NBER WORKING PAPER SERIES

VARIATION IN PERFORMANCE OF COMMONLY USED STATISTICAL METHODS FOR ESTIMATING EFFECTIVENESS OF STATE-LEVEL OPIOID POLICIES ON OPIOID-RELATED MORTALITY

Beth Ann Griffin
Megan S. Schuler
Elizabeth A. Stuart
Stephen Patrick
Elizabeth McNeer
Rosanna Smart
David Powell
Bradley Stein
Terry Schell
Rosalie Liccardo Pacula

Working Paper 27029
http://www.nber.org/papers/w27029

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020

This research was financially supported through an NIH grant (P50DA046351) to RAND (PI: Stein). 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Variation in Performance of Commonly Used Statistical Methods for Estimating Effectiveness of State-Level Opioid Policies on Opioid-Related Mortality
Beth Ann Griffin, Megan S. Schuler, Elizabeth A. Stuart, Stephen Patrick, Elizabeth McNeer, Rosanna Smart, David Powell, Bradley Stein, Terry Schell, and Rosalie Liccardo Pacula
NBER Working Paper No. 27029
April 2020
JEL No. C1,C12,C15,C18,C22,C23,C4,C52,C54,H0,H51,H7,I18,I28,K32

ABSTRACT

Over the last two decades, there has been a surge of opioid-related overdose deaths resulting in a myriad of state policy responses. Researchers have evaluated the effectiveness of such policies using a wide-range of statistical models, each of which requires multiple design choices that can influence the accuracy and precision of the estimated policy effects. This simulation study used real-world data to compare model performance across a range of important statistical constructs to better understand which methods are appropriate for measuring the impacts of state-level opioid policies on opioid-related mortality. Our findings show that many commonly-used methods have very low statistical power to detect a significant policy effect (< 10%) when the policy effect size is small yet impactful (e.g., 5% reduction in opioid mortality). Many methods yielded high rates of Type I error, raising concerns of spurious conclusions about policy effectiveness. Finally, model performance was reduced when policy effectiveness had incremental, rather than instantaneous, onset. These findings highlight the limitations of existing statistical methods under scenarios that are likely to affect real-world policy studies. Given the necessity of identifying and implementing effective opioid-related policies, researchers and policymakers should be mindful of evaluation study statistical design.

Beth Ann Griffin
RAND Corporation
1200 South Hayes
Arlington, VA 22202
bethg@rand.org

Megan S. Schuler
RAND Corporation
20 Park Plaza, Suite 920
Boston, MA 02116
mschuler@rand.org

Elizabeth A. Stuart
Johns Hopkins University
624 N Broadway, Room 804
Baltimore, MD 21205
estuart@jhsph.edu

Stephen Patrick
Vanderbilt University Medical Center
2200 Children's Way
11111 Doctors' Office Tower
Nashville, TN 37232
stephen.patrick@vumc.org

Elizabeth McNeer
Vanderbilt University Medical Center
2200 Children's Way
11111 Doctors' Office Tower
Nashville, TN 37232
elizabeth.mcneer@vumc.org

Rosanna Smart
RAND Corporation
1776 Main Street
Santa Monica, CA 90407
rsmart@rand.org

David Powell
RAND Corporation
1200 South Hayes
Arlington, VA 22202
dpowell@rand.org

Bradley Stein
RAND Corporation
4570 Fifth Avenue
Pittsburgh, PA 15213
stein@rand.org

Terry Schell
RAND Corporation
1776 Main Street
Santa Monica, CA 90401
tschell@rand.org

Rosalie Liccardo Pacula
Sol Price School of Public Policy and
Schaeffer Center for Health Policy & Economics
University of Southern California
635 Downey Way, VPD 514J
Los Angeles, CA 90089
and NBER
rmp_302@usc.edu

An appendix is available at: http://www.nber.org/data-appendix/w27029

1. INTRODUCTION

In 2017, over 47,000 Americans died from opioid overdoses, a rate of more than 130 fatalities per day (Centers for Disease Control and Prevention 2018). In response, states have adopted a myriad of policies and initiatives to curtail the crisis, including those designed to decrease opioid analgesic use, increase access to effective treatment for opioid use disorder, and increase utilization of naloxone, an overdose reversal medication (Mauri, Townsend, and Haffajee 2019, Haegerich et al. 2019). As states sought to make timely policy decisions in the face of limited resources, opioid-policy researchers evaluated the effectiveness of implemented policies using a range of statistical and econometric models.

Studies using these methods often inform policymakers' decisions, but to date, there have been no comprehensive examinations of the relative performance of these methods across settings realistic for state policy evaluation, although various reviews of policy evaluations have been done (Schuler Under review , Mauri, Townsend, and Haffajee 2019, Haegerich et al. 2019). A recent review of the opioid-policy literature suggests that approximately 75% of opioid policy evaluation studies using longitudinal data estimate a policy's effectiveness using comparative case study approaches (commonly referred to as difference-in-difference [DID], comparative interrupted time series [CITS], group panel data, or event studies; (Schuler Under review)). Longitudinal data implies data where one has repeated measures of the outcomes of interest, measured at an aggregate geographic area (e.g., states) with some "exposed" (treated) and "unexposed" (control) locations, where the exposed locations may implement the policy of interest at different times. We assumed state-level implementation (so maximum N=50) and annual measurement of outcomes. While some publications provide guidance on how best to handle policy evaluations using longitudinal data (Basu, Meghani, and Siddiqi 2017, O'Neill et al. 2016, Wing, Simon, and Bello-Gomez 2018, Abadie and Cattaneo 2018, Blundell and Costa Dias 2009), methodological best practices have not been fully adopted by applied opioid policy researchers, and there may be unique considerations in the opioid context that influence practice guidelines.

There are numerous aspects to a study design examining the impact of opioid policies on mortality (or other outcomes) that can influence the accuracy and precision of estimates generated, including: very low outcome occurrence rate (mortality), sample size (both in terms of number of treated states when evaluating a particular opioid policy as well as longitudinal time points), differences across states prior to opioid policy adoption, and specification of regression models (including accounting for repeated measures and assumptions regarding timing of the policy effect). Each choice influences how well the study can estimate the policy effect both in terms of accuracy (e.g., how close to the truth the estimated policy effect is) and statistical precision (e.g., the width of the resulting confidence intervals). The urgency of the opioid crisis necessitates that researchers and policymakers bring accurate, robust statistical methods to bear on identifying and implementing effective state policies; to our knowledge, the influence of each of these data elements and design choices on statistical performance of a particular model is not well-known within the context of opioid-policy evaluations.

There are a growing number of studies highlighting challenges and limitations to using DID models, particularly when the key assumptions of the DID model do not hold (Blundell and Costa Dias 2009, Daw and Hatfield 2018a, b, Ryan, Burgess, and Dimick 2015)or when a study involves small sample sizes (Brewer, Crossley, and Joyce 2017). Additionally, it has been well-established that standard error corrections that attempt to adjust for violations of the assumed independence of the repeated measures in longitudinal datasets are needed to obtain accurate Type I error rates (Abhay, Donohue III, and Zhang 2014, Bertrand, Duflo, and Mullainathan 2004, Donald and Lang 2007, Helland and Tabarrok 2004, Schell, Griffin, and Morral 2019). Despite the wealth of knowledge concerning best practices and the challenges of using DID models in various settings, there have been no comphrenensive examinations of their performance across settings realistic for state opioid policy evaluation. In fact, we were aware of only one other study considering the appropriateness of methodological approaches applied to another area of state public health policy; namely, Schell, Griffin, and Morral (2019) considers the appropriateness of using various analytic approaches for evaluating state gun policy laws on outcomes. This study aims to help fill this gap and the notable dearth of rigorous studies examining the appropriateness of analytic methods being commonly used to evaluate important public health problems and make policy recommendations.

More specifically, this study seeks to provide needed guidance about which statistical methods are most appropriate for measuring the impacts of state-level opioid policies on opioid-related mortality, with lessons that apply to state policy evaluations more broadly. Using a simulation study based on observed state-level opioid related outcomes, we assessed the relative performance of multiple statistical methods commonly used in evaluation studies of state-level opioid policies. Specifically, we compared their performance regarding Type I error rates, bias, statistical power, and root mean squared error to facilitate a more informed consideration of the existing literature and to support sound future state policy evaluations. We found that some commonly-used methods have poor statistical performance, and we provide important insights to statisticians and researchers regarding methods to estimate policy effects. We show that there is still work to be done regarding development of methods that provide accurate results in these complex policy settings.

