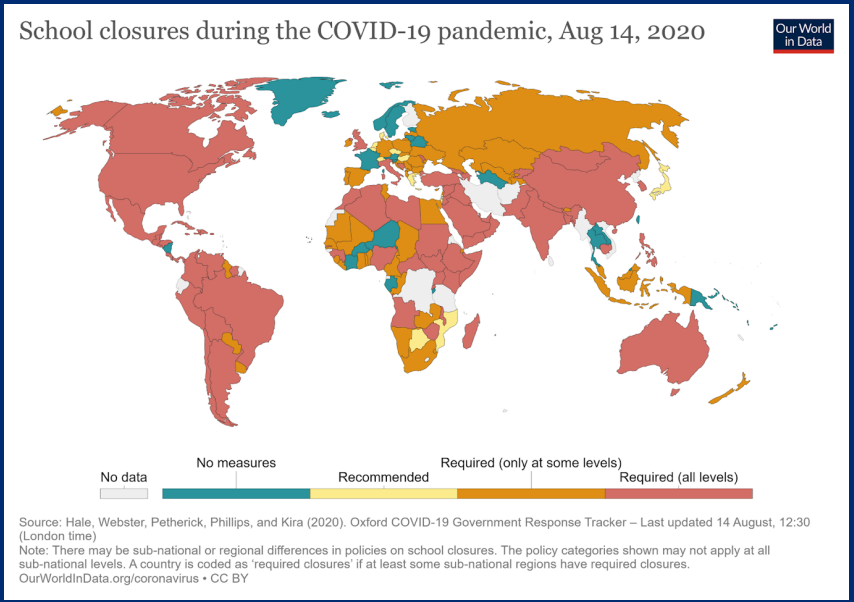
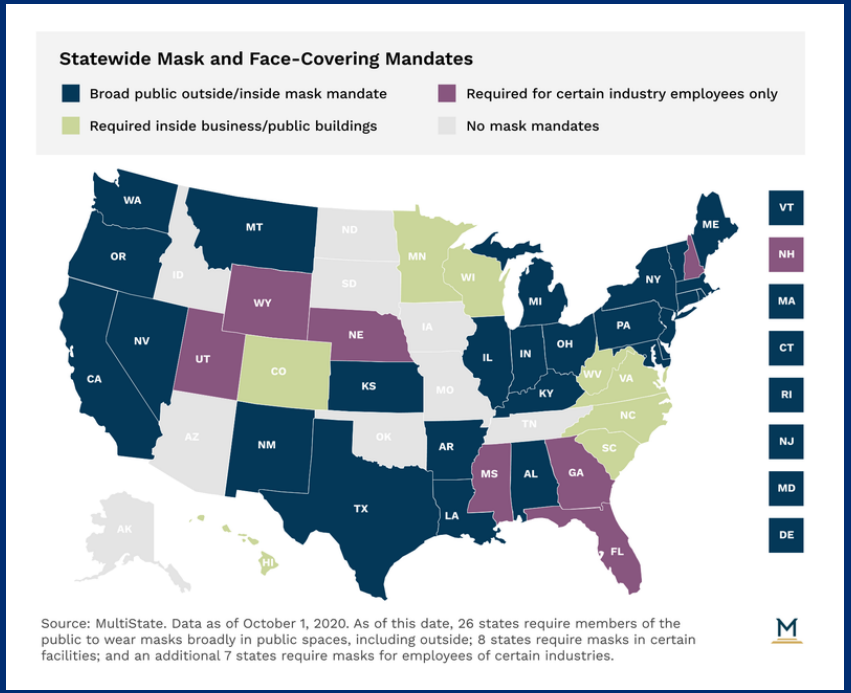
The role of design in policy evaluation: Applications to COVID-related and opioid policies

Elizabeth A. Stuart, PhD
Bloomberg Professor of American Health
www.elizabethstuart.org
@lizstuartdc





How can we learn about the effects of programs and policies?

- The country, and world, is full of variation in local policies being used to address the COVID-19 pandemic
- Could create an opportunity to learn about the effects of those policies, to inform future decision-making
- We have data on policies, outcomes, etc….what could be the problem?

Causal inference is hard

- Need to be able to compare potential outcomes for a well defined population:
    - Y(1): Outcome if treated (exposed)
    - Y(0): Outcome if control (not exposed)

- e.g., Difference in infection rates if a community has a mask mandate vs. does not have a mask mandate

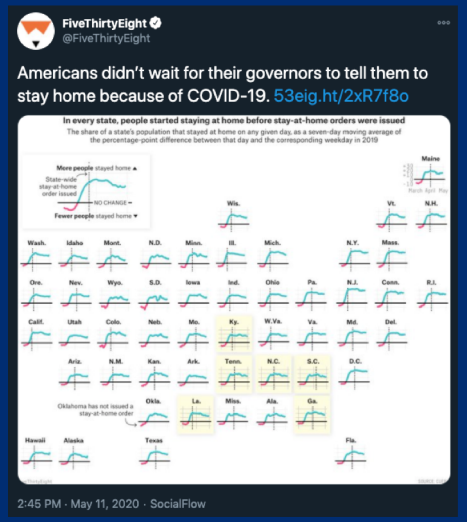
- The “fundamental problem of causal inference” is that we only see one of these potential outcomes for each unit (community)

Causal inference for policy evaluation is really hard

- Can’t randomize to exposure conditions
- Often relatively few units (states, countries)
- Implementation hard to measure (does the policy mean the same thing everywhere?)
- Hard to tease out effects from other things happening, including multiple policy responses

Causal inference for policy evaluation during COVID-19 is *really really* hard

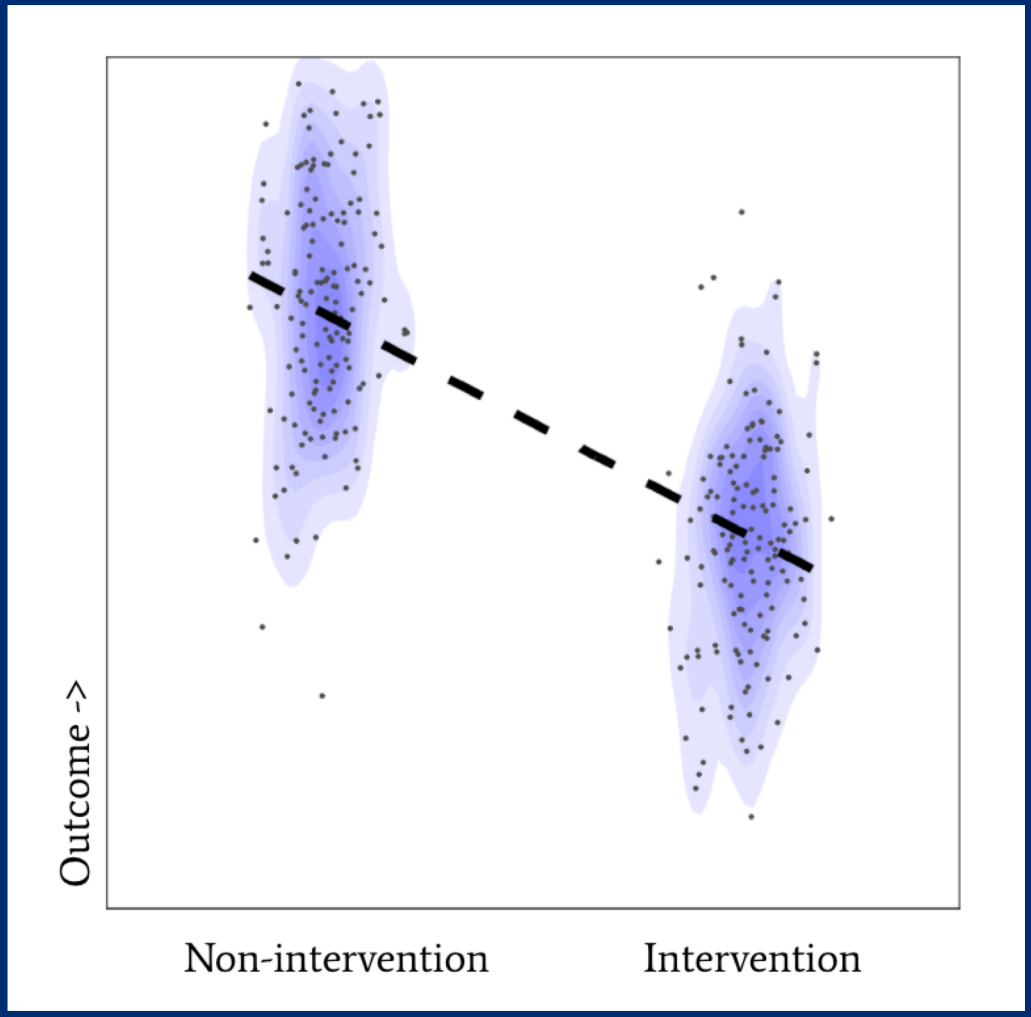
- Infectious diseases spread exponentially and have incubation periods
  - Small differences in model assumptions can have dramatic effects on results
- LOTS of policies and programs being put in place
  - Masks, schools, workplaces, stay at home, rapidly evolving treatments
- Anticipatory actions
  - e.g., staying at home before official orders to do so
- Data challenges
  - e.g., changing test availability and use
- Interactions across communities matter a lot
  - e.g., Sturgis rally



