Learning what works in populations for public health and public policy: The role of careful study design

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

Background and big picture

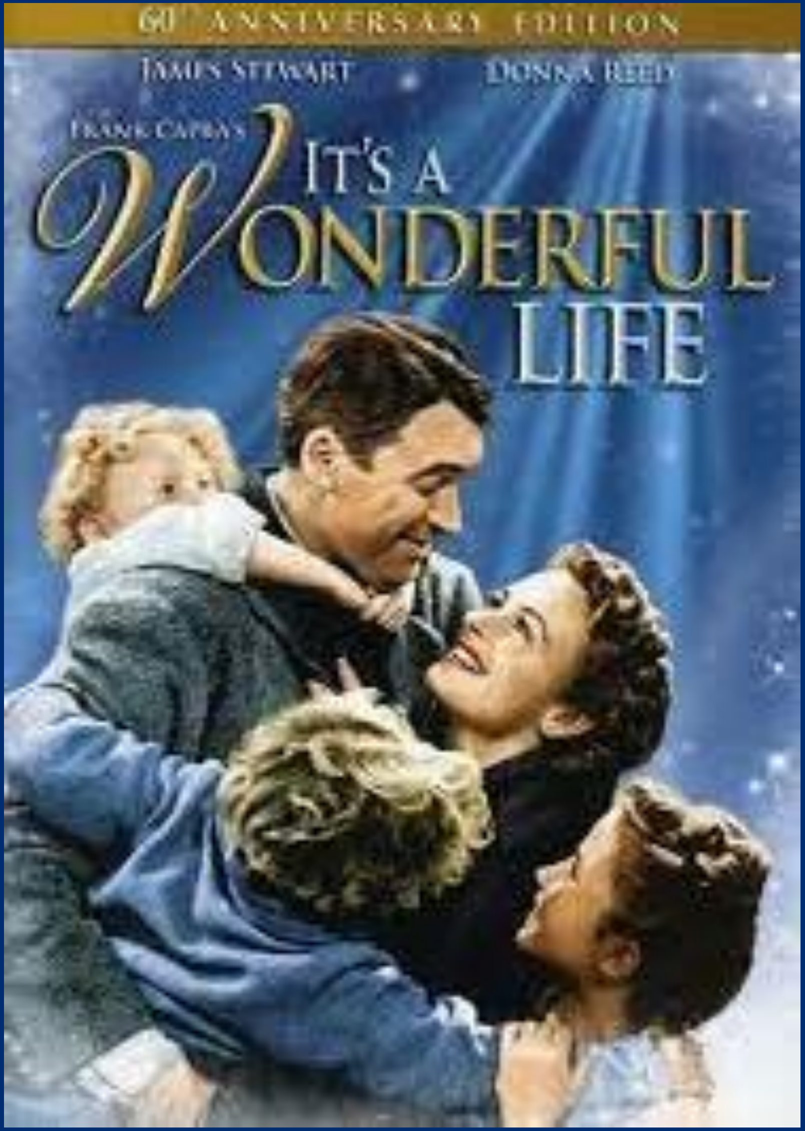
Estimating causal effects

- Does mental health parity increase service utilization of kids with autism (Stuart et al., 2017)?

- Do state opioid policies reduce opioid overdose deaths (McGinty et al., 2022)?

- Can firearm purchaser licensing laws help reduce homicides and suicides (McCourt et al., 2020)?

- Is in-person schooling linked to COVID risk in households, and can mitigation strategies help reduce the risk (Lessler et al., 2021)?



Why are these questions hard to answer?

- Causal questions inherently involve unobserved quantities:
  - What would have happened under some other state of the world?

- [Note that today I am focusing on estimating causal effects, rather than identifying the causes of effects, which is even harder!]

What is a causal effect?

- Comparison of potential outcomes for THE SAME well defined population:
  - Y(1): Outcome if treated (exposed)
  - Y(0): Outcome if control (not exposed)
- An association compares some outcome in two different groups

