

RESEARCH

Open Access



Moving beyond the classic difference-in-differences model: a simulation study comparing statistical methods for estimating effectiveness of state-level policies

Beth Ann Griffin^{1*}, Megan S. Schuler¹, Elizabeth A. Stuart², Stephen Patrick³, Elizabeth McNeer³, Rosanna Smart⁴, David Powell¹, Bradley D. Stein⁵, Terry L. Schell⁴ and Rosalie Liccardo Pacula⁶

Abstract

Background: Reliable evaluations of state-level policies are essential for identifying effective policies and informing policymakers' decisions. State-level policy evaluations commonly use a difference-in-differences (DID) study design; yet within this framework, statistical model specification varies notably across studies. More guidance is needed about which set of statistical models perform best when estimating how state-level policies affect outcomes.

Methods: Motivated by applied state-level opioid policy evaluations, we implemented an extensive simulation study to compare the statistical performance of multiple variations of the two-way fixed effect models traditionally used for DID under a range of simulation conditions. We also explored the performance of autoregressive (AR) and GEE models. We simulated policy effects on annual state-level opioid mortality rates and assessed statistical performance using various metrics, including directional bias, magnitude bias, and root mean squared error. We also reported Type I error rates and the rate of correctly rejecting the null hypothesis (e.g., power), given the prevalence of frequentist null hypothesis significance testing in the applied literature.

Results: Most linear models resulted in minimal bias. However, non-linear models and population-weighted versions of classic linear two-way fixed effect and linear GEE models yielded considerable bias (60 to 160%). Further, root mean square error was minimized by linear AR models when we examined crude mortality rates and by negative binomial models when we examined raw death counts. In the context of frequentist hypothesis testing, many models yielded high Type I error rates and very low rates of correctly rejecting the null hypothesis (< 10%), raising concerns of spurious conclusions about policy effectiveness in the opioid literature. When considering performance across models, the linear AR models were optimal in terms of directional bias, root mean squared error, Type I error, and correct rejection rates.

Conclusions: The findings highlight notable limitations of commonly used statistical models for DID designs, which are widely used in opioid policy studies and in state policy evaluations more broadly. In contrast, the optimal model we identified--the AR model--is rarely used in state policy evaluation. We urge applied researchers to move beyond the classic DID paradigm and adopt use of AR models.

*Correspondence: bethg@rand.org

¹ RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202, USA
Full list of author information is available at the end of the article



Keywords: Difference-in-differences, State-level policy, Policy evaluations, Opioid, Overdose, Simulation

Background

Reliable evaluations of state-level policies are essential to identifying effective policies and informing policymakers' decisions, yet the methodological rigor of published studies varies (see Schuler, et al. (2020) for a review of the opioid policy literature). State-level policy evaluations commonly use a difference-in-differences (DID) study design; yet within this framework, statistical model specification varies notably across studies. The choice of model specification as well as other factors – including low outcome occurrence rates (e.g., opioid mortality), sample size (both the number of policy states as well as the number of time points available), and differences across states prior to policy adoption – can impact the accuracy and precision of effect estimates. Although numerous publications provide analytic guidance for policy evaluations using longitudinal data [1–5], applied researchers have not fully adopted methodological best practices. Furthermore, there have been no comprehensive examinations of the relative performance of commonly used statistical models under conditions that mimic those encountered in actual state policy evaluation settings.

A DID study design, broadly defined, has become dominant in the health care policy literature when longitudinal data are being used to evaluate the impact of state-level policies [6, 7]. A DID design compares the outcomes observed among a group exposed to the policy of interest (treatment group) and an unexposed comparison group across time points prior to policy implementation (first difference) and after policy implementation (second difference). The policy effect is estimated as the difference between the first and second differences, hence “difference-in-differences” [6]. However, a growing number of studies highlight challenges and limitations of a DID design, particularly when the key DID assumptions do not hold [3, 6, 8–10] or when sample size is limited [11]. Additionally, it has been well established that standard error corrections that adjust for violations of the assumed independence of the repeated measures in longitudinal datasets are needed to obtain accurate Type I error rates [12–16]. Despite the wealth of knowledge concerning challenges of and best practices for DID designs in various settings, the applied literature largely does not reflect these insights [17–20].

To promote adoption of more robust statistical methods in health policy research, our study empirically compares the performance of multiple variations

of the two-way fixed effect model traditionally used in the context of a DID design for state-level policy evaluations. Our motivating context is the ongoing U.S. opioid crisis, which claimed over 50,000 lives in 2019 alone [21] and has spurred states to adopt a myriad of opioid-related policies and initiatives. The urgency of the opioid crisis demands that accurate, robust statistical methods are used to identify effective state policies, yet our recent review of the “state of the science” of the opioid-policy literature highlighted that methodological rigor varied notably across studies [22]. Applied researchers would benefit from additional, accessible guidance regarding the multitude of analytic choices both in the context of opioid-policy evaluations and state-level policy evaluations more generally. We are aware of only one other study considering relative performance across statistical methods in the context of health policy – that study compared analytic approaches for evaluating state gun policy laws on gun-related mortality, another high-stakes health policy setting [18]. In some ways the settings are similar in terms of longitudinal state-level outcomes; however, the conclusions may differ due to differences in the underlying outcome distributions (e.g., opioid related mortality is a more highly skewed outcome than total firearm deaths).

Our study seeks to provide needed guidance about which set of statistical models commonly used in evaluations of state-level opioid policies with a DID study design perform best when estimating the impacts of state-level opioid policies on opioid-related mortality. We also seek to provide lessons that may be applicable to state policy evaluations more broadly. Using a simulation study based on observed state-level opioid mortality, we assessed statistical performance using various metrics, including directional bias, magnitude bias, and root mean squared error; we also reported Type I error and the rate of correctly rejecting the null hypothesis, given the prevalence of frequentist null hypothesis significance testing (NHST) in the applied literature.

Our findings indicate that some commonly used methods have poor statistical performance—an outcome that has implications for interpreting the existing literature as well as conducting rigorous future evaluation studies. Our discussion provides important insights for statisticians and researchers regarding ways to estimate policy effects, and highlights the need for methodological development to address the challenges of rigorously estimating policy effects in the context of complex policy settings.

Methods

Below we describe the data structure, simulation conditions, and empirical models considered in our simulation study.

Data structure

The data structure we considered in this study was longitudinal, repeated annualized measures at the state level. The outcome considered was opioid-related mortality, measured annually in each state over 18 years, providing $50 \times 18 = 900$ total observations, clustered within states. We did not consider individual-level data within the aggregate state level data.

Empirical models considered

The focus of our simulation study was to compare performance of multiple statistical models for estimating policy impact using annual state-level outcomes, given a policy landscape in which states implemented a given policy at different times. We compared the classic two-way fixed effects DID model to three additional models, based on the previous gun policy simulation study [18] as well as a review of methods commonly used in opioid policy evaluations [22]. Specifically, we considered: (1) a “detrended” extension of the classic DID model that includes state-specific linear slopes; (2) a one-period lagged autoregressive (AR) model; and (3) generalized estimating equations (GEE) with an autoregressive correlation structure.

To formalize the setting and inferential goal, we use potential outcomes notation for repeated measures data such that Y_{it1} denotes the potential outcome (e.g., opioid-related mortality rate) for state i ($i = 1, \dots, 50$) if the policy was in effect at time t while Y_{it0} denotes the potential outcome for state i if the policy was not in effect at time t . Thus, each state has two potential outcomes at each time point, representing the outcomes that would be achieved with and without the policy in effect. Our primary treatment effect of interest is $E[Y_1 - Y_0]$, averaging across both states and times, with each state and each time point equally weighted. Let $A_{it} = \{0, 1\}$ denote an indicator for whether or not state i had the policy in effect at time t (where $t = 1, \dots, T$). Then, $Y_{it}^{obs} = Y_{it1} * A_{it} + Y_{it0} * (1 - A_{it})$ denotes the observed outcome for state i at time t as measured longitudinally for state i over time $t = 1, \dots, T$.

Classic DID estimation compares the pre-policy to post-policy change in the treated group to the corresponding pre-period to post-period change in the comparison group. This approach provides an estimate of the average policy effect, while controlling for time-invariant differences between treated and untreated states and for

time-varying exogenous factors (i.e., those that affect both treated and untreated states equally). The classic DID specification is generally implemented as a two-way fixed effects model that includes both state- and time-fixed effects, expressed as:

$$g\left(Y_{it}^{obs}\right) = \alpha \bullet A_{it} + \beta \bullet X_{it} + \rho_i + \sigma_t + \varepsilon_{it} \quad (1)$$

where $g(\cdot)$ denotes the generalized linear model (GLM) link function (e.g., linear, log), X_{it} denotes a vector of time-varying state-level covariates, and ε_{it} denotes the error term. State fixed effects, ρ_i , quantify potential differences in the outcome across states, and time fixed effects, σ_t , quantify temporal national trends. The coefficient estimate \hat{a} represents the DID estimator, namely the policy effect of A after accounting for differences between states implementing and not implementing a policy and time trends.

Standard DID models assume that the difference in the outcomes of the treated and untreated groups would remain constant in the absence of the policy intervention (with magnitude equal to that observed pre-policy). In practice, this assumption is often referred to as the “parallel trends” assumption, although we note that “parallelism” is actually a stronger assumption than necessary, as trajectories need only be equivalent, not necessarily parallel in the linear sense [23]. The outcome levels themselves are not assumed to be equivalent across groups; level differences are accounted for by the state fixed effects. A common misperception is that this assumption can be tested by assessing whether pre-policy period trends are parallel; however, this assumption is inherently untestable as it involves the unobservable counterfactual trends in the post-period. Indeed, conducting “tests of parallel trends” in the pre-period can lead to bias and misleading results [23].

The second model we evaluated is an extension of the classic DID model that includes state-specific slopes (referred to as “detrending” the data). The detrended model can be expressed as:

$$g\left(Y_{it}^{obs}\right) = \alpha \bullet A_{it} + \beta \bullet X_{it} + \rho_i + \sigma_t + \sum_{s=1}^{50} (\omega_s \bullet t) + v_{it} \quad (2)$$

where ω_s denotes the state-specific linear slope over time and v_{it} denotes the error term. This model expands on Eq. (1) by adding state-specific linear trends ($\omega_s \cdot t$). In this model, each state has its own fixed effect to account for its mean as well as a unique linear slope over time. Because the model also includes a national time trend (fit via year fixed effects), the state-specific linear trend is interpreted as the difference between the national time trend and the state trend. This model may be used as a

robustness check to rule out differential state trajectories over time – i.e., if Eqs. (1) and (2) yield similar policy effects, this suggests the absence of differential trajectories (see Bilinski and Hatfield (2020) for a discussion of this approach).

In the presence of differential trajectories that are additive, Eq. (2) should offer an improvement over Eq. (1). However, caution must be used: the time trend terms may functionally “over control” and absorb part of the treatment effect in addition to pre-existing differential trends, particularly in the presence of a time-varying treatment effect [24].