2. METHODS

In this section, we describe the data structure and general empirical models considered in our simulation study as well as the different features of the statistical models that we explored in our simulation study.

2.1 Data Structure

The data structure we considered in this study was longitudinal, repeated measures data measured on an annual basis where states were the units of interest. We used repeated measures of the outcome (opioid-related mortality) measured annually over 18 years, clustering time points within state, providing a total number of observations is 50*18 = 900. We did not consider individual-level data.

2.2 General Empirical Models Considered

The focus of our simulation study was to compare statistical methods for estimating policy impact using annual state-level outcomes, given a policy landscape in which states implemented a given policy at different times. A common analytic approach for such aggregate longitudinal data in the health services literature, including opioid policy studies, is broadly termed the DID approach. A DID study design seeks to account for both selection bias (by controlling for differences in pre-policy outcomes between policy and comparison states) and historical trends (by controlling for temporal trends that are unrelated to the policy). Essentially, 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 “difference in differences” provides an estimate of the policy effect, while subtracting out potential confounding arising from systematic differences between states implementing versus not implementing a policy, temporal trends, and other exogenous factors.

Below, we introduce key notation for the general DID model framework our simulation study focuses on. Let A_{it} = 0 or 1 denote an indicator for whether state *i* (where $i = 1, \dots, N$) has implemented the generic policy of interest at time *t* (where $t = 1, \dots, T$). Let Y_{it} denote the outcome of interest, namely opioid-related mortality rates, as measured longitudinally for state *i* over time $t = 1, \dots, T$.

The DID specification is commonly implemented as the two-way fixed effects model which controls for both state- and time-fixed effects as model covariates, expressed as:

$$g(Y_{it}) = \alpha \cdot A_{it} + \boldsymbol{\beta} \cdot \boldsymbol{X}_{it} + \rho_{\boldsymbol{i}} + \sigma_t + \varepsilon_{it} \tag{1}$$

where $g(.)$ denotes the general linear model (GLM) link function (e.g., linear, log), $\boldsymbol{X}_{it}$ denotes a vector of time-varying state-level confounders, 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.

Equation 1 is generally estimated using ordinary least squares. There are 3 commonly used ways to estimate the standard error of the point estimate $(\hat{a})$: (i) no adjustment; (ii) Huber adjustment = robust estimators (also known as sandwich estimators, or Huber corrected estimates) that attempt to adjust the standard error for violations of distributional assumptions (White 1980, Zeileis 2004); or (iii) cluster adjustment = adjustments to account for possible violations of the assumed independence of observations within states (White 1980, Zeileis 2004, 2006).

When implemented using regression analysis, this basic DID estimation is subject to the standard statistical assumptions for linear regression (Wooldridge 2009). Additionally, the DID estimator relies on two additional assumptions: the “common shocks” assumption and the “parallel counterfactual trends” assumption (Angrist and Pischke 2009). The parallel counterfactual trends assumption states that the intervention and comparison groups would have the same trends in the outcome over time, had the intervention not been implemented. Note that outcome levels themselves are not assumed to be equivalent across groups, but rather changes in the outcome; state and time differences are accounted for by fixed effects. When multiple pre-period observations are available, the parallel counterfactual trends assumption can be partially assessed by statistically testing whether the pre-intervention trends differ across intervention groups (Ryan, Burgess, and Dimick 2015). However, the parallel counterfactual trends assumption is not fully testable, since this assumption cannot be assessed for the post-period and in the post-period the assumption involves unobserved counterfactual outcomes.

Additionally, the common shocks assumption states that both groups were subject to similar exogenous factors that may affect the outcome of interest (e.g., market trends, regulatory climate, etc.) and had similar reactions to such factors (Ryan, Burgess, and Dimick 2015). Similarly, this assumption also implies that there is no "anticipatory" effects on the outcome in the intervention group during the pre-period, as individuals are reacting to knowledge of the impending intervention.

On top of fitting the two-way fixed effects model in equation (1), we also considered three additional candidate statistical models in our simulation, based on a literature review identifying the most commonly applied statistical models used in studies trying to assess the causal impacts of opioid policies (Schuler et al., Under Review). Each of these alternative models vary the methods used to control for average differences in the outcome across states by either (a) detrending the data with state-specific linear slopes over time; (b) fitting a one-period lagged autoregressive (AR) model, and (c) estimating a flexible generalized estimating equations (GEE). The statistical assumptions underlying each of these approaches are important to consider.

While equation (1) adjusts for state variability with respect to average outcome levels, another frequently utilized model additionally adjusts for state variability in average outcome trends with state-specific linear slopes over time (referred to as "detrending" the data). Thus, we considered the potential benefits of detrending in our simulation study. Our detrended model can be expressed as:

$$g(Y_{it}) = \alpha \cdot A_{it} + \boldsymbol{\beta} \cdot \boldsymbol{X}_{it} + \rho_{\boldsymbol{i}} + \sigma_t + \eta \cdot \mathrm{t} + \sum_{s=1}^{50}(\omega_s \cdot t \cdot 1(state_i = state_s)) + \upsilon_{it} \qquad (2)$$

This detrended model expands on model (1) by adding in a linear main effect for time (*t*) along with state-specific linear trends. The detrended model can be seen as an improvement over equation (1) when it is likely that states will have different time trends in the outcome. The model aims to capture these time trends using a linear interaction between continuous time (t) and state-level indicators in an effort to avoid over-fitting the data. However, if the assumption of linearity in the time trends does not hold, this model will be misspecified and it is of interest to understand the impact such misspecification might have on our real world state-level data of opioid-related mortality.

Second, we explored whether including a lagged outcome measure (i.e., the outcome at the prior year $Y_{i(t-1)}$) in the model improves policy effect estimation when studying opioid-related mortality by fitting AR models to our simulated data. The motivation for including lagged outcome measures is to both help control for average differences across states implementing versus not implementing a policy and to better predict future outcomes when there is high autocorrelation. Such high correlation is expected in repeated, annual measures of state-level opioid-related outcomes like opioid-related mortality. Thus, it is of interest in this study to assess how controlling for this lag, assuming a linear relationship between the GLM link and the lagged value of the outcome might serve to improve the model in equation (1). AR models include lagged measures of the outcome (e.g., $Y_{i(t-1)}$) as covariates to control for potential average differences in outcome trends and have been shown to yield more accurate effect estimates in the recent gun policy simulation study when examining total firearms deaths (Schell et al., 2018). In particular, we tested the relative performance of the best performing AR model identified from the gun policy study in which we estimated the following specification of our outcome model:

$$g(Y_{it}) = \alpha \cdot (A_{it} - A_{i,t-1}) + \boldsymbol{\beta} \cdot \boldsymbol{X}_{it} + \gamma \cdot Y_{it-1} + \sigma_t + \epsilon_{it} \qquad (3)$$

Notable, this model includes time fixed effects, σ_t, to quantify temporal trends across time but does not include state fixed effects to quantify potential average differences in the outcome across states. Instead, the model attempts to control for confounding by state using the lagged term ($\gamma \cdot Y_{it-1}$).

Including this autoregressive predictor creates a "change" model in which the policy's effect is indicated by the extent to which the opioid-related death rate in a given year is higher or lower than expected given the prior year's rate in the same state (closely related to first-differences models depending on value of autoregressive term – here $\hat{\gamma}$). As such, we coded the policy variable (*A*) using *change coding* $(A_{it} - A_{i,t-1})$ rather than standard *effect coding* (A_{it}) since early work with autoregressive models (e.g., Cochrane and Orcutt (2012)) demonstrated that effect size estimates can be substantially biased AR models that use standard effects coding. The coding of the policy variable needs to take into account the biasing effect of the AR relationship which occurs because controlling for the prior value of the outcome often, indirectly, controls for exactly the effect you are trying to measure. The coefficient estimate $\hat{\alpha}$ on the change coding term represents the AR estimator of the policy effect of A. While commonly used in the broader comparative case study literature, this approach is uncommon in in opioid policy evaluations.

Finally, we considered the use of generalized estimating equations (GEE) which represent a commonly used alternative to the DID models specified in (1) and (2) in the context of longitudinal analyse (Fitzmaurice, Laird, and Ware 2011, Hardin and Hilbe 2003, Liang and Zeger 1986). GEE methods estimate the parameters of a model in which there are correlations between the observations within states by specifying a covariance structure for the clustered outcomes. Here we estimated the following regression model

$$g(Y_{it}) = \alpha \cdot A_{it} + \boldsymbol{\beta} \cdot \boldsymbol{X}_{it} + \sigma_t + \zeta_{it} \tag{4}$$

where we had time fixed effects σ_t , effects coding for the policy variable and time-varying state-level confounders measured in $\boldsymbol{X}_{it}$. 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 the correlation structure **R** for the repeated measures within each state is

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

for each the *t, m* element of **R.** GEE estimates of regression parameters are obtained using an iterative algorithm.