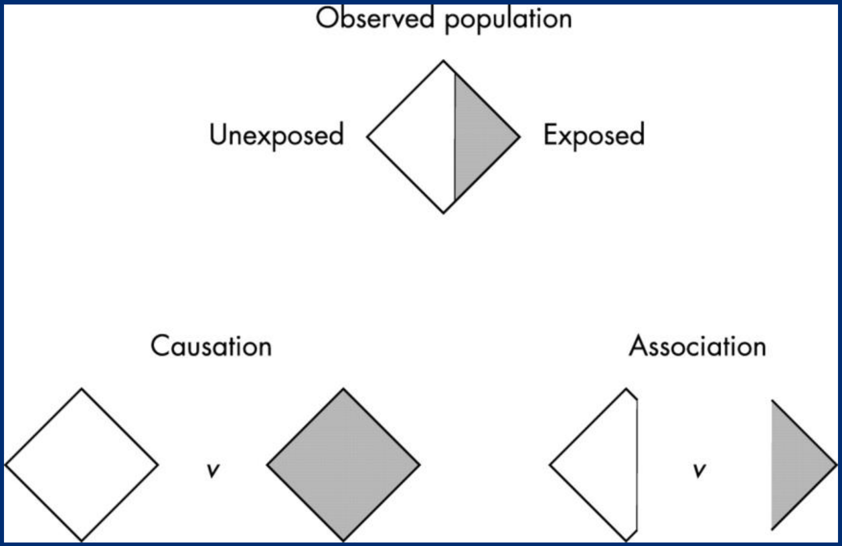


Figure from Hernán (2004): https://jech.bmj.com/content/58/4/265.info

This is what distinguishes causal inference from standard statistical inference: Trying to learn about things we don’t directly observe

For causal inference we need to rely on smart designs to help us learn about the missing potential outcomes, and thus the causal effect

Much of my focus: Estimating population effects

Internal and external validity

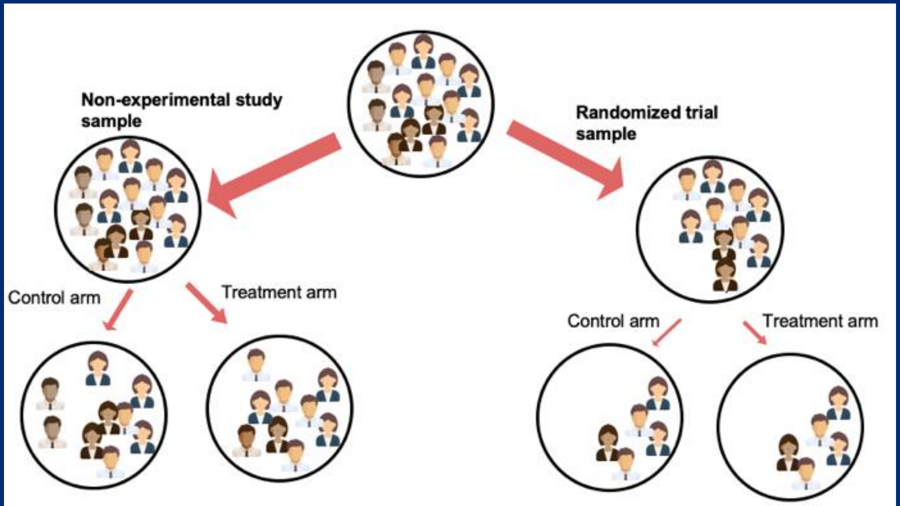
Internal validity reflects ability to estimate causal effects well in the sample at hand (e.g., through randomization of treatment assignment)

External validity reflects how well the effects are estimated for some potentially somewhat different/target population

Traditional study designs have prioritized internal validity

Much of my work tries to look at the balance of the two, and think about the pros and cons of different designs

Imai, King, and Stuart (2008) presents a formal framework for these trade-offs



The role of design

Randomization is a key tool to help us with both of these goals

Random sampling from the population gives external validity – e.g., complex survey designs with a well defined sampling frame and sampling procedure

Random assignment to treatment and control groups gives internal validity

When randomization not feasible at either stage, important to design the study as well as possible to avoid observed and unobserved bias

Important: “Big” data does not necessarily help!

Example 1:

Estimating the effects of suicide prevention centers in Denmark

Erlangsen et al. (*Lancet Psychiatry,* 2014)

Propensity score and related methods

- Main idea: Equate the treatment and comparison groups on a set of observed covariates

- Propensity scores help with that goal -- their statistical properties make them a particularly useful tool for creating covariate balance (Stuart, 2010)
  - Propensity scores used to match, weight, or subclassify the treatment and control groups

- Core ideas from Rosenbaum and Rubin (1983) but need for many extensions

“The planner of an observational study should always ask him[them]self: How would the study be conducted if it were possible to do it by controlled experimentation?” - W. Cochran (1965)

Studying suicide prevention centers

- Suicide prevention programs generally hard to study in a randomized design
- Require large samples, long follow up, and often ethical concerns

- Denmark began rolling out suicide prevention centers around the country in 1992 (now nationwide)

- Causal question: “What is the effect of these centers on the people who go to them, in terms of repeat suicide attempts and death up to 20 years later, compared to if they hadn’t gone?”