Additionally, we considered an AR model. The gun policy simulation study found that AR models performed especially well when estimating the policy effect on total firearms deaths [16]. AR models include one or more lagged measures of the outcome (e.g., Y_{it-1}^{obs}) as covariates to control for potential average differences in outcome trends across treated and comparison states. These models can improve prediction when outcomes are highly autocorrelated, as is the case with annual measures of state-level opioid-related mortality. The AR model examined here included a single lagged value of the outcome (as this was identified as the top performing AR model in the prior gun policy simulation study), expressed as:

$$g(Y_{it}^{obs}) = \alpha \cdot (A_{it} - A_{i,t-1}) + \beta \cdot X_{it} + \gamma \cdot Y_{it-1}^{obs} + \sigma_t + \epsilon_{it} \quad (3)$$

Akin to Eq. (1), this model includes time fixed effects, σ_t , to quantify temporal trends across time, but adjusts for state-specific variability through the use of the AR term ($\gamma \cdot Y_{it-1}^{obs}$) rather than state fixed effects. Notably, inclusion of the AR term creates a “change” model, as the policy effect is defined as the expected difference in the outcome, given the prior year’s outcome. As such, we coded the policy variable (A) using *change coding* ($A_{it} - A_{i,t-1}$), based on early work demonstrating that effect size estimates from AR models can be substantially biased when using standard *effect coding* (A_{it}) [25]. An AR model with a single lagged outcome is very closely related to the first-difference estimator, a commonly used alternative to the fixed effects estimator (e.g., Eq. (1)). Indeed, when there are only 2 time periods, a first-difference estimator and fixed effects estimator are identical; with 3 or more time periods, the relative performance of these estimators depends on the degree of autocorrelation in the outcome [17, 26].

Finally, we considered a fixed effect model using GEE. In the context of correlated outcomes (e.g., within states), GEE model parameters are estimated by specifying a covariance structure for the clustered outcomes [27]. This model can be expressed as:

$$g(Y_{it}^{obs}) = \alpha \cdot A_{it} + \beta \cdot X_{it} + \sigma_t + \zeta_{it}, \quad (4)$$

which includes time fixed effects σ_t and time-varying state-level confounders measured in X_{it} . GEE is a semi-parametric method that requires specification of the covariance matrix for within-subject observations (e.g., exchangeable, autoregressive, unstructured). We assume an autocorrelation structure of order 1 (AR1), which means that the correlation structure \mathbf{R} for the repeated measures within each state is

$$R_{t,m} = \begin{cases} 1 & \text{if } t = m \\ |\rho^{t-m}| & \text{if } t \neq m \end{cases}$$

for the t, m element of \mathbf{R} .

In the context of a longitudinal policy evaluation study, the central challenge is determining to what degree, if any, the observed heterogeneity in outcomes across states is due to a true policy effect versus other factors. All models we considered included time fixed effects to account for state-invariant (i.e., national) temporal trends. Additionally, the classic DID and detrended DID both included state fixed effects in order to reduce bias due to time-invariant factors that vary across states.

In contrast to fixed effects, the AR model adjusted for state-specific variability through the use of the lagged outcome term, and a GEE approach used an AR correlation structure to account for correlation at the state level. The optimal model should be the one in which the underlying assumptions of the model match the true processes of generating the data. As it is impossible to test model assumptions in practice, we used a simulation study with a known data-generating process to assess the relative performance of these statistical models.

Statistical models tested via simulation

Within our four primary DID variations (i.e., classic two-way fixed effect model, detrended model, AR model, and GEE model), we considered three other estimation aspects: GLM link function specification, standard error estimation, and weighting to account for state population. We detail each below and summarize all candidate models in Table 1.

- (1) *GLM specifications*: As opioid-related deaths are discrete and historically rare events, count models or models accounting for the skewed nature of the outcome may be more appropriate than traditional linear models that assume normality. We tested the relative performance of the following GLMs: linear, log-linear (a linear model with log-transformed

Table 1 Overview of statistical models evaluated in simulation study

<i>Regression specification</i>	<i>Link function</i>	<i>SE estimation</i>	<i>Population weighting</i>
Classic 2-way Fixed Effects	Linear	none; Huber; cluster	Population weighted; unweighted
	Log-linear	none; Huber; cluster	Population weighted; unweighted
	Negative Binomial	none; Huber; cluster	Unweighted, with log(population) used as an offset
	Poisson	none; Huber; cluster	Unweighted, with log(population) used as an offset
Detrended	Linear	none; Huber; cluster	Population weighted; unweighted
	Negative Binomial	none; Huber; cluster	Unweighted, with log(population) used as an offset
Autoregressive	Linear	none; Huber; cluster	Population weighted; unweighted
	Log-linear	none; Huber; cluster	Population weighted; unweighted
	Negative Binomial	none; Huber; cluster	Unweighted, with log(population) used as an offset
	Poisson	none; Huber; cluster	Unweighted, with log(population) used as an offset
GEE	Linear	AR1 structure	Population weighted; unweighted

outcome), and two log-link models (negative binomial and Poisson).

- (2) *Standard error (SE) estimation*: There are 3 common ways to estimate the SE of the effect estimate: [22] no adjustment [1]; Huber adjustment: robust estimators (also known as sandwich estimators, or Huber corrected estimates) that attempt to adjust the SE for violations of distributional assumptions [28, 29]; and [2] cluster adjustment: adjustments to account for possible violations of the assumed independence of observations within states [28–30]. For each model (except the GEE model), we estimated the SE in these three ways. For the GEE models, we used the AR [22] covariance structure for our SE estimation. We also considered the Arellano method [31] as implemented in R's `vcovHC` package; we do not report these results for parsimony, as they were very similar to the Huber method (see our Shiny tool for full details).
- (3) *Use of state population weights*: Finally, we explored the impact of using state population as an analytic weight in the linear and log-linear models, an approach commonly used in state-level policy evaluations (e.g., within opioid-related policy studies [32–35]). For state-level analyses of opioid-mortality rates, using population weights puts equal weight on each death, regardless of the state in which the death occurred; unweighted analyses put equal weight on each state, so a death in a small state will have much greater weight than a death in a larger state. Because we generated data so that the policy effects were constant across all states regardless of size or other characteristics, we did not expect weighting to affect bias; however, it could have substantial effects on the SE estimates. Given

that log-link models (e.g., negative binomial, Poisson) are estimated using mortality counts (rather than rates) and do not need to be weighted to be nationally representative, we did not examine the impact of weighting in these models. Instead, these models include the logarithm of state population size as an offset, resulting in a model that is effectively predicting the opioid-related death rate, such that exponentiated model coefficients can be interpreted as incident risk ratios.

Simulation details

This section describes our simulation study in detail, including the data sources used in the study, the data generation scheme, and the performance metrics used to compare the approaches.

Data sources and measures

The outcome of interest was the annual state-specific opioid mortality rate per 100,000 state residents, obtained from the 1999–2016 National Vital Statistics System (NVSS) Multiple Cause of Death mortality files. Consistent with other studies [36–38], we identified opioid related overdose deaths based on ICD10-CM-external cause of injury codes X40–X44, X60–64, X85, and Y10–Y14, indicating accidental and intentional poisoning, with opioid overdose based on the presence of one of the following diagnosis codes: T40.1 poisoning by heroin, T40.2 poisoning by natural and semisynthetic opioids (e.g., oxycodone, hydrocodone), T40.3 poisoning by methadone, and T40.4 poisoning by synthetic opioids excluding methadone (e.g., fentanyl, tramadol).

Given concerns about model overfitting in the presence of numerous covariates [39], we included only a

single covariate: state-level unemployment rate [40]. This covariate was selected because it is frequently used in opioid policy studies [22]. Sensitivity analyses including a broader set of covariates (e.g., poverty rates, income levels, and percentages in defined race/ethnicity and age groups) resulted in no meaningful change to the general findings, with a slight increase in precision; as such, we present findings from the more parsimonious model.

Simulation data generation

The simulation design builds directly from prior work that compared statistical methods for evaluating how state laws affect firearms deaths [18]. For each simulation iteration we generated 5000 simulated datasets.

In each simulated dataset, we selected a random subset of k states to be the policy/treated group, with remaining states serving as the comparison/untreated. This simulation represents the simplified scenario in which there is no confounding by observed or unobserved covariates or by lagged values of the outcome, Y_{it-1}^{obs} . For each state and year, we generated a time-varying indicator A_{it} to denote whether the hypothetical policy was in effect. For comparison states, $A_{it}=0$ for the entire study period. For policy states, the month and year of policy enactment were randomly generated, with year restricted to 2002–2013 (inclusive) to ensure at least three years of outcome data both before and after enactment. In the first year of implementation, A_{it} was coded as fractional value between 0 and 1, indicating the percentage of the year the policy was in effect. Once a policy was implemented, it remained in effect throughout the study period; thus, $A_{it}=1$ for all remaining years.

As we were considering models with different log links, we evaluated their performance using simulated data for which each model was correctly specified to facilitate comparison across models. We simulated outcome data as follows: For untreated states, outcome values were set equal to the actual observed state-specific, year-specific opioid overdose rates for all times t , namely $Y_{it}^{obs} = Y_{it0}$. Similarly, for treated states in the pre-policy period, values were also equal to the actual observed values ($Y_{it}^{obs} = Y_{it0}$). For treated states in the post-policy period, outcomes Y_{it1} were generated by augmenting the observed value Y_{it0} with an effect size of magnitude α as follows: $Y_{it1} = Y_{it0} + \alpha_{linear}$ for linear models; $Y_{it1} = Y_{it0} + \log(\alpha_{log})$ for log-linear models; and $Y_{it1} = Y_{it0} * (\alpha_{log} - 1)$ for log link models.

Simulation conditions varied the following factors:

- (1) *Effect size*. We considered settings when the policy had a null effect, as well as a non-null effect of small, medium and large magnitude. For null effect

conditions ($\alpha = 0$), post-policy observations were equal to actual observed values, $Y_{it}^{obs} = Y_{it0}$, for both treatment groups. When generating non-null effects, we tailored the magnitude of α with respect to link function (i.e., α_{linear} , α_{log}) to ensure that the magnitude of the resulting effect, calculated in terms of the mean number of additional deaths nationally (per 100,000 people), was comparable across models. Specifically, we started by generating data with an $\alpha_{log} = \pm 5\%$ (small), $\pm 15\%$ (medium), and $\pm 25\%$ (large) on the multiplicative scale and then empirically calculated the average excess mortality count across simulated datasets for each effect size. We then specified the corresponding α values for the linear models such that they would yield an effect size of the same magnitude (i.e., $\alpha_{linear} = \pm 0.23$, ± 0.70 , and ± 1.16).

- (2) *Number of treated units*. We also investigated the role of the number of policy states, simulating data in which 1, 5, 15 and 30 states implemented the policy. Note that the total sample size of treated and untreated states is always 50.
- (3) *Timing of policy effect*. State policies often do not become 100% effective immediately after implementation, making it important to consider variation in the onset of policy effectiveness. We considered two possible conditions: an instantaneous effect and a 3-year linear phase-in effect. In both the data generating and analytic models, we specified an instantaneous effect as a simple step-function that has a value of zero when the policy is not in effect and a value of one when the policy is in effect (as described above). The gradual policy effect allows for the effect to grow linearly in the first 3 years after implementation, with values starting at zero and reaching 1 after 3 years of implementation.

