State-level opioid policy analyses:
Moving beyond the classic difference-in-differences model

Beth Ann Griffin, Megan Schuler, Elizabeth Stuart

OPTIC Webinar Series
Wednesday, April 28, 2021
State-level opioid policy analyses: Moving beyond the classic difference-in-differences model

Beth Ann Griffin, Megan Schuler, Elizabeth Stuart

OPTIC Webinar Series
Wednesday, April 28, 2021
Introductions

- **Beth Ann Griffin** is a senior statistician at the RAND Corporation and Co-Director of the Opioid Policy Tools and Information Center (OPTIC). Her research career has been devoted to identifying optimal methods for the estimation of causal effects when using observational data.

- **Megan Schuler** is an applied statistician at the RAND Corporation and Co-Investigator at OPTIC. Her research focuses substantively on substance use and mental health disorders and methodologically on causal inference methods that facilitate rigorous analysis of observational health data.
Elizabeth Stuart is Bloomberg Professor of American Health in the Department of Mental Health at the Johns Hopkins Bloomberg School of Public Health, with joint appointments in the Department of Biostatistics and the Department of Health Policy and Management. She is a leading expert in the field of causal inference and policy evaluation with applications in multiple fields, including addiction, mental health, violence, and education.
Roadmap for today’s presentation

I Motivating policy context
II Overview of the classic difference-in-differences (DID) method
III Overview of the autoregressive model: An alternative approach
IV Comparing relative performance: Results from a simulation study
Motivating Policy Context
Myriad of policy approaches have been adopted to address opioid crisis

**INITIATION**
- Day supply limits
- Unused Rx disposal
- Physician education
- Clinical guidelines
- Insurance coverage for non-opioid pain management options

**MISUSE**
- PDMPs
- Drug reformulation
- Pain clinic regulations
- High dose limits
- Insurance pre-authorization

**TREATMENT & RECOVERY**
- Expanded MOUD coverage
- Mandated SUD benefit coverage
- Physician waivers
- Patient limits raised
- Low barrier MOUD access

**HARM REDUCTION**
- Naloxone access laws
- Naloxone distribution campaigns
- Good Samaritan laws
- Fentanyl test strips
Rapid growth in publication of opioid policy studies

Motivating policy context

Policy research often seeks to identify the causal effect of a local, state, or federal level policy

- Our primary focus in this workshop will be on state-level policies
- Our assumed data structure will be state-level longitudinal data
- Our motivating examples will focus on opioid-related policies and outcomes

Many of the analytic techniques and ensuing challenges apply more broadly to non-opioid policy evaluations!
Motivating policy case study

- Our motivating case study will focus on estimating the effect of prescription drug monitoring programs (PDMPs), systems that track and monitor prescribing and dispensation (of certain drugs)
  - What effect does state implementation of any PDMP have on prescription opioid mortality rate?
- Numerous different PDMP policy definitions have been used in the literature (e.g., any PDMP, electronic PDMP, modern PDMP, operational PDMP, “must access” PDMP)
- As highlighted in our previous OPTIC AHSR workshop, the choice of how the policy of interest is defined matters!

Policy state: We observe outcome both before and after policy adoption
Conceptually defining policy effect

Policy effect: how much the outcome changed, compared to if this state had not adopted the policy
For a state that adopts the policy of interest, we only observe their outcomes under the policy.

Ideally, would like to observe outcomes for this state under both conditions: policy adoption and no policy adoption.

Unfortunately, we don’t have a time machine into a parallel universe!

Instead, we need to identify a comparison group to serve as proxy for how the policy state would have fared if they hadn’t adopted policy.
Basic Difference-in-Difference model
Analytic context for DID

Key components of our analytic context:

- “Sudden” policy onset in some states, creating variation in exposure across states
- A treatment group that is affected by the policy of interest
- A (suitable) comparison group that is not affected by the policy of interest
  - Validity of DID lies in the choice of an appropriate comparison group
- Measurable outcome that is a direct target of the policy
Potential Outcome Notation

- Let $Z_t$ denote policy status at time $t$: $Z_t = 0$ or $Z_t = 1$
- For each state, two “potential” (i.e., possible) outcomes at each time $t$:
  - Outcome with policy: $Y(1)_t$
  - Outcome without policy: $Y(0)_t$
Causal estimand of interest