The data: Danish registries

- Amazing large-scale and comprehensive data on residents of Denmark
- Linked registers: Danish civil register, national registry of patients, psychiatric central registry, registry of causes of death
- Allows longitudinal data on individuals (and their families!!), including extensive social and health information
- Data on individuals 10+ from 1992-2011

A non-design based approach

- Typical analysis approach (especially years ago) would be to just fit a big regression model

- Use data from everyone in Denmark

- Model like: f(Y) ~ Treatment + Covariates

- Interpret coefficient on Treatment as the estimated effect

- Why isn't this great?
    - Not a careful comparison
    - Doesn't anchor time
    - Relies on extrapolation if treatment and control groups dissimilar

The treatment and comparison groups

- Treatment group: ~ 6,000 people who went to one of the suicide prevention centers after a suicide attempt

- Comparison group data, used to estimate what would have happened to the treatment group members had they not gone to one of the centers:
  - ~ 60,000 people who had an (index) suicide attempt but then did not go to one of the suicide prevention centers – way to define their "baseline"
  - (Wouldn't want to use people without an index suicide attempt; this is part of careful design)

The design

- Use 3:1 propensity score matching to find 3 comparison group individuals for each treated subject
    - Propensity scores estimated using 31 covariates, including demographics, previous suicide attempts, method of attempt, family history, and psychiatric disorders
    - Also require “exact match” on two particularly important confounders: any psychiatric disorder and previous attempts

- Importantly, can check how well this worked!

- Also importantly, done without using the outcome data!

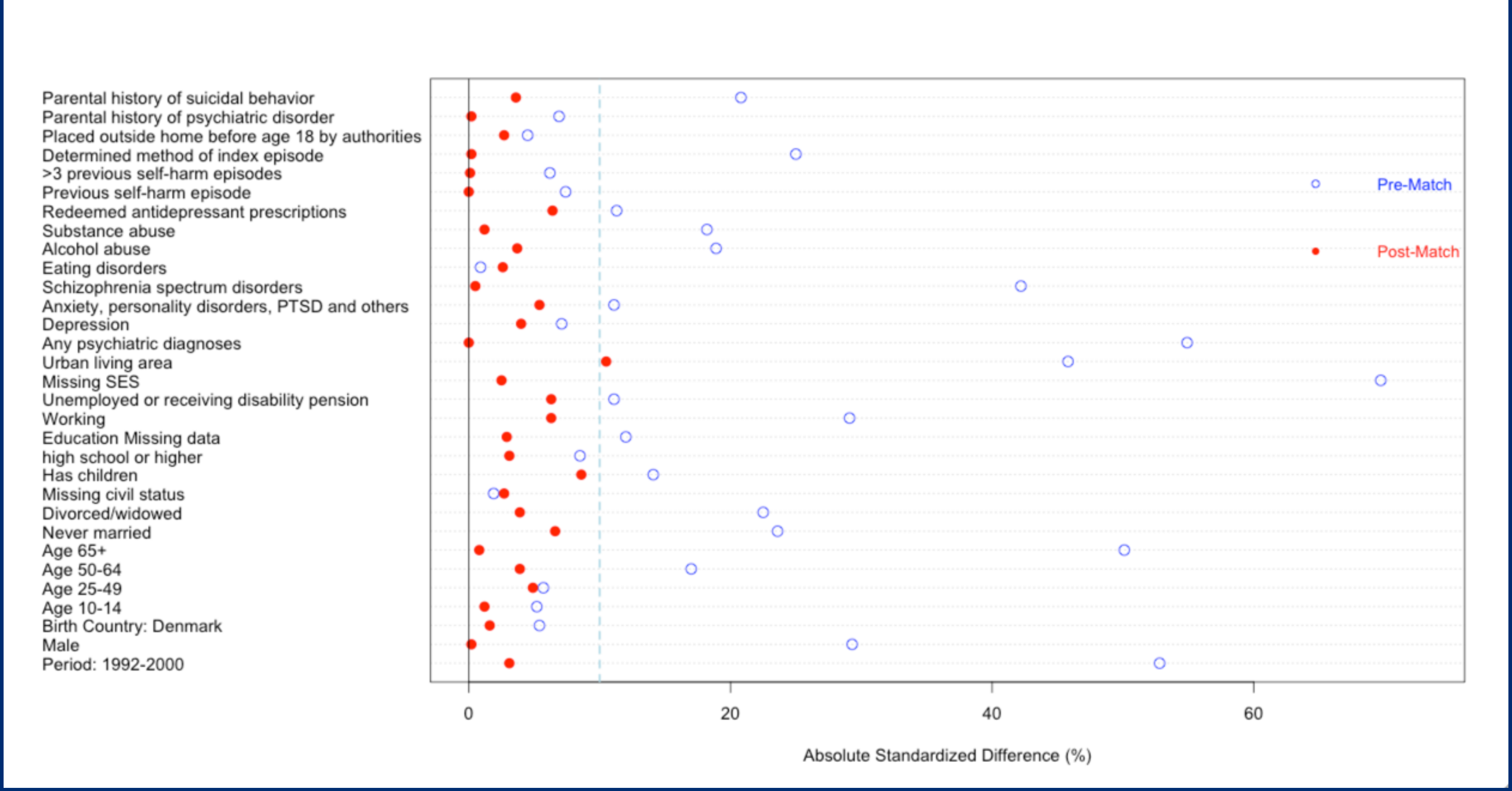
What does the data look like before matching?

