Justify your model: Identifying the best model assumptions

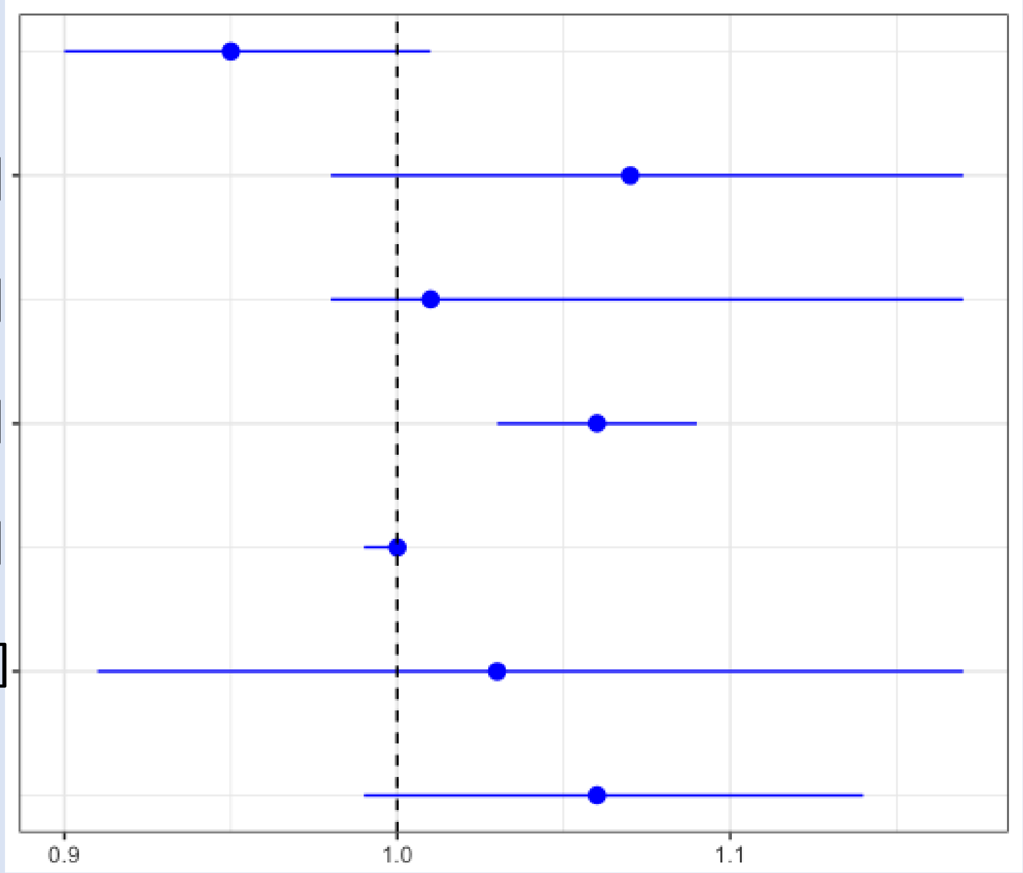
Terry Schell, Andrew Morral, Beth-Ann Griffin, and Matt Cefalu

This research was supported by internal RAND funding and Arnold Ventures

Disagreement about the effect of RTC laws on murder

Incidence rate ratios with 95% confidence interval

Martin & Legault (2005)	0.95 [0.90,1.01]
Rosengart et al. (2005)	1.07 [0.98,1.17]
Grambsch (2008; random effects)	1.01 [0.98,1.17]
Grambsch (2008; fixed effects)	1.06 [1.03,1.09]
Kendall & Tamura (2010)	1.00 [0.99,1.00]
Aneja, Donohue, Zhang (2014)	1.03 [0.91,1.17]
Webster, Crifasi, Vernick (2014)	1.06 [0.99,1.14]



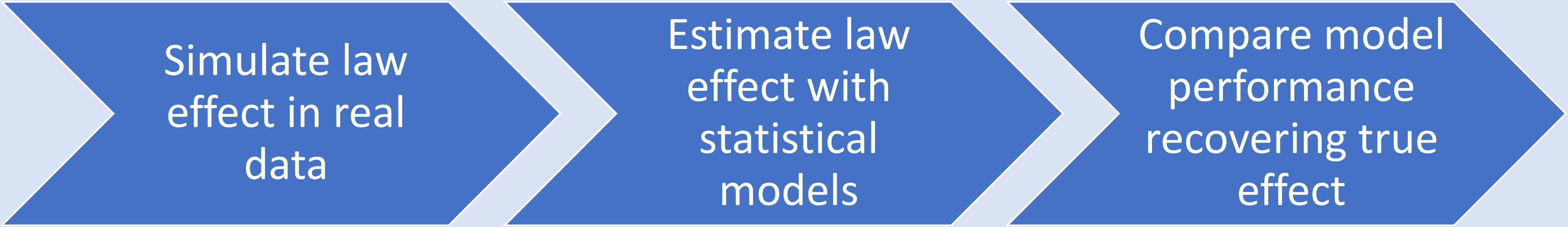
How do we synthesize these results?

- These are not independent estimates, but rely on similar (or identical) underlying data.
- These papers are making very different statistical assumptions in both their models and variance estimation techniques
- We would like to rely on the estimates from studies that made the most appropriate assumptions for the data, and ignore the other studies.
- Whose assumptions are most appropriate?

