JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

The use of synthetic control and other covariate adjustment strategies for policy evaluation

Elizabeth Stuart, PhD
Professor, Associate Dean for Education
Johns Hopkins Bloomberg School of Public Health
estuart@jh.edu; @lizstuartdc; www.elizabethstuart.org

Acknowledgments

- HUGE thanks to **Avi Feller** (Berkeley) and **Luke Miratrix** (Harvard) for help with the slides
- Big thanks to other statisticians doing important work in this area, including **Beth Ann Griffin, Megan Schuler, and Laura Hatfield**
  - Check out Laura's amazing website: https://diff.healthpolicydatascience.org/
  - It makes sense of the RAPIDLY growing area
- Thanks to **Daniel Webster, Alex McCourt, Beth McGinty, Lainie Rutkow, Colleen Barry, Alene Kennedy-Hendricks, Sachini Bandara, Brendan Saloner, …** for involving me in interesting policy evaluations in gun violence and opioids

Warning:

This talk is going to raise more questions than it answers!

Goal:

Give an overview of thinking and issues when trying to estimate policy effects using aggregate or individual-level data

A monthly conversation

- Collaborator: Liz, we want to study the effects of XXX state policy [marijuana legalization, gun background checks, etc.]. Should we use synthetic controls?
- Liz: Sounds interesting! What does your data look like?
- Collaborator: Well we have individual level data, e.g., healthcare claims, for people in and outside the state
- Liz: Hmm….I have no idea what to do!
  - Synthetic control? Augmented synthetic control?
  - ”Traditional” difference in difference design/comparative interrupted time series?
  - “Trial emulation” approach?
  - Marginal structural model?
  - How do/can we use the individual level data?
  - How should we deal with confounding?
  - How should we code time?
  - Ack; my head hurts and I have no idea what to recommend!

AUTISM & MENTAL HEALTH PARITY

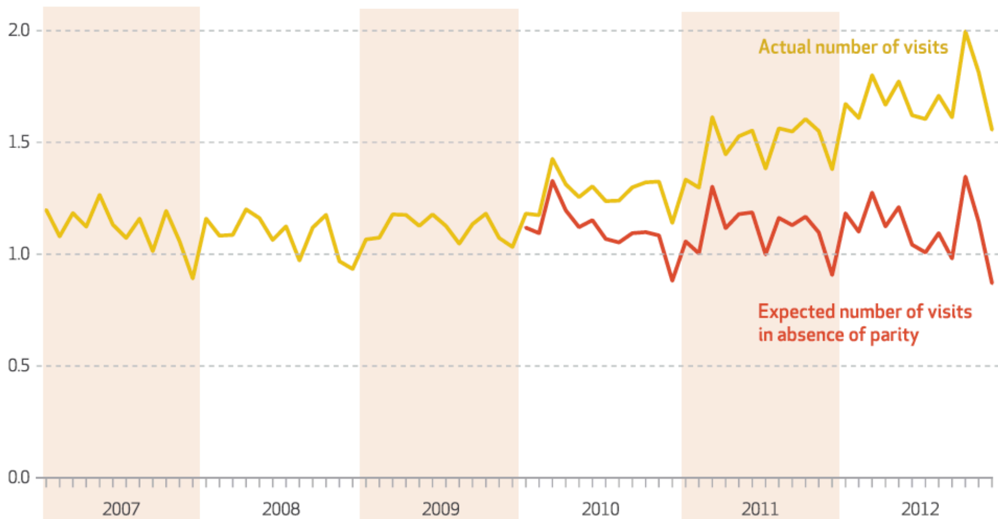
By Elizabeth A. Stuart, Emma E. McGinty, Luther Kalb, Haiden A. Huskamp, Susan H. Busch, Teresa B. Gibson, Howard Goldman, and Colleen L. Barry

Increased Service Use Among Children With Autism Spectrum Disorder Associated With Mental Health Parity Law

DOI: 10.1377/hlthaff.2016.0824
HEALTH AFFAIRS 36, NO. 2 (2017): 337–345

EXHIBIT 3

Associations between the federal parity law and outpatient mental health and functional therapy visits, among youth using those services