Characteristic	Therapy group	Comparison group	Standardized mean difference
Male	31%	45%	0.29
Born in Denmark	90%	91%	0.05
Age 65+	2%	9%	0.50
Has children	39%	46%	0.14
Working	40%	25%	0.29
Any psychiatric diagnosis	72%	48%	0.55
> 3 previous suicide attempts	1.5%	2.3%	0.06

And after?

Characteristic	Therapy group	Matched comparison group	Comparison group	(After) standardized mean difference
Male	31%	31%	45%	0.00
Born in Denmark	90%	90%	91%	0.02
Age 65+	2%	2%	9%	0.01
Has children	39%	43%	46%	0.09
Working	40%	37%	25%	0.06
Any psychiatric diagnosis	72%	72%	48%	0.00
> 3 previous suicide attempts	1.5%	1.5%	2.3%	0.00

Matched groups similar on all 31 covariates



Now can compare outcomes

Outcome	Odds ratio	Confidence interval
Repeat attempt in 5 years	0.80	(0.73, 0.87)
Death by suicide in 5 years	0.74	(0.57, 0.97)
Death from any cause in 5 years	0.66	(0.56, 0.77)

The Achilles Heel: Unobserved confounding

- Still concern there may be an unobserved confounder related to going to one of the Centers and outcomes
- Sensitivity analysis can assess how strong such an unobserved variable would have to be to change study conclusions
- Turn a broad qualitative worry into a more quantitative concrete statement

- For one of the weaker effects (repeated self-harm after 20 years) a binary unobserved confounder with prevalence 0.5 would have to have a 1.8-fold association with participation in the program and a two-fold association with the outcome in order to explain the results

- Substantive experts felt this is unlikely

Lessons

- It is possible to use large-scale data to estimate causal effects in non-experimental studies
- Helps to have extensive covariates measured
- Use design elements, such as strategic selection of comparison subjects and approaches to help equate the treatment and comparison subjects on observed characteristics
- And important to acknowledge potential for unobserved confounding

Example 2:
Policy trial emulation

Ben-Michael, Feller, and Stuart (*Epidemiology,* 2021)
McGinty et al. (*Annals of Internal Medicine*, 2022, 2023)

Policy evaluation studies

- Lots of interest in understanding the effects of local policies, e.g., state opioid policies, local stay at home orders, etc.
- Can't randomize to exposure conditions
- Often relatively few units (e.g., 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 – have to be very careful to not attribute changes over time as causal effects

State opioid policies

Question: 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

Big picture

- Design based approach to do a difference-in-differences design, comparing trends in outcomes in treated and comparison states

- Use large-scale health insurance claims data (IBM MarketScan commercial claims)

- Rather than rely on models, use smart design choices (e.g., careful selection of states, individuals)

