Target Trial Emulation for Evaluating Mental Health Policy

Nicholas J. Seewald, PhD

Assistant Professor of Biostatistics
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania Perelman School of Medicine

28 June 2024
Thomas R. Ten Have Symposium on Statistics in Mental Health
Joint with E.M. Stone, E.E. McGinty, E.A. Stuart

Perelman SCHOOL OF MEDICINE
University of Pennsylvania

Slides are Online

slides.nickseewald.com/tenhave2024.pdf

Disclosures

The views expressed in this presentation are my own and do not necessarily reflect the views of the University of Pennsylvania or the National Institutes of Health.

- NIDA R01DA049789 (PI: McGinty)

Thanks for letting me be here today.

Case Study 1: Texas Mental Health Agency Integration

- 1 in 3 people with an intellectual and developmental disability (IDD; e.g., Down, ASD) have a co-occurring mental health (MH) condition.
- 1 in 4 people in the general population have a mental health condition.
- Fragmented IDD & MH care systems lead to significant unmet care needs among individuals with co-occurrence
- In 2017, Texas began integrating its state disability and mental health agencies to “better integrate similar programs and services together”.

Stone, E. M. et al. (To Appear). *Community Ment Hlt J*.

Image generated by Google Gemini

Case Study 1: Texas Mental Health Agency Integration

Two key components of the Texas integration:

1. Eliminate Department of Aging and Disability Services
2. Integrate programming and regulatory responsibilities into the department of mental health within the Health and Human Services Commission.

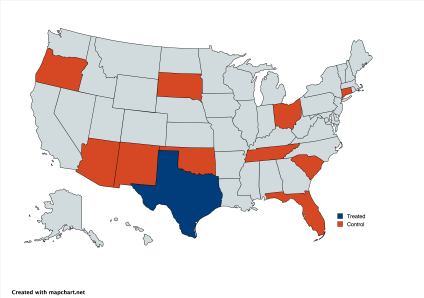
Question: Did integration improve individual mental health service outcomes?

Stone, E. M. et al. (To Appear). *Community Ment Hlt J*.

Case Study 1: Texas Mental Health Agency Integration

Our sample:

- 1 *treated* state (Texas)
- 10 *comparison* states that had separate disability and MH agencies 2014-2020



Stone, E. M. et al. (To Appear). *Community Ment Hlt J*.

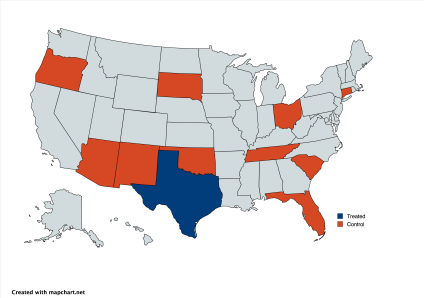
Case Study 1: Texas Mental Health Agency Integration

Our sample:

- 1 *treated* state (Texas)
- 10 *comparison* states that had separate disability and MH agencies 2014-2020

Goal: Estimate the effect, in Texas, of merging IDD and MH agencies on mental health care utilization.

Stone, E. M. et al. (To Appear). *Community Ment Hlt J*.



Case Study 2: Medical Cannabis Laws and Opioid Prescribing

- Cannabis is a potentially effective treatment for chronic non-cancer pain, but evidence is limited.
- Patients with chronic non-cancer pain are eligible to use cannabis under all existing state medical cannabis laws
- Some evidence of substitution of cannabis for opioids among adults with chronic non-cancer pain

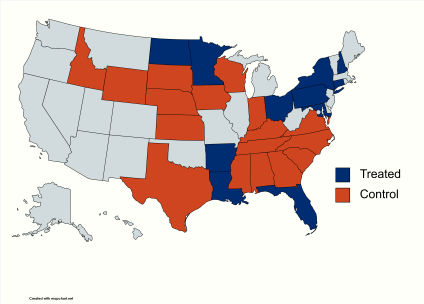
Question: What are the effects of state medical cannabis laws on receipt of opioid and non-opioid pain treatment among patients with chronic non-cancer pain, relative to what would have happened in the absence of such a law?

Bicket, M. C., Stone, E. M., and McGinty, E. E. (2023). *JAMA Netw Open*.