How to justify your methods

- Identifying all of the different statistical assumptions across two different statistical methods is difficult, testing each of assumption is more difficult, and integrating those tests results across multiple assumptions is nearly impossible.
- Rather than test individual statistical assumptions, we can simulate the statistical properties of the estimator of interest for each method within the actual dataset of interest.
  - Does the method yield Type 1 error ≈ alpha under the null?
  - Does the method have a high correct rejection rate (low MSE) under alternative?
  - Does the method have bias toward positive or negative findings?
  - Does the method underestimate or overestimate the true magnitude of an effect?

Simulation Logic



Simulate law effect in real data

Estimate law effect with statistical models

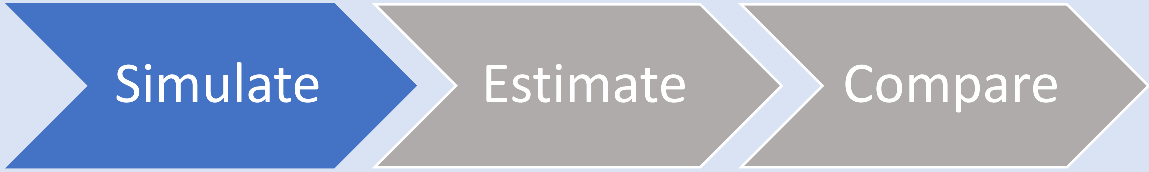
Compare model performance recovering true effect

5000 trials for each of 18 conditions

Dozens of combinations of modeling assumptions

Four performance measures

Illustration of simulation


Simulate

Estimate

Compare

1. Real US State firearms death rates (1979 to 2014)

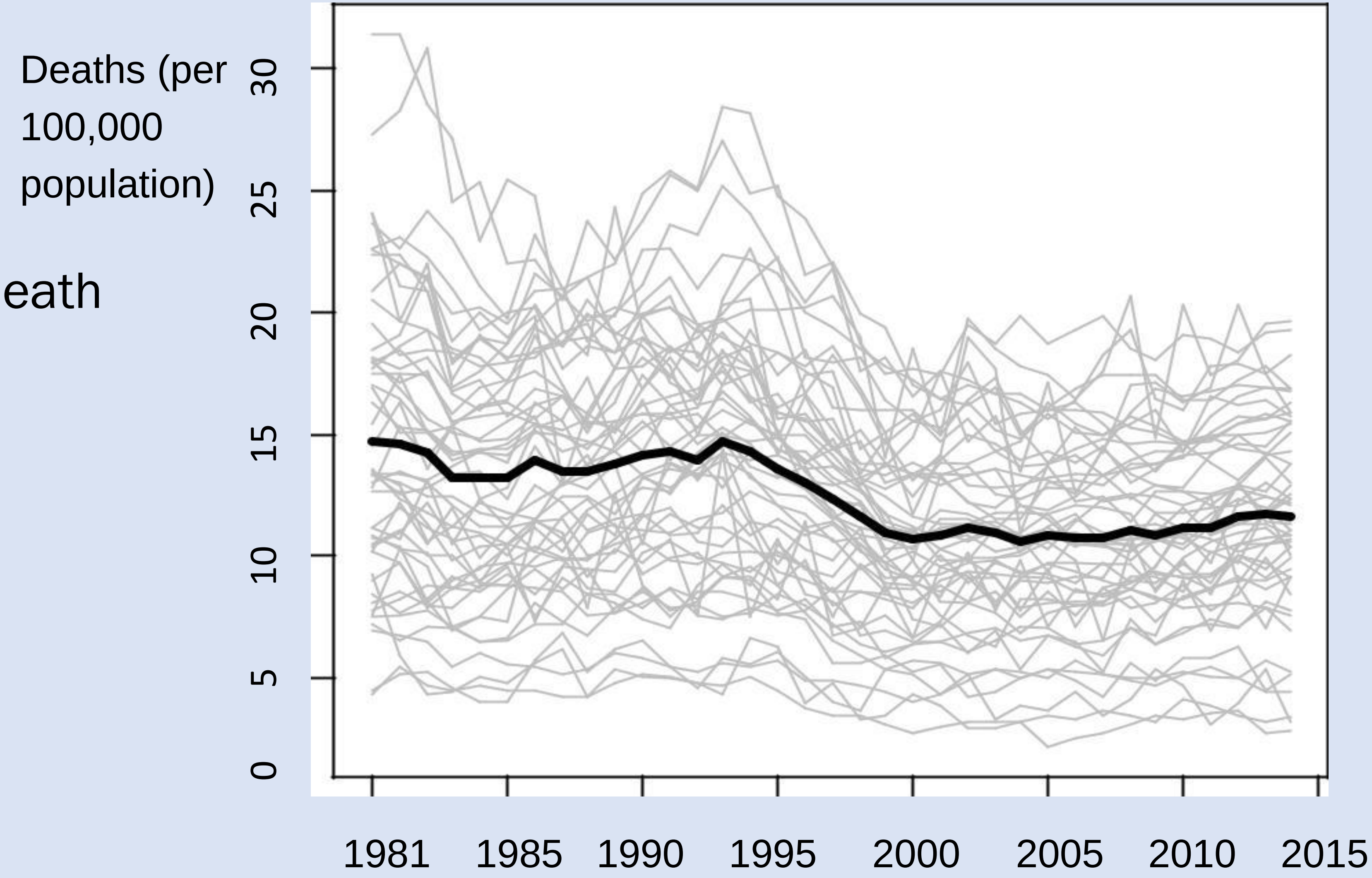
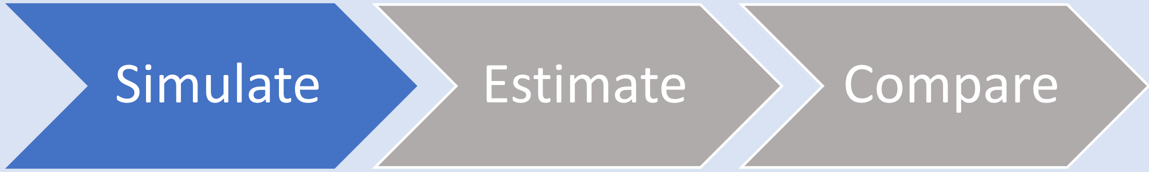