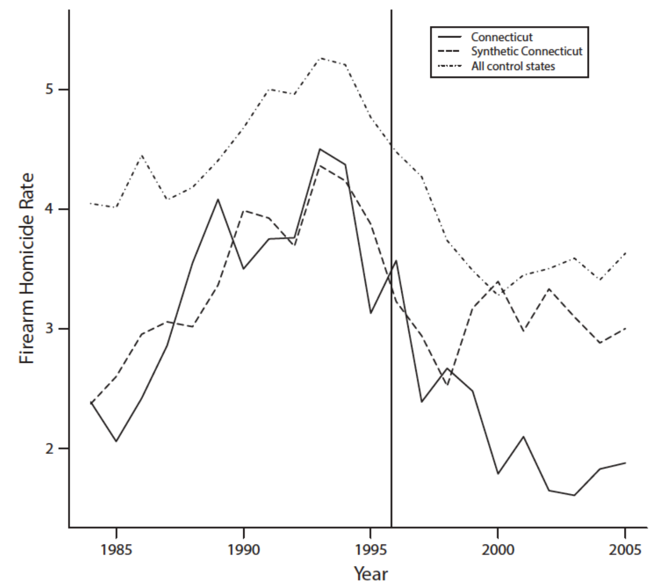
SOURCE Authors' analysis of data from the Truven Health MarketScan Research Database, 2007–12. NOTE Functional therapy visits include speech and language therapy and occupational and physical therapy.

Association Between Connecticut's Permit-to-Purchase Handgun Law and Homicides

Kara E. Rudolph PhD, MPH, MHS, **Elizabeth A. Stuart** PhD, **Jon S. Vernick** JD, and **Daniel W. Webster** ScD, MPH

[+] Author affiliations, information, and correspondence details

Accepted: April 04, 2015 Published Online: July 09, 2015

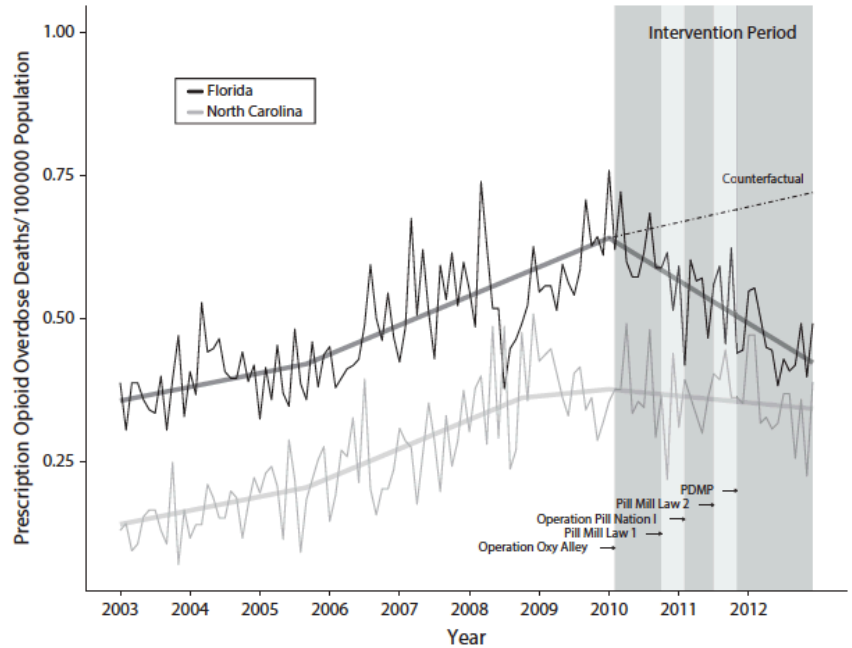


Note. Connecticut (solid line) compared with synthetic Connecticut (dashed line) and all states in the control pool, equally weighted (dotted dashed line). The vertical line indicates when Connecticut's permit-to-purchase law was implemented.

FIGURE 1—Firearm homicide rates: Connecticut, 1996–2005.