We emphasize that the statistical properties and underlying assumptions in these model differ in how we were specifiying the right-hand side of the outcome model (for all three models) and in how we are handling the correlation between annual measures of the outcome within a state (for the AR and GEE models). The optimal model should be the one for which the underlying assumptions of the model match the state-level data we had available to us for the analysis. It is often difficult to test model assumptions in real world applications and as such we used a simulation to understand the relative performance of these different models on our data.

Given that only a few published opioid policy studies have utilized random effects to control for unobservable state variation (Schuler Under review), we did not consider random effect models in this study.

2.3 Statistical Models Tested via Simulation

In total, we identified 17 candidate models, detailed below and summarized in **Table 1**, drawing from our review of the opioid policy literature (Schuler Under review) as well as the best models identified from the recent gun policy simulation study by Schell, Griffin, and Morral (2019). (Note that those results have not been widely implemented in opioid policy research, and, importantly, the conclusions may also differ for the different setting and type of outcome). The different candidate model specifications were chosen in such a way to allow us to identify the preferred specification within each of the following domains:

Table 1. Overview of candidate statistical models evaluated in simulation study

	GLM	*Regression Specification*	*Weighting*
1	Linear	Fixed effects (FE)	Population weighted
2			Unweighted
3		FE + Detrended	Population weighted
4			Unweighted
5		Autoregressive	Population weighted
6			Unweighted
7		GEE model	Population weighted
8			Unweighted
9	Log-linear	Fixed effects (FE)	Population weighted
10			Unweighted
11		Autoregressive	Population weighted
12			Unweighted
13	Negative Binomial	Fixed effects (FE)	Unweighted; log(population) used as an offset
14		FE + Detrended	Unweighted; log(population) used as an offset
15		Autoregressive	Unweighted; log(population) used as an offset
16	Poisson	Fixed effects (FE)	Unweighted; log(population) used as an offset
17		Autoregressive	Unweighted; log(population) used as an offset

(1) *GLM specifications*: In statistics, GLM is a flexible generalization of ordinary linear regression that allows for the outcome being modeled to have an error distribution other than the normal distribution. 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 assuming normality. GLMs are advantageous in that alternative transformations and/or assumptions about opioid-related mortality distribution can be tested

within the same class of models by varying both the link function ($g(.)$) and the assumed distribution function of the outcome. We tested the relative performance of the following GLMs: linear, log-linear (i.e., a linear model with log-transformed outcome), and two log-link models, (negative binomial and Poisson).

(2) *Regression specification:* As noted, we considered four different ways to specify our regression models: only using two-way fixed effects of time and state in the model, using two-way fixed effects plus state-specific linear trends (detrended models), AR models, and GEE models.

(3) *Standard error estimation:* For each model (except the GEE model), we explored the impact of various methods for estimating standard errors (SE), including robust SE estimators that adjust for violations of homoskedasticity assumptions in the data or cluster adjustments adjusting for non-independence in the observations within states. More specifically, for each model run, we estimated the standard error in three ways: no adjustment; Huber adjustment; and cluster adjustment.

(4) *Use of state-level 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 opioid policy evaluations [e.g., Ali et al. (2017), Buchmueller and Carey (2018), McInerney (2017), Paulozzi, Kilbourne, and Desai (2011)]. Given that log-link models (e.g., negative binomial, Poisson) are conducted directly on the opioid-related death counts (rather than the rates) and do not need to be weighted to be nationally-representative, we did not examine the impact of weighting in these models. Using population weights in state-level analyses of opioid-mortality rates results in models for which each opioid-related death is treated as equally important regardless of which state it occurred in. Unweighted analyses treats each state, rather than each person, as equally important (see for example; Bachhuber et al. (2014), Birk and Waddell (2017),

Chang et al. (2016), Dowell et al. (2016), Xu et al. (2018), Yarbrough (2018)). This will result in an estimation in which a death in small states will have much greater weight than deaths in larger states. We note that data was generated such that policy effects are constant across all states regardless of size or other characteristics, so weighting is not expected to bias the asymptotic values of the effect estimates but may have substantial effects on their SEs.

3. SIMULATION DETAILS

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

3.1 Metrics for Assessing Relative Performance of Candidate Statistical Methods

To guide selection of the preferred statistical methods, we relied on several statistical metrics commonly used to judge model performance.

(1) *Type I error rate*. This is the rate by which a null hypothesis that is true (i.e., there truly is no policy effect) is rejected based on the model estimated coefficients and standard errors. 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 α =.05 level of significance.

(2) *Power.* Power refers to the ability of the model to correctly identify that the null hypothesis is false. Typically, studies are considered to have good power for a given effect size when they have 80% or higher power. When comparing statistical power across candidate models, we needed to ensure that we were properly penalizing models with high Type I error rates (see Section 4.1). For all models, we computed a correction factor based on the null runs that can be applied to the estimated standard errors from the given model to ensure a Type I error rate of 0.05; we then used this correction factor when

calculating the appropriate level of power for a given model in the non-null runs. For virtually all the models, this correction factor inflates the estimated standard errors in a given model and helps us ensure power is being accurately captured for the models.

(3) ***Bias.*** Bias assesses the average difference between the estimated effect and true effect over all simulations showing the tendency of the estimated effects of a given model to fall closer or further from the true effect on average.

(4) *Root Mean Squared Error (RMSE).* RMSE provides us with a broader measure of the difference between the estimated policy effects in an individual simulated data set and the true policy effect by taking the square root of the sum of the mean squared errors (e.g., $\sqrt{\sum_{k=1}^{5000}(\hat{\alpha}_k-\alpha)^2}$) where α represents the true policy effect and $\hat{\alpha}$ represents the estimated policy effect from a given simulation and model. It gives us a sense of how much error exists occurs for a given model specification, and takes into account both bias and variance.

3.2 Data Sources and Measures

Our outcome of interest is the annual, state-specific opioid mortality rate per 100,000 state residents, using the 1999-2016 National Vital Statistics System (NVSS) Multiple Cause of Death mortality files. Consistent with other studies (Abouk, Pacula, and Powell 2019, Chan, Burkhardt, and Flyr In press, Kilby 2015), opioid related overdose deaths were identified 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, we included a single covariate: state-level unemployment rate (U.S. Department of Labor 2019). This covariate was selected both because of the frequency of its use in opioid policy studies (Schuler Under review), and its potential confounding associations with both opioid-related mortality and state policy responses. Annual state-level unemployment data comes from the U.S. Department of Labor, Bureau of Labor Statistics. Sensitivity analyses including a broader set of covariates (e.g., poverty rate, income level and percent race/ethnicity and age groups) resulted in no meaningful change to the general findings with a slight increase in model precision. Thus, we present findings from the more parsimonious model.

We also 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.

3.3 Simulation Data Generation

The simulation design builds directly from prior work that compared statistical methods for evaluating the impact of state laws on firearms deaths (Schell, Griffin, and Morral 2019). For each simulation iteration, we selected a random subset of *k* states to be the policy/treated group (i.e., A_{it} = 1 at some point in the study period), with remaining states serving as the comparison group (i.e., A_{it} = 0 for the entire study period). We generated a time-varying indicator for whether a state has implemented the hypothetical opioid policy (A_{it}) in a given year.

When generating the implementation date for the treated states who enact the policy, we randomly selected both the month and year of policy enactment so we had fractional values for the amount of time a law was in effect during the first year, restricting the year to be between 2002 and 2013 to ensure we had three years of outcome data before and after enactment. This simulation represents the simplified scenario in which there is no confounding by observed or unobserved covariates or by lagged values of the outcome, $Y_{i(t-1)}$. Once a policy is implemented, it remains in effect throughout the study period. In the first year of implementation the intervention variablve is coded as a variable between 0 and 1, indicating the percentage of the year the policy was in effect.

For control states and treated (policy-enacting) states in the pre-policy period, generated Y_{it} values are equal to the actual observed state-specific, year-specific opioid overdose rates. For treated (policy-enacted) states, we generated synthetic overdose outcomes (Y_{it}) for the time periods following the randomly assigned policy implementation date, allowing us to systematically vary the magnitude of the true effect of A on Y.