Illustration of simulation



1. Real US State firearms death rates (1979 to 2014)
2. Randomly select three states

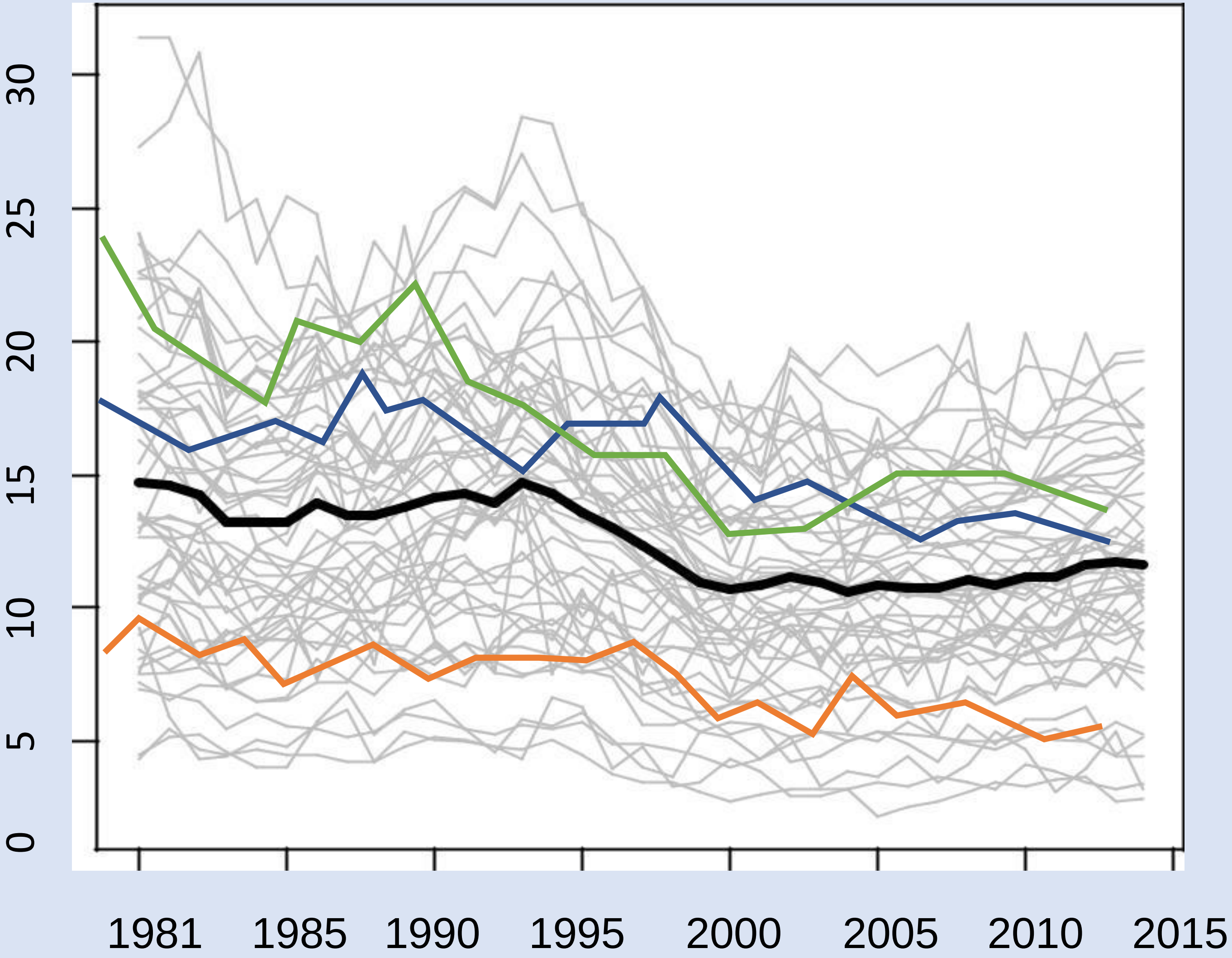
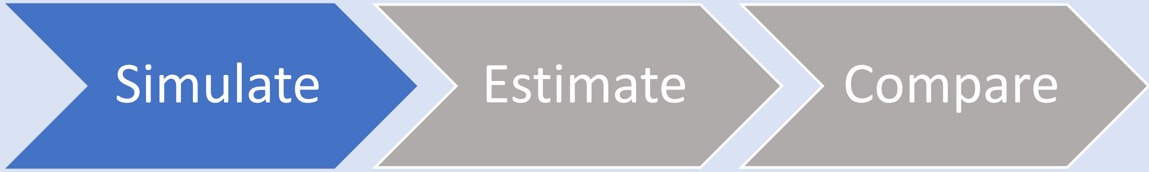