So what can we do?

What does a good (or bad) policy evaluation look like?

Cross sectional two group comparison



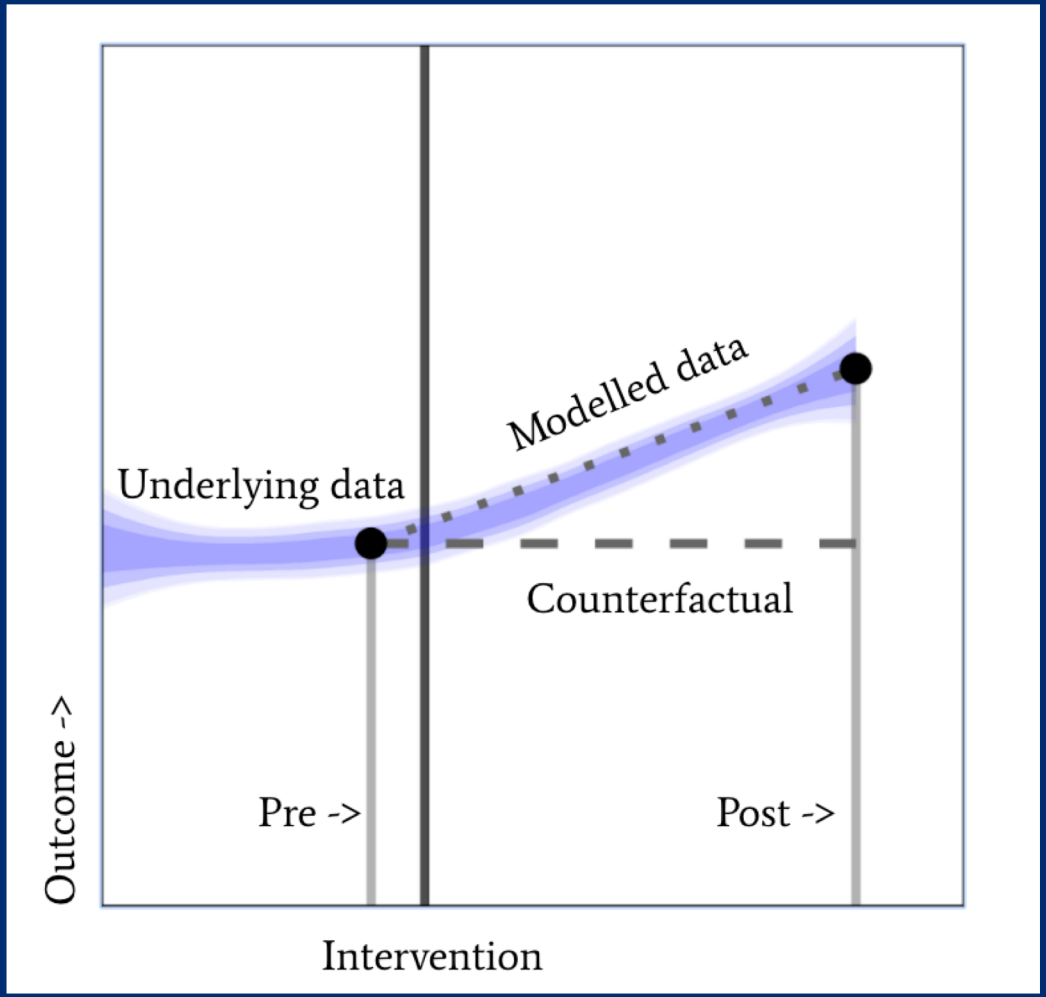
Compares outcomes between exposed and unexposed groups

Assumed counterfactual: Non-intervention location provides estimate of Y(0) for intervention location

Highly susceptible to confounding: locations that implement a policy likely quite different from those that don't

And subtle differences (e.g., in R0) may make a big difference in policy impact estimates, esp. given exponential growth

Simple pre/post

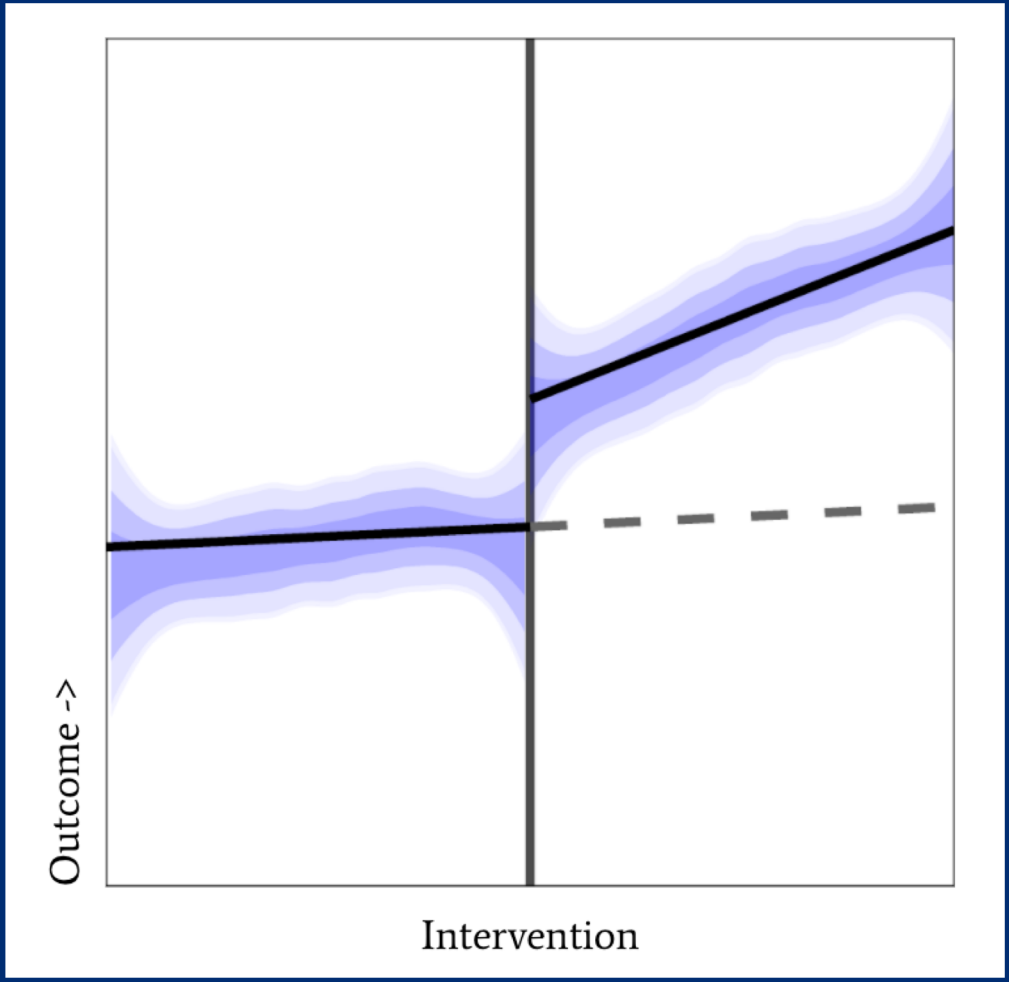


Compares outcome levels at two time points

Assumed counterfactual:
All change over time is due to the intervention

VERY questionable in infectious diseases (and many other areas without stable outcomes in the pre period)