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

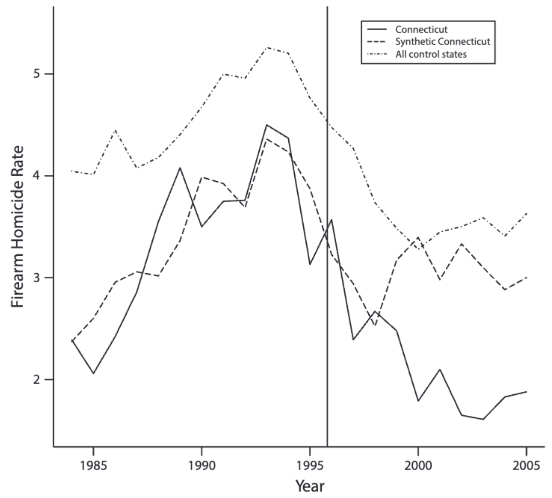
State Law	Law Date	Study Period	Comparison States[1]
Opioid Prescribing Cap Law			
Delaware	4/1/17	4/1/15-3/31/19	AL, IA, KS, MT, MS, ND, NM, OR, TN, WY
Kentucky	7/1/17	7/1/15-6/31/19	AL, IA, KS, MS, MT, ND, NM, OR, WY
New York	7/22/16	8/1/14-7/31/18	AL, IA, KS, MS, MT, ND, OR, WY
Ohio	8/31/17	9/1/15-8/31/19	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 (e.g., Rudolph et al., 2015)
- 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

- 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



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.

In this application…

Compare changes in outcome measures pre/post law in treated 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)

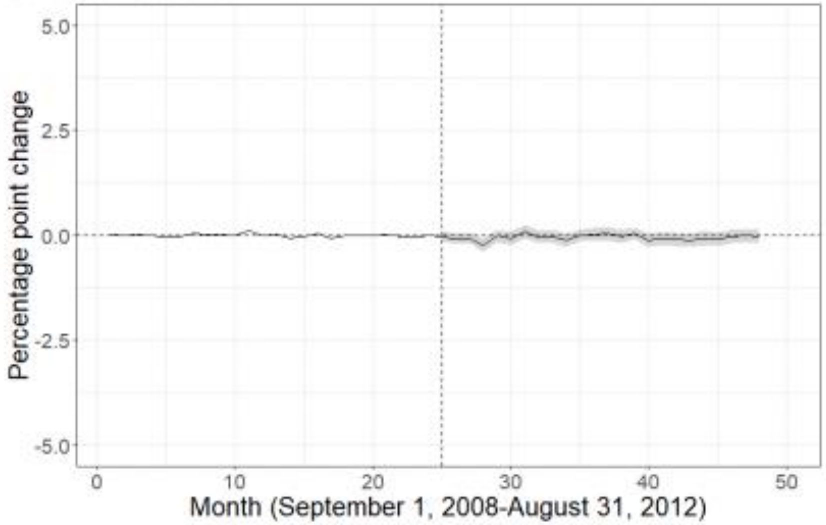
- Not directly testable because it involves counterfactual outcomes!
- (The pre-treatment trends analog is testable, although often low power and arguably equivalence testing better than traditional hypothesis testing)

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

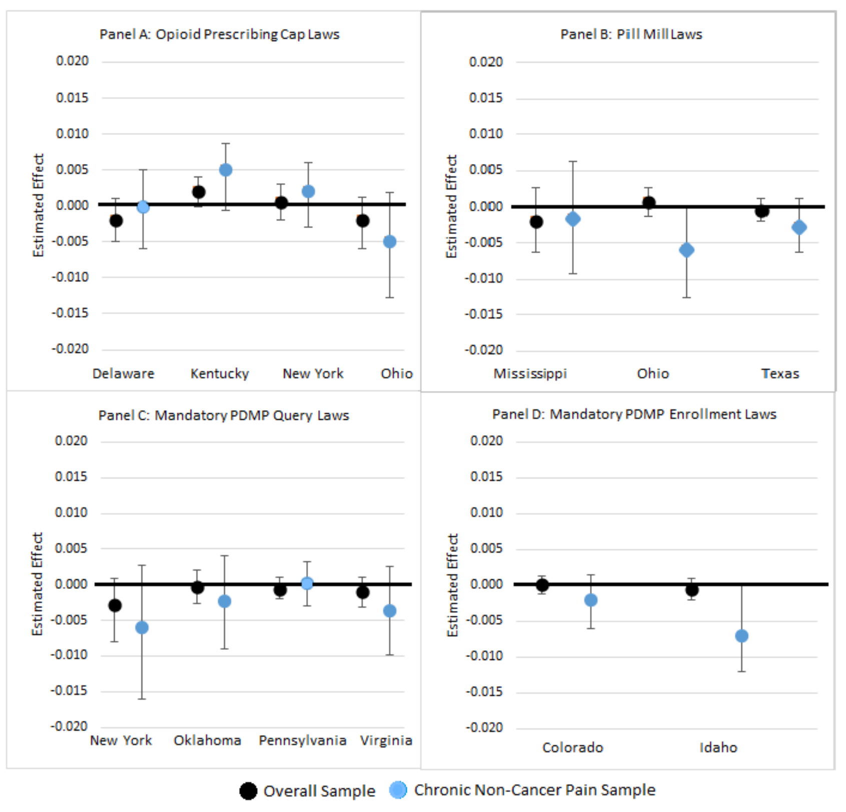
- This is what motivates the (augmented) synthetic control approach
- But this is no guarantee of the actual underlying assumption!
    - “The quality of our match historically is what makes us comfortable with extrapolation” (Luke Miratrix, Harvard)