Opioid Overdose Deaths and Florida's Crackdown on Pill Mills

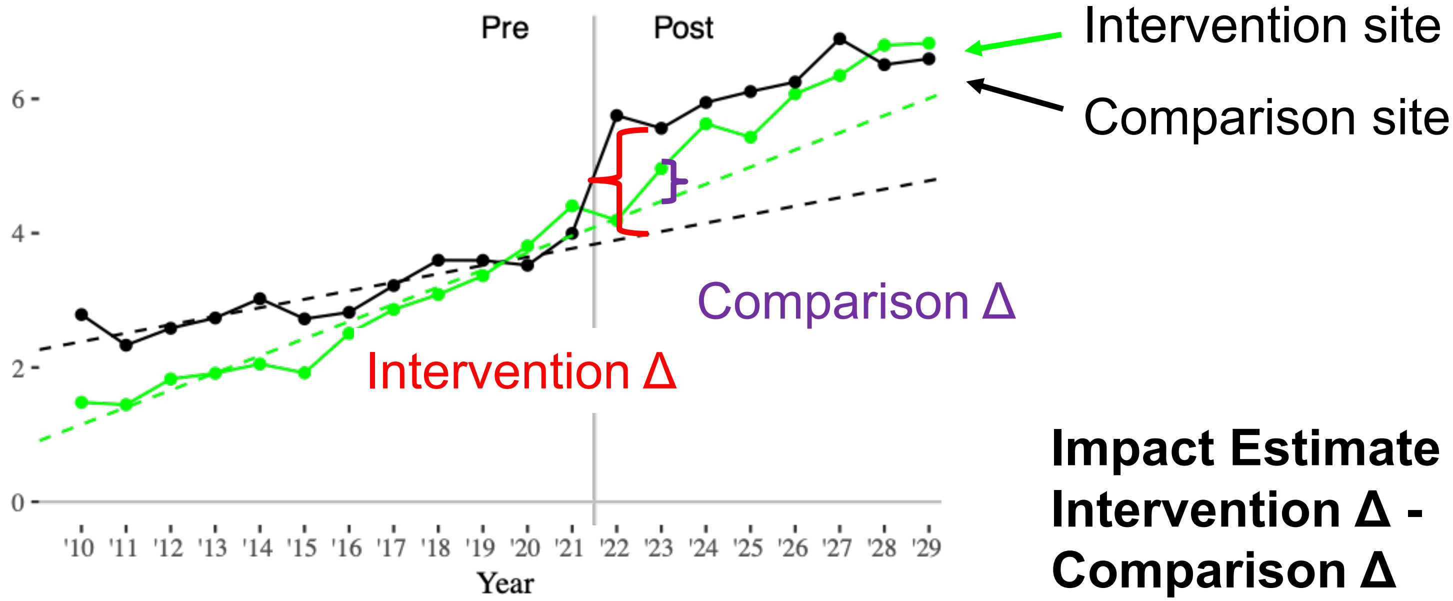
Alene Kennedy-Hendricks, PhD, Matthew Richey, PhD, Emma E. McGinty, PhD, MS, Elizabeth A. Stuart, PhD, Colleen L. Barry, PhD, MPP, and Daniel W. Webster, ScD, MPH



Note. PDMP = prescription drug monitoring program. The figure overlays the fitted multivariate adaptive regression spline (MARS) model on the observed prescription opioid overdose mortality rates in Florida and North Carolina. Under the counterfactual, we assumed that the rate of change in Florida's prescription opioid overdose mortality rates would have changed by the same amount that North Carolina's slope changed at the change point most proximate to the intervention period (January 2010). Note that the year labels indicate the start of each year (January) rather than the midpoint of each year.

FIGURE 1—Changes in Prescription Opioid Overdose Mortality Rates: Florida and North Carolina, 2003–2012

The general set up



Impact Estimate = Intervention Δ - Comparison Δ

Some common methods used in this context

Method	Typical data structure	Attempt to control confounding?
Synthetic controls and extensions	Aggregate One intervention site; multiple comparison sites	Yes, by trying to equate trends in pre-period on outcome and covariates of interest
Traditional comparative interrupted time series models, e.g., “two-way fixed effects”	Aggregate or individual-level	By including covariates in model, or some work on combining matching/weighting with these models Note: Augmented synthetic controls combines synth ideas and outcome regression model
AutoRegressive Integrated Moving Average (ARIMA) models	Few units, many time points	By including covariates in model
Marginal structural models	Individual level data, with staggered implementation dates (e.g., individuals going on treatment at different times)	Through propensity score type weights
“Trial emulation” idea	Individual level data, with staggered implementation dates	Through propensity score type weights

Note: See also Megan Schuler’s presentation!