Interrupted time series

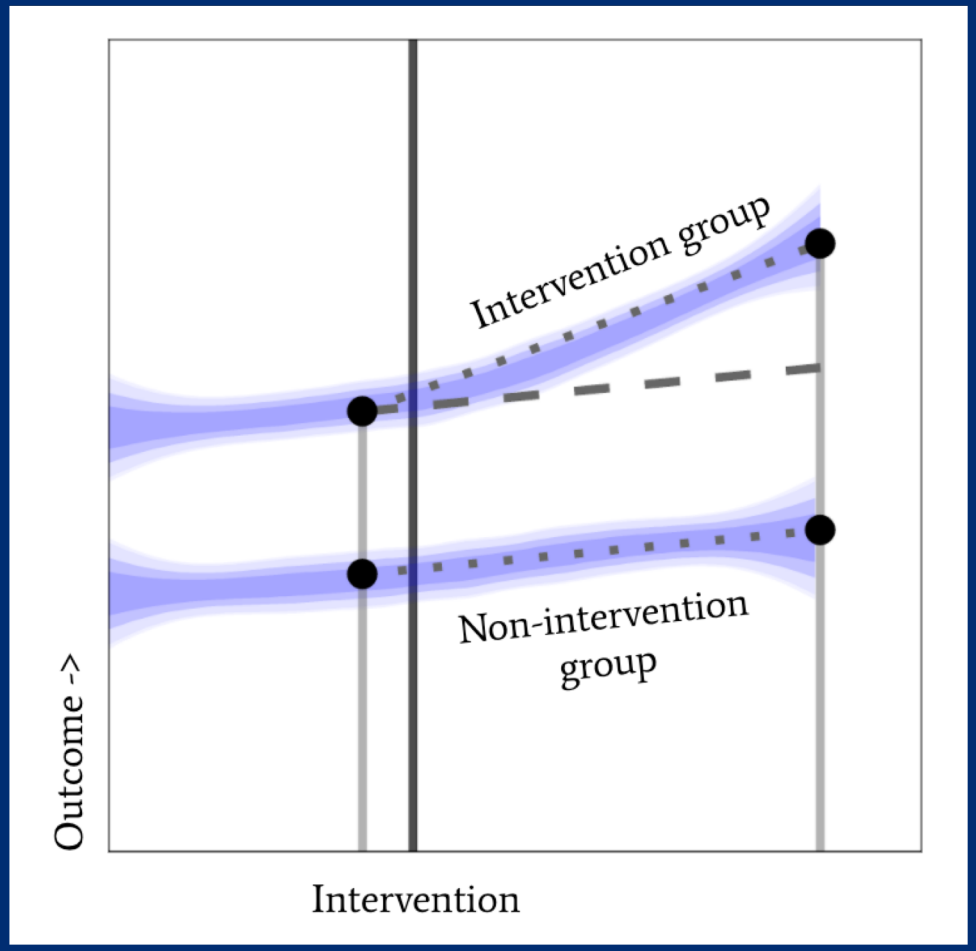


Models outcome in “pre” period and projects that out into the post period

Assumed counterfactual: outcome would have continued on the same (modeled) trajectory, if not for the intervention

Possibly better than pre/post, but relies on ability of the model to project accurately into the future

Comparative interrupted time series/Difference-in-differences



Models trends over time and across exposed and unexposed groups

Assumed counterfactual: Outcome would have changed in the same way pre to post as it did in the comparison group, if not for the intervention

Relies on a “parallel counterfactual trends” assumption

Note: Parallel pre-trends “make us feel better” but do not directly assess the key assumptions

Lots of nuances in CITS/DiD

- How to select comparison locations
- How to take advantage of individual level data
- Statistical power concerns (RAND, 2018)
- Challenges with staggered implementation across locations

- BTW rapidly growing methods area, and lots of variations on this; also known as event study designs, group panel data, segmented regression, …

Case Study:
State Opioid Policies

My goal today…

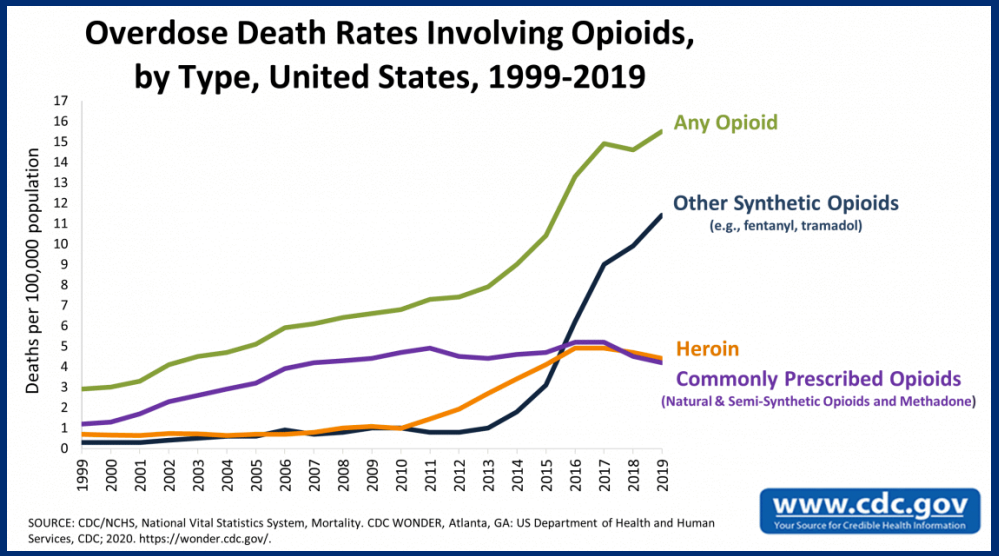
Not planning to talk about the specifics of statistical methods

Focus more on the applied questions and the challenges that come up in practice

How can we use basic design elements to help increase the rigor and relevance of studies?

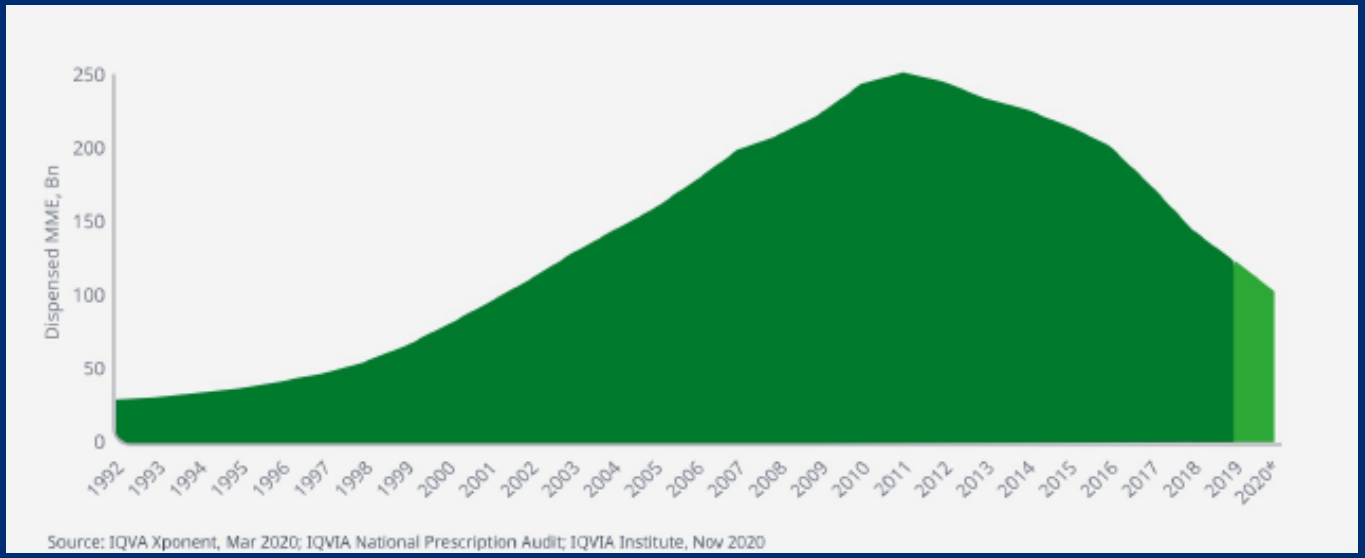
Opioid overdoses in the US

- High volume of opioid prescribing a key driver of the ongoing US opioid crisis
- State opioid prescribing laws implemented to try to curb inappropriate prescribing
  - Mandatory PDMP enrollment laws
  - Mandatory PDMP query laws
  - Pill mill laws
  - Opioid prescribing cap laws