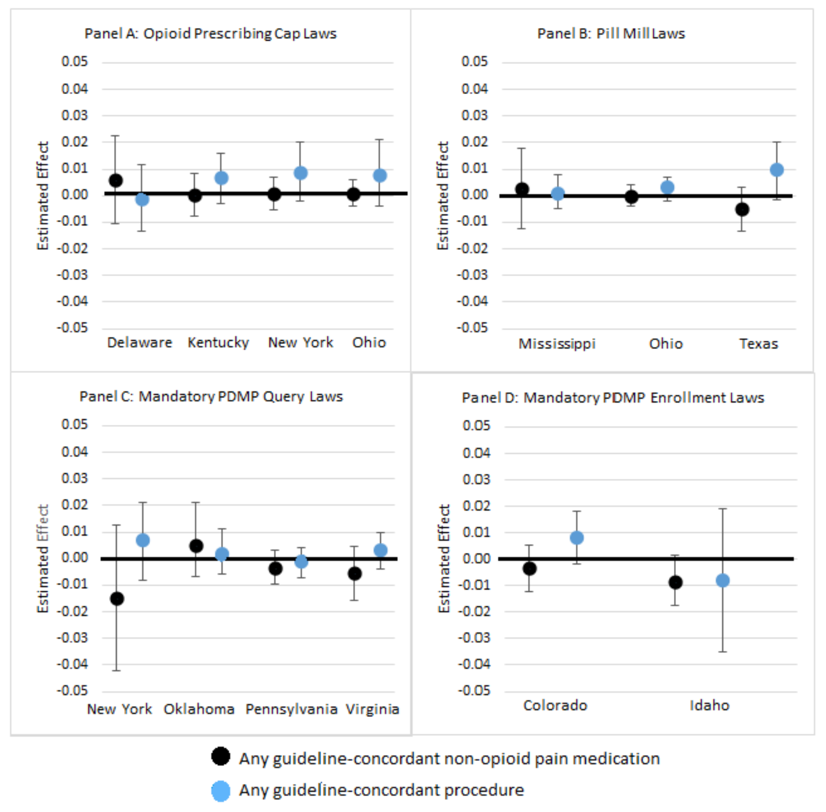
Augmented synth diagnostics

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



Basically no effects on opioid prescriptions (left) or on the probability of receiving guideline concordant non-opioid treatments (right) among people with chronic non-cancer pain





Lessons

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

- Engage with subject matter experts!!

- Note: MANY existing policy evaluations very poorly done
  - 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)

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

Example 3:

Enhancing the external validity of experiments

Stuart et al. (*Journal of the Royal Statistical Society – Series B,* 2011)
Cole and Stuart (*American Journal of Epidemiology*, 2010)
Olsen et al. (*Journal of Policy Analysis and Management*, 2013)