Case Study 2: Medical Cannabis Laws and Opioid Prescribing

Our sample:

- 12 *treated* states that implemented a medical cannabis law between 2012 and 2019 and *do not also have recreational cannabis laws*
- 17 *comparison* states without medical or recreational cannabis laws



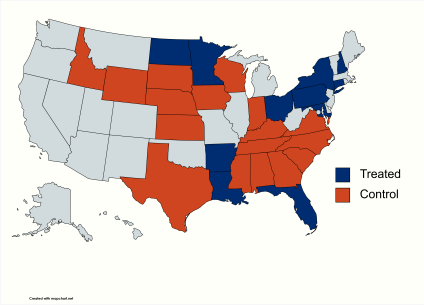
McGinty, E. E. et al. (2023). *Ann Intern Med*.

Case Study 2: Medical Cannabis Laws and Opioid Prescribing

Our sample:

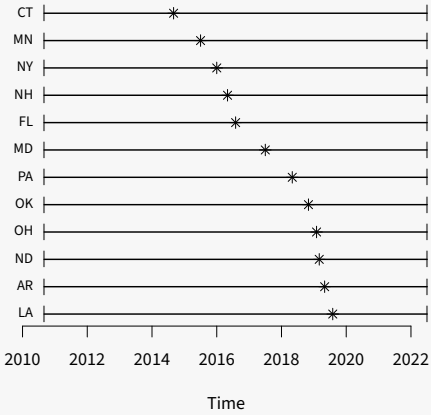
- 12 *treated* states that implemented a medical cannabis law between 2012 and 2019 and *do not also have recreational cannabis laws*
- 17 *comparison* states without medical or recreational cannabis laws

Goal: Estimate the effect of implementing a medical cannabis law on opioid prescribing outcomes, relative to what would have happened in the absence of treatment, among states that implemented such a law (an ATT).



McGinty, E. E. et al. (2023). *Ann Intern Med*.

"Staggered Adoption" of the Policy across States



States implemented medical cannabis laws at different times

Causal Inference for Policy Evaluation is Hard

We're asking causal questions about the effects of a policy. But:

- Necessarily limited sample size
- Can't randomize
- Often high variability in "treatment" definitions
- Hard to isolate a particular policy's effects when other policies are in place.

Target Trial Emulation

A framework for thinking about non-experimental studies that enables stronger designs and facilitates causal inference.

- **Key Idea:** Think about the trial you would run if you could, then design a non-experimental study that gets as close to that as possible.
- Commonly used in epi, but broadly applicable.
- Not magic! “Trial emulation” *per se* does not guarantee quality.

Target Trial Emulation

A framework for thinking about non-experimental studies that enables stronger designs and facilitates causal inference.

- **Key Idea:** Think about the trial you would run if you could, then design a non-experimental study that gets as close to that as possible.
- Commonly used in epi, but broadly applicable.
- Not magic! “Trial emulation” *per se* does not guarantee quality.

Target trial emulation is a way to talk about non-experimental study design in a way familiar to trialists.

A Warning!

Health policy applications sometimes require different considerations than epidemiologic ones. Crucially:

1. Policies are cluster-level interventions
2. Policy evaluations require natural experiments
3. Policy evaluations often have small sample sizes and policy-level units are non-exchangeable (e.g., states)

We have to make trade-offs. Keep this in mind throughout the talk!

The Scientific Question

In general:

> *"What is the effect of a policy on outcomes of interest over a defined period of time, relative to what would have happened in the absence of the policy?"*

The question is operationalized through clear definitions of 7 components of a policy trial emulation.

Components of a Policy Trial Emulation

1. Units and Eligibility Criteria
2. Definitions of Exposure & Comparison Conditions
3. Assignment Mechanism
4. Baseline / Time Zero
5. Outcomes and Follow-Up
6. Causal Estimand
7. Statistical Analysis & Assumptions

Components of a Policy Trial Emulation

1. Units and Eligibility Criteria
2. Definitions of Exposure & Comparison Conditions
3. Assignment Mechanism
4. Baseline / Time Zero
5. Outcomes and Follow-Up
6. Causal Estimand
7. Statistical Analysis & Assumptions

We recommend explicit comparisons of the non-experimental study to a target trial on these 7 components.

Units and Eligibility Criteria

Policy evaluations must consider

1. Units that could implement the policy or comparison condition (“policy-level”)
2. Units that the policy is designed to affect and on which outcomes are measured (“impact-level”)

If policy-level and impact-level units are different, policy evaluations will emulate *cluster-randomized* trials.