Metrics for assessing relative performance of candidate statistical methods

Performance metrics included directional bias, magnitude bias, and root mean squared error, as well as Type I error and rate of correctly rejecting the null hypothesis, given the prevalence of frequentist NHST in the applied literature.

- (1) *Directional bias*. Directional bias assesses the average difference between the estimated effect and true effect over all simulations for a given effect size (e.g., $\pm 5\%$), showing the tendency of the estimated effects from a given model to fall closer or further from the

true effect. We report directional bias summarized over both the positive effect size conditions (e.g., +5%) as well as the negative effect size conditions (e.g., -5%) to quantify how the models are doing on average for a fixed effect size α , regardless of the direction. We define directional bias as the average of the sum of the bias across positive and negative effect simulations, as follows:

$$\text{DirectionalBias}_\alpha = \left(\sum_{k=1}^{5000} \frac{\hat{\alpha}_{k, \text{pos}} - \alpha_{\text{pos}}}{5000} + \sum_{k=1}^{5000} \frac{\hat{\alpha}_{k, \text{neg}} - \alpha_{\text{neg}}}{5000} \right) / 2$$

Additionally, we standardized bias by reporting it with respect to the mortality count for both linear and nonlinear models to facilitate comparison across models. Then, we converted the standardized directional bias into percent directional bias, dividing it by the expected change in mortality count that corresponds to the given α (e.g., when $\alpha = \pm 5\%$ the expected change in deaths nationally will equal ± 700 , respectively).

- (2) *Magnitude bias*. Magnitude bias assesses whether the estimated effects are systematically too small or too large, relative to the true effect. Magnitude bias is computed by taking the average of the bias across the positive and negative effect simulations, after multiplying the bias from the negative effect simulations by negative one.

$$\text{MagnitudeBias}_\alpha = \left(\sum_{k=1}^{5000} \frac{\hat{\alpha}_{k, \text{pos}} - \alpha_{\text{pos}}}{5000} - \sum_{k=1}^{5000} \frac{\hat{\alpha}_{k, \text{neg}} - \alpha_{\text{neg}}}{5000} \right) / 2$$

For example, with a model that shows a magnitude bias of +0.1 with a true effect size of ± 0.30 , the model typically gives estimates of +0.4 or -0.4 for the positive and negative effect versions of the simulation, respectively, exaggerating the true effect size in both cases. Conversely, a model that shows a magnitude bias of -0.1 would give estimates of +0.3 or -0.2 for the positive and negative effect simulation, respectively, underestimating the true effect size. As with directional bias, we standardized magnitude bias so it represents mortality count and report percent magnitude bias below, dividing it by the corresponding expected change in deaths nationally that would correspond to the given α .

- (3) *Root mean squared error (RMSE)*. RMSE is calculated by taking the square root of the sum of the mean squared errors (e.g., $\sqrt{\sum_{k=1}^{5000} (\hat{\alpha}_k - \alpha)^2 / 5000}$). RMSE quantifies error for a given model specification, taking into account both directional bias and variance.
- (4) *Type I error rate*. In the context of traditional NHST, Type I error rate is the frequency of incorrectly

rejecting the null hypothesis (i.e., there truly is no policy effect). When data are generated such that there is no true policy effect (i.e., the null hypothesis is true), the model should identify a statistically significant effect (i.e., reject the null hypothesis) no more than 5% of the time if tested with an 0.05 level of significance.

- (5) *Correct NHST rejection rates*. We also assessed the ability of the model to correctly identify that the null hypothesis is false in the context of traditional NHST. We quantified the “rate of correct rejections” for each model by calculating the proportion of estimates that were both statistically significant and in the same direction as the true effect. When conducting this significance test, we used a SE correction factor to ensure comparability of correct NHST rejection rates across models with the exact same Type I error rate. Without applying the SE correction factor, models that underestimate the true error in their estimates would appear to have excellent statistical correct rejection rates, even though the actual sampling variability in their estimates may be quite high, in which case the model may not actually be sensitive to detecting a true effect. Typically, analyses are considered to have adequate statistical correct rejection rates/power if the likelihood that they correctly reject the null hypothesis is 80% or higher.

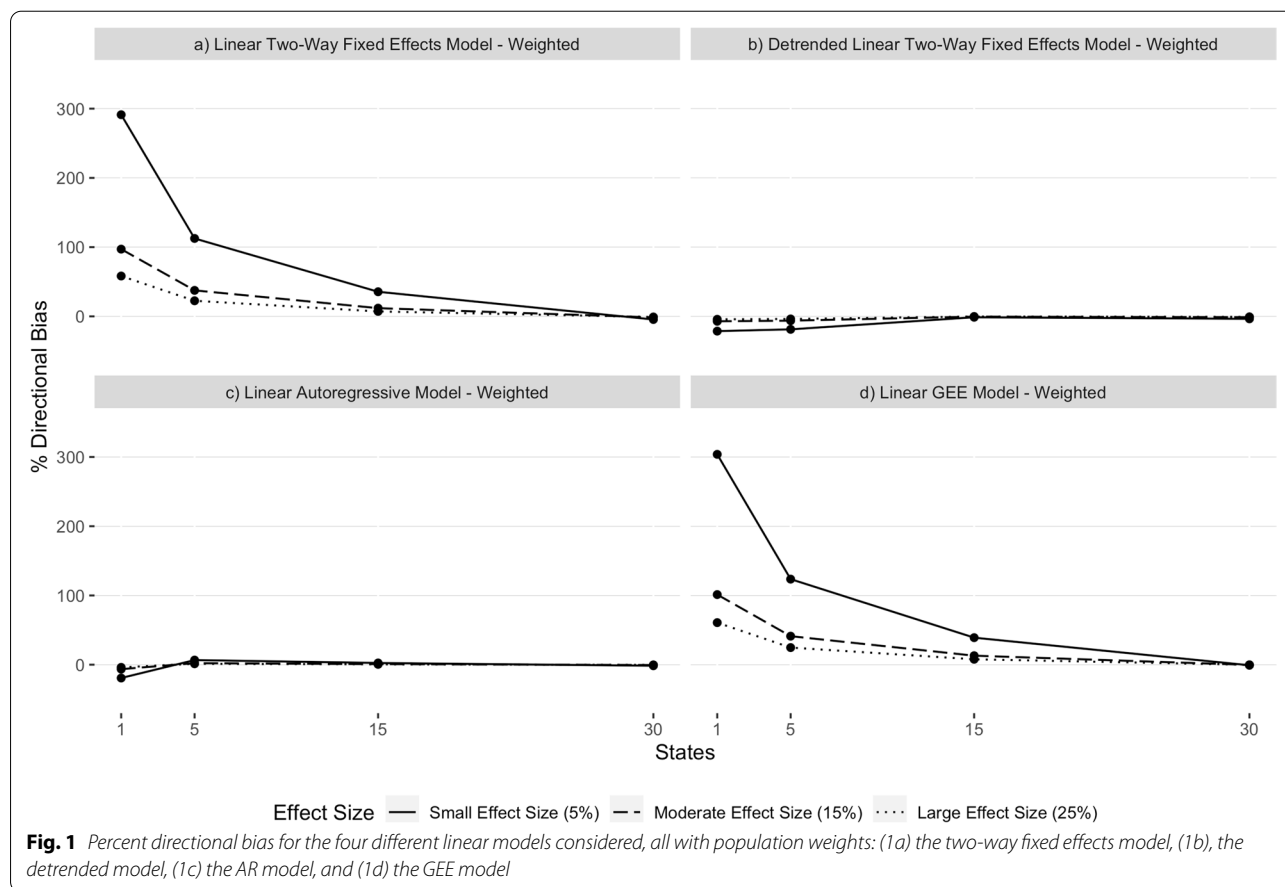
We conducted all simulations in R, using the OPTIC. simRM package which implements our simulation code on user-provided outcome data. The package is currently available on github <https://github.com/aescherling/optic-core>, including code to create the figures. Extensive results for all statistical models considered in our simulation are available via a Shiny tool (<https://elizabethmcneer.shinyapps.io/statmodelsim/>).

Results

In each section below, we first compare results for the set of four linear models (i.e., linear two-way fixed effects, linear detrended, linear AR, and linear GEE models). We then discuss the relative performance across different GLMs (i.e., negative binomial, Poisson, and log-linear models). For parsimony, all summary statistics are averaged across simulation conditions with a gradual policy effect and an instantaneous policy effect.

Directional bias

Figure 1 shows percent directional bias as a function of both effect size magnitude and the number of policy



states for the four different linear models (using population weights). In all cases, percent directional bias decreased both as effect size increased and the number of policy states increased. Most notably, the linear two-way fixed effects model (the classic DID model) had high percent directional bias when the number of treated states was lower than 15 (e.g., ranged from 22 to 291%) (Fig. 1a). The linear GEE had similar directional bias to the linear two-way fixed effects (ranged from 0 to 305%) (Fig. 1d). Directional bias was much lower for the detrended model and AR models compared to the two-way fixed effects and GEE models (ranging from ±3% to -21%) (Fig. 1b and c).

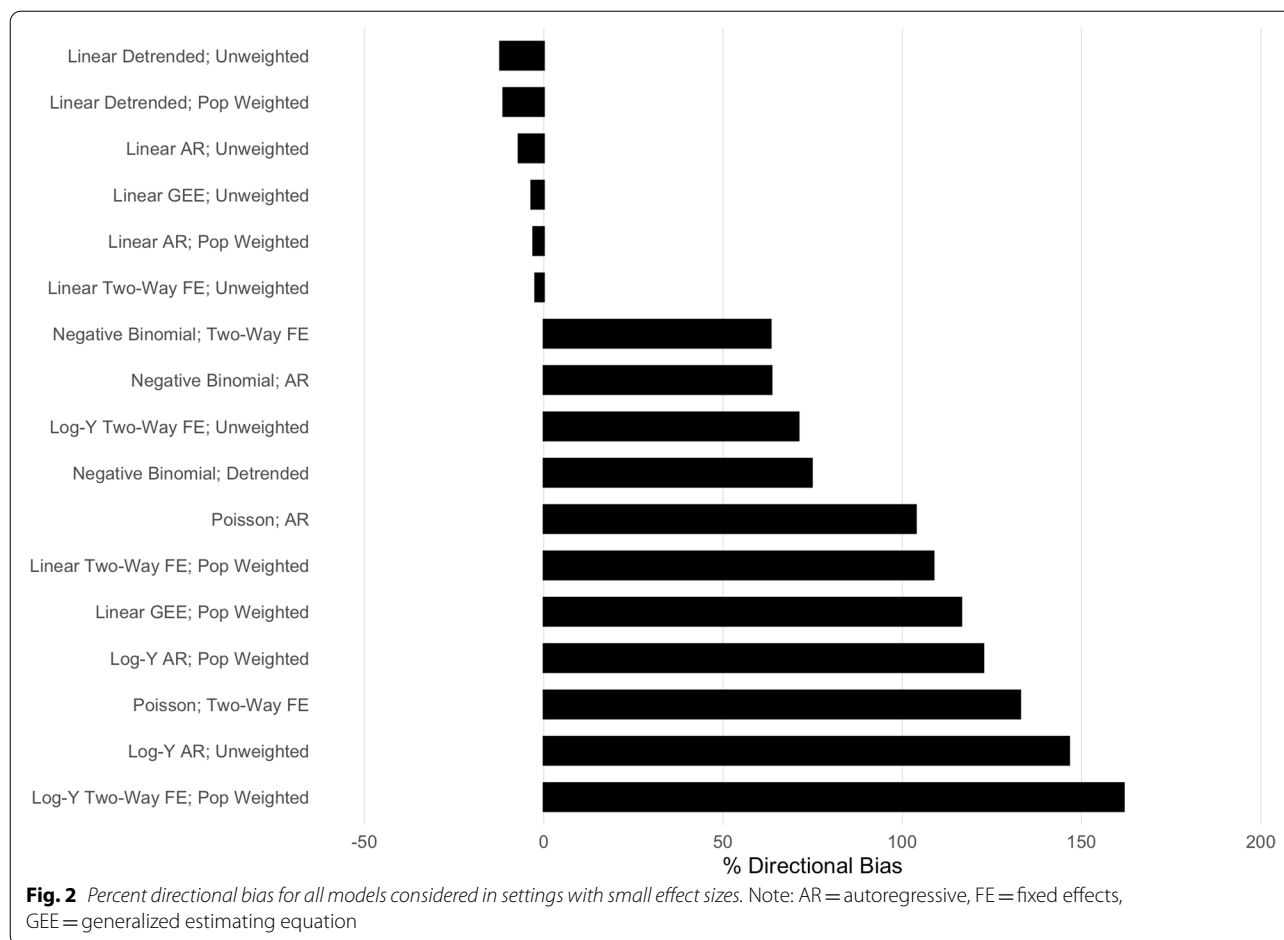
Figure 2 shows the percent directional bias for all models under the small effect size condition. Notably, the majority of models had positive directional bias, suggesting that estimated effects tend to be numerically larger in a positive way on average, regardless of the direction of the true policy effect. The large majority of models had very high rates of directional bias. For example, non-linear models yielded directional bias ranging from 64 to 162%, which translates into excess mortality estimates that are off by 448 to 1134 more deaths. Directional bias was smallest in the linear models (ranging from -2% to