- Target causal estimand Average Treatment Effect on the Treated (ATT):
  \[ ATT = \mathbb{E}(Y(1)_2 - Y(0)_2 | Z = 1) \]
- In words: the difference in outcomes, at the post-period, for the treated state if it had adopted the policy versus not adopted the policy
Causal estimand of interest

Target causal estimand in DID context: $ATT = \mathbb{E}(Y(1)_2 - Y(0)_2 | Z = 1)$

- Estimand is written in terms of potential outcomes
- Need to identify appropriate estimator that relies on observed values
- Insight of DID method: can use observations from an (appropriate) comparison group to impute unobserved potential outcome for treated group.
- **Key point:** doing so requires strong assumptions!!
Quick overview of DID estimator

- Original DID estimand $ATT = \mathbb{E}(Y(1)_2 - Y(0)_2 | Z = 1)$

- Add and subtract $Y(1)_1$

\[
ATT = \mathbb{E}(Y(1)_2 - Y(1)_1 - Y(0)_2 + Y(1)_1) | Z = 1) \\
= \mathbb{E}(Y(1)_2 - Y(1)_1 | Z = 1) - \mathbb{E}(Y(0)_2 - Y(1)_1) | Z = 1)
\]

- No policy effect in pre-period: $Y(1)_1 = Y(0)_1$

\[
ATT = \mathbb{E}(Y(1)_2 - Y(1)_1 | Z = 1) - \mathbb{E}(Y(0)_2 - Y(0)_1 | Z = 1)
\]
Quick overview of DID estimator

\[ ATT = \mathbb{E}(Y(1)_2 - Y(1)_1|Z = 1) - \mathbb{E}(Y(0)_2 - Y(0)_1|Z = 1) \]

"Parallel trends” assumption: \[ \mathbb{E}(Y(0)_2 - Y(0)_1|Z = 1) = \mathbb{E}(Y(0)_2 - Y(0)_1|Z = 0) \]

\[ ATT = \mathbb{E}(Y(1)_2 - Y(1)_1|Z = 1) - \mathbb{E}(Y(0)_2 - Y(0)_1|Z = 0) \]

Consistency assumption: \[ Y_{z,t}^{obs} = (1 - Z) \cdot Y(0)_t + Z \cdot Y(1)_t \]

\[ ATT = \mathbb{E}(Y_{1,2}^{obs} - Y_{1,1}^{obs}) - \mathbb{E}(Y_{0,2}^{obs} - Y_{0,1}^{obs}) \]
DID Estimator

DID estimator is a “difference” of “differences:” compares the differences, before and after the policy, between policy and comparison groups.

\[ ATT = \mathbb{E}(Y_{1,2}^{obs} - Y_{1,1}^{obs}) - \mathbb{E}(Y_{0,2}^{obs} - Y_{0,1}^{obs}) \]

Before-After change in treated group

\[ \mathbb{E}(Y_{1,2}^{obs} - Y_{1,1}^{obs}) \]

Before-After change in comparison group

\[ \mathbb{E}(Y_{0,2}^{obs} - Y_{0,1}^{obs}) \]
Understanding DID Estimator

Policy state
- Policy starts

Comparison state
- pre-policy difference
- post-policy difference

Outcome
- Treated: pre-post difference

State-level opioid policy analyses

Wednesday, April 28, 2021 21 / 57
Understanding DID Estimator

- Policy state
- Comparison state

Policy starts

Comparison: pre-post difference
“Parallel trends” Assumption

Assumption: The change in outcomes from pre- to post-policy in the comparison group is a good proxy for the counterfactual change in the treated group, if they were untreated.
Understanding DID Estimator

Comparison group provides estimate of potential outcome for treated state

Impact = difference in difference
2 groups, multiple time periods

Policy starts

Treatment state

Control state
Visualizing 2 group, multiple time period DID model

![Graph showing visualizing 2 group, multiple time period DID model](image)

Image credit: Andrew Goodman-Bacon
DID Estimation: 2-way fixed effect model

DID model commonly estimated with 2-way fixed effect model:

\[ Y_{it}^{obs} = \alpha Z_{it} + \rho_i + \sigma_t + \epsilon_{it} \]

where \( \rho_i \) are state-level fixed effects and \( \sigma_t \) are time fixed effects.