Policy-Level Units

In a **hypothetical policy trial**, policy-level units would be

- units that *could* implement the policy (e.g., states, organizations) and
- monitored longitudinally

Eligibility criteria would be based only on pre-policy information:

- “Has not implemented the policy before” or more complex (“has not previously implemented policies X, Y, Z”)

Policy-Level Units

In a **policy trial emulation**, policy-level units would be

- units that *did* implement the policy (e.g., states, organizations) or *did* implement the comparison condition
- at “time zero” / “study entry” (*ideally*), and
- monitored longitudinally

Eligibility criteria *should* be based only on pre-policy information:

- “Has not implemented the policy before” or more complex (“has not previously implemented policies X, Y, Z”)

Impact-Level Units

In a **hypothetical policy trial**, impact-level units are those that the policy is designed to affect. Possibly

- the policy-level units themselves, OR
- units nested in policy-level units on which outcomes are measured, ideally from the population the policy is designed to affect

“Eligibility” criteria would be based only on pre-policy information:

- “Lives in state X” for policies that apply to everyone
- “Lives in state X and has diagnosis Y pre-policy” for more targeted policies

Describe retention efforts if data-level units are “followed” longitudinally.

Impact-Level Units

In a **policy trial emulation**, impact-level units are those that the policy is designed to affect. Possibly

- the policy-level units themselves, OR
- units nested in policy-level units on which outcomes are measured, ideally from the population the policy is designed to affect

“Eligibility” criteria would be based only on pre-policy information:

- “Lives in state X” for policies that apply to everyone
- “Lives in state X and has diagnosis Y pre-policy” for more targeted policies

Describe retention efforts if data-level units are “followed” longitudinally.

Impact-Level Units

In a **policy trial emulation**, outcome data is ideally available from the impact-level units.

For example,

- Individuals with co-occurring IDD and mental health conditions in Texas and comparison states
- Individuals with chronic non-cancer pain diagnoses in eligible states

Available Data Affects Emulation Quality

Quality of emulation is partially determined by available data.

- "Group panel" data aggregated to policy-level is common
  - Might not be possible to restrict to target population $\rightarrow$ weaker trial emulation
  - Okay if aggregated from target population (e.g., everyone in a state) or in some contexts (e.g., state-month homicide counts)
- Impact-level data enables additional eligibility criteria
  - Ability to restrict to target population strengthens trial emulation

Case Studies: Data Availability

- Texas agency integration study
  - Medical Expenditure Panel Survey (MEPS) from Agency for Healthcare Research and Quality (AHRQ)
  - Households surveyed 5 or 7 times over 2 years + data from employers and medical providers
  - No specific IDD measure – had to assemble from ICD-9/10 codes. (Key for sensitivity analyses!)
- Medical cannabis laws study
  - Insurance claims data from large national carrier
  - Rich, detailed information to identify individuals with chronic non-cancer pain diagnoses
  - Only from commercially-insured adults

Stone, E. M. et al. (To Appear). *Community Ment Hlt J*.
McGinty, E. E. et al. (2023). *Ann Intern Med*.

Longitudinal Follow-Up for Impact-Level Units

In a policy trial emulation, following individuals longitudinally vs. in repeated cross-sections changes the “sampling frame”.

May choose to mimic high-quality retention efforts in an RCT by requiring “continuous presence” (e.g., continuous enrollment in health insurance)

- Maybe not appropriate if exposure affects probability of continuous presence.
- For insurance claims data, individual churn rates vary (highest in Medicaid).

Longitudinal Follow-Up for Impact-Level Units

In a policy trial emulation, following individuals longitudinally vs. in repeated cross-sections changes the “sampling frame”.

May choose to mimic high-quality retention efforts in an RCT by requiring “continuous presence” (e.g., continuous enrollment in health insurance)

- Maybe not appropriate if exposure affects probability of continuous presence.
- For insurance claims data, individual churn rates vary (highest in Medicaid).
- Not doing this probably leads to missing service use and allows patient case-mix to change over time, threatening internal validity (but improving external); weighting can help.
- BUT, impacts generalizability.

Longitudinal Follow-Up for Impact-Level Units

In a policy trial emulation, following individuals longitudinally vs. in repeated cross-sections changes the “sampling frame”.