-12%), with the exception of the weighted linear two-way fixed effects and weighted GEE models, where directional bias was quite large (116 and 109%, respectively).

The directional bias was relatively similar between weighted and unweighted versions of both the linear AR and linear detrended models. In contrast, directional bias was significantly larger for the weighted version, compared to the unweighted version, for both the traditional DID model (unweighted = -2%; weighted = 109%) and linear GEE model (unweighted = -3%; weighted = 116%). Further, directional bias was notably larger when there was a gradual versus an instantaneous policy effect, although the magnitude of this difference varied by model.

Magnitude bias

Broadly, as seen with directional bias, magnitude bias decreased as both effect size and number of policy states increased. We present magnitude bias results for all models under the small effect size condition (Fig. 3). Magnitude bias was less than 10% for most models, with the exception of the four non-linear AR models (14-25% for the negative binomial, Poisson, and log-linear AR

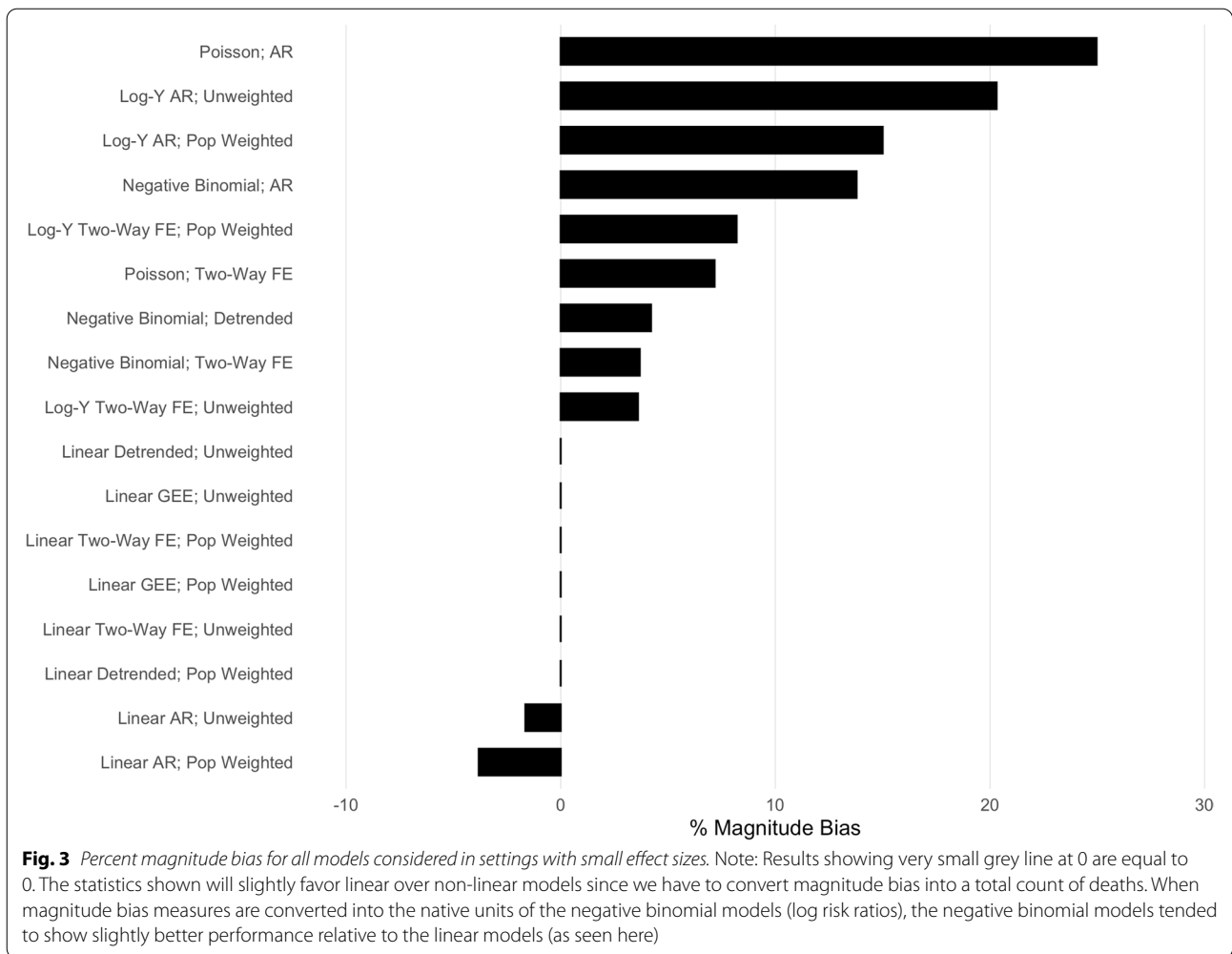


models). Most of the models with non-zero magnitude bias had positive magnitude bias (i.e., overestimating the true policy effect), ranging from 4% (negative binomial 2-way fixed effects and detrended models) to 25% (Poisson AR model). In contrast, the linear AR model had negative magnitude bias (i.e., underestimating the true policy effect), ranging from -4% (with population weights) to -2% (no population weights). For each GLM type, magnitude bias was greater for the AR model compared to the two-way fixed effect or detrended models.

The use of population weights in the linear and log-linear models did not consistently or notably influence magnitude bias. Furthermore, the magnitude bias remained essentially 0% for the linear two-way fixed effects, linear GEE, and linear detrended models for both the gradual and instantaneous policy effect conditions. For all the other models, magnitude bias was consistently higher for the gradual versus the instantaneous effect conditions (e.g., for the negative binomial AR model, magnitude bias was 10% for the instantaneous condition and 23% for the gradual condition).

Root mean square error

Figure 4 shows the average RMSE for simulation conditions with a null treatment effect. Among linear models, AR models had the lowest RMSE (1.08-1.12) compared to the two-way fixed effects models (1.67-1.78), detrended models (1.63-1.69), and GEE models (1.37-1.92) (Fig. 4a). For the two-way fixed effects, detrended, and GEE models, RMSE was lower for the unweighted models than the corresponding weighted models; however, for the AR models, population weighting yielded slightly lower RMSE. Among non-linear models, the negative binomial models had consistently lower RMSE compared to the Poisson and log-linear models (Fig. 4b). For the negative binomial model, the detrended and two-way fixed effects models had the lowest RMSE (0.22) while the AR model had the highest (0.31). Finally, as expected, RMSE was larger for simulation conditions with a gradual policy effect relative to an instantaneous effect (e.g., for the linear population two-way fixed effects model, RMSE = 1.58 for instantaneous and RMSE = 1.95 for gradual).