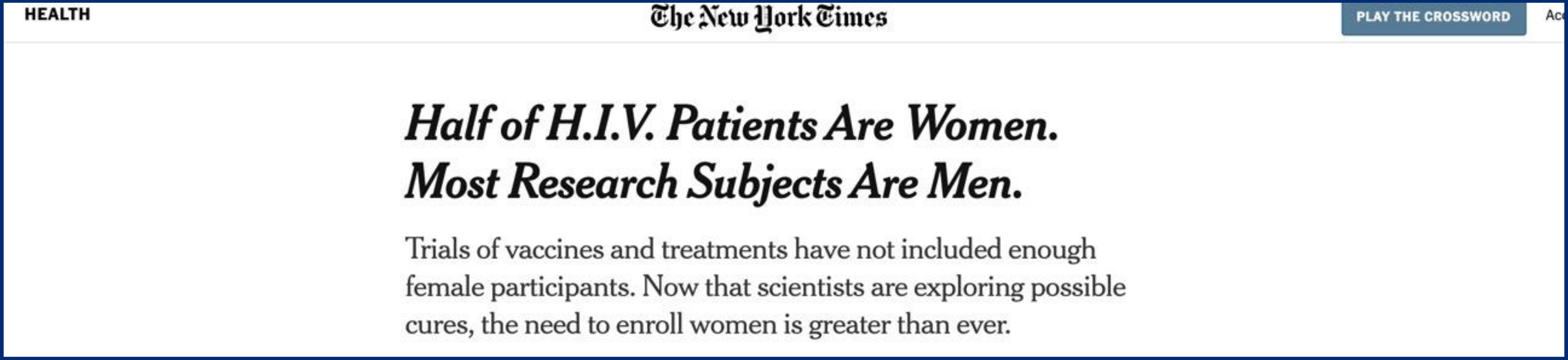
The value of randomization

Randomized experiments particularly useful for causal inference

Treatment and control groups only randomly different from one another, and so the outcomes observed in each group are an excellent proxy for (in fact an unbiased estimate of) the missing potential outcomes in the other group

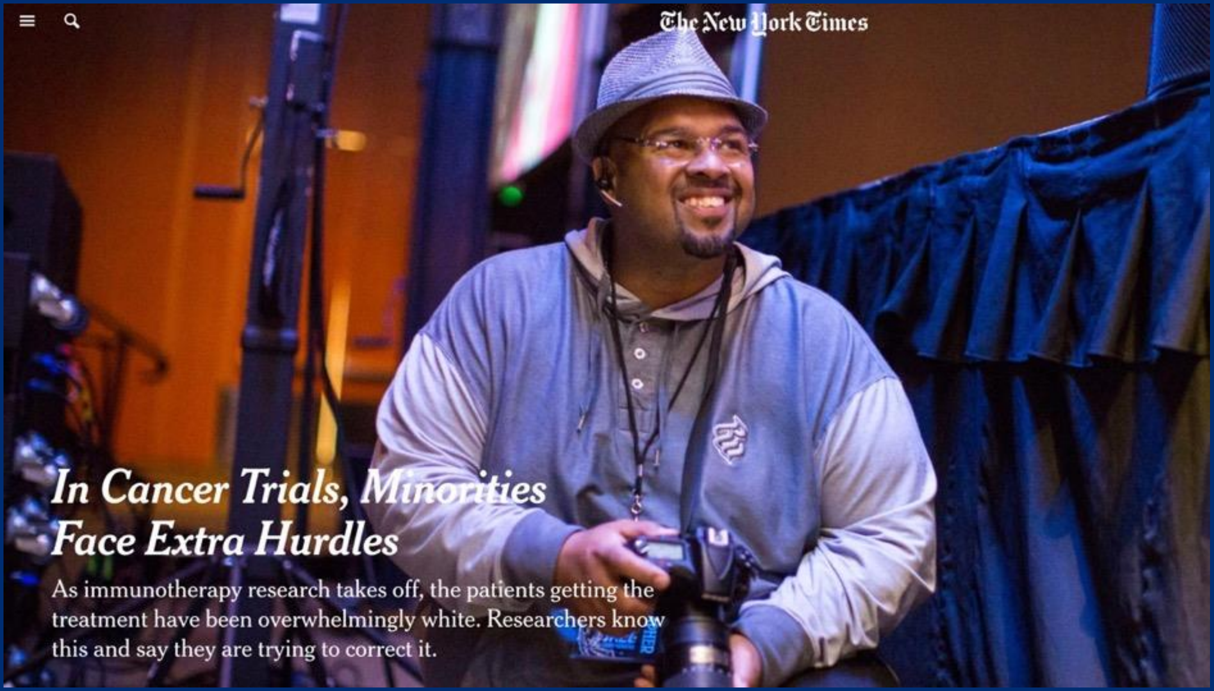
Formally: Can provide an unbiased effect estimate with (generally) essentially no assumptions

But….



Half of H.I.V. Patients Are Women. Most Research Subjects Are Men.