Simulation conditions varied the following factors:

(1) *Effect size*. We considered the performance of our candidate statistical models when the policy had null effect, as well as an effect size (ES) of ±5%, ±15% and ±25%. For null effect conditions, post-policy Y_{it} values for the policy states were generated such that they were not correlated with A_{it} (i.e., setting α = 0 in equation 1). For conditions with a true policy effect, synthetic post-policy Y_{it} values were generated to reflect a fixed increase/decrease relative to the true state-specific, year-specific opioid overdose rate. For a given candidate model, this data generation process was tailored according to the specific link function used in the model, such that the absolute magnitude of the generated ES would be equivalent on both the linear (i.e., additive) and log linear (i.e., multiplicative) scale. For each ES, we generated data in which the policy has both a positive and negative effect on the outcome in order to assess whether candidate models were biased towards estimating positive or negative effects, as well as assessing bias.

(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, respectively. We note that the maximum total sample size 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 consider two possible conditions: an instantaneous effect and a 3-year linear phase-in effect. In both the data generating and analytic models, an instantaneous effect was specified 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. The gradual policy effect was specified as a linear spline with values starting at zero and reaching an asymptote 3 years after implementation.

We assess performance of each candidate model across the same set of 5000 randomly generated datasets. Simulations were conducted in R; code is available in the appendix. Extensive results for all statistical models considered in our simulation are available via our Shiny tool (https://elizabethmcneer.shinyapps.io/statmodelsim/).

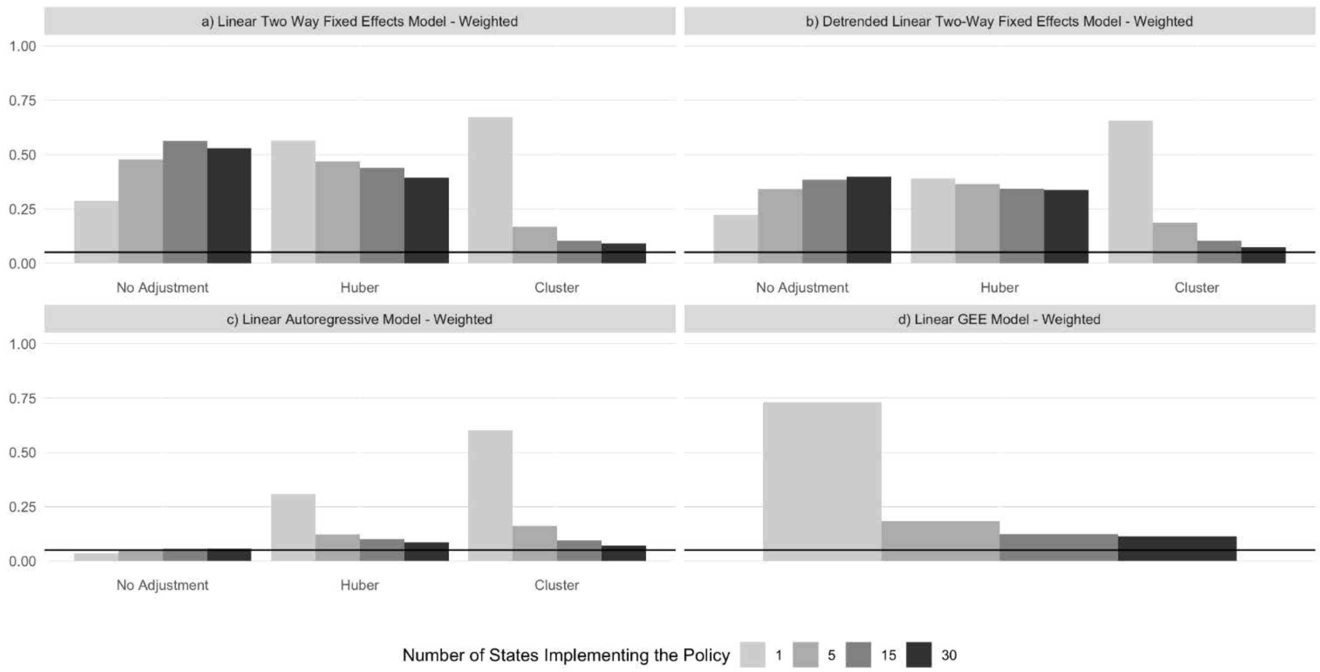
4. RESULTS

In the results below, we first present findings regarding Type 1 error rates, highlighting cases where models have unreasonably high Type I errors. Next, we discuss our findings regarding power over a range of different effect sizes, highlighting analytic ways in which models can be optimized to increase power in a study. Third, we discuss bias for each model, assessing which analytic approaches introduce bias versus those that do not. Finally, we discuss root mean squared error (RMSE) in an effort to identify the best models that minimize the bias-variance trade off. In each section, we first discuss findings for the linear model, comparing the overall performance of the two-way fixed effects, detrended, AR, and GEE models and briefly highlighting the impact of using population weights in the linear model. Then we discuss the relative performance of the different GLMs for a given regression specification (e.g., two-way fixed effects), comparing the relative performance of the linear, log-linear, Poisson, and negative binomial models. In all cases, we report our summary statistics averaging across the simulations where the policy has an instantaneous and 3-year phase-in, noting here that performance for all metrics gets worse the longer it takes for a policy to become effective.

4.1 Type I Error Rates

Figure 1 shows the Type I error rates for the four different linear models considered in our simulation, all of which use population weights: (1a) the two-way fixed effects model, (1b), the detrended model, (1c) the AR model, and (1d) the GEE model. Of note, as shown in **Figure 1a**, we saw very high Type I error rates for most of the models (ranging from 0.29 to 0.67). However, use of cluster adjustment to the standard error greatly reduced the Type I error rates when 5 or more states are implementing a policy, but they generally were still 2 to 3 times larger than the traditional target of 0.05 (e.g., range between 0.09 and 0.17). The detrended model, shown in Figure 1b, was similar though we noted that Type I error rates for the detrended model tends to be better (e.g., lower; mostly less than 0.4). In contrast, **Figure 1c** highlights the general finding that AR models do not require use of any standard error adjustment to obtain Type I error rates that are reasonable when the number of states implementing a policy is greater than or equal to 5 (e.g., Type I error rates for the linear, population weighted AR model range from 0.04 to 0.06). In fact, use of standard error adjustments in the AR models tends to inflate the Type I error rates. Finally, **Figure 1d** shows the Type I error rates for the linear GEE model with population weights. As shown, Type I error rates reach 0.18 or less once at least 5 states are implementing a policy, though the rates are also still two to three times higher than the traditional target of 0.05. The findings concerning the relative performance of the different regression specifications shown (two-way fixed effects vs detrended vs AR) holds across all GLM specifications (linear, log-linear, Poisson, and negative binomial).

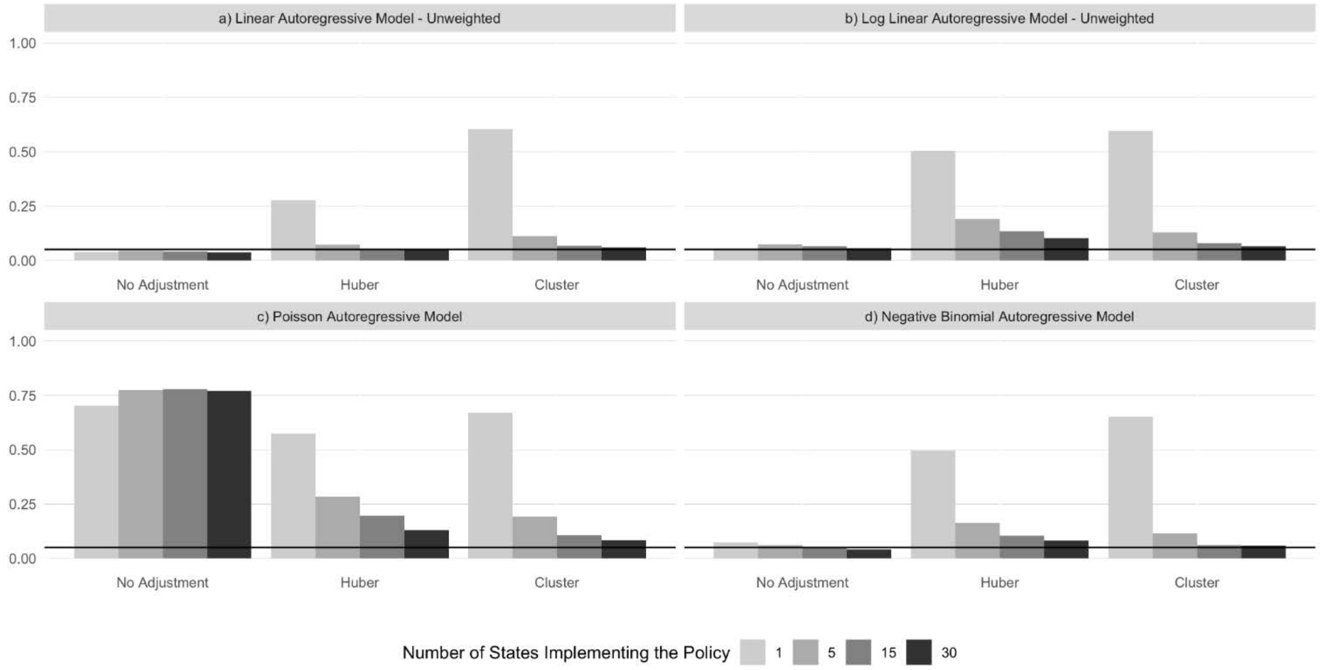
Figure 1. Type I error rates for the four different linear models considered, all with population weights: (1a) the two-way fixed effects model, (1b), the detrended model, (1c) the AR model, and (1d) the GEE model. Black horizontal line denotes the target Type I error rate value of 0.05.