Illustration of simulation



1. Real US State firearms death rates (1979 to 2014)
2. Randomly select three states
3. Randomly select law implementation date

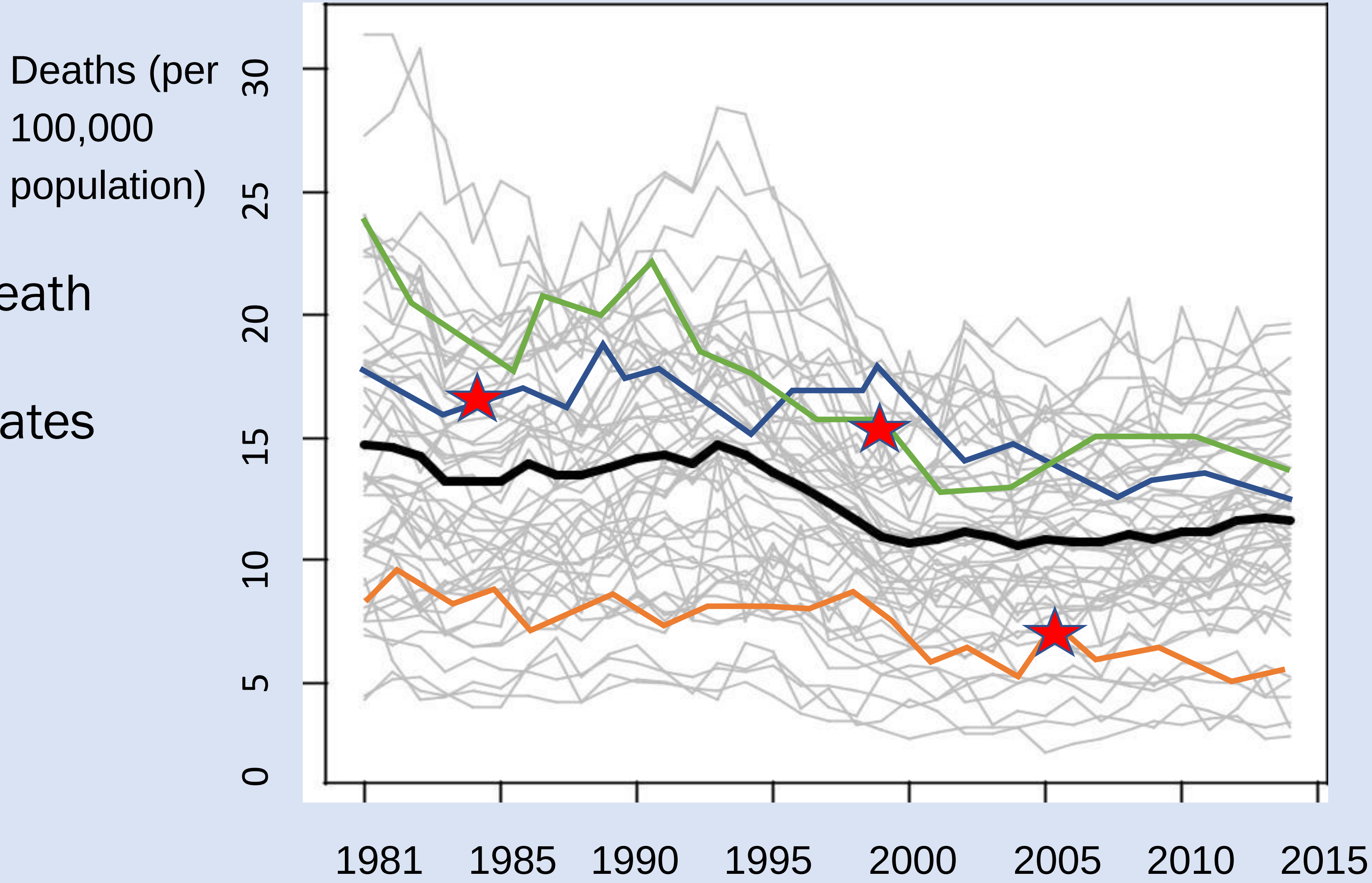
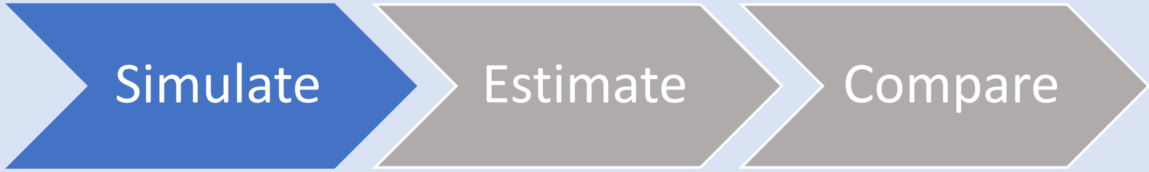
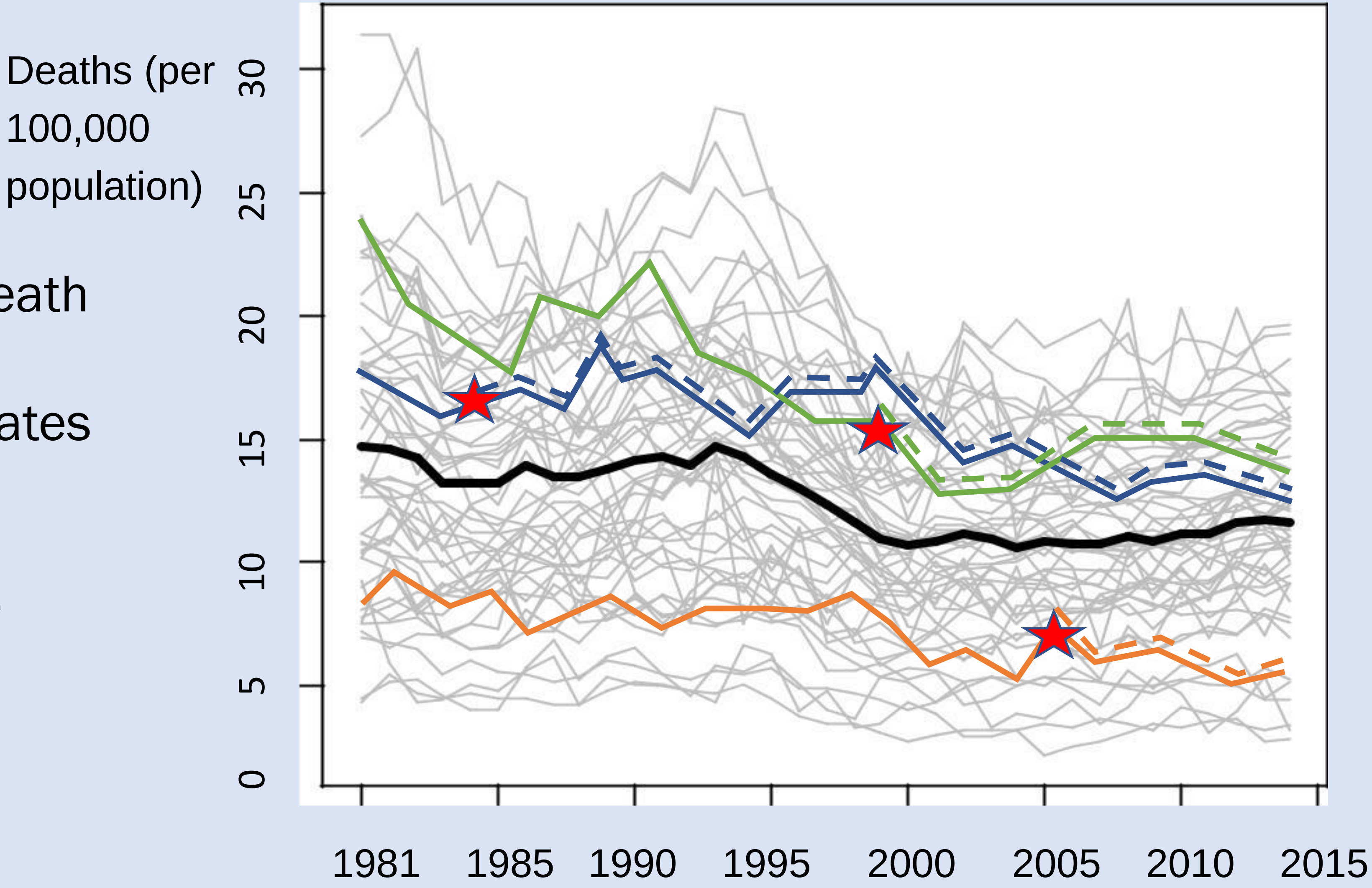