- The policy indicator is coded as time-varying – for the treated group, \( Z_{it} = 0 \) for \( t \) during pre-policy period and \( Z_{it} = 1 \) for \( t \) during post-policy period
- DID estimate is \( \alpha \), coefficient on \( Z_{it} \)
DID Estimation: 2-way fixed effect model

- Can easily control for time-varying covariates $X_{it}$
  - Covariates can help control for confounding (i.e., measured differences across groups) as well as increase precision of estimates by reducing residual variance
- Accommodates multiple groups, multiple time points
  - DID is calculated as the difference of the average across the “post” period and the average across the “pre” period
  - **Fundamental assumption: policy effect is not time-varying**
- Policy can be implemented at different times for different states
  - **Fundamental assumption: policy effect is not time-varying**
Case study: Implementing DID in STATA

Overview of dataset:

- Annual, state-level data from 1999-2016
- Policy of interest: Any PDMP
- Outcome of interest: State prescription opioid-related mortality rate
**Case study: Implementing DID in STATA**

* Fit basic 2-way FE DID model

```stata
reg OpioidRate AnyPDMP unemploy i.state i.year
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpioidRate</td>
<td>Prescription opioid-related mortality rate</td>
</tr>
<tr>
<td>AnyPDMP</td>
<td>Time-varying indicator of any PDMP</td>
</tr>
<tr>
<td>year</td>
<td>Specify year as factor for fixed effects</td>
</tr>
<tr>
<td>state</td>
<td>Specify state as factor for fixed effects (must be numeric)</td>
</tr>
<tr>
<td>unemploy</td>
<td>Covariate - state-level unemployment rate</td>
</tr>
</tbody>
</table>
### DID model results

```
.reg opioid_rate AnyPDMP unemploy i.state i.year
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 900</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2672.22668</td>
<td>68</td>
<td>39.2974512</td>
<td>F(68, 831) = 39.16</td>
</tr>
<tr>
<td>Residual</td>
<td>833.923484</td>
<td>831</td>
<td>1.00351803</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R-squared = 0.7622</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.7427</td>
</tr>
<tr>
<td>Total</td>
<td>3506.15016</td>
<td>899</td>
<td>3.9000558</td>
<td>Root MSE = 1.0018</td>
</tr>
</tbody>
</table>

| opioid_rate    | Coef.    | Std. Err. | t      | P>|t|  | [95% Conf. Interval] |
|----------------|----------|-----------|--------|------|----------------------|
| AnyPDMP unemploy | -.5497076 | .1243021 | -4.42  | 0.000 | -.7936906 -.3057245 |
| state ALASKA   | 1.456051  | .3353507 | 4.34   | 0.000 | .7978175 2.114285   |
| ARIZONA        | 2.05307   | .3342182 | 6.14   | 0.000 | 1.397059 2.709081   |
| ARKANSAS       | .8370077  | .338725  | 2.47   | 0.014 | .1721505 1.501865   |
Case study: Implementing DID in STATA

Variations on basic DID model

* State-population weighted 2-way FE model
  * Weight variable = ‘pop’
  \[
  \text{reg OpioidRate AnyPDMP unemploy i.state i.year [aw=pop]}
  \]

* Detrended DID model: include state-specific linear time trends
  \[
  \text{reg OpioidRate AnyPDMP unemploy i.state i.year i.state#c.year}
  \]

STATA syntax notes:

<table>
<thead>
<tr>
<th>i.var</th>
<th>Treated as factor (categorical) variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>c.var</td>
<td>Treated as continuous variable</td>
</tr>
</tbody>
</table>
Use of population weights

- For state-level analyses, the use of **population weights** results in models for which each individual is weighted equally
  - Each outcome is given equal weight
  - Overall sample characteristics are nationally representative
- Unweighted analyses put equal weight on each state
  - Outcomes in small states effectively upweighted
  - Sample is not nationally representative (e.g., more rural and White)
Inference in context of DID model