Additional findings from our simulation studies show that weighting using the population size in the linear models results in slightly higher Type I error rates in the two-way fixed effects, detrended, and GEE models than for the corresponding unweighted versions of these models (see Shiny Application). Conversely, for the AR models, use of population weights do not consistently perform better or worse than unweighted models.

Next, we explored the relative performance of the different GLMs. We note here that the best four models in terms of Type I error when using the best performing standard error adjustment and comparing the maximum Type I error rates across the four different sample sizes include: the linear AR weighted model (0.06), the linear AR unweighted model (0.07), the log-linear AR unweighted model (0.07), and the negative binomial AR model (0.07). **Figure 2** shows the Type I error rates for the AR models for our four different GLMs: (2a) linear (unweighted), (2b) log linear (unweighted), (2c) Poisson, and (2d) negative binomial. This figure highlights how well the AR models do in terms of Type I error rates, regardless of the GLM model. It also highlights that generally the log-linear model performed slightly worse than the linear model and that the Poisson model can produce rather poor performance relative to the other GLMs.

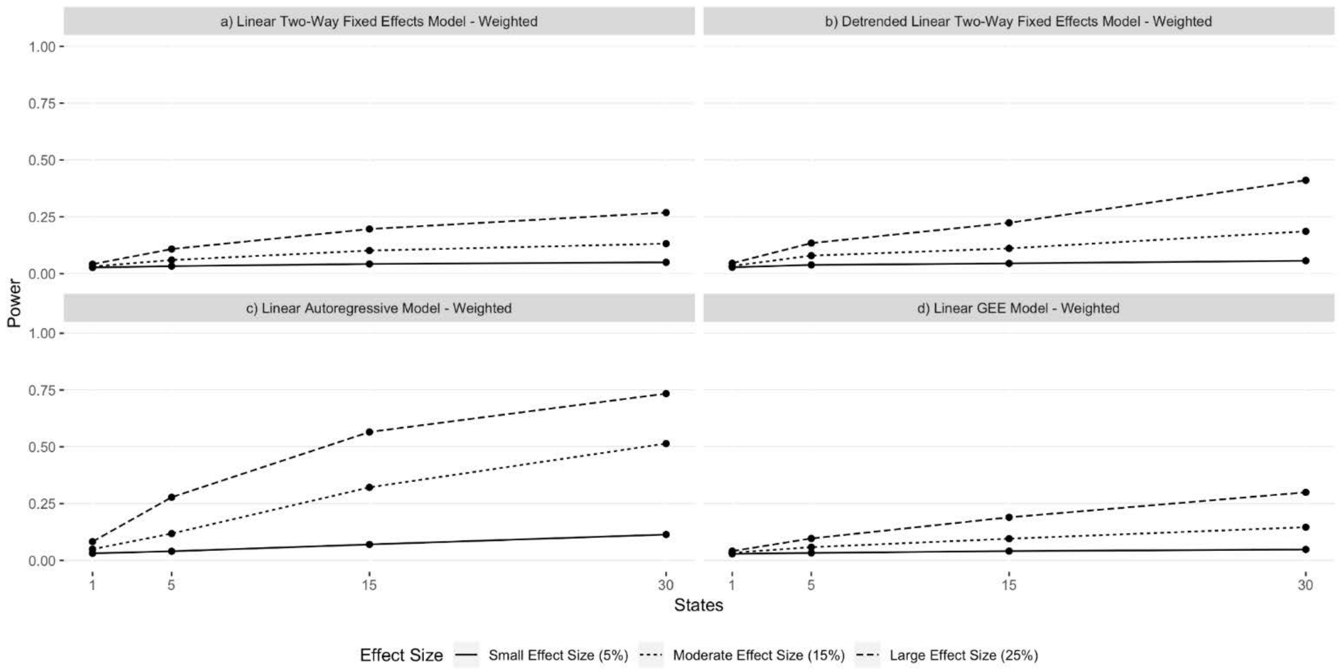
Figure 2. Type I error rates for the AR models for four different GLMs: (2a) linear (unweighted), (2b) log linear (unweighted), (2c) Poisson, and (2d) negative binomial. Black horizontal line denotes the target Type I error rate value of 0.05.



4.2 Power

Figure 3 shows power (i.e., correct rejection rates) as a function of the number of states implementing a policy and the effect size impact of the policy for the four different linear models considered in our simulation, all of which use population weights: (2a) the two-way fixed effects model, (2b) the detrended model, (2c) the AR model, and (2d) the GEE model. Power is shown for the model that uses the best method for SE adjustment (namely, the SE adjustment that produces a Type I error rate closest to 0.05).

Figure 3. Power as a function of the number of states implementing a policy and the effect size impact of the policy for four different linear models, all with population weights: (2a) the two-way fixed effects model, (2b), the detrended model, (2c) the AR model, and (2d) the GEE model. Solid line = 5% effect size; dotted line = 15% effect size; dashed line = 25% effect size.

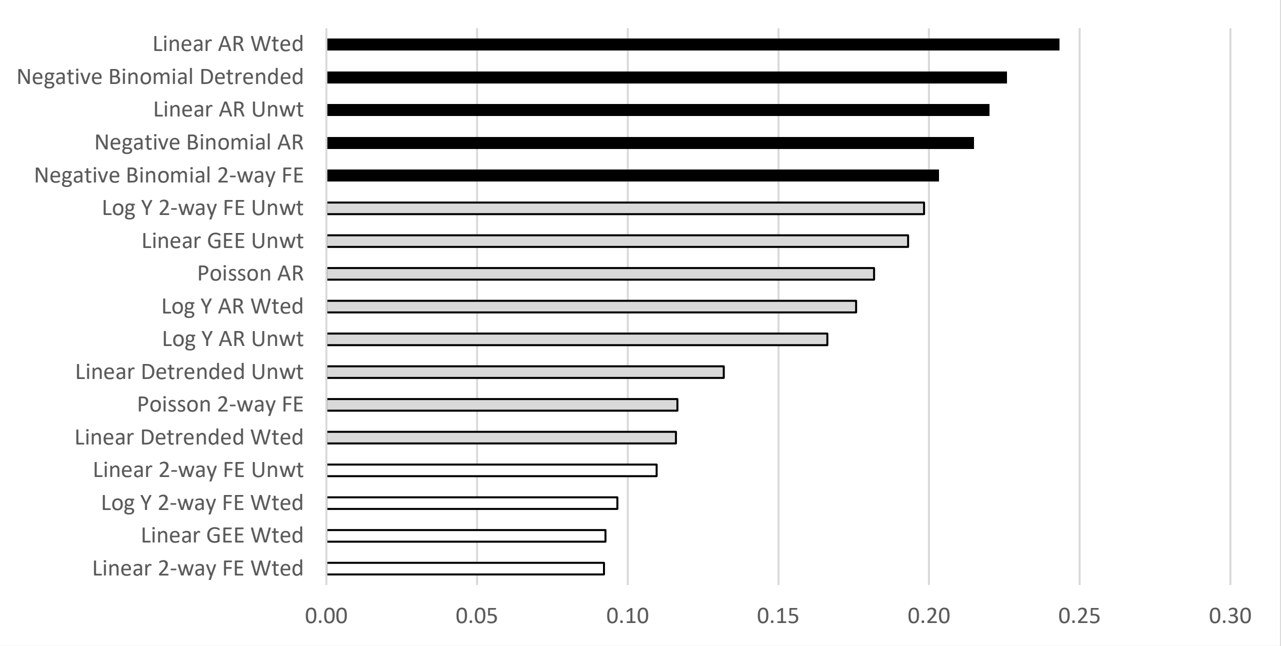


In all cases, the power increases as either the sample size of states implementing a policy increases or as the effect size of the policy increases for the range of sample sizes considered here. Most notably, the commonly used two-way fixed effects model (**Figure 3a**) has poor power for all effect sizes, reaching a maximum power of 0.27 when the number of states implementing a policy reached 30 and the effect size of the policy is 25%. **Figure 3b** shows that power tends to increase for the detrended model when compared to the linear two-way fixed effects model that does not control for state specific time trends. In contrast, to the linear two-way fixed effects model, the detrended model reaches a maximum power of 0.41 when the number of states implementing a policy reached 30 and the effect size of the policy is 25%. Still, power is highest for the AR model (**Figure 3c**) in comparison to either the linear two-fixed effects model or the linear detrended model, with maximum power of 0.73 when the number of states implementing a policy reached 30 and the effect size of the policy is 25%. In the case of using population weights, the linear GEE has similar power to the linear two-way fixed effects with the weighted GEE model reaching a maximum power of 0.30 in **Figure 3d**.