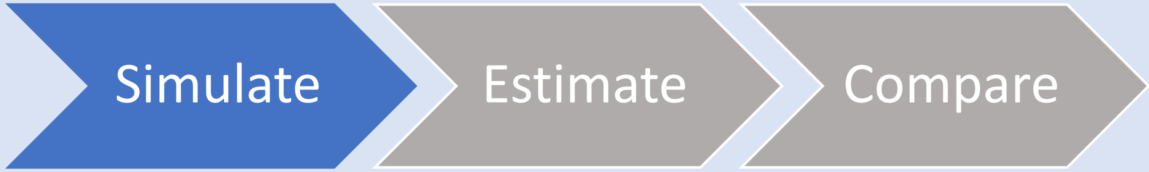
Illustration of simulation





1. Real US State firearms death rates (1979 to 2014)
2. Randomly select three states
3. Randomly select law implementation date
4. Introduce law effect after implementation date

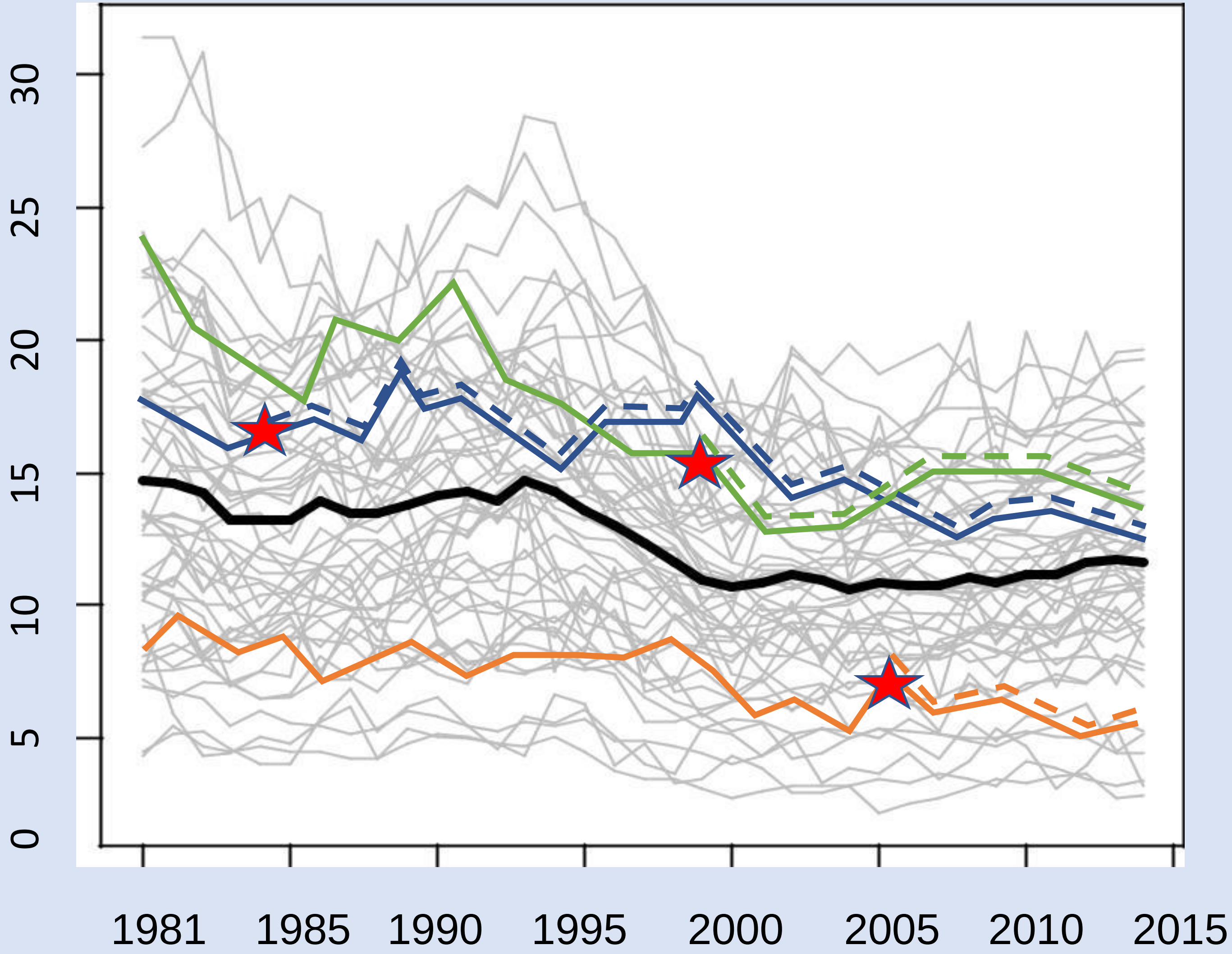
Illustration of simulation



Simulate

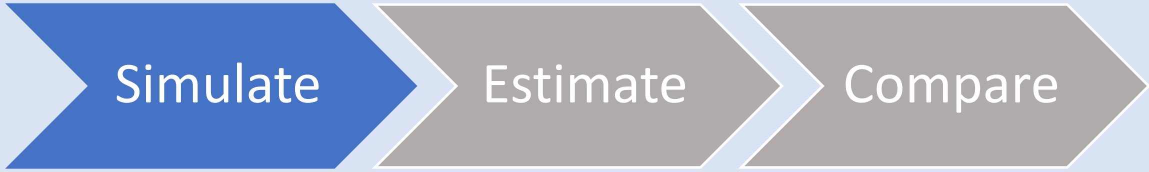
Estimate

Compare



1. Real US State firearms death rates (1979 to 2014)
2. Randomly select three states
3. Randomly select law implementation date
4. Introduce law effect after implementation date
5. Do steps 1-4, 5000 times

5000 simulated datasets constructed for each of 18 conditions



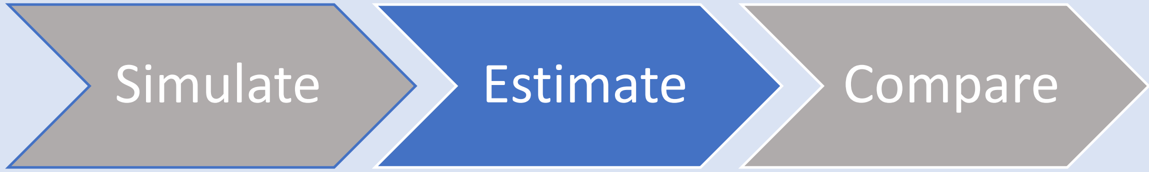
Simulate

Estimate

Compare

- Three “law effect” conditions.
  - Null: State outcomes are unchanged from real (observed) outcomes
  - Negative: State outcomes reduced by effect size equivalent to 1000 fewer deaths nationally
  - Positive: State outcomes increased by effect size equivalent to 1000 more deaths nationally
- Three “law prevalence” conditions
  - Randomly select 3, 15 or 35 states as “implementing” a law
- Two “law phase-in” conditions
  - 5-year phase in to full effect vs. Instant phase in