Opioid prescribing decreasing

- Due to state policies?
- Could there be unintended consequences of the laws, e.g., for individuals with chronic pain?
- Prior research limited by methodological limitations – biases from two-way fixed effects approaches, policy changes happening close in time, unclear comparison groups



Two case studies

- Both aiming to estimate the effects of state opioid prescribing policies
- Both use large-scale medical claims data, aggregated to state/time level (at least for now)
- Both have strong focus on study design
- One (studying prescribing cap laws) uses a stacked comparative interrupted time series/Callaway & Sant'Anna approach to deal with 3 cohorts of states implementing one specific law at different points in time
- The other estimates the effects of 13 law changes (4 types of laws) in 13 treated states, with careful attention to isolating one particular law change and defining the comparison states (and timing) for each, using an augmented synthetic control approach
    - Original idea was to then correlate those state-specific effects with implementation information at the state level, but that became not very interesting, for reasons you will see

Study 1: Scientific question

What are the effects of prescribing cap laws on treatment of non-cancer chronic pain?

- Clinical guidelines have concluded opioid risks often outweigh benefits for these conditions
- Expect reductions in opioid prescriptions and increases in non-opioid pharmacologic and non-pharmacologic treatments (steroid injections, physical therapy)

(Note: Related work also doing qualitative work in states to understand policy implementation)

How to approach this?

Individual-level health insurance claims data from OptumLabs® Data Warehouse

- De-identified retrospective administrative claims data, including medical and pharmacy claims and eligibility information

Outcomes

- Proportion with an opioid Rx in a given year
- Average # of opioid Rx per year

Adults with continuous enrollment in commercial insurance from 2013 to 2019

- Diagnosis of one of five chronic pain non-cancer conditions: low back pain, fibromyalgia, chronic headaches, arthritis, or neuropathic pain
- ~ 1.4 million individuals

Treatment and comparison states

3 cohorts of states that implemented opioid prescribing caps

- 2017 (12 states)
- 2018 (12 states)
- 2019 (9 states)

Will consider “never-treated” states as the comparison group; can also use “not yet treated”

Adjust for aggregate values of individual level covariates: sex, age, indicator of any mental illness, indicator of any substance use disorder, Elixhauser co-morbidity index

- All measured during the pre-period
- (This is in part why the continuous enrollment requirement is useful)

Traditional regression approach

- Standard approach is to fit a regression model using aggregate longitudinal ("panel") data
    - e.g., two-way fixed effects model
    - Basically:
        - Y ~ (Time + Post) + (Time + Post)*Treatment + [State_i + Year_t]
    - Lots of names: Difference-in-differences, comparative interrupted time series, etc.
    - What is actually quite common is for papers to not show the model used and just to use terms like diff-in-diff or event study, without clearly describing what was being fit and what the parameters that relate to effects even are

Common extensions

With multiple units and time points common to add:

- Time fixed effects
- Unit fixed effects
- Covariates (sometimes time-varying, sometimes not)
- Multilevel context if individual-level data available

Can potentially also add weighting or matching on top of this

BUT this adds complications

- See work by Andrew Goodman-Bacon and others about challenges with staggered implementation and interpretation of the overall effect: Can lead to the wrong sign! [Fundamentally about controlling for post-treatment data!]
- And Daw & Hatfield (2018) have shown that matching/weighting in the pre period can cause bias (regression to the mean) if not done appropriately!
- And have to be very careful to avoid conditioning on post-treatment variables (inc. in the fixed effects!)

HUGE and evolving literature!

What about a design-based approach?

Yes! Nested policy trial emulation ("stacked CITS")

- Main idea: Think of staggered implementation of policies as nested implementation of hypothetical "target trials"
- At each policy implementation date, equate states that implement the policy with those that don't yet have the policy on a set of baseline characteristics, inc. baseline measures of the outcome
- Use traditional non-experimental study methods for each target trial
    - Ensures clear temporal ordering of covariates –> exposure -> outcomes
    - Can use a variety of design tools to be thoughtful about who is being compared
    - And clear diagnostics on similarity ion the pre period

- Extends ideas of Don Rubin, Paul Rosenbaum, Miguel Hernan on replicating a randomized trial using non-experimental data
- Ben-Michael, Feller, and Stuart (2021, *Epidemiology*)

Analysis strategy

- Aggregate the individual level data to be at the state-year level
- Continuously enrolled cohort implies that this will be a consistent group over time; may need other strategies if individuals in sample changing

- Basically, fit standard regression approach for each cohort and the "never-treated" comparison states, with time anchored at that cohort's policy start date

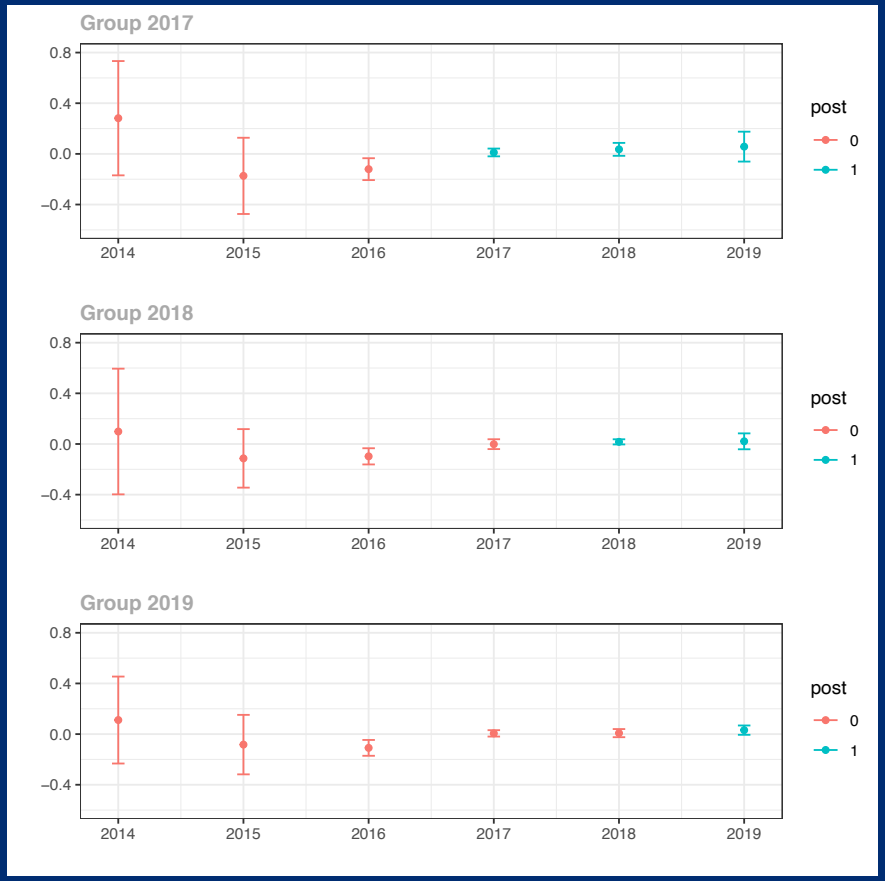
- Then "stack" (aggregate) the individual cohort effects to obtain an overall estimate

