) whereas DID/CITS with temporal dynamics uses Equation (4) to do so. The data have been simulated 500 times and the results are averaged across these replications

the generalized SCM was robust to the introduction of a few invalid controls. In scenarios in which most controls are invalid, the generalized SCM is also expected to provide biased estimates, although this is likely to be true also for all methods considered here that rely on controls.

By single-group vs multiple-group designs

- Across all of the multiple-group designs evaluated in our simulation study, the estimates from the generalized SCM were generally less biased compared with other multiple-group designs.
- Across all the single-group designs evaluated in our simulation study, the ITS was generally less biased compared with the pre-post design.

By assumptions

- When the common shock assumption was violated, i.e., the common shock did not hold (in the case of multiple-group designs) as well as when there was history bias (in the case of single-group designs), all methods in our simulations yielded biased estimates.
- When the parallel trend assumption was violated (only relevant for multiple-group designs), only estimates from the generalized SCM were unbiased in our simulations.
- When there were invalid controls (only relevant for multiple-group designs), only estimates from the generalized SCM were unbiased in our simulations. As mentioned, in scenarios in which most controls are invalid or when no one suitable control exists, the generalized SCM will also provide biased estimates.
- When the outcome trends were non-linear (but modelled as if they were linear), estimates obtained from all multiple-group designs evaluated in this study were unbiased. However, estimates obtained using single-group designs (ITS and pre-post) were biased.

Discussion

Summary of the findings

Overall, we found that when data for multiple time points and for multiple control groups are available (multiple-group designs), the generalized SCM was generally less biased than other methods. The traditional SCM and the DID/CITS with and without temporal dynamics also generated low bias in our simulation setting. In particular, the traditional SCM tended to perform less well than the generalized SCM. The latter avoids bias since unlike the traditional SCM, it nests two-way fixed effects estimation within its factor structure, and cross-validation allows the method to determine when this structure is appropriate.[25] Additionally, Xu[19] (see Appendix B5 in[19] for more details) noted that when overlap of support between the treated unit and control unit diminishes (as can be the case when the parallel trend assumption is violated), there could be significantly more bias in the traditional SCM compared with the generalized SCM provided the underlying model is correctly specified.[19]

More generally, having access to (i) data for an extended period before an intervention begins and (ii) unexposed units and/or outcomes that are unaffected by the intervention allows analysts to account for richer forms of unobserved confounding [through using methods such as the generalized SCM, interactive fixed effects (IFE) models or the DID/CITS with temporal dynamics] and to assess the plausibility of some identifying assumptions. In addition, when all included units were exposed to treatment (i.e. single-group designs) but there exist data for a sufficiently long pre-intervention period then the ITS performs very well, provided the underlying model is correctly specified. Furthermore, when all assumptions evaluated in this study were met (i.e. parallel trends, common shock, valid controls, linear trends/correct model specification), all methods yielded unbiased estimates except for the pre-post design, which was consistently biased throughout.

Regardless, it is critical to adjust for any unit-time-varying observed confounders to reduce both bias and improve precision, provided they do not lie on the causal pathway. It is important to note that quasi-experimental evaluators are often not in control of the intervention, the scenarios in which it is implemented and the kinds of data available. As a result, the decision to use a pre-post design for instance (or a controlled pre-post design) is often out of necessity rather than because it is deemed the 'best design'. Nonetheless, researchers should be aware of the limitations implicit in a particular research design. Additional advantages and disadvantages for each analytical method are also presented in Supplementary Table S1 (available as Supplementary data at *IJE* online).