Multiple analytic approaches for standard error (SE) estimation in context of clustering, serial correlation:

- Naive approach: Aggregate data into one pre- and one post-intervention period
- Clustered standard errors at state-level
- Robust SE
- Nonparametric block bootstrap: resample at the group level
Case study: Standard error estimation in STATA

* Clustered SEs by state
reg OpioidRate AnyPDMP unemploy i.state i.year, vce(cluster state)

* Robust SEs
reg OpioidRate AnyPDMP unemploy i.state i.year, vce(robust)

* Bootstrapped SEs
reg OpioidRate AnyPDMP unemploy i.state i.year, vce(boot)

Notes:
- All of these SE methods perform well when number of groups is reasonably large
- Bootstrap not allowed with population-weighting
### Comparing across DID models

<table>
<thead>
<tr>
<th>Model</th>
<th>Coef</th>
<th>Unadj 95% CI</th>
<th>Robust 95% CI</th>
<th>Cluster 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard DID</td>
<td>-0.55</td>
<td>[-0.79, -0.31]∗</td>
<td>[-0.80, -0.30]∗</td>
<td>[-1.27, 0.17]</td>
</tr>
<tr>
<td>Detrended DID</td>
<td>-0.25</td>
<td>[-0.50, 0.00]∗</td>
<td>[-0.45, -0.05]∗</td>
<td>[-0.58, 0.08]</td>
</tr>
</tbody>
</table>
Note of Caution 1: “Testing” DID Assumptions

- In applied work, it’s become commonplace to “test” for parallel trends by examining the pre-treatment period.

- Showing that trajectories are similar in pre-policy period does not establish that the parallel trends assumption is upheld.

- Fundamentally, cannot test this assumption, as it is an assumption regarding the (unobservable) potential outcomes.

Note of Caution 2: Beware of overfitting

- An overfit model contains more parameters than can be supported by the observed data
- Will create a model with unreliable regression coefficients and standard errors
- **1 in 10 rule** is typically considered to be the bare minimum rule to follow when fitting your data
  - Checking case study: 50 states over 18 years, so we should not have models with more than 90 parameters
  - DID model: 49 state fixed effects + 17 time fixed effects + 1 policy variable + 1 control covariate = 68 parameters
  - Detrended model: adds 49 more parameters!!! It is subject to dangers of overfitting
Note of Caution 3: Staggered policy adoption

Image credit: Andrew Goodman-Bacon
Chaisemartin, Clément de, and Xavier D’Haultfoeuille. 2019. Two-way fixed effects estimators with heterogeneous treatment effects.
Overview of Autoregressive Models
Autocorrelation in Time Series Data

- Autocorrelation exists within a time series if values in the series are associated in some way with earlier values
  - Also known as "autoregressive (AR) data"
- When observations are autocorrelated, assumptions of traditional OLS regression are violated
- Regression coefficients are still unbiased, but no longer have minimum variance property
- Standard error estimates may be biased
Examples of autocorrelated data

Positive Autocorrelation

Negative Autocorrelation
Estimation

- **Big picture:** We need to adjust “ordinary” regression estimates for the fact that the data have an AR structure.
- AR model uses parametric correction that assumes an AR process and includes lagged outcome terms in regression model
- Need adequate length time series: may perform poorly with short time series
Intro to the Autoregressive (AR) Model

- Model in which outcome (from a time series) is regressed on previous values from that same time series: e.g., $y_t$ regressed on $y_{t-1}$
- The order of an autoregressive model is the number of lagged values that are used to predict the outcome.

AR(1) fixed effects model:

$$Y_{it}^{obs} = \beta_0 + \beta_1 Y_{i,t-1}^{obs} + \sigma_t + \epsilon_{it}$$

AR(2) fixed effects model:

$$Y_{it}^{obs} = \beta_0 + \beta_1 Y_{i,t-1}^{obs} + \beta_2 Y_{i,t-2}^{obs} + \sigma_t + \epsilon_{it}$$

where $\sigma_t$ denotes time fixed effects.