May choose to mimic high-quality retention efforts in an RCT by requiring “continuous presence” (e.g., continuous enrollment in health insurance)

- Maybe not appropriate if exposure affects probability of continuous presence.
- For insurance claims data, individual churn rates vary (highest in Medicaid).
- Not doing this probably leads to missing service use and allows patient case-mix to change over time, threatening internal validity (but improving external); weighting can help.
- BUT, impacts generalizability.
- Done in medical cannabis laws study because infeasible that law would impact insurance enrollment.

Definitions of Exposure & Comparison Conditions

Trials require clear definitions of what each randomized arm receives and would (try to) ensure consistent treatment delivery.

In a policy trial, we would

- have one policy that all implementing units are assigned to implement, and
- do the same for controls (if control is a specific alternative policy) or “business as usual”

Definitions of Exposure & Comparison Conditions

Trials require clear definitions of what each randomized arm receives and would (try to) ensure consistent treatment delivery.

In a policy trial, we would

- have one policy that all implementing units are assigned to implement, and
- do the same for controls (if control is a specific alternative policy) or “business as usual”

In non-experimental policy evaluation, specifics of each policy can be quite heterogeneous.

Definition of Exposure

Goal: Identify a class (or small number of classes) of qualitatively similar policies that will be the exposure(s).

- “Policy mapping”/“legal epidemiology”: systematic approach to understanding timing of policies and the granular rules within them
- Understand different versions and core components of the policy, then decide which are qualitatively similar.
- Could emulate a multi-arm trial under high heterogeneity.

Definition of Exposure

In our medical cannabis law study, the exposure was

> *"A state medical cannabis law permitting cannabis use among individuals with chronic non-cancer pain with cannabis available for patient purchase through dispensaries"*

McGinty, E. E. et al. (2023). *Ann Intern Med*.

Refining the Definition of Exposure

"Confounding" policies may offer an alternative explanation for any observed effect.

- A strong policy trial emulation will precisely define exposure and comparison conditions to disentangle effect of interest.
- In medical cannabis law study, exposure was refined to a state medical cannabis law *and* absence of a recreational cannabis law throughout the entire study period.

Simultaneously-implemented policy bundles can only be studied in aggregate.

Defining the Comparison Condition

In a hypothetical policy trial, the comparison condition could be

- a specific comparator policy
- business as usual

The way we operationalize this in a policy trial emulation may be complex and involves tradeoffs.

Defining the Comparison Condition

Best practices for trial emulation:

1. At time zero, the comparison group is every unit that has not been exposed at that time
2. If unexposed units become exposed later, censor their outcomes at their exposure time

This ideal design isn't always practical for policy evaluations.

Choosing Comparators for Policy Evaluation

Choosing appropriate comparators is critical.

- Should consider contextual factors that may affect policy adoption and outcomes *differently over time*.
- Selecting geographically distant controls could help with spillover concerns, *but* near-neighbor comparators are probably more similar to the exposed unit
- Could use all units unexposed at baseline (ideal), but this leads to changing composition of the comparison group over time.

Choosing Comparators for Policy Evaluation

Let's examine the tradeoffs between two types of comparison groups.

- **"Unexposed at Baseline" Comparators**
  - Avoids conditioning on post-treatment information
  - BUT, allows the comparison group to change (possibly meaningfully) over time. Is an observed effect due to the policy or the changing comparison group?
- **"Never-Exposed" Comparators**
  - Chosen using knowledge of policy status later in time – could lead to bias!
  - Clearly not ideal in the target trials framework, BUT
  - the comparison group remains constant over time

Never-Exposed Comparators

To be clear:

Studies that choose to use never-exposed comparators are subject to additional assumptions about the comparability of ever- and never-exposed units and are subject to bias. *This choice deviates from the TTE framework.*

Redesign options:

- Change policy-level eligibility criteria to *de facto* exclude likely bad comparators (geography, urbanicity, etc.). Pay attention to remaining sample size.
- Limit the follow-up period to one in which good comparators exist.

Assignment Mechanism

Hypothetical Target Trial	Policy Trial Emulation Analogue
• Cluster-randomized • Possibly stratified • Almost certainly unblinded • Unconfounded on average	• Not randomized • Emulates cluster randomization • Almost certainly unblinded • *Affected by known and unknown characteristics of policy-level units*

Baseline / Time Zero

Hypothetical Target Trial	Policy Trial Emulation Analogue
The time of randomization. • Recruitment & prep done prior, so policy can be implemented immediately.	When policy could start impacting outcomes. • e.g., when first cannabis dispensary opens in a state. *Each policy-level unit could have its own unique time zero*

Baseline / Time Zero

Without randomization, baseline is complicated for comparison units. When could they have implemented the policy but did not?