Quasi-experimental methods are particularly useful for epidemiologists interested in understanding whether changes in outcomes can be attributed to an event or policy. Indeed, the use of quasi-experimental methods has drastically increased in the epidemiological literature in recent years with the vast majority of papers focusing on the evaluation of the effectiveness of health policies. It is worth mentioning that these methods can also be used to address non-policy-related etiological questions such as effects of acute shocks/exposures including natural disasters for example.[26]

Limitations, extensions and future work

Different limitations and potential issues related to each of the methods are presented in Supplementary Table S2 (available as Supplementary data at *IJE* online). Furthermore, given space constraints, it was not possible to consider an exhaustive set of methods that use data before and after an intervention. For instance, we did not explore extensions of DID methods using propensity score methods such as matching[27,28] or inverse probability weighting, although we note there are similarities between these methods and the synthetic control methods.[29] Whereas we find that the generalized SCM performs very well here, other related methods not discussed here may also perform well in the contexts considered here, since they also relax the parallel trends assumption. These include for instance methods such as the matrix completion methods,[30] the augmented synthetic method[29] and Bayesian interactive factor models.[31]

We focused on a single treated unit/group with treatment commencing in a single period and did not allow units to transition into and out of treatment. However, it is important to acknowledge that in some settings in which multiple groups receive the treatment of interest, the timing of this treatment may differ between groups. In this setting, some groups will be both considered as potential control groups and then will become a treated group. A growing literature focuses on this issue. For instance, Goodman–Bacon highlight that a simple two-way fixed effects DID model can provide biased estimates when treatment commences at different time points for treated units and effects are heterogeneous.[32] Callaway and Sant'Anna also proposed an analytical solution for such a case in which there are more than two time periods and units that can become treated at different points in time while relaxing the time-invariant treatment effects assumption.[23] New estimators by Borusyak *et al.* have also been proposed to deal with such problems.[33] Effect heterogeneity may also be of interest even if an average treatment effect in the total population (ATE) is of interest but multiple units received the intervention with effects that may vary across units. Besides the approach proposed by Callaway and Sant'Anna,[23] other approaches have been proposed such as by de Chaisemartin and D'Haultfœuille[34] to address this issue.

Conclusion

In this paper, we described different quasi-experimental methods that use data before and after an intervention, and compared their performance within a simulation framework.

Quasi-experimental methods are a powerful tool for epidemiologists, although they require careful consideration of the appropriateness of their identifying assumptions and an awareness of the forms of unobserved confounding that they can and cannot account for. Overall, when using a quasi-experimental method using data before and after an intervention, epidemiologists should strive to use, whenever feasible, data-adaptive methods that learn patterns from the data without overfitting and nest alternative identifying assumptions including relaxing the parallel trend assumption (e.g. generalized SCM). More generally, simulations studies can be helpful in understanding when a method is likely to perform well/poorly. Here, we found that the data-adaptive methods such the generalized SCM could constitute a suitable approach across a range of settings when multiple control units and a moderate number of time points before the intervention of interest are available.

Ethics approval

Ethics approval was not required since we simulated hypothetical data for this study.

Data availability

All data relevant to the study are included in the article or uploaded as Supplementary information. We have also provided a link to the DGP and analysis codes at https://github.com/nianogo/quasiexperimentalmethods.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

R.A.N. and T.B. conceptualized the study and led the problem definition. R.A.N. and S.O.N. implemented the data simulation and analysis in R. T.B. and S.O.N. heavily contributed to the analysis and interpretation of the results. R.A.N., T.B. and S.O.N. wrote the first draft of the manuscript. All authors provided critical input and insights into the development and writing of the article and approved the final manuscript as submitted.

Funding

R.N. was funded through the National Institute on Minority Health and Health Disparities (1K01MD014163-01A1), National Institutes of Health, Bethesda, MD. The funder had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. The corresponding author (R.N.) had full access to all the data in the study and takes final responsibility for the paper.

Conflict of interest

None declared.