For each simulated dataset, we compared a range of modeling choices



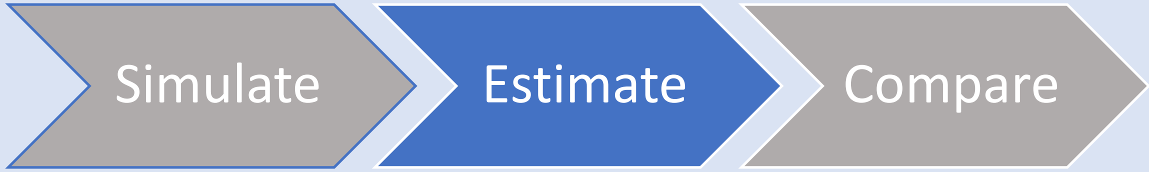
Simulate

Estimate

Compare

- link function (linear or log-link) and likelihood function
  - OLS, log(Y) OLS, Poisson or Negative Binomial
- population weights vs not
- inclusion of lagged effects vs not
- inclusion of state-fixed or random effects
- inclusion of state-specific linear trends vs not
- use of general estimating equations
- use of standard error adjustments for clustering by state
- use of robustness adjustments to the standard error
- type of coding used for the law's effect: *effect* versus *change coding*

Estimating effects with a lagged outcome



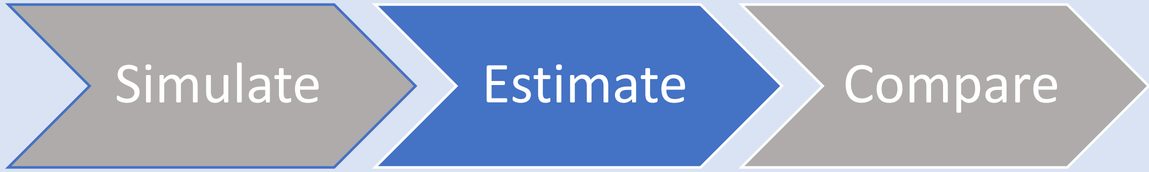
$Deaths_T = b_1 Deaths_{T-1} + b_2 LawEffect + \ldots$

Year	-2	-1	0	1	2	3
Effect Coding	0	0	1	1	1	1
Change Coding	0	0	1	0	0	0

Law Implemented

- Coding for time series analyses:
  - Levels models use simple effect coding
  - First-differences models use change coding
- Autoregressive models are a cross between a levels model and a first-difference model.
  - Is a levels model when autoregression = 0
  - Is first-difference model when autoregression = 1
  - Autoregression coefficient in data ≈ 0.9
- All studies in this field that include a lagged outcomes use *effect coding*

Basic features of all models



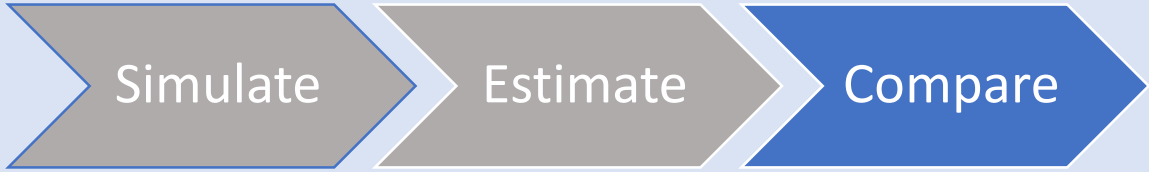
Simulate

Estimate

Compare

- Law effect (*change* or *effect* coded; instant or 5-year phase in)
- 36 time-varying state covariates, reduced to 17 principal components that explain 95% of variance
- Year fixed effects

We compared the performance of statistical modeling choices



Simulate

Estimate

Compare

- We estimate each model/method for getting an effect within each of 5000 x 18 datasets.
- For each model/method we then compute:
  1. Type I error rates: false alarm probability when no true effect
  2. Correct rejection rate: probability of detecting real effects
  3. Directional bias: estimates are too positive or negative
  4. Magnitude bias: absolute size of estimates are too large or small