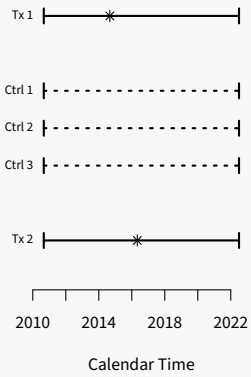
Poor definition of baseline for comparators can lead to bias from conditioning on post-treatment information.

Especially complicated under staggered adoption. One solution is **serial trial emulation**:

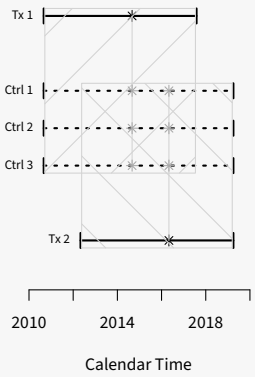
- Define baseline for each treated unit, then use those calendar times to define a series of baselines for comparators.
- Creates multiple trial emulations, one per unique policy implementation date.

Serial Trial Emulation

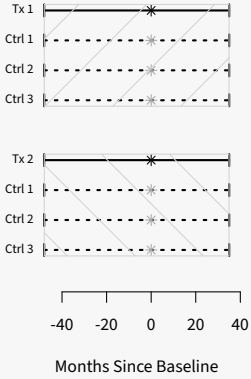
1. Identify Implementation Dates



2. Map Implementation Dates and Study Periods onto Controls

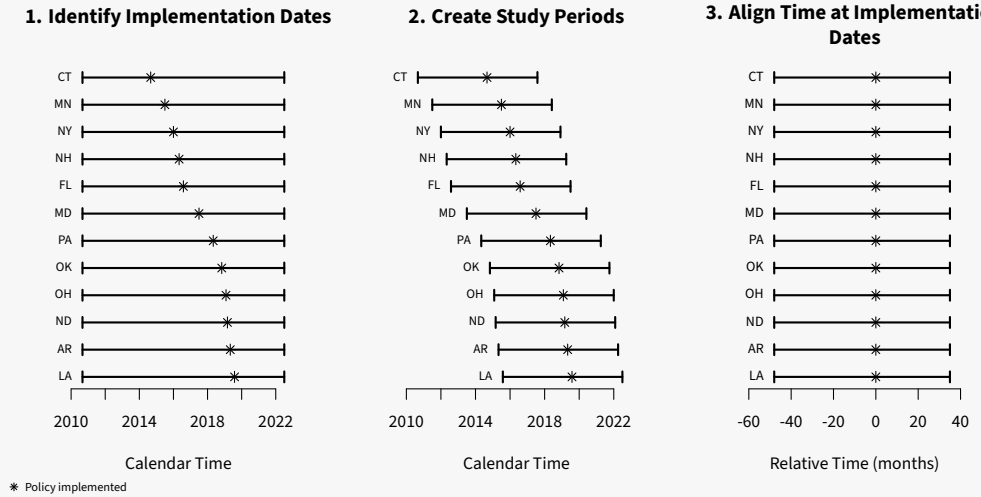


3. Create Unique Trials Aligned in Relative Time



✳ Policy implemented

Serial Trial Emulation in Medical Cannabis Study



Outcomes

Outcomes are interpreted at the policy level: they'll be proportions, means, etc. for each policy-level unit.

- Natural for group-panel data
- Individual-level data will be aggregated to the policy level

Can be prospectively designed in an RCT, but non-experimental policy evaluations are retrospective by nature.

Follow-Up

- RCTs typically have one (or few) pre-exposure measurements.
- In non-experimental context, validity of causal estimate relies on reasonably large number of pre-treatment measurement occasions.
  - Need to establish pre-policy outcome trends & anticipation effects
  - 4 years in medical cannabis law study
- Post-exposure follow-up should capture meaningful effects & changes therein (e.g., ramp-up)
  - 3 years in medical cannabis law study: balance need to look for ramp-up effects against need to avoid confounding laws

Causal Estimand

An *estimand* is a population-level quantity that statistically describes the treatment effect of interest.

Here, a causal quantity that describes the average difference between counterfactual outcomes in policy-level units under exposure and comparison conditions.