Trials of vaccines and treatments have not included enough female participants. Now that scientists are exploring possible cures, the need to enroll women is greater than ever.



In Cancer Trials, Minorities Face Extra Hurdles

As immunotherapy research takes off, the patients getting the treatment have been overwhelmingly white. Researchers know this and say they are trying to correct it.

External validity bias

- External validity bias is a function of three factors (Cole & Stuart, 2010; Olsen, Orr, Bell, & Stuart, 2013):
  - Variation in treatment effects
  - Variation in probabilities of participation in a study
  - Correlation between effects and those probabilities
- Almost never know these quantities
- So how can we design and analyze studies to minimize external validity bias?
  - Will focus on post-study adjustments today

Generalizing trial results to target populations

- New(ish) research area -- methods to generalize trial results to target populations by combining trial and population data

- Crucial to adjust for factors that moderate treatment effects and differ between trial and population

- Main approaches
  - Weight trial and population on observed characteristics (Stuart et al., 2010)
  - Model outcome in trial, use that model to project outcomes in population (Kern, Stuart, Hill, & Green, 2016)
  - Doubly robust methods that combine these two (Dahabreh et al., 2018, 2020, 2022)

Assumptions

- Experiment was randomized
- Sample ignorability for treatment effects: selection into the trial independent of effects given the observed covariates

$$(Y_i(1) - Y_i(0)) \perp S_i | X_i$$

  (Note: Can be relaxed somewhat if Y(0) observed in the population)

- Overlap: All individuals in the population had a positive probability of participating in the trial

$$0 < P(S_i = 1 | X_i) < 1 \text{ for all } X_i$$

ACTG trial

- ACTG trial randomized ~ 1200 adults to HAART or standard combination therapy
- Policy question (Cole & Stuart, 2010): What would be the population effects if HAART could be given to all newly infected adults in the US?
- Weight trial participants to match the US population on age, sex, and race

$$W_i = \frac{P(S_i = 1)}{P(S_i = 1|X_i)}$$

- [Limitation: Don't observe key potential moderator CD4 cell count in the population]

	Hazard ratio	95% confidence interval
Crude trial result	0.51	(0.33, 0.77)
Age weighted	0.68	(0.39, 1.17)
Sex weighted	0.53	(0.34, 0.82)
Race weighted	0.46	(0.29, 0.72)
Age-sex-race weighted	0.57	(0.33, 1.00)

Lessons

- We can use quantitative methods to help understand how well randomized trial results might carry over to target populations

- But we need to design our trials to allow generalization

- Common measures a key part – need consistent measures of moderators in trials and populations

Discussion

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

The role of statistics

- New strategies for harnessing large-scale data and for combining data sources
- Ability to utilize large-scale data sources with extensive information on confounding factors and other variables
- Need to think carefully about where the data comes from, what biases may exist, and then have methods to deal with those
- HUGE value in multiple fields talking to each other and collaborating – engage with subject matter experts [Dobbs example in SD]

Lessons for data science

- Remember that causal inference inherently involves trying to learn about things we don't directly observe

- Think carefully about time and ensure temporal ordering

- Find a devil's advocate/”hostile critic”

- Non-experimental studies will always involve some untestable assumptions
  - [And if someone claims they can test them there must be some other assumption underlying the test!]

EVER GIVEN
BAD STUDY DESIGN
REALLY FANCY STATISTICS
imgflip.com

Building a body of evidence

- “In conclusion, observational studies are an interesting and challenging field which demands a good deal of humility, since we can claim only to be groping toward the truth.” (Cochran, 1972): **No one study will be definitive!**

- **“Cochran’s Causal Crossword”** (Rosenbaum, 2015): "To take Cochran’s advice seriously is to be skeptical of investigations that derive stout conclusions from slender evidence. It is to be skeptical of grand studies and grand conclusions, the suggestion that a single proposed entry settles a major issue, that consistent completion of the puzzle is inevitable given this one entry, and hence consistent completion is not needed and not worth the effort.”

“If only the proponents of big data for causal purposes would take the time to read Cochran’s 1972 paper with care!” (Feinberg, 2015)

Acknowledgements

Many co-authors, students, colleagues, mentors
https://www.elizabethstuart.org/people/

Funding
Bloomberg American Health Initiative
National Institutes of Health
National Science Foundation
US Department of Education
WT Grant Foundation
PCORI
Meta

FAMILY
Brian, Clara, Paul