The assumptions of nearly all of these approaches

- “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 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)

What are the general challenges?

- Almost never have randomization
  - Confounding may be strong
  - May be hard to believe any comparison units are appropriate (e.g., states)
  - Confounding may be based on observable and unobserved factors, including the pre-intervention measures of the outcome(s) of interest
- Sometimes have very few units, e.g., even just one intervention site!
  - Of course ideally also have comparison sites
- Sometimes have few time periods, e.g., 6 years of annual data in education settings!
- Roll-out of the intervention may not be sudden
- Effects of interest may change over time (e.g., immediate, gradual, etc.)
  - What is the estimand of interest? Effect at a particular point in time? Averaged over multiple times? Change in intercept? Change in slope?

Another Challenge: Terminology!

- Difference in differences
- Event study methods
- Synthetic control
- Augmented synthetic control
- Marginal structural model
- Interrupted time series
- Comparative interrupted time series
- Two-way fixed effects
- Panel data methods
- ….

Are these even all different things???

The basic regression formulation

$$Y_{it} = \beta_0 + \beta_1 \text{post}_t + \beta_2 \text{treat}_i + \beta_3 \left(\text{post}_t \times \text{treat}_i\right) + \epsilon_{it}$$

- Typically estimate DD via linear regression:
  - *Post:* dummy variable for treated ***time period***
  - *Treat:* dummy variable for treated ***group***

- Does this look familiar? Regression with interactions!

- Useful because we can extend in many ways:
  - Covariates
  - Fixed effects
  - Event studies

Source: Avi Feller (UC Berkeley)

Common extensions

- With multiple units and time points common to add:
  - Time fixed effects
  - Unit fixed effects
  - Covariates (sometimes time-varying, sometimes not)

- 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
  - And Daw & Hatfield (2018) have shown that matching/weighting in the pre period can cause bias if not done appropriately!
  - Nice summary of alternatives: https://andrewcbaker.netlify.com/2019/09/25/difference-in-differences-methodology/

Some of the considerations that often aren't talked about…

- Time-scale to use
  - More time points may make it easier (or harder?) to model time trends
- Best methods if data aggregated already (e.g., monthly statewide homicide counts) vs. individual level data available (e.g., claims data)?
- How do the "CITS" methods compare to things like marginal structural models?
- How to deal with confounding?
  - Challenge especially because some variables will be "post treatment" (this is a challenge with the standard fixed effects approaches)
- What if there are multiple intervention states/sites?
  - Run a synthetic control for each and then aggregate (e.g., meta-analysis)? This usually doesn't work well though!
  - Run a combined approach?

Okay but so what is this synthetic control thing?

1) Build a "synthetic comparison group" by taking a weighted average of other similar "donor" units so that our synthetic comparison group is "as much like" our "treatment" unit as possible before the policy change or event occurred.

and then

2) Use the observed outcome trajectory of our synthetic comparison group to represent the counterfactual outcome trajectory for our treatment unit.

Thanks to Luke Miratrix!