References

1. Basu S, Meghani A, Siddiqi A. Evaluating the health impact of large-scale public policy changes: classical and novel approaches. *Annu Rev Public Health* 2017;**38**:351–70.
2. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;**66**:688–701.
3. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986;**81**:945–60.
4. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Soc Sci Med* 2018;**210**:2–21.
5. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA* 2004; **291**:2720–26.
6. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol* 2017;**46**:348–55.
7. Jandoc R, Burden AM, Mamdani M *et al.* Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. In. *J Clin Epidemiol* 2015;**68**:950–56.
8. Linden A. Conducting interrupted time-series analysis for single- and multiple-group comparisons. *Stata J* 2015;**15**:480–500.
9. Caniglia EC, Murray EJ. Difference-in-difference in the time of cholera: a gentle introduction for epidemiologists. *Curr Epidemiol Rep* 2020;**7**:203–11.
10. Bouttell J, Craig P, Lewsey J *et al.* Synthetic control methodology as a tool for evaluating population-level health interventions. *J Epidemiol Commun Health* 2018;**0**:673–78.
11. Rehkopf DH, Basu S. A new tool for case studies in epidemiology: the synthetic control method. *Epidemiology* 2018;**29**:503–05.
12. Bonander C, Humphreys D, Esposti MD. Synthetic control methods for the evaluation of single-unit interventions in epidemiology: a tutorial. *Am J Epidemiol* 2021;2700–11.
13. Degli Esposti M, Spreckelsen T, Gasparrini A *et al.* Can synthetic controls improve causal inference in interrupted time series evaluations of public health interventions? *Int J Epidemiol* 2021;**49**: 2010–20.
14. 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;**55**:328–38. doi: 10.1111/1475-6773.13274.
15. 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 Methodol* 2016;**16**:1.
16. Handley MA, Lyles CR, McCulloch C, Cattamanchi A. Selecting and improving quasi-experimental designs in effectiveness and implementation research. *Annu Rev Public Health* 2018;**39**:5–25.
17. Craig P, Katikireddi S, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annu Rev Public Health* 2017;**38**:39–56.
18. 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**:493–505.

19. Xu Y. Generalized synthetic control method: causal inference with interactive fixed effects models. *Polit Anal* 2017;**25**: 57–76.
20. Roth J. Pretest with caution: Event-study estimates after testing for parallel trends. *AER: Insights* 2022;**4**:305–22.
21. Matthay EC, Gottlieb LM, Rehkopf D, Tan ML, Vlahov D, Glymour MM. What to do when everything happens at once: analytic approaches to estimate the health effects of co-occurring social policies. *Epidemiol Rev* 2021;**43**:33–47.
22. Roth J, Sant'Anna PHC. Efficient estimation for staggered rollout designs. *arXiv preprint arXiv:2102.01291*; doi:10.48550, 2 February 2021, preprint: not peer reviewed.
23. Callaway B, Sant'Anna PHC. Difference-in-differences with multiple time periods. *J Econ* 2020;**225**:200–30.
24. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2022. https://www.R-project.org/ (25 March 2023, date last accessed).
25. Ferman B, Pinto C. Synthetic controls with imperfect pretreatment fit. *Quant Econ* 2021;**12**:1197–221. doi:10.3982/qe1596
26. Sheridan P, McElroy S, Casey J, Benmarhnia T. Using the generalized synthetic control method to estimate the impact of extreme weather events on population health. *Epidemiology* 2022; **33**:788–96. doi:10.1097/EDE.0000000000001539
27. Stuart EA. Matching methods for causal inference. *Stat Sci* 2010; **25**:1–21. doi:10.1214/09-STS313T4
28. Stuart EA, Huskamp HA, Duckworth K *et al.* Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Serv Outcomes Res Methodol* 2014; **14**:166–82.
29. Ben-Michael E, Feller A, Rothstein J. The augmented synthetic control method. *J Am Stat Assoc* 2021;**116**:1789–1803.
30. Athey S, Bayati M, Doudchenko N *et al.* Matrix completion methods for causal panel data models. *J Am Stat Assoc* 2021; **116**:1716–30.
31. Pang X, Liu L, Xu Y. A Bayesian alternative to synthetic control for comparative case studies. *Polit Anal* 2022;**30**:269–88.
32. Goodman-Bacon A. Difference-in-differences with variation in treatment timing. *J Econ* 2021;**225**:254–77.
33. Borusyak K, Jaravel X, Spiess J, Revisiting Event Study Designs: Robust and Efficient Estimation. *arXiv preprint arXiv:2108.12419*; doi:10.48550, 27 August 2021, preprint: not peer reviewed.
34. de Chaisemartin C, D'Haultfœuille X. Two-way fixed effects estimators with heterogeneous treatment effects. *Am Econ Rev* 2020;**110**:2964–96. doi:10.1257/aer.20181169