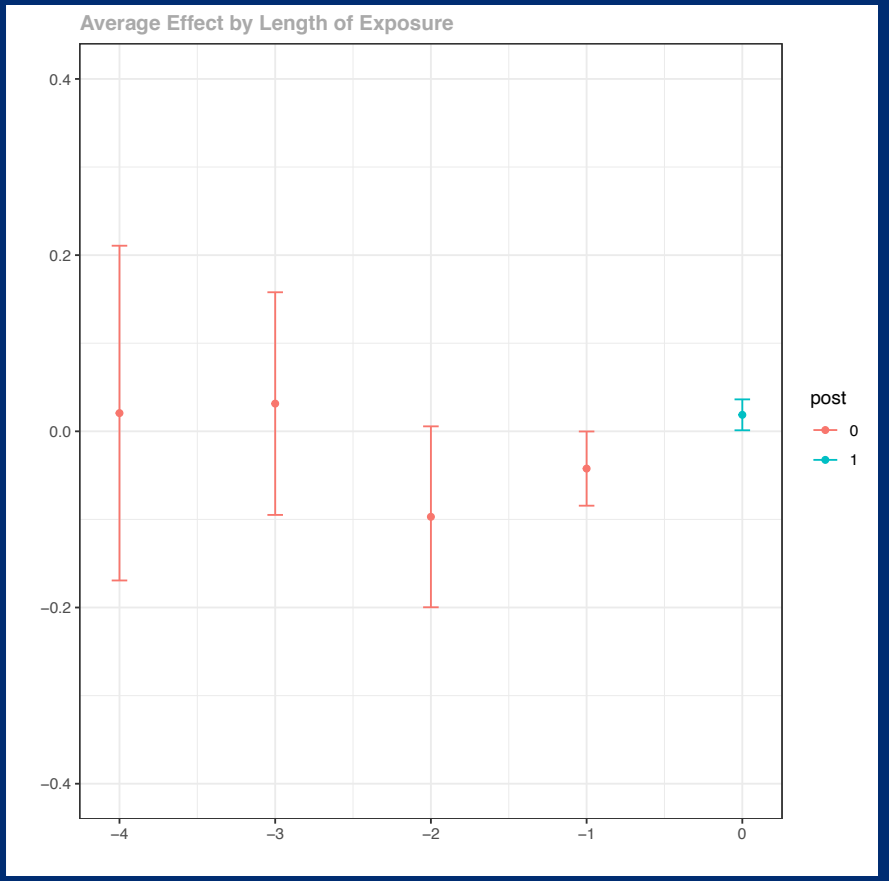
- Benefits: Get estimate for each treatment cohort and post-treatment year
- Can then aggregate as desired (or not)

- Clarity of the estimand, and the units used in the estimation

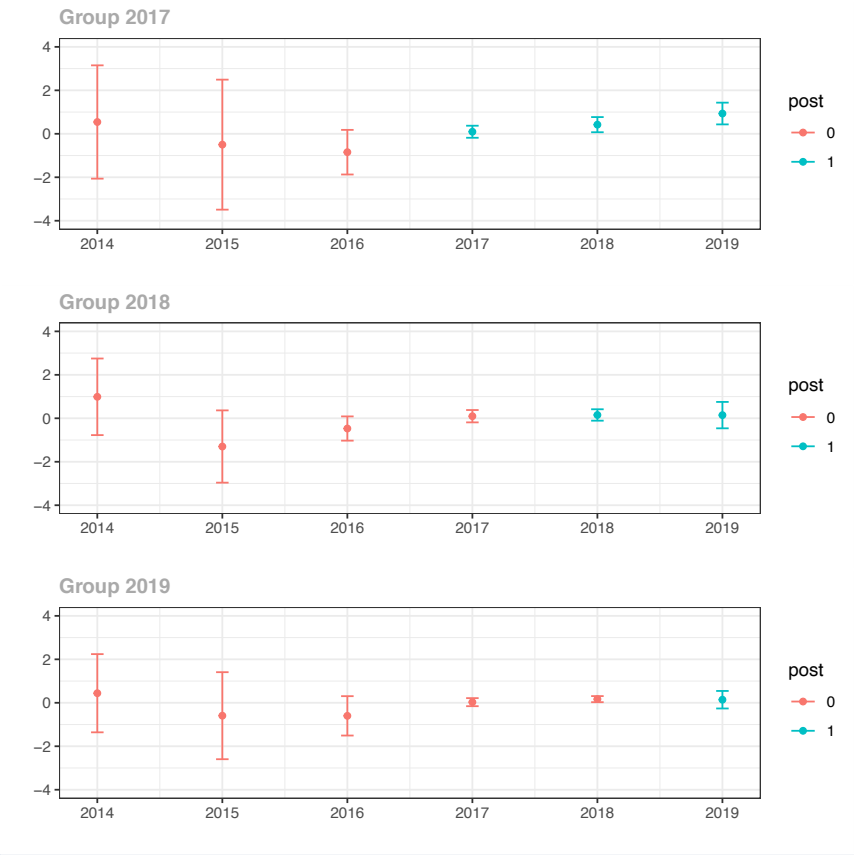
- Approach detailed in Callaway & Sant'Anna (2020); did package in R

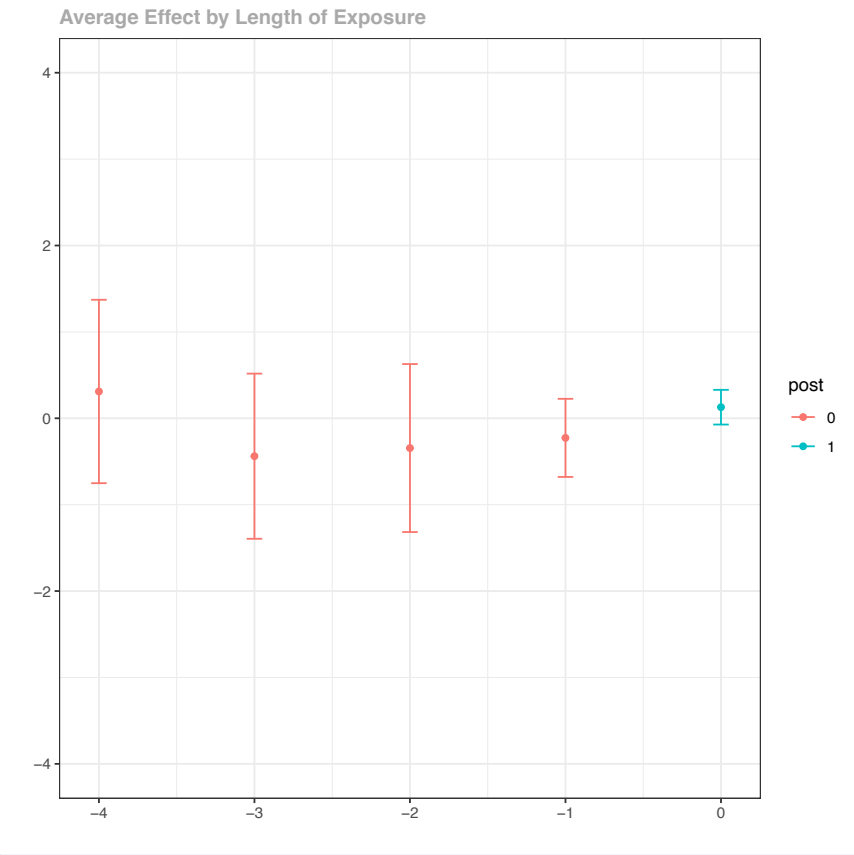
Any opioid Rx





of opioid Rx


Group 2017
Group 2018
Group 2019
post
0
1
2014
2015
2016
2017
2018
2019


Average Effect by Length of Exposure
post
0
1
−4
−3
−2
−1
0

Results

- These weight the cohort effects by the # of treated states in each cohort
- No big effects seen in opioid Rx for this chronic pain sample
- Consistent with other work looking at the effects of these policies on the general population
- Currently looking into subgroups defined by specific diagnoses, and by chronicity of chronic pain [this is one benefit of the individual-level data]

Study 2: Research questions

What are the effects of mandatory PDMP enrollment, mandatory PDMP query, pill mill, and opioid prescribing cap laws on patterns in receipt of opioid prescriptions among patients overall, and among a subset of patients with chronic non-cancer pain conditions?