Model Type	Autoreg Effect	State Effect	SE Adj	Instant 3	Instant 15	Instant 35	5-year 3	5-year 15	5-year 35	Avg	Worst
Neg binomial	AR-change	None	None	0.04	0.04	0.04	0.03	0.02	0.02	**0.03**	0.02*
Neg binomial	AR-change	None	Cluster	0.23	0.09	0.07	0.22	0.08	0.06	0.12	0.23
Neg binomial	AR-effect	None	None	0.10	0.10	0.09	0.10	0.10	0.10	0.10	0.10*
Neg binomial	AR-effect	None	Cluster	0.22	0.12	0.11	0.23	0.13	0.12	0.15	0.23
Neg binomial	AR-change	Fixed	None	0.06	0.05	0.06	0.09	0.11	0.13	**0.08**	0.13*
Neg binomial	AR-change	Fixed	Cluster	0.22	0.09	0.07	0.21	0.09	0.09	0.13	0.22
Neg binomial	AR-effect	Fixed	None	0.19	0.28	0.26	0.21	0.30	0.27	0.25	0.30
Neg binomial	AR-effect	Fixed	Cluster	0.21	0.12	0.11	0.21	0.13	0.11	0.15	0.21*
Neg binomial	None	Fixed	None	0.48	0.53	0.51	0.50	0.56	0.53	0.52	0.56
Neg binomial	None	Fixed	Cluster	0.19	0.09	0.07	0.20	0.09	0.07	0.12	0.20*
Neg binomial	None	Fixed and trend	None	0.37	0.39	0.41	0.41	0.44	0.44	0.41	0.44
Neg binomial	Non										
Log lin-wgt	Non										
Log lin-wgt	Non										
Log lin-wgt	Non										
Log lin-wgt	Non										
Linear-wgt	AR-cha										
Linear-wgt	AR-cha										
Linear-wgt	AR-change	None	Cluster	0.30	0.12	0.09	0.27	0.13	0.10	0.17	0.30
Linear-wgt	AR-change	None	Both	0.21	0.09	0.07	0.08	0.08	0.06	0.10	0.21
Linear-wgt	AR-effect	None	None	0.12	0.16	0.16	0.11	0.16	0.16	0.15	0.16

Lots of interesting results, but too much to cover here.
I will focus on highlights only

Key observations about Type 1 error

- Almost all of the models commonly used to estimate gun law effects have poor type one error rates
- For example, averaged across 3 implementing states and both slow and fast phase-in conditions, with alpha=.05:

Model	SE adjustment	Type 1 error
Standard 2-way linear fixed effects (diff in diff), population weighted	None	.62
	Huber	.60
	Cluster	.20
	Huber & Cluster	.59

Key observations about Type 1 error

- Averaged over all simulation conditions, only 8 methods had Type 1 error rates below .10
- The four models with average Type 1 error rates closest to 0.05

Model	SE adjustment	Type 1 error
Negative binomial, autoregressive, change coded, no state fixed effect	None	.03
Linear, unweighted, autoregressive, change coded, state fixed effects	None	.05
Linear, unweighted, autoregressive, effect coded, no state fixed effects	None	.05
Linear, unweighted, autoregressive, change coded, state random effects	None	.04

All models had very low correct rejection rates for the effect size we examined

- All models had low power to detect an effect that corresponds to 1000 more or fewer deaths nationally, a small but important effect.
- 7 had power < 0.1; 24 had power < 0.2, only 1 had power > 0.2
- Among the models with good Type 1 error rates:

Model	SE adjustment	Power
Negative binomial, autoregressive, change coded, no state fixed effect	None	.21
Linear, unweighted, autoregressive, change coded, state fixed effects	None	.10
Linear, unweighted, autoregressive, effect coded, no state fixed effects	None	.11
Linear, unweighted, autoregressive, change coded, state random effects	None	.11

Few models showed directional bias

- An exception is one of the most commonly used models in the field
- Linear two-way fixed effects with log transformed outcome had a positive bias
- Especially a problem when few states (3) implemented the law
    - Negative effects were estimated as less than half the true effect
    - Positive effects were 50% greater than true effect

Magnitude bias was a problem for autoregressive models

- On average, autoregressive models using effect coding underestimated true effect magnitude by 73%
- Autoregressive models using change coding showed minimal magnitude bias
- The model that performed best on power, and had appropriate Type 1 error, had magnitude bias of just 5%:

negative binomial, autoregressive, change coding, no SE adjustment

- This model also has some other desirable features
  - Minimally sensitive to exclusion of covariates
  - Does not show artifacts due to regression to the mean

Conclusions from simulation

- Some of the most commonly used models are poorly suited for evaluating the effects of state laws on firearm death rates
  - Researchers in this area should be wary of using ”robust” SE adjustments
  - *Change coding* worked dramatically better than *effect coding* when including an autoregressive term in these data
  - Model performance is often unreliable when only a few states have implemented a law
- Statistical power is very low – frequentist hypothesis testing framework may not be a useful way to summarize the evidence
- Full report on the simulation (along with all code) is published at:

  https://www.rand.org/pubs/research_reports/RR2685.html

Applying the selected model to real laws

We have begun modeling the effects of gun laws, but have made some changes from what was used in the simulation:

- Changes in how we code the laws.
  - Since we do not know how much of effect is instant vs phased-in over time, we include terms for both an instant and a 6-year phase-in of the effect
  - We code laws using both change and effect coding. Thus, the “optimal” model from the simulations is a special case of the model we are using.
  - To get the total effect of the law in a given year we compute the marginal effect, integrating over:
    - the instant and slow phase in terms
    - the change and effect coding terms
    - time through the autoregressive term

Applying the selected model to real laws

- Changes in how we estimate the model
  - We use Bayesian estimation (using STAN)
  - Our priors are that the log IRR of the total effect has mean zero, SD = .10, which corresponds to assuming 95% probability that the true IRR of each law is between 0.82 and 1.22.
- We estimate the effects for three laws simultaneously.
  - Child-access prevention laws
  - Stand-your-ground laws
  - Right to carry laws
  - We also estimate combination of these laws comparing restrictive regime (CAP, but no SYG or RTC) to a permissive regime (no CAP, but with SYG and RTC)

RAND
CORPORATION