- Note: state FEs not included; use lagged term(s) to account for state-specific variability.
Important: Must code policy variable using change coding

- Generally, policy variable is coded using effect coding, $Z_{it}$, which denotes the policy status at time $t$.
- Alternatively, under change coding, the policy variable $Z_{it}$ denotes the change in policy status from time $t-1$ to $t$.
  - Change coding variable defined as $Z_{it} - Z_{i,t-1}$.
- Prior statistical theory work has shown that AR models should use change coding, as policy effect estimates will be biased when using standard effect coding.
Choosing the number of lags

- **Autocorrelation plot:** plots autocorrelation value (ranging from −1 to 1) as a function of lag timing
  - Helps to diagnose whether autocorrelation is present in your dataset, as well as the number of lags that autocorrelation is present
Choosing the number of lags

- **Partial autocorrelation plot**: plots autocorrelation value at a given lag $k$ that has not been accounted for by lags 1 through $k-1$. 
Choosing the number of lags

- Alternative approach: Use model fit statistics like Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC)
  - Estimate series of models with increased number of lag terms (these are nested models)
  - Information criterion can guide model selection: want model that minimizes BIC or AIC
  - AIC has smaller penalty term than BIC, so will generally indicate more lagged terms than BIC
Case study: Initial diagnostics in STATA

Observed state-specific trends in opioid-related mortality rate
Case study: Autocorrelation plot diagnostics

* Generate autocorrelation plot for OpioidRate variable
ac OpioidRate
Case study: Partial autocorrelation plot diagnostics

* Generate autocorrelation plot for OpioidRate variable
  pac OpioidRate
Case study: Implementing AR model in STATA

Can easily generate lagged terms in STATA

* Generate lagged terms for OpioidRate variable
sort state year
by state: gen opioid_lag1 = OpioidRate[_n-1]

* Generate new policy variable with change coding
by state: gen anyChange = AnyPDMP[_n] - AnyPDMP[_n-1]