We found that weighted linear and log-linear models tend to have lower power than the unweighted versions of these models. The maximum power for the weighted versus unweighted linear two-way fixed effects model is 0.27 versus 0.40, respectively. For the linear AR model, it is 0.81 versus 0.72, respectively, when 30 states are implementing a policy and the effect size is large (25%). Use of population weights in the GEE model has the greatest impact with maximum power going from 0.30 to 0.67 when one removes the population weights from the model.

Figure 4 shows the average power across all scenarios for each type of model to highlight the relative performance of the different models considered. Power is poor across all methods, but, of note, we saw clear superiority of both the linear AR models (weighted or unweighted; average power = 0.24 and 0.22, respectively) and the negative binomial models (power ranges from 0.20 to 0.23). The worst performing models are the linear and log-linear two-way fixed effects models and the lienar weighted GEE model (power ranges from 0.09 to 0.11). The Poisson models considered have power that ranges from 0.12 for the two-way fixed effects model to 0.18 for the AR model; for each type of model, the negative binomial is clearly more powerful.

Figure 4. Average power for all models considered in this simulation

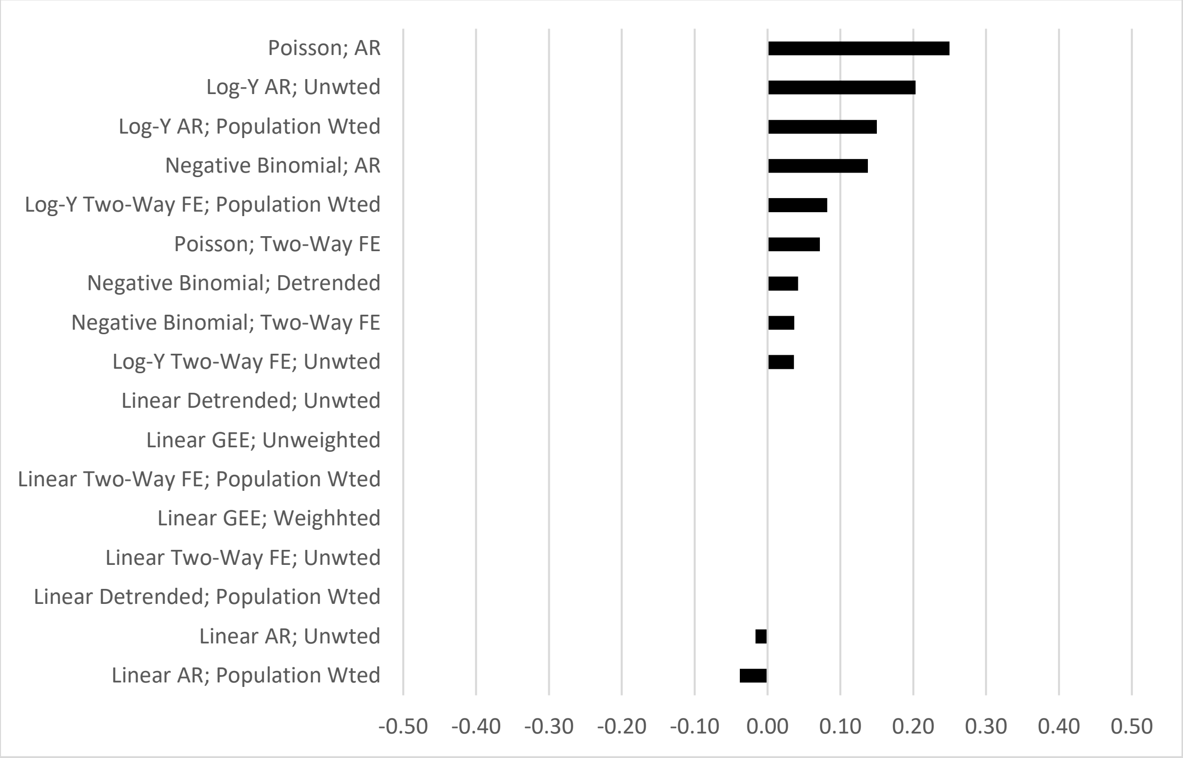


Most notably for all models considered in our simulation, the models had extremely low power to detect small effect sizes of 5% for a policy, an effect size that would equate to a reduction in 700 opioid-related mortalities per year (e.g., maximum power for the negative binomial models was approximately 0.08 and power for the linear models ranged from 0.04 to 0.11).

4.3. Bias

Figure 5 shows the percent bias for all models considered in this simulation when the true effect of a law is 5% (or ~700 deaths per year). We first standardized the results so quantities from the linear and nonlinear models represent the count of deaths by which the model is biased on average over all the simulations. Then, we converted bias into percent bias by dividing by 700. In general, bias is low for most models (e.g., less than 10%) with the exception of the non-linear AR models (14-25% percent); however, care needs to be taken when comparing the linear models to the log-linear, Poisson, and negative binomial models. In addition, we found that converting all bias metrics into linear units (count of deaths) provides a small advantage to the linear models when making comparisons to nonlinear models because this metric is directly proportional to the model coefficients in a linear model. Thus, direct comparisons across link functions here (e.g, linear versus log-linear) is not recommended. For example, it is possible that a negative binomial model that has unbiased model coefficients in their native units will show a small bias on the exponentiated coefficients used to convert the bias to a total count of deaths. There is no method to compare bias across linear and nonlinear models that allows both to be in their native units. When bias measures are converted into the native units of the negative binomial models (log risk ratios), the negative binomial models tended to show slightly better performance relative to the linear models than is evident in **Figure 5**. Taken together, we found that the greatest bias for each type of type of GLMs occurs in the AR models as would be expected. Further, as the number of states implementing a law or the size of the effect of the law increases, bias decreases to less than 5% for all models considered (see Shiny app).

Figure 5. Percent bias for all models considered in this simulation when true effect is 5% (or 700 deaths)