How did law implementation contribute to those effects (or lack thereof?)? [Original plan….]

Chronic non-cancer pain conditions: low back pain, headache, fibromyalgia, arthritis, neuropathic pain

Methods: Quantitative analysis of claims data; qualitative interviews of individuals involved in the laws' implementation in each of the 13 treatment states

McGinty et al. (2022). Effects of state opioid prescribing laws on use of opioid and other pain treatments among commercially insured U.S. adults. *Annals of Internal Medicine*.

Data

IBM MarketScan commercial claims data – 350 commercial payers, approximately 25% of individuals with commercial insurance and their families in the U.S. 2013-2019.

Sample

Continuously enrolled adults aged 18+ overall and a sub-sample diagnosed with arthritis, low back pain, headache, fibromyalgia, or neuropathic pain in the pre-law period (two outpatient claims or one inpatient discharge diagnosis)

People with cancer diagnoses were excluded

Outcomes

Opioid prescribing measures, per state-month (% receiving a prescription, days' supply, prescription length, etc.)

Methods

Augmented Synthetic Control Approach

Study designed to address the problem of inability to disentangle effects of state laws implemented at or around the same time.

Treatment states: States that implemented one of the four laws of interest, and no other laws of interest or potentially confounding laws, in a four-year period: 2 years pre-, 2 years post-law (each Tx state has its own 4-year study period).

Control pool states: States that implemented no laws of interest or potentially confounding laws during a treatment state's 4-year study period AND had the exact same underlying opioid prescribing law environment as the treatment state, minus the law of interest in the treatment state, for the entire 4-year period (each Tx state has its own control pool).

Potentially confounding laws: Voluntary PDMP, doctor-shopping, physical exam, and pharmacy ID laws

State Law	Law Date	Study Period	Comparison States[1]
Opioid Prescribing Cap Law			
Delaware	4/1/17	4/1/15-3/31/19	AL, IA, KS, MT, MS, ND, NM, OR, TN, WY
Kentucky	7/1/17	7/1/15-6/31/19	AL, IA, KS, MS, MT, ND, NM, OR, WY
New York	7/22/16	8/1/14-7/31/18	AL, IA, KS, MS, MT, ND, OR, WY
Ohio	8/31/17	9/1/15-8/31/19	AL, IA, KS, MS, MT, ND, NM, OR, WY
Pill Mill Law			
Mississippi	3/1/11	3/1/09-2/28/13	AL, AZ, CO, IA, ID, IL, IN, LA, MI, MO, NC, NV, NY, ND, OK, PA, RI, SC, VA, WY
Ohio	7/1/11	7/1/09-6/30/13	AL, AZ, CO, ID, IN, IA, IL, LA, MA, MI, MO, NC, NV, NY, ND, OK, PA, RI, SC, VA, WY
Texas	9/1/10	9/1/08-8/31/12	AL, AZ, CO, CT, ID, IL, IN, LA, MA, MI, MO, NC, NV, NY, OK, PA, RI, SC, TN, VA, WV, WY
Mandatory PDMP Query Law			
New York	8/27/13	9/1/11-8/31/15	AK, AZ, CA, CO, IA, FL, LA, KS, MO, MI, MN, NC, ND, OR, SD, UT, WA, WY
Oklahoma	11/1/15	11/1/13-10/31/17	FL, GA, IA, KS, KY, LA, MI, MO, MS, MT, ND, NE, NM, OR, SD, TN, WV, WY
Pennsylvania	6/30/15	7/1/13-6/30/17	FL, GA, IA, KS, KY, LA, MI, MO, MS, MT, ND, NE, NM, OR, SD, TN, WV, WY
Virginia	7/1/15	7/1/13-6/30/17	FL, GA, IA, KS, KY, MI, MO, MS, MT, ND, NE, NM, OR, SD, TN, WV, WY
Mandatory PDMP Enrollment Law			
Colorado	1/1/15	1/1/13-12/31/16	AK, AZ, FL, IA, KS, KY, LA, MI, MO, MS, MT, NC, ND, NE, NM, OR, SC, SD, TN, UT, WA, WY
Idaho	7/1/14	7/1/12-6/30/16	AK, CA, AZ, DE, FL, IA, KS, KY, LA, MI, MN, MO, MT, NC, ND, NE, OR, SC, SD, UT, WA, WV, WY

An aside on synthetic control methods

- Method became more popular in past 10 or so years
- Basically, weight the control states to look like the policy state in the pre-policy time period

- Ben-Michael, Feller, and Rothstein showed that standard synthetic controls is not ideal
    - Like a weird type of propensity score weighting for one treated unit
    - No easy inferences (permutation test based)

- Generalized to augmented synthetic controls, which adds in a regularized regression model
    - Better performance
    - More straightforward inferences
    - Like a doubly robust version of synthetic controls

In this application…

Compare changes in outcome measures pre/post law in Tx states to changes in outcomes in a weighted group of comparison states, or "synthetic control"

Vector of state-specific weights that minimizes the mean squared prediction error between pre-law trends in the outcome of interest and covariates in the treatment and control pool states

- Covariates:
    - Individual: sex, age, co-morbid mental health diagnoses, substance use diagnoses, Elixhauser co-morbidity index
    - State: % Black, % Hispanic, % employed, % below FPL, % with no post high-school degree

Augmented with a ridge regression outcome model including the same covariates above + state fixed-effects

Single state analyses, state-month is unit of analysis

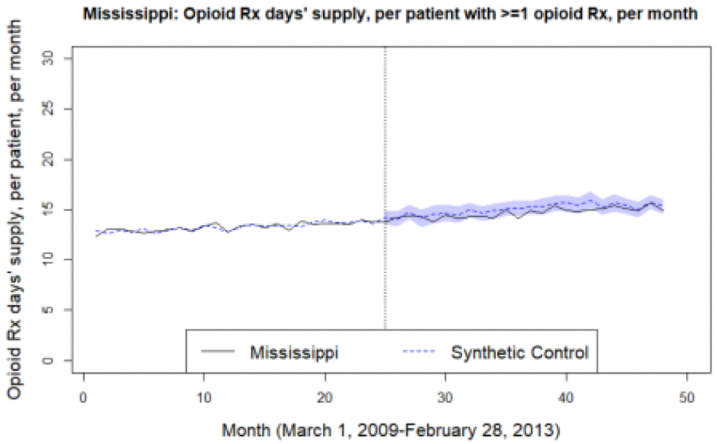
The key assumption in DiD

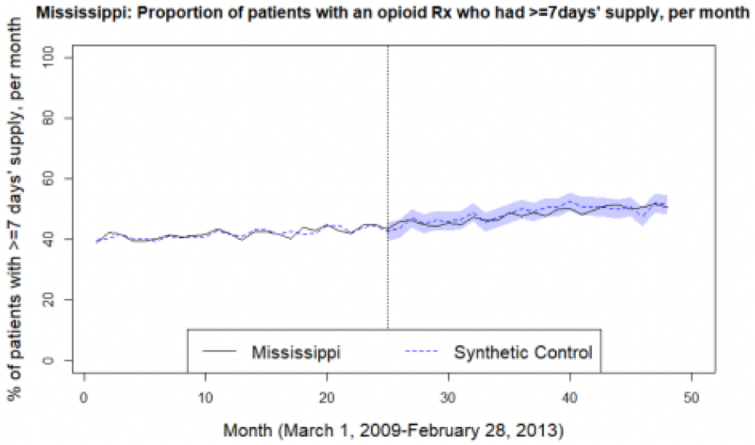
"Parallel counterfactual trends": "We assume that the change in outcomes from pre- to post-intervention in the control group is a good proxy for the *counterfactual* change in untreated potential outcomes in the treated group" (Hatfield website)