Fixed effect model that includes autoregressive term

* AR1 model
* Includes year fixed effects, but not state fixed effects
reg OpioidRate anyChange unemploy opioid_lag1 i.year
AR1 model results (Unweighted)

```
.reg opioid_rate anyChange unemploy opioid_lag1 i.year
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 850</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2821.55974</td>
<td>19</td>
<td>148.503144</td>
<td>F(19, 830) = 238.43</td>
</tr>
<tr>
<td>Residual</td>
<td>516.961556</td>
<td>830</td>
<td>.622845248</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3338.52129</td>
<td>849</td>
<td>3.93229834</td>
<td>R-squared = 0.8452</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.8416</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 0.78921</td>
</tr>
</tbody>
</table>

| opioid_rate     | Coef.    | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----------------|----------|-----------|-------|-----|---------------------|
| anyChange       | -.1619135| .184565   | -0.88 | 0.381| -.5241826 to 0.2003555 |
| unemploy        | .02384   | .0206637  | 1.15  | 0.249| -.0167192 to 0.0643993 |
| opioid_lag1     | .9134081 | .0155804  | 58.63 | 0.000| .8828266 to .9439897 |
| year            |          |           |       |     |                     |
| 2001            | .0625888 | .1583639  | 0.40  | 0.693| -.2482521 to 0.3734297 |
| 2002            | -.031379 | .1607177  | -0.20 | 0.845| -.3468399 to .2840818 |
| 2003            | .0041792 | .1619141  | 0.03  | 0.979| -.3136301 to .3219884 |
Inference in context of AR model

Similar to DID models, various choices for standard error (SE) estimation in context of clustering, serial correlation:

- Clustered standard errors at state-level
- Robust SE
- Nonparametric block bootstrap: resample at the group level
Case study: Standard error estimation in STATA

* Clustered SEs by state
reg OpioidRate anyChange unemploy opioid_lag1 i.year, vce(cluster state)

* Robust SEs
reg OpioidRate anyChange unemploy opioid_lag1 i.year, vce(robust)

* Bootstrapped SEs
reg OpioidRate anyChange unemploy opioid_lag1 i.year, vce(boot)

Notes:
- All of these SE methods perform well when number of groups is reasonably large
- Bootstrap not allowed with population-weighting
### Comparing across DID models

<table>
<thead>
<tr>
<th>Model</th>
<th>Coef</th>
<th>Unadj 95% CI</th>
<th>Robust 95% CI</th>
<th>Cluster 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard DID</td>
<td>-0.55</td>
<td>[-0.79, -0.31]*</td>
<td>[-0.80, -0.30]*</td>
<td>[-1.27, 0.17]</td>
</tr>
<tr>
<td>Detrended DID</td>
<td>-0.25</td>
<td>[-0.50, 0.00]*</td>
<td>[-0.45, -0.05]*</td>
<td>[-0.58, 0.08]</td>
</tr>
<tr>
<td>AR(1) Model</td>
<td>-0.16</td>
<td>[-0.52, 0.20]</td>
<td>[-0.51, 0.18]</td>
<td>[-0.48, 0.16]</td>
</tr>
</tbody>
</table>
Traditional view of AR models

- Objective of AR model is prediction (forecasting) not causal inference
- In prediction, $R^2$ matters a lot
- Omitted variable bias is not a concern
- Not interested in interpreting regression coefficients
- External validity is key: model based on historical data must accurately predict future data
Comparing relative performance: Results from a simulation study
Simulation Study

• Overarching goals:
  • Identify the *best* causal inference *methods* for assessing the effects of opioid policies under several different scenarios
  • Improve the state of the science and provide guidelines for best practices
• How:
  • Using existing data (state-level for now), simulate the effect of a policy on outcomes and determine which method(s) most accurately detect the effect
  • Methods considered:
    • Common approaches like DID, detrended, autoregressive, and GEE models
• Key challenge:
  • How to simulate causal effects of policies using data structures such as those used to evaluate opioid policies and with “realistic” effect structures and sizes?
We evaluated model performance using a simulation

Simulate policy effect in real data

Estimate effect with statistical models

Compare model performance

5,000 trials for each of a number of conditions; N = 50 states

Dozens of combinations of modeling assumptions; Using annual and state aggregate data; Using one additional covariate (unemployment rate)

Four performance measures: Type 1 error rates (“false positives”), power, bias, root mean squared error (RMSE)
Illustration of how we simulate policy effects

1. Real U.S. state opioid-related death rates
Illustration of how we simulate policy effects

1. Real U.S. state opioid-related death rates
2. Randomly select 5 states

Opioid Deaths (per 100,000 population)
Illustration of how we simulate policy effects

1. Real U.S. state opioid-related death rates
2. Randomly select 5 states