Core ideas

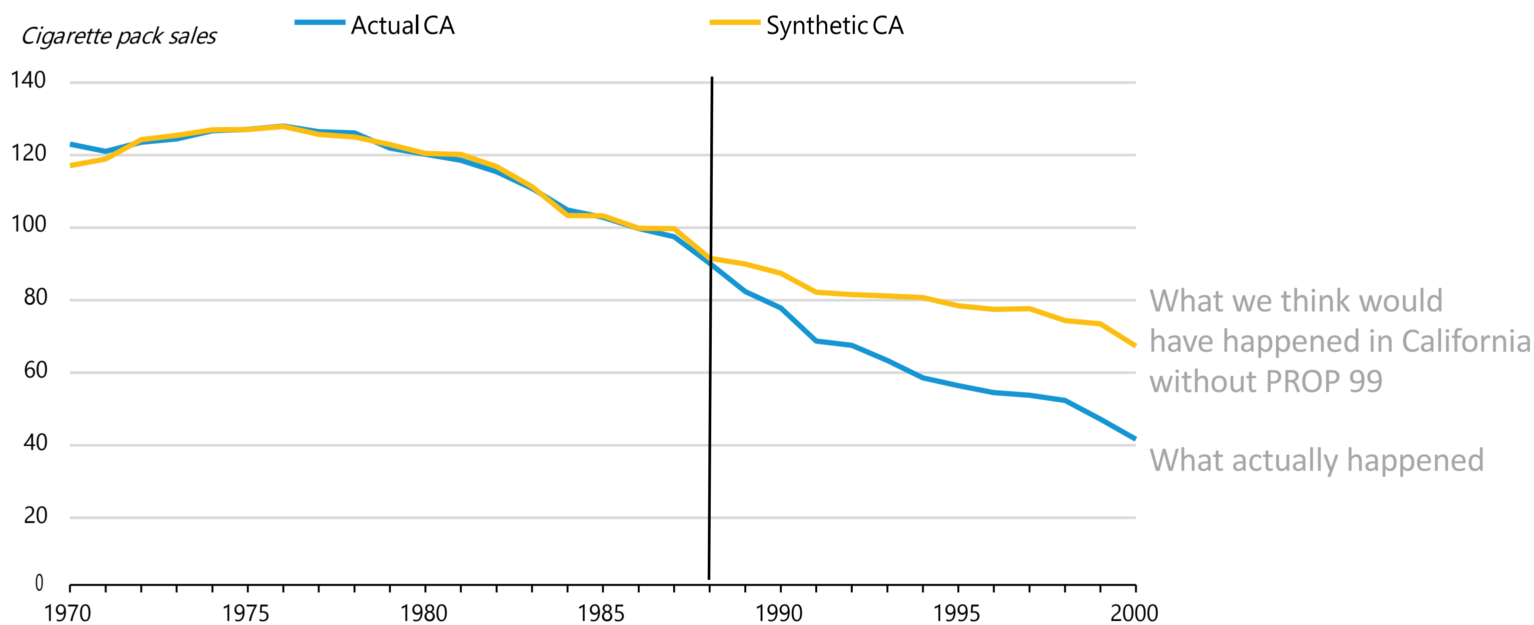
- Underlying structure:
  - A time-series of outcome measures plus some baseline covariates for
    - A *treatment unit* (e.g. a school, firm, city, county, state or country) and
    - A matched *synthetic comparison group* (a weighted average of other similar units)
  - A well-defined pre-intervention period and post-intervention period
  - A “donor pool” of eligible units for building the synthetic comparison group
- Guiding principles
  - A weighted composite of multiple matched comparison units is usually preferable to a single matched comparison unit.
  - Matching comparison units on past outcomes and baseline covariates that predict those outcomes is usually preferable to matching solely on past outcomes.
  - A data-driven process is usually the best (and most transparent) way to construct a matched synthetic comparison group
  - Works best when obtain EXCELLENT pre-treatment fit/balance

Thanks to Luke Miratrix and Avi Feller!

The role of weighting and outcome models

- Standard synthetic control essentially is trying to deal with the confounding by matching (weighting) in the pre period
- Other strategies tackle the problem using parametric outcome models (e.g., the two-way fixed effects approaches referenced earlier)
- In fact it's best to combine the two!
  - Similar to the long literature in single time point non-experimental studies and the benefits of double robust approaches

Figure 1: The Seminal Application: Annual Cigarette Sales for California and its Synthetic Comparison Group Before and After PROP 99