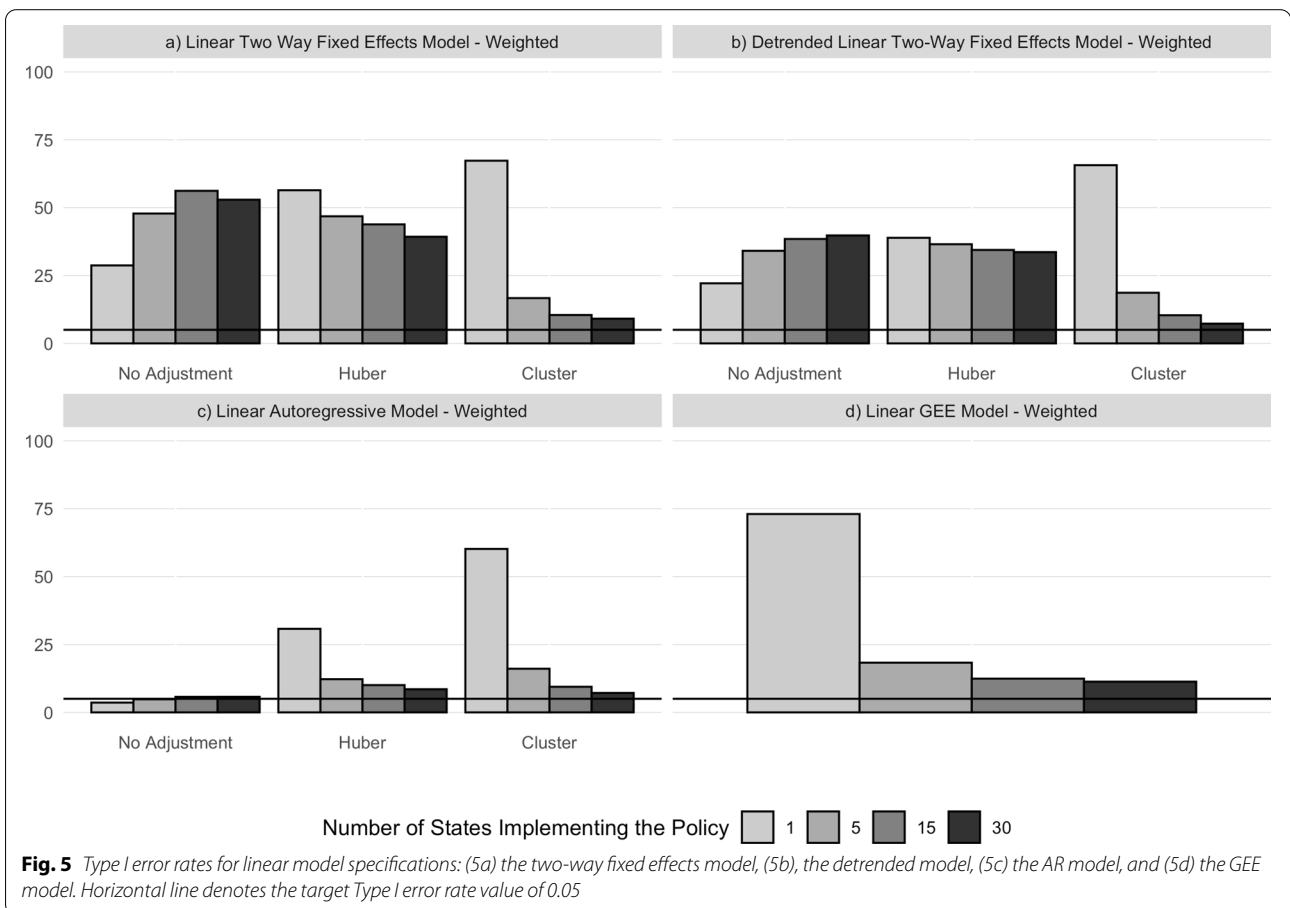
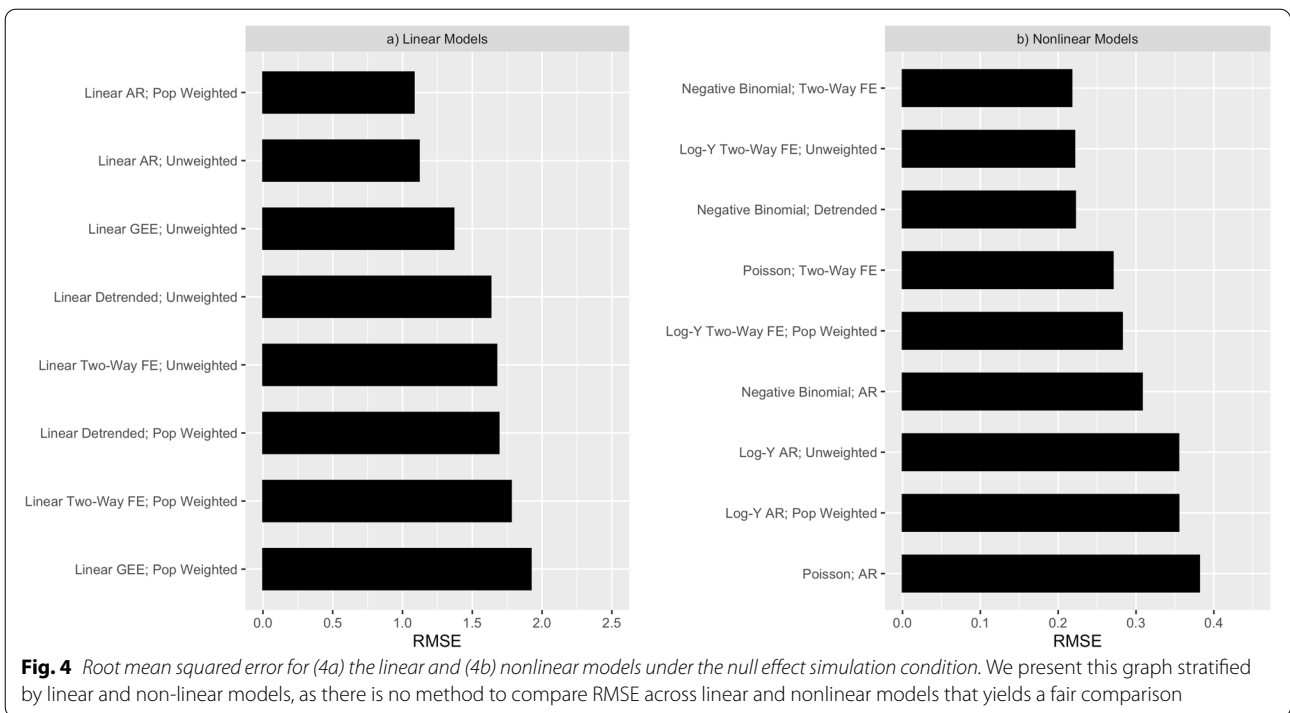
Type I error rates

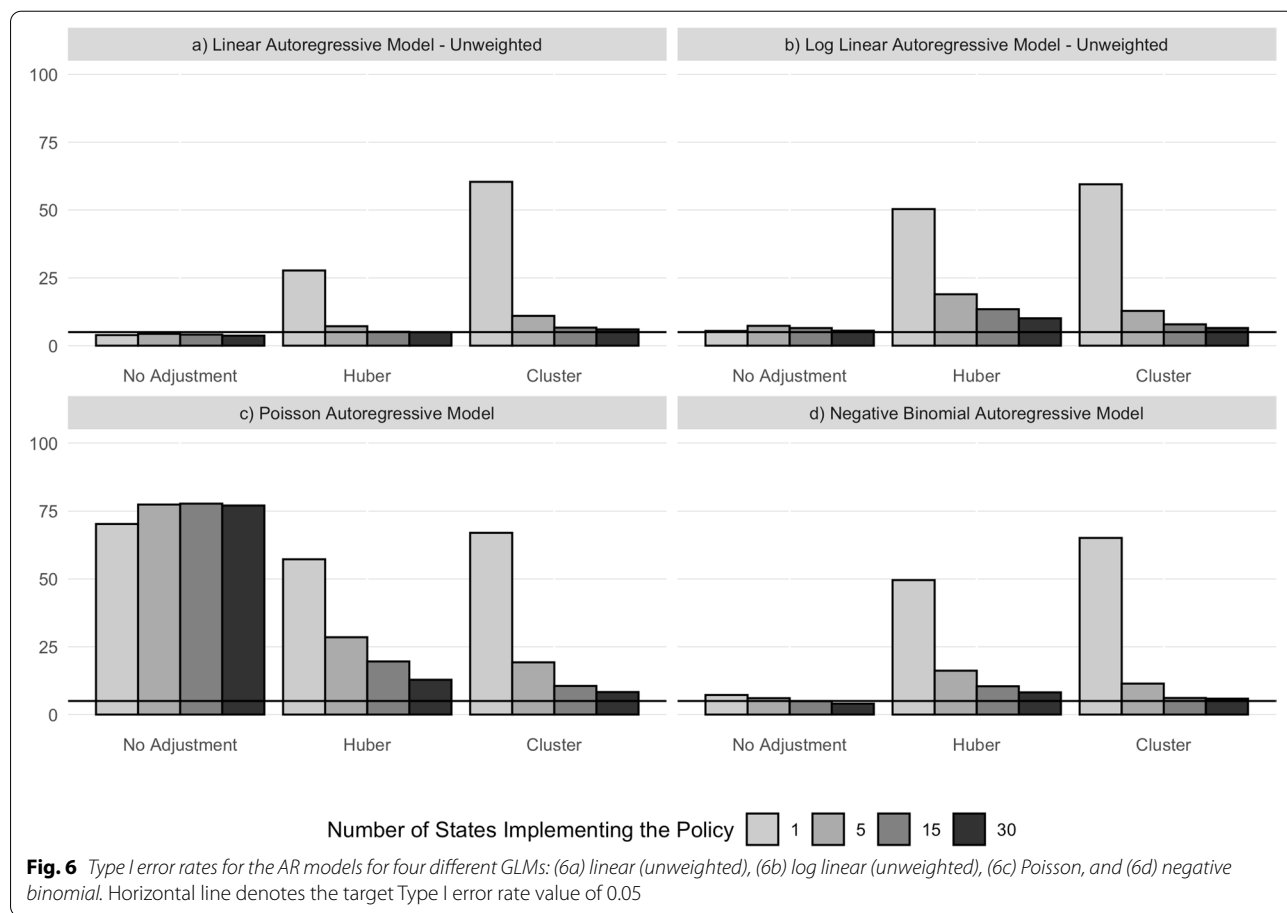
Figure 5 presents the Type I error rates for the four linear models using population weights. Type I error rates were very high for the classic DID two-way fixed effects model (Fig. 5a), ranging up to 67%. Cluster SE adjustment greatly reduced the Type I error rates for this model when 5 or more states implemented a policy, but they were still 2 to 3 times larger than the traditional target of 5%, ranging from 9 to 17%. The detrended model (Fig. 5b) generally had slightly lower Type I error rates than the two-way fixed effects model, with Type I error rates mostly less than 40%. Notably, the AR model (Fig. 5c) did not require use of any SE adjustment to obtain appropriate Type I error rates for conditions with 5 or greater policy states (e.g., Type I error rates ranged from 4 to 6%); in fact, SE adjustments in the AR models tended to inflate the Type I error rates. For linear GEE models (Fig. 5d), Type I error rates were 18% or less for simulation conditions with at least 5 policy states, though rates were still 2-3 times higher than the traditional target of 5%. As in the case of

linear models, AR models performed best, followed by detrended models, then two-way fixed effects models.

For linear models, population weighting yielded slightly higher Type I error rates for the two-way fixed effects, detrended, and GEE models compared to the corresponding unweighted models (see Shiny Application). In contrast, for the AR models, population weighted models did not consistently perform better or worse than unweighted models. Additionally, Type I error rates were higher (by approximately 8 percentage points) for simulation conditions with a gradual relative to an instantaneous effect.

Given the top performance of the AR model, we also present the relative performance of the AR model across four different GLMs: linear (unweighted), log-linear (unweighted), Poisson, and negative binomial (Fig. 6). Similar to the results seen for the linear AR weighted model (Fig. 4), very good Type I error rates are obtained in the absence of SE adjustment for linear AR unweighted model, the log-linear AR unweighted model, and the





negative binomial AR model, regardless of the number of policy states. We note that this does not hold for the Poisson AR model.