3. Randomly select policy implementation date

Opioid Deaths (per 100,000 population)
Illustration of how we simulate policy effects

1. Real U.S. state opioid-related death rates
2. Randomly select 5 states
3. Randomly select policy implementation date
4. Introduce policy effect after implementation date
Illustration of how we simulate policy effects

1. Real U.S. state opioid-related death rates
2. Randomly select 5 states
3. Randomly select policy implementation date
4. Introduce policy effect after implementation date
5. Repeat 5000 times

Opioid Deaths
(per 100,000 population)
Data Generation Scenarios

- 5,000 simulated datasets constructed for each of 56 different simulation settings

- Four policy prevalence conditions (i.e., # of treated states)
  - Randomly select 1, 5, 15, or 30 states as “implementing” a law

- Seven policy effect conditions after policy implemented
  - Null: State outcomes are unchanged from real (observed) outcomes
  - Negative: State outcomes reduced by effect sizes of 5%, 15%, and 25%
  - Positive: State outcomes increased by effect sizes of 5%, 15%, and 25%

- Two policy effect phase-in conditions
  - 3-year phase-in to full effect vs. instant phase in
<table>
<thead>
<tr>
<th>GLM Specification</th>
<th>Regression Specification</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Linear</td>
<td>Fixed effects (FE)</td>
<td>Population weighted</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>3 FE + Detrended</td>
<td></td>
<td>Population weighted</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>5 Autoregressive</td>
<td></td>
<td>Population weighted</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>7 GEE model</td>
<td></td>
<td>Population weighted</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>9 Log-linear</td>
<td>Fixed effects (FE)</td>
<td>Population weighted</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>11 Autoregressive</td>
<td></td>
<td>Population weighted</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>13 Negative Binomial</td>
<td>Fixed effects (FE)</td>
<td>Unweighted</td>
</tr>
<tr>
<td>14 FE + Detrended</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>15 Autoregressive</td>
<td></td>
<td>Unweighted</td>
</tr>
<tr>
<td>16 Poisson</td>
<td>Fixed effects (FE)</td>
<td>Unweighted</td>
</tr>
<tr>
<td>17 Autoregressive</td>
<td></td>
<td>Unweighted</td>
</tr>
</tbody>
</table>

Compare findings from 17 different models

Simulation Project
Alternative statistical methods to address differences across states

1. **Detrended model** = allows for state-specific linear slopes over time
   - adding in state-specific linear trends with continuous time \( (t) \)
   \[
g(\text{\textit{Y}}_{it}) = \alpha \cdot Z_{it} + \beta \cdot X_{it} + \rho_i + \sigma_t + \sum_{s=1}^{50}(\omega_s \cdot t \cdot 1(\text{\textit{state}}_i = \text{\textit{state}}_s)) + \nu_{it}
   \]

2. **Autoregressive (AR) model** = includes lagged measures of outcome
   - Controlling for lagged outcome adjusts for average differences across policy and comparison states and improves prediction of future outcomes when there is high autocorrelation
   \[
g(\text{\textit{Y}}_{it}) = \gamma \cdot (Z_{it} - Z_{i,t-1}) + \beta \cdot X_{it} + \eta \cdot \text{\textit{Y}}_{it-1} + \sigma_t + \epsilon_{it}
   \]

3. **Generalized estimating equations (GEE) model** = specifies a covariance structure for within-state clustered outcomes; assume an autocorrelation structure of order 1 (AR1)
   \[
g(\text{\textit{Y}}_{it}) = \alpha \cdot Z_{it} + \beta \cdot X_{it} + \sigma_t + \zeta_{it}
   \]

Simulate    Estimate    Compare
We examined which modeling approach best describes the true effect of the simulated policy by capturing:

1. **Directional bias**: Does model consistently over or under estimate effect?
2. **Magnitude bias**: Is absolute value of effect consistently over/under estimated?
3. **Root Mean Squared Error**: How much error exists, taking into account both directional bias and variance?

We also capture traditional metrics of statistical significance related to null hypothesis testing (given the literature still solely focused on these)

1. **Type I error rates**: False alarm probability when no true effect
2. **Statistical power**: Probability of detecting real effects
<table>
<thead>
<tr>
<th>Model Type</th>
<th>Autoreg Effect</th>
<th>State Effect</th>
<th>SE Adj</th>
<th>Instant 3</th>
<th>Instant 15</th>
<th>Instant 35</th>
<th>5-year 3</th>
<th>5-year 15</th>
<th>5-year 35</th>
<th>Avg</th>
<th>Worst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neg binomial</td>
<td>AR-change</td>
<td>None</td>
<td>None</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td><strong>0.03</strong></td>