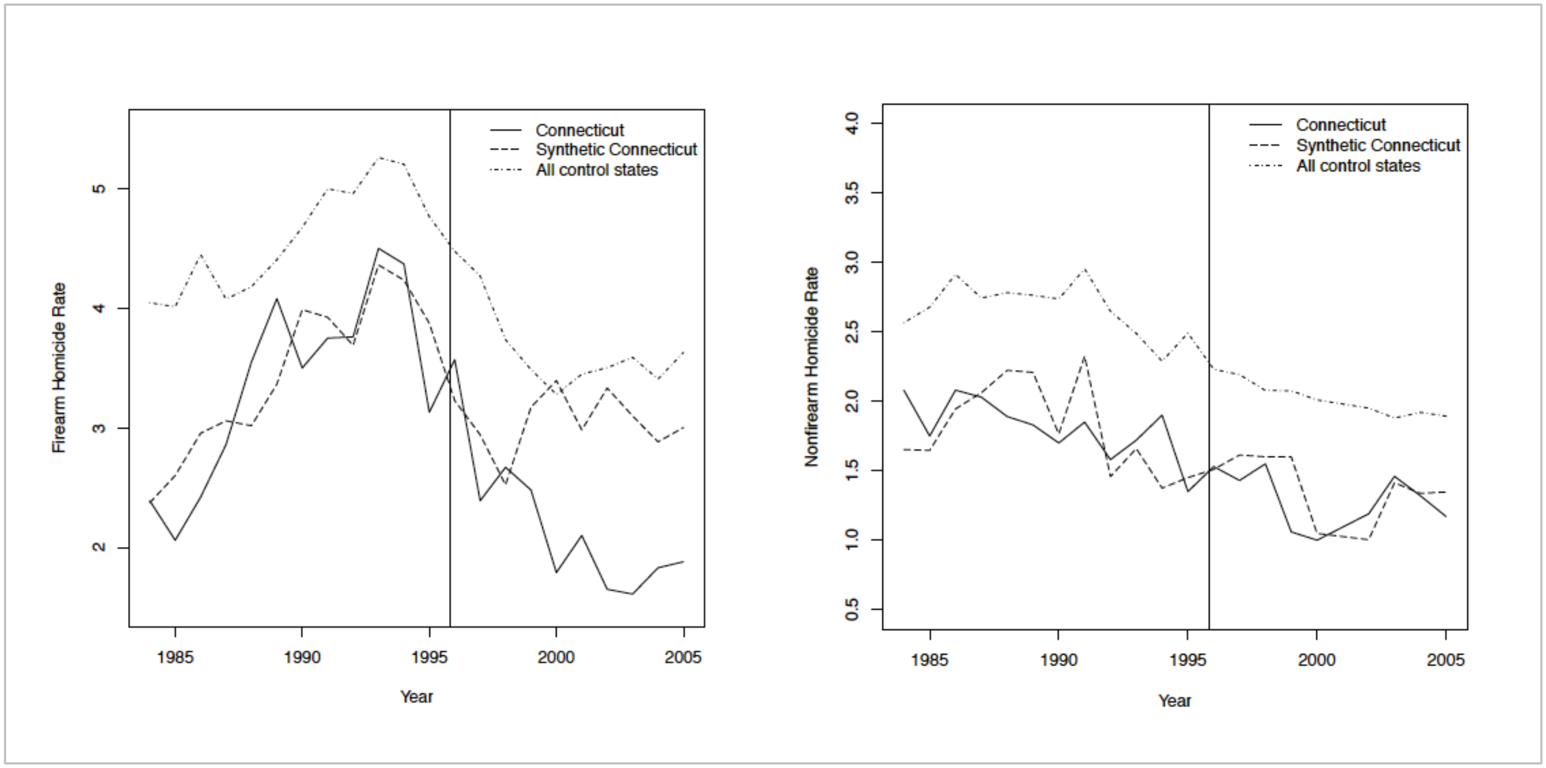


Graph from Robert McClelland and Sarah Gault, 2017, *The Synthetic Control Method as a Tool to Understand State Policy*, Washington, DC: Urban Institute.