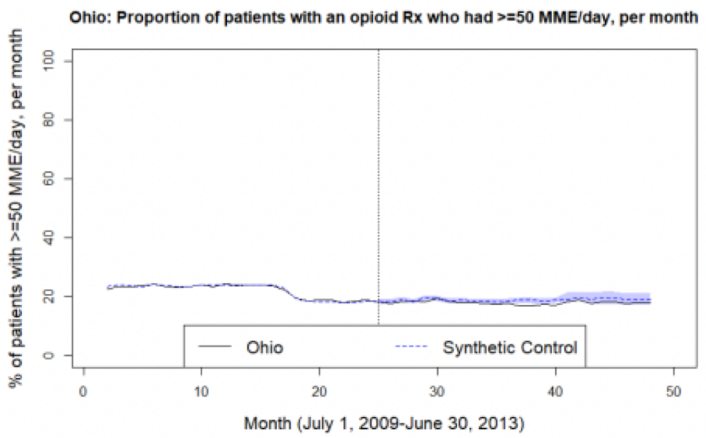
- Not directly testable because it involves counterfactual outcomes!
- (The pre-treatment trends analog is testable, although often low power and arguably equivalence testing better than traditional hypothesis testing)

We feel better about this if the trends in intervention and comparison sites are similar in the pre period

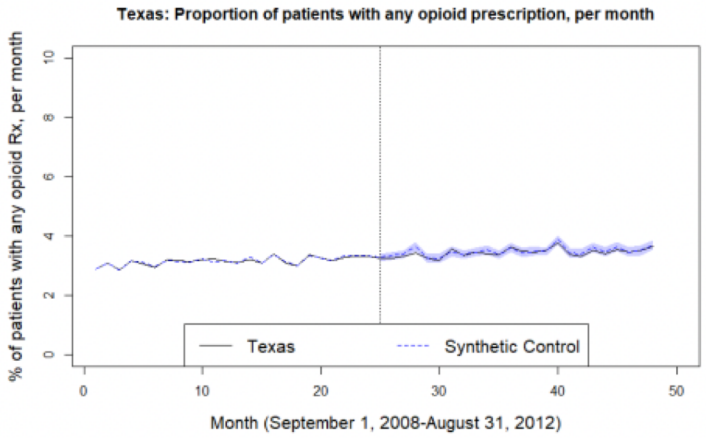
- This is what motivates the (augmented) synthetic control approach
- But this is no guarantee of the actual underlying assumption!
  - "The quality of our match historically is what makes us comfortable with extrapolation" (Luke Miratrix, Harvard)





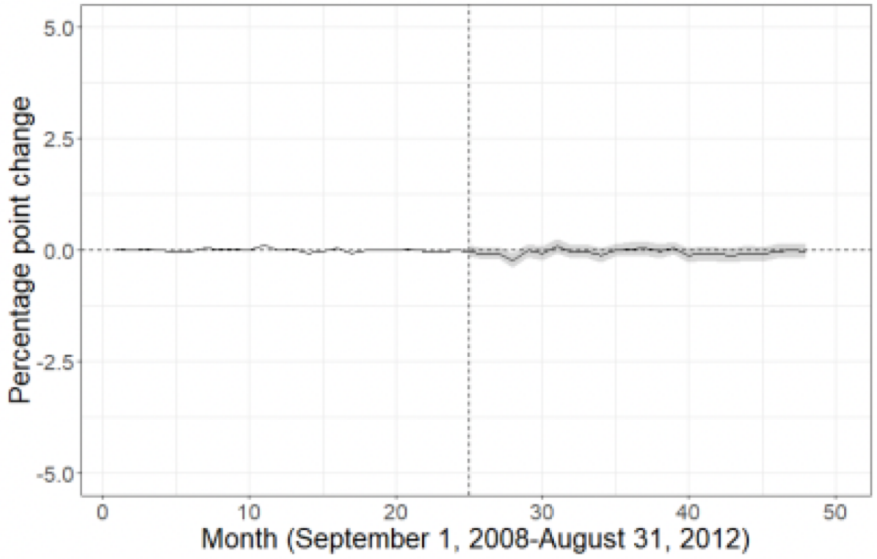


Texas Pill Mill Law, Overall Adult Sample (N=2,998,858)