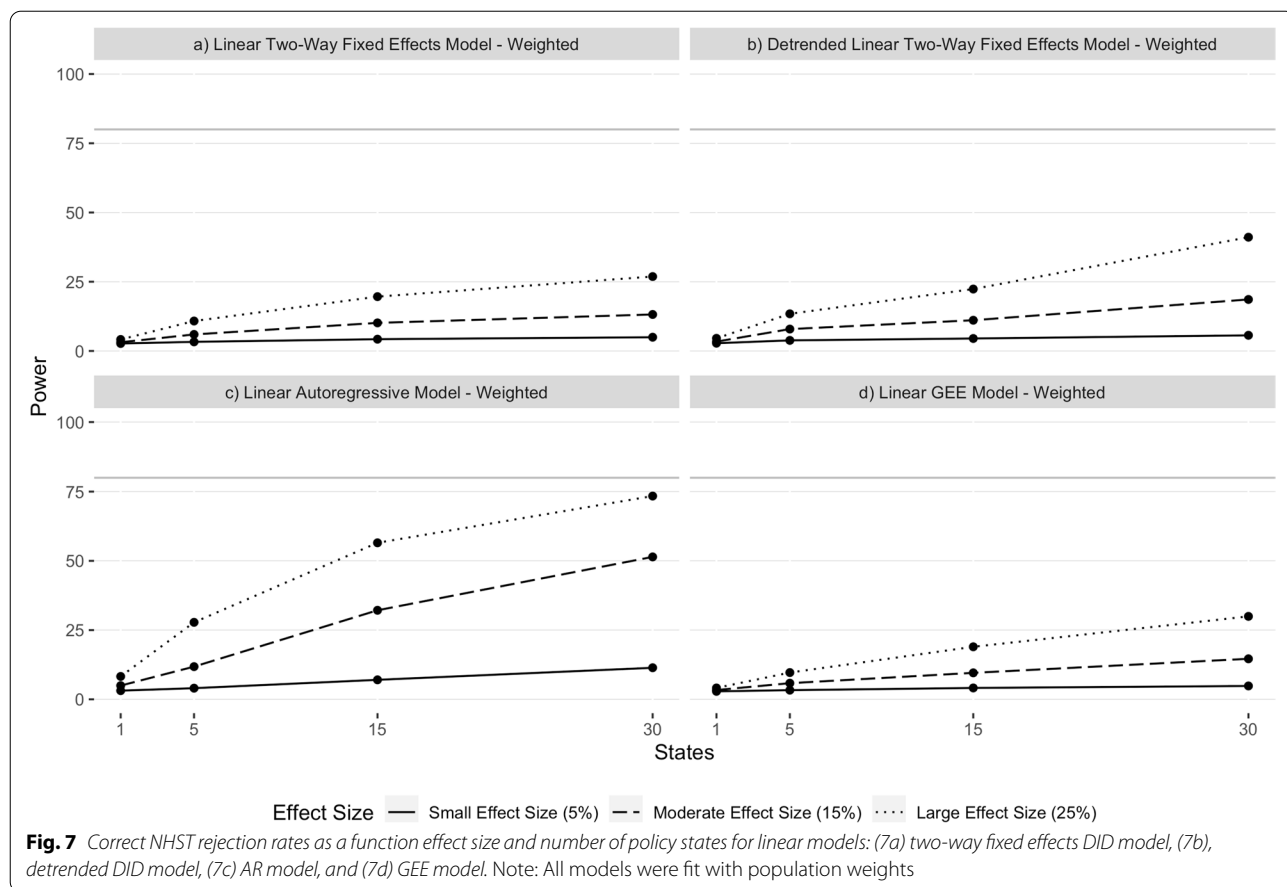
Correct NHST rejection rates

Figure 7 shows correct NHST rejection rates as a function of both the effect size and the number of policy states for the linear models (using population weights). In all cases, as expected, correct rejection rates increased both as the effect size increased and the number of policy states increased, with maximum values obtained for the simulation condition with 30 policy states and a large effect size. For the two-way fixed effects model (Fig. 7a), correct rejection rates were low across all effect sizes, with a maximum value of 27%. In contrast, correct rejection rates were highest for the AR model (Fig. 7c), which achieved a maximum value of 73% (nearly the desired 80% rate). Relative to the two-way fixed effects model, correct rejection rates were similar for the GEE model (maximum value = 30%) and slightly higher for the detrended model (maximum value = 41%). Importantly, all models considered had extremely low correct

rejection rates for simulation conditions with a small effect size – e.g., the rate of correctly rejecting the null hypothesis was 8% for negative binomial models and ranged from 4 to 11% across linear models.

For linear and log-linear models, correct rejection rates tended to be higher for unweighted models relative to weighted models. Specifically, the linear two-way fixed effects model yielded a correct rejection rate of 40% for the unweighted model compared to 27% for the unweighted model for the simulation condition with 30 policy states and a large effect size. Similarly, the unweighted linear AR model yielded the correct rejection rate of 81% (compared to 72% for weighted) and the unweighted GEE model yielded the correct rejection rate of 67% (compared to 30% for weighted). Correct rejection rates were consistently smaller (by 3 percentage points on average) for simulation conditions with a gradual relative to an instantaneous policy effect.

Figure 8 presents correct rejection rates averaged across all simulation conditions in order to highlight relative performance across models. Correct rejection rates were low across all models but were highest for linear AR models (ranging from 22 to 24%) and negative binomial models



(ranging from 20 to 23%). The worst performing models were the linear and log-linear two-way fixed effects models and the linear weighted GEE model (correct rejection rates ranged from 9 to 11%). Correct rejection rates for the Poisson models ranged from 12% (two-way fixed effects model) to 18% (AR model); we note that for all specifications, the Poisson model was outperformed by the corresponding negative binomial model.

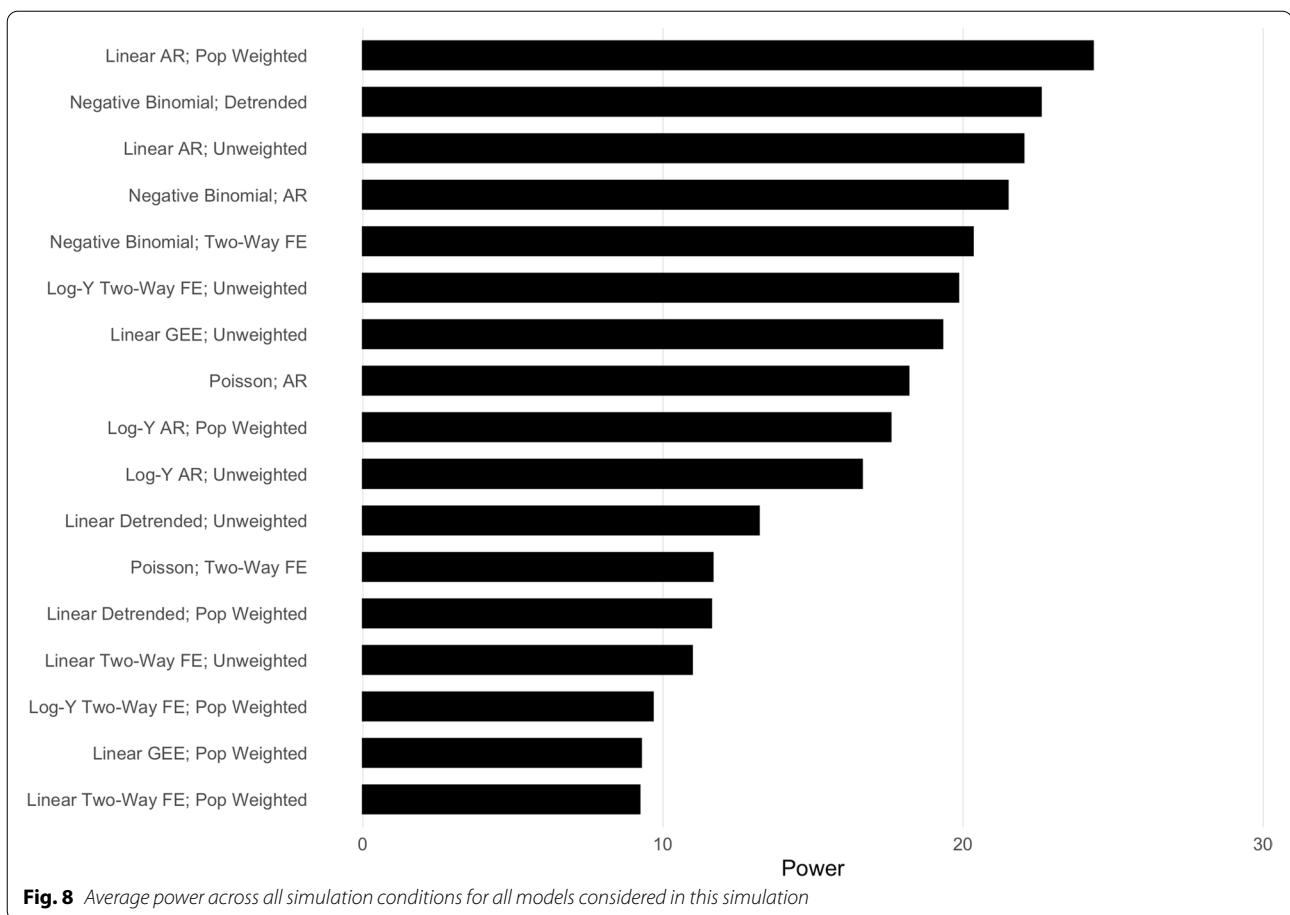
Discussion

State-level policy evaluations commonly use a DID study design; however, model specification varies notably across studies, and the field lacks clear guidance on which models are optimal. We conducted a novel simulation study to compare the relative performance of multiple variations of the two-way fixed effect model traditionally used for DID, using simulated data based on actual national opioid mortality data so as to mirror data features encountered in practice. Specifically, we compared the classic, linear two-way fixed effects DID model to three alternative models: a detrended model, an AR model, and a fixed effect model estimated with GEE

with an AR correlation structure. Within these classes of models, we additionally compared link function specifications, SE estimation methods, and the use of population weighting.

We found that the linear AR model was optimal when the outcome was specified as a mortality rate, and a negative binomial model was optimal when the outcome was specified as a mortality count. Our results highlighted that two widely-used linear DID models – two-way fixed effect and detrended – were consistently outperformed by the less commonly-used AR linear model, which was consistently optimal in terms of directional bias, RMSE, Type I error, and power. We urge applied researchers to move beyond the classic linear two-way fixed effect DID paradigm and consider the use of AR models. Overall, our results indicated notable differences in the performance of the models considered—differences that have substantial implications for conducting and interpreting state-level policy evaluations.

Results from our study are highly consistent with findings from a prior gun policy simulation study [18], as both studies identified AR models as top-performing for



estimating state-level policy effects. We note that our current study considers a broader range of simulation conditions than the gun policy study (e.g., a range of policy effect sizes (5 to 25%) compared to a single effect size (3%)), strengthening the generalizability of the results. We emphasize that the optimal choice of the link function may vary by the characteristics of the outcome variable: the gun policy simulations study, which examined firearm-related mortality, found that the negative binomial AR model was optimal. Our study, which examined opioid-related mortality, identified the linear AR models as optimal, as the negative binomial AR model yielded higher directional and magnitude bias (relative to the linear AR model). This was likely due to the greater relative skew in the distribution of state-level opioid-related deaths compared to firearm-related deaths, suggesting the benefit of running these types of simulations on specific outcome data to ensure selection of the optimal model for a given outcome. To this end, we have created an R package (OPTIC.simRM) for executing these simulations on any repeated measures levels data.

We make recommendations for practice in Table 2. Many of these results have emerged from other studies.

However, they have not been well appreciated in the statistical or applied literature, and questions have remained regarding best practices with real-world data like opioid-related mortality rates. For example, with regard to standard error corrections, prior simulation studies [12, 15] show that cluster adjustments are needed to reduce Type I error rates. Bertrand, et al. (2004) showed that the classic sandwich estimator does poorly with small samples; that paper also shows DID without adjustment has high Type I errors (approximately 45%) in their case study data, where they randomly simulated random “placebo” laws, as done here. Our work extends prior work by highlighting the challenges specific to the context of evaluating state-level opioid policies with respect to opioid-related mortality, a widely-used outcome in the field.