<td><strong>0.02</strong></td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-change</td>
<td>None</td>
<td>Cluster</td>
<td>0.23</td>
<td>0.09</td>
<td>0.07</td>
<td>0.22</td>
<td>0.08</td>
<td>0.06</td>
<td>0.12</td>
<td>0.23</td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-effect</td>
<td>None</td>
<td>None</td>
<td>0.10</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td><strong>0.10</strong></td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-effect</td>
<td>None</td>
<td>Cluster</td>
<td>0.22</td>
<td>0.12</td>
<td>0.11</td>
<td>0.23</td>
<td>0.13</td>
<td>0.12</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-change</td>
<td>Fixed</td>
<td>None</td>
<td>0.08</td>
<td>0.05</td>
<td>0.06</td>
<td>0.09</td>
<td>0.11</td>
<td>0.13</td>
<td><strong>0.08</strong></td>
<td><strong>0.13</strong></td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-change</td>
<td>Fixed</td>
<td>Cluster</td>
<td>0.22</td>
<td>0.09</td>
<td>0.07</td>
<td>0.21</td>
<td>0.09</td>
<td>0.09</td>
<td>0.13</td>
<td>0.22</td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-effect</td>
<td>Fixed</td>
<td>None</td>
<td>0.19</td>
<td>0.28</td>
<td>0.26</td>
<td>0.21</td>
<td>0.30</td>
<td>0.27</td>
<td>0.25</td>
<td>0.30</td>
</tr>
<tr>
<td>Neg binomial</td>
<td>AR-effect</td>
<td>Fixed</td>
<td>Cluster</td>
<td>0.28</td>
<td>0.18</td>
<td>0.16</td>
<td>0.29</td>
<td>0.18</td>
<td>0.16</td>
<td>0.21</td>
<td>0.29*</td>
</tr>
<tr>
<td>Log lin-wgt</td>
<td>AR-change</td>
<td>None</td>
<td>None</td>
<td>0.52</td>
<td>0.70</td>
<td>0.70</td>
<td>0.54</td>
<td>0.73</td>
<td>0.71</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>Log lin-wgt</td>
<td>AR-change</td>
<td>None</td>
<td>Huber</td>
<td>0.59</td>
<td>0.66</td>
<td>0.62</td>
<td>0.61</td>
<td>0.69</td>
<td>0.64</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>Log lin-wgt</td>
<td>AR-change</td>
<td>None</td>
<td>Cluster</td>
<td>0.28</td>
<td>0.18</td>
<td>0.16</td>
<td>0.29</td>
<td>0.18</td>
<td>0.16</td>
<td>0.21</td>
<td>0.29*</td>
</tr>
<tr>
<td>Linear-wgt</td>
<td>AR-change</td>
<td>None</td>
<td>Both</td>
<td>0.30</td>
<td>0.11</td>
<td>0.09</td>
<td>0.27</td>
<td>0.13</td>
<td>0.16</td>
<td>0.17</td>
<td>0.30</td>
</tr>
<tr>
<td>Linear-wgt</td>
<td>AR-change</td>
<td>None</td>
<td>Cluster</td>
<td>0.21</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>0.06</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>Linear-wgt</td>
<td>AR-effect</td>
<td>None</td>
<td>None</td>
<td>0.12</td>
<td>0.16</td>
<td>0.16</td>
<td>0.11</td>
<td>0.16</td>
<td>0.16</td>
<td>0.15</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Lots of interesting results!!!!!
Results for classic DID and AR
Classic DID model has high Type I error rates, highlighting potential to say policies have an effect when they do not.

Goal: To have models with Type 1 error = 0.05
AR model has much more reasonable Type I error rates, much less likely to have false positives saying policies have an effect when they do not.

Goal: To have models with Type 1 error $= 0.05$
Classic DID (linear 2-way fixed effects) model has low power across policy effect sizes

Goal: To have models with power = 0.80
AR model has much better power across different policy effect sizes

Goal: To have models with power = 0.80
Results for all models
Best performing models include lagged outcomes or use negative binomial specification
Final note: Performance always worse when it takes policy longer to become effective
Summary of key findings

- Type I error rates are unreasonably high when number of states implementing new policy is low (< 15)
- Power is very low for all models; need to find better approaches to account for level of uncertainty
  - Use of lagged outcomes as control covariates is helpful in the linear model
  - Use of negative binomial link performs better than Poisson

Several important guidelines for best practice

• Caution needed when such studies report “statistically significant” findings - may be a false positive (e.g., saying a law has an effect when it truly does not)

• Critical to use cluster adjustments to standard errors when using state and year fixed effects in linear or log-linear models

• Recommend use of AR models over classic two-way fixed effects DID models
Acknowledgments & Disclosures

• This work has been supported by the National Institute on Drug Abuse of the NIH through P50-DA046351
• No conflicts of interest
• Contributors:
  • Bradley Stein
  • Rosalie Pacula
  • Megan Schuler
  • Rosanna Smart
  • Sara Heins
  • Courtney Kase
  • Elizabeth Stuart
  • Stephen Patrick
  • Sierra Smucker
  • Joseph Pane
  • David Powell
  • Mary Vaiana
  • Hilary Peterson
Thanks!!

Beth Ann Griffin bethg@rand.org
Megan Schuler mschuler@rand.org
Elizabeth Stuart estuart@jhu.edu