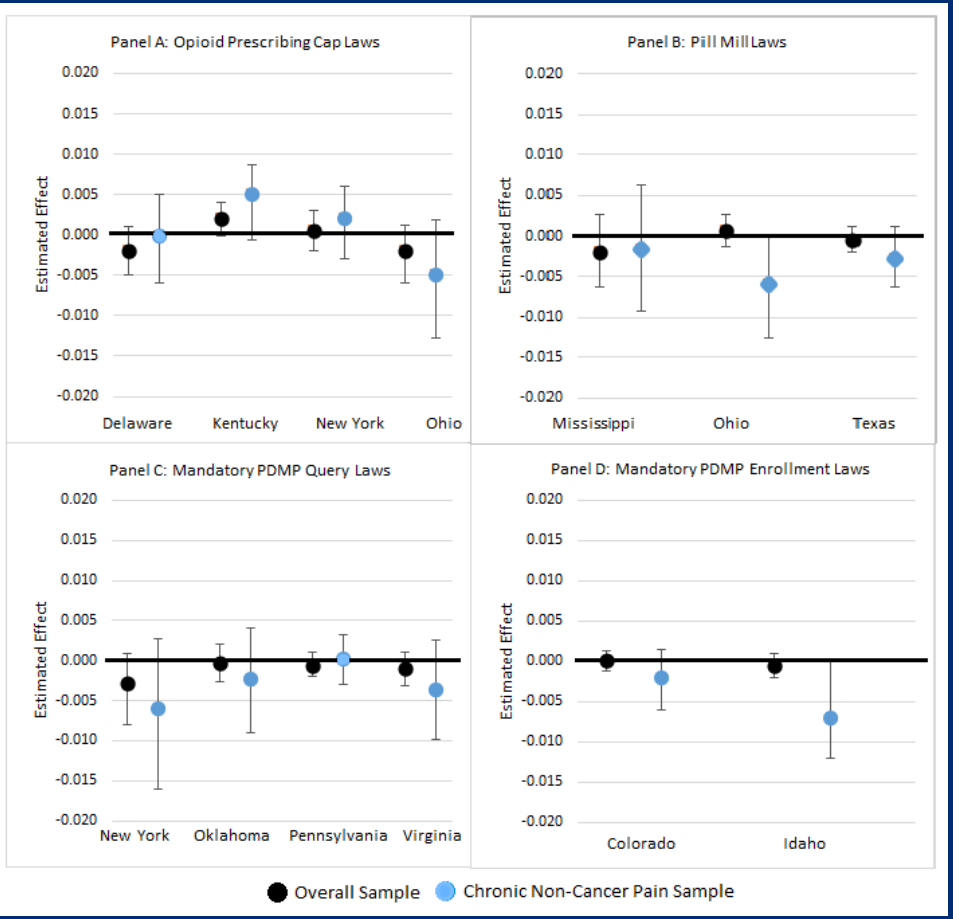


Augmented synth diagnostics

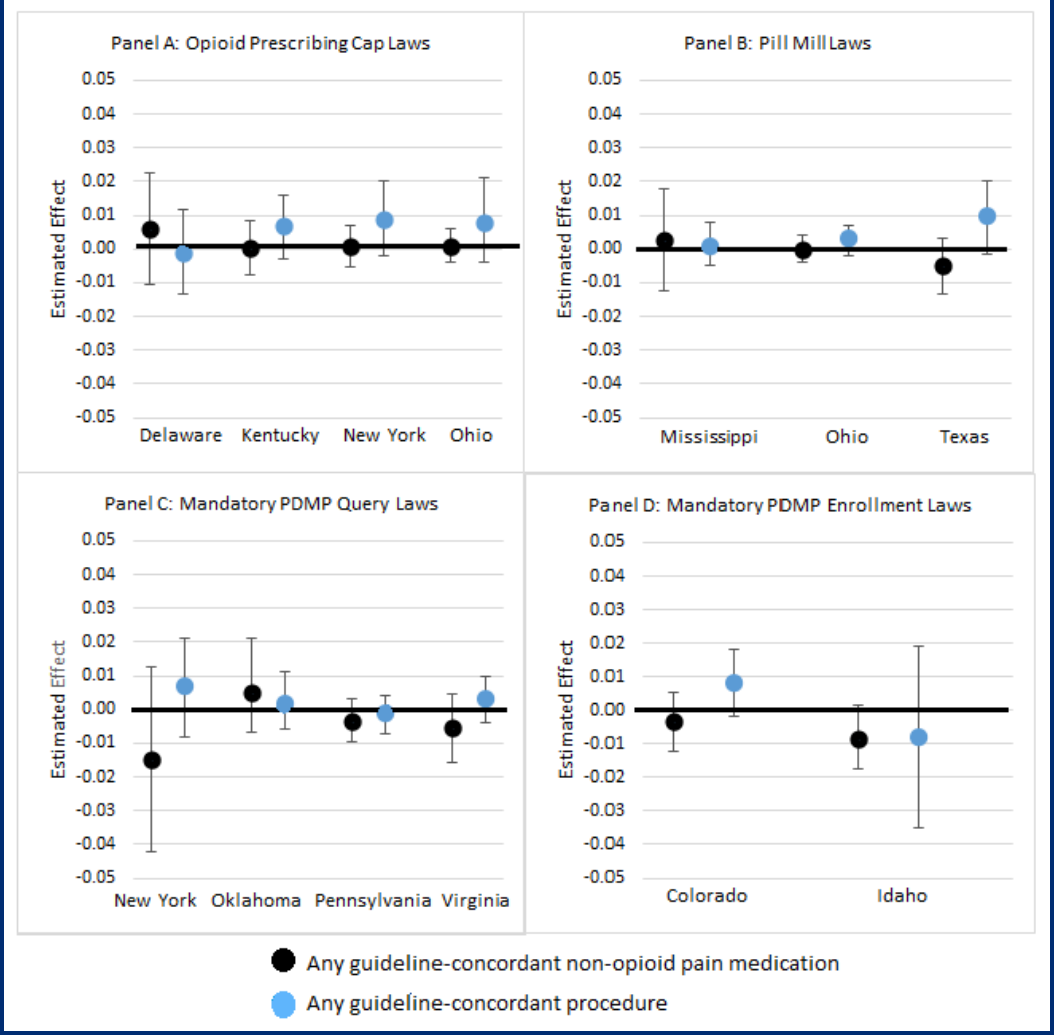
Texas Pill Mill Law, Overall Adult Sample: Change in the proportion of patients receiving any opioid Rx, per month, attributable to the law



Effects on P(receiving an opioid prescription)



Effects on monthly prob. of receiving guideline concordant non-opioid Tx among people with chronic non-cancer pain conditions



Summary of results

“For adults overall and those with chronic noncancer pain, the 13 state laws were each associated with a change of less than 1 percentage point in the proportion of patients receiving any opioid prescription and a change of less than 2 percentage points in the proportion receiving any guideline-concordant nonopioid treatment, per month. The laws were associated with a change of less than 1 in days' supply of opioid prescriptions and a change of less than 4 in average monthly MME per day per patient prescribed opioids.”

Sensitivity analyses

- Standard difference-in-differences
- Analyses examining whether laws' effects changed over time (e.g., ramp up due to implementation)
- Stratified analysis by chronic pain condition
- Analyses limited to people who used prescription opioids in the pre-law period
- Analyses excluding states that changed their cannabis laws during the study period

(140 page supplement to the paper!)

Substantive conclusions

- Results suggest that secular trends related to changing standards in pain medicine may be driving declines in opioid prescribing, as opposed to state laws

- Findings do not support the narrative that state opioid prescribing laws have significantly reduced dose or duration of opioid prescriptions among patients with chronic non-cancer pain

- While there was some variation in key in key implementation/ enforcement domains across states, this did not correlate with variation in laws' effects on outcomes

- Null findings may be driven by exemptions in state opioid prescribing laws and/or implementation and enforcement challenges, which are well-documented in qualitative research, including in the qualitative component of this study.

Methods conclusions

- Important to take a design-based approach for policy evaluation, just like in other fields
- Clear temporal ordering
- Avoid concurrent treatments/policies
- Use methods to help make the parallel counterfactual trends assumption more believable
- But also still lots of open research questions!

Discussion

Additional thoughts…

- Important to be thoughtful and careful with policy evaluation
- Potentially highly impactful

- Policy trial emulation/stacked CITS allows careful thought of the comparisons being made, and care regarding pre and post time periods, confounding, etc.
    - Transparent comparisons and diagnostics

- Recommend avoiding models that simply fit regressions to the longitudinal data, with fixed effects for state and time
    - No clear "design," unclear comparisons and diagnostics, potential bias
    - "Design" the policy evaluation by thinking about the target trial that you would implement if possible

- Still open research questions
    - How to take advantage of individual-level data
    - Advantages of different ways of dealing with covariates
    - How well can we estimate state-specific effects?

What does policy evaluation look like now?

- Most research has a LONG way to go
- Not uncommon for published evaluations (in COVID, gun policy, opioid policy) to be a simple cross-sectional two group comparison of treated and untreated sites
- COVID: Only 4/36 studies met even a relatively low bar for temporality, attention to time trends, display of outcomes over time (Haber et al., 2021)
- Opioids: "…only 29 (20 % of studies) met each of three key criteria for rigorous design: analysis of longitudinal data with a comparison group design, adjustment for difference between policy-enacting and comparison states, and adjustment for potentially confounding co-occurring policies." (Schuler et al., 2020)

How do we balance these challenges with the need to generate answers to important policy questions?

- “Be clear about what is knowable” – Goodman-Bacon and Marcus (2020)
- Acknowledge the challenges
- Conduct diagnostics and sensitivity analyses
- Choose methods that allow transparency and a design orientation
- Collaborate across fields
- Build a body of evidence: don’t just rely on one study
- Point out problems with very basic policy evaluations that do not yield rigorous results


BAD
STUDY DESIGN
REALLY
FANCY STATISTICS
imgflip.com

Acknowledgements

Augmented synthetic controls and stacked CITS: Avi Feller (Berkeley), Eli Ben-Michael (Harvard)

JHSPH colleagues: Beth McGinty, Alex McCourt, Kayla Tormohlen, Elizabeth Stone, Ian Schmid

Network for Public Health Law: Corey Davis

Funding:
Arnold Ventures
Bloomberg American Health Initiative
National Institutes of Health funded RAND Opioid Policy Tools and Information Center
National Institute of Drug Abuse

To learn more…

Ben-Michael, Feller, and Stuart (2021). A trial emulation approach for policy evaluations with group-level longitudinal data. *Epidemiology.* https://arxiv.org/abs/2011.05826

French and Stuart (2020). Study designs and statistical methods for studies of child and adolescent health policies. *JAMA Pediatrics.*

Goodman-Bacon and Marcus (2020): Using difference-in-differences to identify causal effects of COVID-19 policies. https://cdn.vanderbilt.edu/vu-my/wp-content/uploads/sites/2318/2020/05/11154933/Covid-DD_v2.pdf

Haber, Clarke-Deelder, Salomon, Feller, and Stuart (2020): Policy evaluation in COVID-10: A graphical guide to common design issues. https://arxiv.org/abs/2009.01940

Haber et al. (2021). Problems with Evidence Assessment in COVID-19 Health Policy Impact Evaluation (PEACHPIE): A systematic strength of methods review. https://www.medrxiv.org/content/10.1101/2021.01.21.21250243v1.full

https://diff.healthpolicydatascience.org/

https://coronavirus.jhu.edu/from-our-experts/evaluating-the-effectiveness-of-covid-19-policies-a-q-and-a-with-dr-elizabeth-stuart

https://github.com/ebenmichael/policy-trial-emulation

https://causalinf.substack.com/p/callaway-and-santanna-dd-estimator