Furthermore, researchers and policymakers must recognize the inherent implications of a fundamentally limited sample size of 50 states (of which perhaps only a few, or even a single state, implemented the policy of interest) for continued reliance on p-values to determine statistical significance. Under traditional NHST, correct rejection rates for the majority of scenarios were

Table 2 Key Takeaways for the Practice

When modeling opioid-related mortality as a crude rate in a linear model, including an AR term significantly improves estimation performance with regard to RMSE.

When modeling counts of opioid-related mortality, a negative binomial model performs better than a Poisson model.

Linear AR models performed optimally with respect to bias, RMSE, Type I error, and correct rejection rates in the context of estimating state-level policy effects of opioid-related mortality

Sample size matters for SE estimation. For linear and log-linear models, clustered SEs significantly improved estimation when the treated group comprised 15+ states, yet they had worse performance than unadjusted SEs in the case of only a single treated state.

extremely low, lower than 25% across all scenarios considered. Rates were only above 50% for the best performing models and when there was a large effect size (25%) and the most balanced allocation to treatment versus control. Additionally, Type I error rates for the majority of models relying on NHST were unreasonably high when fewer than 15 states are implementing a new policy, meaning that these models could yield a significant effect estimate when in fact such an effect does not exist. It is critical for researchers to use models that minimize Type I error rates whenever possible; using standard error corrections to ensure a Type I error rate of 0.05 is necessary in this context when performing NHST. However, echoing repeated calls from statisticians [41] to move beyond inferences based on *p*-values, we strongly recommend that the field overall reduce reliance on traditional NHST, given concerns across a range of scientific areas regarding the use of often arbitrary *p*-value thresholds within that framework [41]. The applied field of state-policy research is still overwhelmingly implementing traditional NHST -- all of the studies in our recent opioid literature review [22] relied on traditional NHST to determine if the effect of the primary policy was “statistically significant.” One alternative approach that holds promise is using Bayesian approaches to estimate state-level policy effects. Bayesian methods can be used to estimate effects that directly correspond to the likely effects of the yes/no decisions facing policymakers considering such legislation (namely, the probability that a given law is associated with an increase or a decrease in firearms death), and can also more accurately reflect the large amount of uncertainty in these analyses. For an illustration of an Bayesian approach in context of gun policy, see Schell, et al. (2018).

Fundamentally, longitudinal and panel data do not conform to the traditional regression assumption of independent and identically-distributed (*iid*) residuals. When considering various modeling approaches, it may be helpful to distinguish between three distinct phenomena that contribute to departures from *iid* residuals and to have diagnostic checks for which deviation might be occurring in a given data set: outcome autocorrelation, clustering at the state-level, and departures from model distributional assumptions.

First, some degree of autocorrelation in the outcome timeseries is likely. Our results from both the current simulation, as well as the prior gun policy simulation, highlight that cause-specific mortality outcomes are likely to be highly autocorrelated. Similarly, autocorrelation is expected for other key health policy outcomes, such as disease-specific incidence rates and healthcare spending measures. The presence of autocorrelation following an AR1 structure can be assessed using the Durbin-Watson test; more generally, an autocorrelation function (ACF) plot, also called a correlogram, can be used to assess the degree of autocorrelation across lagged time periods (Friendly 2002, Durbin and Watson 1971). Autocorrelation is effectively addressed through the use of an AR model or GEE with an AR correlation structure. See Beard, et al. (2019) for a pragmatic discussion of timeseries data analysis in the context of addition research.

With regard to state-level clustering, one can compare cluster-adjusted versus unadjusted standard errors or compute intraclass correlation coefficients (ICC) to understand how strongly clustering will affect the study design. However, sample size is a key consideration and diagnostics such as ICCs are not reliable when sample sizes are less than 30 [42]. Our results indicate that when in the context of only a single treated state, cluster and Huber SE adjustments yield worse performance than no adjustment. While this has been previously demonstrated in the literature [13], these insights are often not reflected in the applied literature.

Table 3 provides additional insights into how applied researchers might more readily adapt their current practice to include use of AR models. It lays out the analytic steps necessary for using an AR model in practice, including creation of lagged variables, assessing diagnostics regarding the optimal number of lags, and recoding the policy indicator using change level coding. Critically, if Step 2 (i.e., using common diagnostics for the AR models) shows that lags are not correlated, the AR model will not be advantageous, as leveraging correlated lagged values is key to the AR model's ability to improve estimation precision in the context of time series data such as opioid mortality rates.

Table 3 Key processing steps for implementing AR models in practice when using state-level data

Step 1: Create the needed lag terms for the outcome in the data; we recommend creating between 5 and 10 lags to assist in Step 2.
Step 2: Use common diagnostics for AR models to determine the optimal number of lags for your model; we have found using the partial autocorrelation plot particularly helpful for detecting the optimal number of lags [43].
Step 3: Ensure that the policy variable/indicator is properly coded using change coding (e.g., $(A_{it} - A_{it-1})$).
Step 4: Ensure that the right-hand side of the regression model controls for the optimal number of lags, change coding of the policy variable, time fixed effects, and other key controls. Note – state fixed effects are not needed.
Step 5: Check to ensure that you have not overfit your model by confirming that you have at least 10 observations per control variable on the right hand side of the regression model.
Step 6: Do not include any cluster adjustment to the standard errors.

Fundamentally, both DID and AR models rely on the parallel counterfactual trend assumption. The two models differ in how they parametrically control for observed differences between the treated and control states -- the DID model relies on state and time fixed effects while the AR model relies on lagged outcomes and time fixed effects. We note that our data generating process only generated synthetic observations for the treated states in the post-period (in order to induce a policy effect of a known magnitude), rather than generating complete trajectories for both treated and untreated states. As such, we (like applied researchers) were not privy to the “truth” about whether the parallel counterfactual trend assumption (the core identifying DID assumption) was upheld. However, since our treated and control states were selected randomly, we do not expect these groups to exhibit systematically differential trajectories. We highlight that the parallel counterfactual trends assumption is untestable, given that this assumption pertains to unobservable counterfactual outcomes. Yet in practice, researchers often conduct a so-called “partial test of parallel trends” by statistically testing whether the pre-intervention trends differ across groups [5, 6]. We discourage this practice: it is not informative regarding the actual underlying counterfactuals and indeed may induce a false sense of confidence in the validity of the common trends assumption. However, a detrended model may be used as a robustness check: if the classic two-way fixed effect model and a detrended model that allows for differential state trajectories over time yield similar policy effects, this provides some evidence in favor of the common trends assumption. See Bilinski and Hatfield (2020) and Rambachan and Roth (2019) for further discussion of these issues and alternative strategies for assessing plausibility of the parallel counterfactual trends assumption. We also note that if the parallel counterfactual trends assumption holds on one model scale (e.g. linear) it may not automatically hold on other scales (e.g., count).

Our simulation design has several limitations and future research is needed to build upon this work. By

randomly selecting states to enact a given policy, this simulation represents the simplified scenario in which there is no confounding by observed or unobserved covariates (including lagged values of the outcome). Future simulation work will consider more complex scenarios, including where such confounding exists given the likelihood that states implementing certain policies differ from states that do not. A growing set of methods aims to deal with potential confounding and need to be considered, including: incorporation of propensity score weighting into the DID framework [44], synthetic control methods [45–47] and augmented synthetic control methods [48], and doubly-robust DID estimators [49], as well as DID extensions that are robust to violations of the parallel trends assumption [50]. More broadly, our simulation study did not exhaustively compare models used in practice: for example, we did not consider random effect models in this study, as prior work indicated that they are not commonly used in opioid policy evaluations [22]. Furthermore, while the timing of policy enactment varied across treated states, our simulated data had a constant policy effect across states and across time, which may be an unlikely assumption in some contexts. Recent work has showed that in the presence of heterogeneity in policy timing and treatment effects, the classic linear two-way fixed effect DID model yields biased treatment effect estimates [10, 51, 52]. Future work is needed to investigate relative model performance in the context of treatment heterogeneity.

Overall, given the consistency of our current findings and those from the previous gun policy simulation study, it is likely that the advantages of AR models may generalize to contexts beyond opioid- and firearm-related mortality. In particular, many of the methodological considerations discussed in this paper are applicable to evaluation studies of Covid-19 policies, particularly state-level policies. However, we highlight that the Covid-19 context may pose unique complexities, including those introduced by the large number of heterogeneous

Covid-19 policies implemented at the city and county levels, in addition to state and federal responses. As such, a thorough understanding of specific policy environments is essential for conducting rigorous policy evaluations, as co-occurring policies and relative policy timing have important analytic implications. We refer readers to a recent methodological review of published Covid-19 policy evaluation studies [53] for additional discussion regarding specific analytic considerations in the Covid-19 context as well as Schuler, et al. (2020) which details additional methodological considerations relevant to the opioid policy context. Broadly, future work should extend this line of simulation research through careful consideration of additional outcomes and policy contexts.

As noted by Schell, et al. (2018): “A scientific field built on studies with such low power (e.g., less than 0.20) will have a large fraction of significant results that are spurious, a substantial proportion of significant effects that are in the wrong direction, and significant effects that substantially overestimate the true effect size (56).” There is an urgent need for the field to develop more robust and powerful methods to help guide state policy decisions in the face of numerous, ongoing public health crises in the United States (e.g., opioid epidemic, gun violence, COVID-19). In particular, development and adoption of statistical approaches that improve accuracy while rigorously acknowledging uncertainty in policy effect estimates are crucially needed to address the needs of policy researchers, key decision makers, and stakeholders.

Conclusions

Study findings highlight notable limitations of commonly used statistical models for DID designs--designs widely used in opioid policy studies and in state policy evaluations more broadly. In contrast, the optimal model identified (the AR model) is rarely used in state policy evaluation. We urge applied researchers to move beyond the classic DID paradigm and adopt use of AR models.

Appendix: Additional technical details and code

Standard error (SE) estimation: As noted, we estimated the SE in the three ways for each model (except GEE): (1) no adjustment; (2) Huber adjustment (robust estimators that attempt to adjust the SE for violations of distributional assumptions (White 1980; Zeileis 2004)); and (3) cluster adjustment (adjustments to account for possible violations of the assumed independence of observations within states (White 1980; Zeileis 2004, 2006)). To illustrate the differences, we can write each adjustment in the following way (STATA Undated):

- (1) No adjustment: This is just the normal ordinary least squares (OLS) estimator. The generic formula for a regression model with covariate matrix \mathbf{X} capturing all the needed data information from the sample of observations can be written:

$$V_{OLS} = s^2 * (\mathbf{X}'\mathbf{X})^{-1}$$

where $s^2 = \left(\frac{1}{N-k}\right) \sum_{i=1}^N e_i^2$, k denotes the number of parameters in a model and e_i denotes the estimated residual for the i -th observation.

- (2) Huber adjustment: This is the typical robust (unclustered) variance estimator and can be written as:

$$V_{huber} = s^2 * (\mathbf{X}'\mathbf{X})^{-1} * \left[\sum_{i=1}^N (e_i * x_i)' * (e_i * x_i) \right] * (\mathbf{X}'\mathbf{X})^{-1}$$

where x_i denotes the i -th observation's vector of covariate values going into the covariate matrix \mathbf{X} .