Example: Effects of CT permit to purchase law

- Research question: How many homicides were prevented by Connecticut's 1995 Permit-to-Purchase handgun law?
  - Permit to purchase laws require individuals to get a permit to purchase a handgun, which involves a background check, fingerprints, safety training
- Use data on homicide rates over time (1984-2005) for CT and 39 states “at-risk" for implementing a PTP law in 1995 (i.e., without one)
- Fundamentally, compare trends and changes in CT with trends and changes in the other states
- Look at firearm homicides and nonfirearm homicides

Results graphically: Firearm (left) and non-firearm (right) homicides



Rudolph et al. (2015; AJPH)

Results numerically

- Law associated with a significant 40% reduction in the firearm homicide rate for the 10 years following implementation (296 deaths prevented)
- Law associated with a nonsignificant 24 nonfirearm homicides prevented for the 10 years following implementation
- Consistent with results looking at repeal of Missouri's permit to purchase law, and with results looking at suicide rates in CT and MO
- CT work cited by President Obama in his 2016 executive actions around gun control

But…what are some complications?

- Still no agreed upon approach for statistical inference
  - Traditional synthetic control approach uses a permutation based inference based on distribution of “placebo effects”
    - This is not intuitive for interpretation and leads to inference challenges
- Also does not take advantage of benefits that may be obtained by modeling the outcome over time
- New augmented synthetic control approach (Ben-Michael et al., 2019) fixes both of these things!
  - Can think of synthetic controls as a form of propensity score weighting
  - So then can combine with an outcome model, as usual!
- Standard implementation only for one site and single implementation
  - Brand new work extending to staggered implementation

What are some other recent advances in this area?

- Using matching methods to ensure comparable cohorts before and after a policy change
  - Used when individual-level data available, but not longitudinal on the same individuals
  - And desire is to separate changes in case-mix from the intervention's effects
  - Stuart et al. (HSORM, 2014)

- Bayesian Structural Time Series methods (Scott and Varian, 2014)
  - See forthcoming application in *American Journal of Epidemiology*, looking at Florida's PDMP program (Feder et al., in press)

- Imai et al. clarification of what the fixed effects models are estimating
  - https://imai.fas.harvard.edu/research/files/FEmatch.pdf

- New methods in economics and political science coming out frequently!

- So much it's hard to keep track!

And to talk up work and opportunities at Hopkins…post-doc openings currently too!

Reach out @lizstuartdc, estuart@jhu.edu

Center for Mental Health and Addiction Policy Research

Bloomberg American Health Initiative

ABOUT FOCUS AREAS DEAN'S MESSAGE FELLOWSHIPS NEWS & MEDIA CONTACT US

https://www.jhsph.edu/about/bloomberg-american-health-initiative/index.html

https://www.jhsph.edu/research/centers-and-institutes/center-for-mental-health-and-addiction-policy-research/index.html

Public health is rising to the challenge… again.

To learn more

- https://diff.healthpolicydatascience.org/
- https://andrewcbaker.netlify.com/2019/09/25/difference-in-differences-methodology/
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105: 493–505. doi:10.1198/jasa.2009.ap08746.
- Athey, S., and Imbens, G. 2018. “Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption.” *arXiv:1808.05293 [Cs, Econ, Math, Stat]*, August. http://arxiv.org/abs/1808.05293.
- Ben-Michael, E., Feller, A., and Rothstein, J. 2019. The augmented synthetic control method. https://arxiv.org/abs/1811.04170
- Ben-Michael, E., Feller, A., and Rothstein, J. 2019b. Synthetic controls and weighted event studies with staggered adoption. https://arxiv.org/abs/1912.03290
- Daw, J., and Hatfield, L. 2018. “Matching in Difference-in-Differences: Between a Rock and a Hard Place.” *Health Services Research* 53 (6): 4111–7. doi:10.1111/1475-6773.13017.
- Goodman-Bacon, Andrew. 2018. “Difference-in-Differences with Variation in Treatment Timing.” 25018. National Bureau of Economic Research. https://www.nber.org/papers/w25018.
- Imai, Kosuke, and In Song Kim. In Press. “When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?” *American Journal of Political Science*. http://web.mit.edu/insong/www/pdf/FEmatch.pdf.
- Stuart, Elizabeth A., Haiden A. Huskamp, Kenneth Duckworth, Jeffrey Simmons, Zirui Song, Michael E. Chernew, and Colleen L. Barry. 2014. “Using Propensity Scores in Difference-in-Differences Models to Estimate the Effects of a Policy Change.” *Health Services and Outcomes Research Methodology* 14 (4): 166–82. doi:10.1007/s10742-014-0123-z.