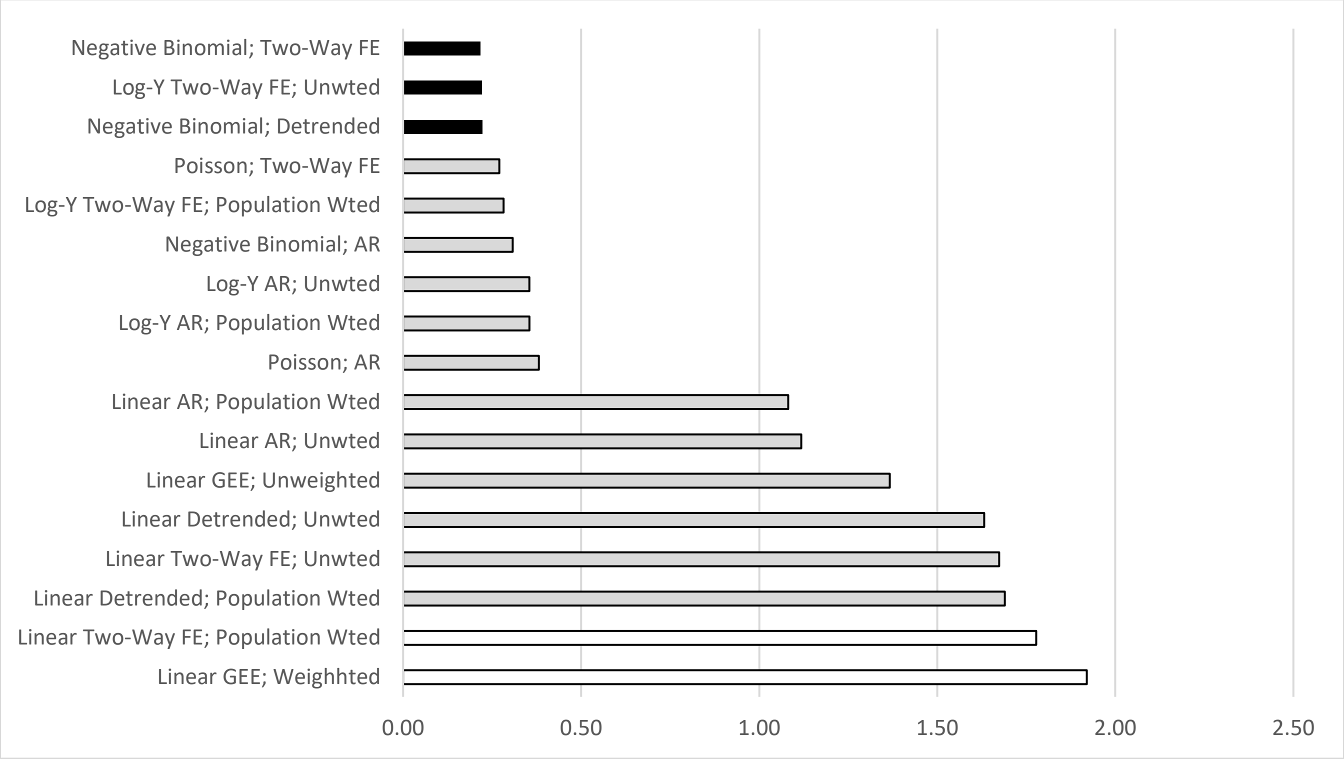
Note: It is not really recommended that we directly compare across link functions here since there is no method to compare bias across linear and nonlinear models that for a one-to-one comparison. The statistics shown will slightly favor linear over non-linear models since we have to convert

4.4 Root Mean Square Error

Figure 6 shows the average RMSE for the simulations where the true policy effect is 0 for each model considered in our simulation. First, we considered findings for the linear model, comparing the overall performance of the two-way fixed effects, detrended, AR, and GEE models. In general, the RMSE of the two-way fixed model can be improved upon by using detrending or AR models (e.g., for the linear two-way fixed effects with population weight model the RMSE = 1.78 versus 1.69 for the detrended model and 1.08 for the AR model). The linear AR models have the lowest RMSE (1.08-1.12). For the linear two-way fixed effects, detrended and GEE models, the unweighted models have lower

RMSE as compared to the weighted versions of the same models, while for the AR models, we saw slightly lower RMSE for the weighted models.

Figure 6. Root Mean squared error for all models considered in this simulation when the policy has no true effect



Note: It is not really recommended that we directly compare across link functions here since there is no method to compare RMSE across linear and nonlinear models that for a one-to-one comparison

Next, we considered the RMSE for the non-linear models. Notably, we found that RMSE for the count models is minimized by using the negative binomial over the Poisson or log-linear models. For the negative binomial model, the detrended and two-way fixed effects models have the lowest RMSE (0.22) while the AR model has the highest RMSE (0.31).

5. DISCUSSION

Our findings highlight several key challenges for opioid-policy research and more broadly state-level policy evaluations. First, policy makers using the results of these models to make decisions must recognize the challenges of such studies given the limited total sample size of 50 states. Power for the majority of scenarios was lower than the typically desired 0.80, even in the most advantageous cases when we had 30 out of 50 states implementing the policy of interest and a large assumed effect size of 25%. Furthermore, Type I error rates for the majority of models when fewer than 15 states are implementing a new policy were unreasonably high, meaning these models could show a notable effect of a policy when in fact such an effect does not exist. It is critical that researchers use models that minimize Type 1 error rates whenever possible and use of standard error corrections to ensure a Type I error rate of 0.05 are needed in this space.

Additionally, we found notably differences in the relative performance of commonly used statistical methods for estimating the effects of state-level opioid policies which has substantial implications for interpretation and application to state-level policy evaluations more generally. In our simulations, we found two classes of models that performed best when estimating the effects of a simulated opioid policy on opioid-related mortality in terms of (i) producing Type I error rates near 0.05, (ii) maximizing power for the scenarios considered, and (iii) minimizing RMSE. These were 1) linear AR models (weighted or unweighted) and 2) negative binomial models. Linear two-way fixed effects models and Poisson models generally performed the worst across the scenarios considered. To maximize power for opioid-related mortality, we highly recommend that researchers utilize either linear autoregressive models or negative binomial model specifications when estimating the effects of state-level policies on opioid-related mortality.

It is unclear how many of these specific recommendations are notable for any specific state policy evaluation, as the findings will necessarily be dependent on the level and distribution of the outcome variable as well as the rate of state-level adoption of the policy. At a minimum, however, we think four general recommendations for practice come from our findings. First, it is critical to utilize cluster adjustments to the standard errors when using state and year fixed effects in a linear or log-linear specification of a model evaluating state policy adoption. Second, use of an autoregressive term is particularly helpful in the linear model in terms of RMSE, when modeling opioid-related mortality as a crude rate. Third, detrending models with state and year fixed effects improves performance so long as not overfitting the data. Fourth, use of a negative binomial model performs better than a Poisson model when modeling counts of opioid-related mortality.

Although these generalizations have been found by others, they have not been well appreciated by 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 (Abhay, Donohue III, and Zhang 2014, Helland and Tabarrok 2004) show that cluster adjustments are needed to reduce Type I error rates. Bertrand, Duflo, and Mullainathan (2004) showed that the classic sandwich estimator does poorly with small samples; this paper also shows DID without adjustment has high Type I errors (approximately 45%) in their case study data where they also randomly simulated random "placebo" laws. Our work builds on this prior work by highlighting the challenges to evaluating state-level policies within the context of the opioid crisis and with respect to opioid-related mortality, a commonly used outcome in evaluations of state-level opioid policies.

Of note in this simulation, we never formally tested the functional form of the model to help us determine the best performing GLM specification. We did this because we aimed to compare commonly used specifications that have been published in the literature. In practice, it is highly recommended that more careful consideration be given to selection of the GLM specification. For example, one way to limit concerns about using an inappropriate model is to do a parks test and boxcox test to determine what is appropriate for the outcome transformation and variance adjustment (Box and Cox 1964). We also note that use of population weights generally had small effects on performance so the choice whether to use them should be based on the inferences desired by the research, although this may be in part be a reflection of the situation considered, with constant effects across states. Finally, we note that we expected bias to be small in our simulations given the policies are randomly assigned; the bias is largely going to be driven by model misspecification.

Findings from our simulations on opioid-related mortality are highly similar to findings for total firearm deaths considered in Schell, Griffin, and Morral (2019). The models considered here are virtually identical to the models considered in the Schell, Griffin, and Morral (2019) study, though we expand the simulation scenarios to consider performance across a wider range of assumed effect sizes of the law (5% to 25% versus 3% in Schell, Griffin, and Morral (2019)). In their case, one model proved optimal across all performance metrics considered: the negative binomial AR model, whereas we found in favor of two set of models of which the negative binomial AR model is one. We suspect this is in part due to the additional consideration of a range of effect sizes and that we only assumed a three-year phase in (versus five in Schell, Griffin, and Morral (2019)) for the slower phase in period of the new policy.

The goal of this study was to assess the relative performance of commonly used statistical methods for evaluating the impacts of state-level opioid policies on opioid-related mortality and provide insights into the limitations and sources of bias introduced by more commonly used methods. Study findings can help statisticians, researchers, and policy makers better gauge the validity of the existing evidence base and to conduct sound evaluations of new policies enacted by states. The use of simulation studies to assess the statistical properties of commonly used methods is an innovative yet underutilized approach for helping researchers examine methods' performance in a specific context (Black et al. 2019).

The simulation design has several limitations and future research is needed to build upon this work and provide best methods for the field. First, by randomly selecting states that will enact a given policy, this simulation represents the simplified scenario in which there is no confounding by observed or unobserved covariates or by lagged values of the outcome. Future work will expand the simulation to consider more complex scenarios where such confounding exists given the reality that states implementing certain policies likely differ from their comparison states in systematic ways. For example, it is often the case that the outcome of concern will be notably increasing among states that choose to enact the policy being evaluated (e.g., states enacting pain management clinic laws have notably higher fatal opioid overdose rates in the years before implementing the law than states without these laws (Popovici et al. 2018)). Second, while there are numerous outcomes of interest when evaluating the impact of an opioid policy, we focused on fatal overdoses given that approximately 1/3 of published evaluation studies of state opioid policies use this outcome. Future work will expand to consider additional outcome domains like prescribing and distribution.