- (3) Cluster adjustment: This can be written as:

$$V_{cluster} = (\mathbf{X}'\mathbf{X})^{-1} * \left[\sum_{j=1}^{n_c} (u_j' * u_j) \right] * (\mathbf{X}'\mathbf{X})^{-1}$$

where $u_j = \sum_{i \in cluster} (e_i * x_i)$ and n_c denotes the number of clusters.

Our simulations were run in R so we used the `vcovHC` package (RDocumentation Undated) to estimate Huber adjusted SEs by running the following command on the model output (here denoted by `m1`): `vcovHC(m1, type="HC0")`. As noted in R, `type = "HC0"` corresponds to the traditional Huber-White robust SE estimator. We confirmed that this Huber SE adjustment is the same as the “robust” SE option used by Stata (`vce(robust)`). To estimate the cluster adjusted SE, we utilized the following code to obtain the needed sandwich adjusted SE:

```
robust.se <- function(model, cluster){
  require(sandwich)
  require(lmtest)
  M <- length(unique(cluster))
  N <- length(cluster)
  K <- bigel$rank
  dfc <- (M/(M - 1)) * ((N - 1)/(N - K))
  uj <- apply(estfun(bigel), 2, function(x) tapply(x, cluster, sum));
  rcse.cov <- dfc * sandwich(bigel, meat = crossprod(uj)/N)
  rcse.se <- coeftest(bigel, rcse.cov)
  return(list(rcse.cov, rcse.se))}
```

```
clustervar <- mapply(paste, "State.", x$STATE, sep = "")
m1$coefficients <- robust.se(m1, clustervar)[[2]]
```

We confirmed that the results from this function produces identical SE estimates as the cluster adjustment command used in Stata (`vce(cluster state)`).

Weighting: State-population weights were treated as analytic weights, not survey weights, within the analyses.

Model differences: Note that since the AR models use lagged outcomes in the regression model, they utilized one less year of data from the time series than the other models considered.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12874-021-01471-y>.

Additional file 1.

Acknowledgements

The authors would like to thank Kosali Simon and the participants of the 2019 ASHEcon session that provided helpful comments on an earlier draft. Finally, the authors want to thank Hilary Peterson for her assistance with manuscript preparation and submission and Mary Vaiana for careful review and editing of the submission.

Authors' contributions

The authors jointly conceived the idea for the study. BG and TS designed, developed and implemented the original simulation code; BG lead all adaptations and runs under consultation with all authors. EM developed needed graphics and the associated Shiny app for this work. BG, MS, and ES drafted the manuscript with input from all authors. All authors extensively edited and provided input on all phases of the study and all authors read and approved the final manuscript.

Funding

This research was financially supported through a National Institutes of Health (NIH) grant (P50DA046351) to RAND (PI: Stein). NIH had no role in the design of the study, analysis, and interpretation of data nor in writing the manuscript.

Availability of data and materials

The data that support the findings of this study are available from National Vital Statistics System (NVSS) Multiple Cause of Death mortality files (1989 through 2018) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The data can be request under a similar license (data use agreement) from the NVSS.

Declarations

Ethics approval and consent to participate

The RAND Corporation Institutional Review Board deemed this study exempt (Human Subjects Assurance Number 00003425 (6/22/2023)).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202, USA. ²Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA. ³Vanderbilt University Medical Center and School of Medicine, Nashville, TN 37232, USA. ⁴RAND Corporation, Santa Monica, CA 90401, USA. ⁵RAND Corporation, Pittsburgh, PA 15213, USA. ⁶University of Southern California, Los Angeles, CA 90089, USA.

Received: 4 May 2021 Accepted: 15 November 2021

Published online: 13 December 2021

References

1. Abadie A, Cattaneo M. Econometric methods for program evaluation. *Annual Review of Economics*. 2018;10:465–503.
2. 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.
3. Blundell R, Costa DM. Alternative approaches to evaluation in empirical microeconomics. *J Hum Resour*. 2009;44(3):565–640.
4. O'Neill S, Kreif N, Grieve R, Sutton M, Sekhon JS. Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Serv Outcome Res Methodol*. 2016;16:1–21.
5. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health*. 2018;39:453–69.
6. Ryan AM, Burgess JF Jr, Dimick JB. Why we should not be indifferent to specification choices for difference-in-differences. *Health Serv Res*. 2015;50(4):1211–35.
7. Chaisemartin Cd, D'Haultfoeuille X. Two-way fixed effects estimators with heterogeneous treatment effects 2019.
8. Daw JR, Hatfield LA. Matching and regression to the mean in difference-in-differences analysis. *Health Serv Res*. 2018;53(6):4138–56.
9. Daw JR, Hatfield LA. Matching in difference-in-differences: between a rock and a hard place. *Health Serv Res*. 2018;53(6):4111–7.
10. Goodman-Bacon A. Difference-in-differences with variation in treatment timing; 2018.
11. Brewer M, Crossley T, Joyce R. Inference with difference-in-differences revisited. *Journal of Economic Methods*. 2017;7(1):2156–6674.
12. Abhay A, Donohue III J, Zhang A. The impact of right to carry laws and the NRC Report: The latest lessons for the empirical evaluation of law and policy. NBER Working Paper No 18294. 2014.
13. Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? *Q J Econ*. 2004;119(1):249–75.
14. Donald SG, Lang K. Inference with difference-in-differences and other panel data. *Rev Econ Stat*. 2007;89(2):221–33.
15. Helland E, Tabarrok A. The fugitive: evidence on public versus private law enforcement from bail jumping. *J Law Econ*. 2004;47(1):93–122.
16. Schell T, Griffin B, Morral A. Evaluating methods to estimate the effect of state Laws on firearm deaths: a simulation study. RR-2685-RC. Santa Monica, CA: RAND Corporation; 2018.
17. Schuler MS, Griffin BA, Cerdá M, McGinty EE, Stuart EA. Methodological challenges and proposed solutions for evaluating opioid policy effectiveness. *Health Serv Outcomes Res Method*. 2020.
18. Schell T, Griffin B, Morral A. Evaluating methods to estimate the effect of state Laws on firearm deaths: a simulation study. Santa Monica, CA: RAND Corporation; 2018.
19. Ioannidis JPA, Stanley TD, Doucouliagos H. The power of Bias in economics research. *Econ J*. 2017;127(605):F236–F65.
20. Haber N, Clarke-Deelder E, Salomon J, Feller A, Stuart EA. Policy evaluation in COVID-19: A guide to common design issues. arXiv:2009.01940v5. arXiv. 2020.
21. Centers for Disease Control and Prevention. Vital Statistics Rapid Release: Provisional Drug Overdose Death Counts 2020 [Available from: <https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm>].
22. Schuler MS, Heins SE, Smart R, Griffin BA, Powell D, Stuart EA, et al. The state of the science in opioid policy research. *Drug Alcohol Depend*. 2020;214:108137.
23. Bilinski A, Hatfield LA. Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions; 2020.
24. Wolfers J. Did unilateral divorce laws raise divorce rates? A reconciliation and new results. *Am Econ Rev*. 2006;96(5):1802–20.
25. Cochrane D, Orcutt GH. Application of least squares regression to relationships containing auto-correlated error terms. *J Am Stat Assoc*. 1949;44(245):32–61.
26. Wooldridge J, Jeffrey M. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, MA: MIT Press; 2010.

27. Liang K-Y, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13–22.
28. White H. A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica*. 1980;48:817.
29. Zeileis A. Econometric computing with HC and HAC covariance matrix estimators. *J Stat Softw*. 2004;11(10):1–17.
30. Zeileis A. Object-oriented computation of sandwich estimators. *J Stat Softw*. 2006;16(9):1–16.
31. Arellano M. Computing robust standard errors for within-groups estimators. *Oxford B Econ Stat*. 1987;49(4):431–4.
32. Ali MM, Dowd WN, Classen T, Mutter R, Novak SP. Prescription drug monitoring programs, nonmedical use of prescription drugs, and heroin use: evidence from the National Survey of drug use and health. *Addict Behav*. 2017;69:65–77.
33. Buchmueller TC, Carey C. The effect of prescription drug monitoring programs on opioid utilization in Medicare. *Am Econ J-Econ Polic*. 2018;10(1):77–112.
34. McInerney M. The affordable care act, public insurance expansion and opioid overdose mortality; 2017.
35. Paulozzi LJ, Kilbourne EM, Desai HA. Prescription drug monitoring programs and death rates from drug overdose. *Pain Med*. 2011;12(5):747–54.
36. Abouk R, Pacula RL, Powell D. Association between state Laws facilitating pharmacy distribution of naloxone and risk of fatal overdose. *JAMA Intern Med*. 2019;179(6):805–11.
37. Chan NW, Burkhardt J, Flyr M. The effects of recreational marijuana legalization and dispensing on opioid mortality. *Econ Inq*. 2020;58(2):589–606.
38. Kilby A. Opioids for the masses: welfare tradeoffs in the regulation of narcotic pain medications. Cambridge: Massachusetts Institute of Technology; 2015.
39. Frost J. Regression analysis: an intuitive guide for using and interpreting linear models: James D. Frost; 2020. Available from: <https://statisticsbyjim.com/regression/regression-analysis-intuitive-guide/>.
40. U.S. Department of Labor. Bureau of Labor Statistics 2019 [Available from: <https://www.bls.gov/>].
41. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2):129–31.
42. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med*. 2002;21(9):1331–5.
43. Brockwell P, Davis R. *Introduction to Time Series and Forecasting*. 2nd ed: Springer-Verlag; 2002.
44. Stuart EA, Huskamp HA, Duckworth K, Simmons J, Song Z, Chernew M, et al. Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Serv Outcome Res Methodol*. 2014;14(4):166–82.
45. Xu YQ. Generalized synthetic control method: causal inference with interactive fixed effects models. *Polit Anal*. 2017;25(1):57–76.
46. Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S. Synthetic difference in differences. 2019.
47. 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.
48. Ben-Michael E, Feller A, Rothstein J. The Augmented Synthetic Control Method. 2019.
49. Sant'Anna PHC, Zhao J. Doubly Robust Difference-in-Differences Estimators. *Journal of Econometrics*. 2020.
50. Ye T, Keele L, Hasegawa R, Small DS. A Negative Correlation Strategy for Bracketing in Difference-in-Differences with Application to the Effect of Voter Identification Laws on Voter Turnout. 2020. Contract No.: arXiv:2006.02423.
51. Callaway B, Sant'Anna PHC. Difference-in-differences with multiple time periods and an application on the minimum wage and employment. 2018.
52. Sun L, Abraham S. Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects 2020 [Available from: <http://economics.mit.edu/files/14964>].
53. Haber NA, Clarke-Deelder E, Feller A, Smith ER, Salomon J, MacCormack-Gelles B, et al. Problems with Evidence Assessment in COVID-19 Health Policy Impact Evaluation (PEACHPIE): A systematic review of study design and evidence strength. *medRxiv*. 2021:2021.01.21.21250243.
54. Beard E, Marsden J, Brown J, Tombor I, Stapleton J, Michie S, et al. Understanding and using time series analyses in addiction research. *Addiction*. 2019;114(10):1866–84.
55. Rambachan A, Roth J. An honest approach to parallel trends. 2019.
56. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect Psychol Sci*. 2014;9(6):641–51.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