More broadly, as noted by Schell, Griffin, and Morral (2019): "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 (Gelman and Carlin 2014)." There is an urgent need for the field to develop more robust and powerful methods that can be used to help guide state-policy. This call is needed to face the current crises in the U.S. (gun violence and the opioid epidemic) but also extends beyond to future crises that will develop (e.g., climate change). One area that holds promise would be to promote the use of Bayesian approaches to estimate state-level policy effects as a way to ensure better representation of the large amount of uncertainty in these analyses (Schell, Griffin, and Morral 2019). Research in this areas is needed to help us ensure we are meeting the needs of applied policy researchers and key decision makers.

Acknowledgements: This research was financially supported through an NIH grant (P50DA046351) to RAND (PI: Stein). 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.

REFERENCES

Abadie, A., and Cattaneo, M. (2018), "Econometric methods for program evaluation." *Annual Review of Economics,* 10, 465-503.

Abhay, A., Donohue III, J., and Zhang, A. (2014), "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*.

Abouk, R., Pacula, R. L., and Powell, D. (2019), "Association Between State Laws Facilitating Pharmacy Distribution of Naloxone and Risk of Fatal Overdose." *JAMA Intern Med,* 179, 805-811.

Ali, M. M., Dowd, W. N., Classen, T., Mutter, R., and Novak, S. P. (2017), "Prescription drug monitoring programs, nonmedical use of prescription drugs, and heroin use: Evidence from the National Survey of Drug Use and Health." *Addict Behav,* 69, 65-77.

Angrist, J., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empericist's Companion*: Princeton University Press.

Bachhuber, M. A., Saloner, B., Cunningham, C. O., and Barry, C. L. (2014), "Medical cannabis laws and opioid analgesic overdose mortality in the United States, 1999-2010." *JAMA Intern Med,* 174, 1668-73.

Basu, S., Meghani, A., and Siddiqi, A. (2017), "Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches." *Annu Rev Public Health,* 38, 351-370.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How much should we trust differences-in-differences estimates?". *The Quarterly Journal of Economics,* 119, 249-275.

Birk, E. G., and Waddell, G. R. *The Mitigating Role of Prescription Drug Monitoring Programs in the Abuse of Prescription Drugs* [Internet Resource; Archival Material]. Bonn: Institute for the Study of Labor (IZA) 2017.

Black, B., Hollingsworth, A., Nunes, L., and Simon, K. (2019), "The Effect of Health Insurance on Mortality: Power Analysis and What We Can Learn from the Affordable Care Act Coverage Expansions. NBER Working Paper No. 25568."

Blundell, R., and Costa Dias, M. (2009), "Alternative approaches to evaluation in empirical microeconomics." *Journal of Human Resources,* 44, 565-640.

Box, G., and Cox, D. (1964), "An analysis of transformations." *Journal of the Royal Statistical Society,* Series B, 211-252.

Brewer, M., Crossley, T., and Joyce, R. (2017), "Inference with difference-in-differences revisited." *Journal of Econmic Methods,* 7, 2156-6674.

Buchmueller, T. C., and Carey, C. (2018), "The Effect of Prescription Drug Monitoring Programs on Opioid Utilization in Medicare." *American Economic Journal-Economic Policy,* 10, 77-112.

Centers for Disease Control and Prevention. *CDC WONDER*, November 19, 2019 2018 [cited November 22, 2019. Available from https://wonder.cdc.gov/.

Chan, N., Burkhardt, J., and Flyr, M. (In press), "The effects of recreational marijuana legalization and dispensing on opioid mortality." *Economic Inquiry*.

Chang, H. Y., Lyapustina, T., Rutkow, L., Daubresse, M., Richey, M., Faul, M., Stuart, E. A., and Alexander, G. C. (2016), "Impact of prescription drug monitoring programs and pill mill laws on high-risk opioid prescribers: A comparative interrupted time series analysis." *Drug Alcohol Depend,* 165, 1-8.

Cochrane, D., and Orcutt, G. (2012), "Application of least squares regression to relationships containing auto-correlated error terms." *Journal of the American Statistical Association,* 44, 32-61.

Daw, J. R., and Hatfield, L. A. (2018a), "Matching and Regression to the Mean in Difference-in-Differences Analysis." *Health Serv Res,* 53, 4138-4156.

Daw, J. R., and Hatfield, L. A. (2018b), "Matching in Difference-in-Differences: between a Rock and a Hard Place." *Health Serv Res,* 53, 4111-4117.

Donald, S. G., and Lang, K. (2007), "Inference with difference-in-differences and other panel data." *Review of Economics and Statistics,* 89, 221-233.

Dowell, D., Zhang, K., Noonan, R. K., and Hockenberry, J. M. (2016), "Mandatory Provider Review And Pain Clinic Laws Reduce The Amounts Of Opioids Prescribed And Overdose Death Rates." *Health Aff (Millwood),* 35, 1876-1883.

Fitzmaurice, G., Laird, N., and Ware, J. (2011), *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons.

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspect Psychol Sci,* 9, 641-51.

Haegerich, T. M., Jones, C. M., Cote, P. O., Robinson, A., and Ross, L. (2019), "Evidence for state, community and systems-level prevention strategies to address the opioid crisis." *Drug Alcohol Depend,* 204, 107563.

Hardin, J., and Hilbe, J. (2003), *Generalized Estimating Equations*. London: Chapman and Hall.

Helland, E., and Tabarrok, A. (2004), "The fugitive: Evidence on public versus private law enforcement from bail jumping." *Journal of Law & Economics,* 47, 93-122.

Kilby, A. (2015), *Opioids for the Masses: Welfare Tradeoffs in the Regulation of Narcotic Pain Medications*. Cambridge: Massachusetts Institute of Technology.

Liang, K.-Y., and Zeger, S. (1986), "Longitudinal data analysis using generalized linear models." *Biometrika,* 73, 13-22.

Mauri, A. I., Townsend, T. N., and Haffajee, R. L. (2019), "The Association of State Opioid Misuse Prevention Policies With Patient- and Provider-Related Outcomes: A Scoping Review." *Milbank Q*.

McInerney, M. 2017. The Affordable Care Act, Public Insurance Expansion and Opioid Overdose Mortality. University of Connecticut, Department of Economics, Working papers: 2017-23.

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., and Sekhon, J. S. (2016), "Estimating causal effects: considering three alternatives to difference-in-differences estimation." *Health Serv Outcomes Res Methodol,* 16, 1-21.

Paulozzi, L. J., Kilbourne, E. M., and Desai, H. A. (2011), "Prescription drug monitoring programs and death rates from drug overdose." *Pain Med,* 12, 747-54.

Popovici, I., Maclean, J. C., Hijazi, B., and Radakrishnan, S. (2018), "The effect of state laws designed to prevent nonmedical prescription opioid use on overdose deaths and treatment." *Health Econ,* 27, 294-305.

Ryan, A. M., Burgess, J. F., Jr., and Dimick, J. B. (2015), "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences." *Health Serv Res,* 50, 1211-35.

Schell, T., Griffin, B., and Morral, A. (2019), *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study. RR-2685-RC*. Santa Monica, CA: RAND Corporation.

Schuler, M. (Under review), "The state of the science in opioid policy research."

U.S. Department of Labor. *Bureau of Labor Statistics* 2019 [cited November 22, 2019. Available from https://www.bls.gov/.

White, H. (1980), "A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity." *Econometrica,* 48, 817.

Wing, C., Simon, K., and Bello-Gomez, R. A. (2018), "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annu Rev Public Health,* 39, 453-469.

Wooldridge, J. (2009), *Introductory Econometrics: A Modern Approach*: South-Western College Publishing.

Xu, J., Davis, C. S., Cruz, M., and Lurie, P. (2018), "State naloxone access laws are associated with an increase in the number of naloxone prescriptions dispensed in retail pharmacies." *Drug Alcohol Depend,* 189, 37-41.

Yarbrough, C. R. (2018), "Prescription Drug Monitoring Programs Produce a Limited Impact on Painkiller Prescribing in Medicare Part D." *Health Serv Res,* 53, 671-689.

Zeileis, A. (2004), "Econometric computing with HC and HAC covariance matrix estimators." *Journal of Statistical Software,* 11, 1-17.

Zeileis, A. (2006), "Object-oriented computation of sandwich estimators." *Journal of Statistical Software,* 16, 1-16.