- Answers questions about what would have happened under different states of the world (e.g., with and without the policy exposure of interest)

"The expected difference in the proportion of individuals receiving any opioid prescription in a given month, averaged over exposed states and over three years, had the law been implemented versus had it not been implemented in those states."

Causal Estimand

An *estimand* is a population-level quantity that statistically describes the treatment effect of interest.

Here, a causal quantity that describes the average difference between counterfactual outcomes in policy-level units under exposure and comparison conditions.

- Answers questions about what would have happened under different states of the world (e.g., with and without the policy exposure of interest)

"The expected difference in the proportion of Texas individuals with at least one mental health related inpatient stay in a given year, averaged over the first four years following agency integration, had the agencies been integrated versus had they not been integrated."

Categories of Causal Estimand

1. **Average treatment effect (ATE)** compares expected counterfactual outcomes under exposure to those under the comparison condition on average over the entire population: $\mathrm{E}\left[Y(1) - Y(0)\right]$.
2. **Average treatment effect among the treated (ATT)** compares observed outcomes in the exposed group to what would have happened had it not been exposed: $\mathrm{E}\left[Y(1) - Y(0) \mid A = 1\right]$.
3. **Average treatment effect among comparators (ATC)** compares observed outcomes in the unexposed group to what would have happened had it been exposed: $\mathrm{E}\left[Y(1) - Y(0) \mid A = 1\right]$

Policy evaluations typically target the ATT: most feasible with fewest big conceptual jumps.

Statistical Analysis

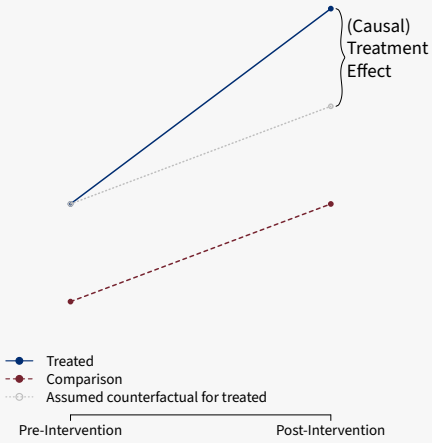
The cluster-randomized target trial can used “standard” analytic tools.

In non-experimental policy trial analogue:

- Methods typically use pre-baseline information from exposed and comparison groups to extrapolate an estimate of exposed group’s counterfactual outcomes under no policy.
- Broad class of methods: difference-in-differences, synthetic controls, etc.
- Analytic approach should estimate the estimand under reasonable assumptions.

Difference-in-Differences

- Compare change in outcome over time between exposed and comparison groups
- Under assumption that exposed group would look like comparison group in absence of the policy, can estimate causal policy effect
  - This is called the *(counterfactual) parallel trends assumption*



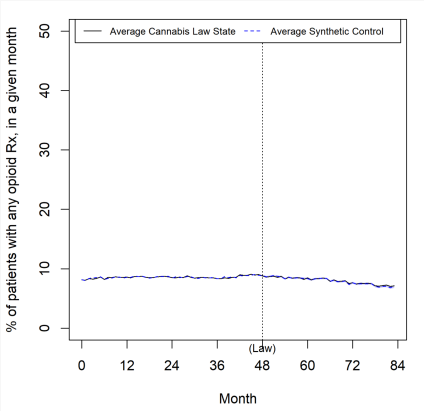
Stacked Difference-in-Differences

Uses diff-in-diff to estimate effects for each serial per-implementation-date trial emulation then aggregates them if appropriate.

Baker, A. C., Larcker, D. F., and Wang, C. C. Y. (2022). *Journal of Financial Economics*.

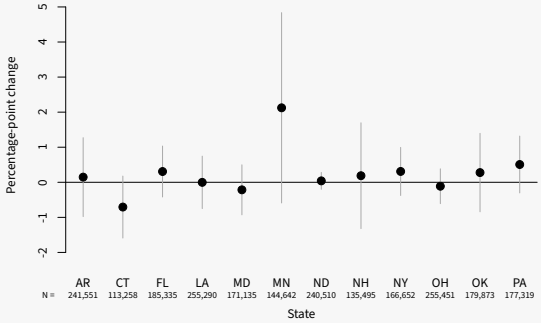
Synthetic Controls

- Construct a weighted combination of control states that mimics the outcome trajectory of the exposed state in the pre-policy period.
- Use the “synthetic control” trajectory to estimate exposed state’s counterfactual under no policy.



Abadie, A., Diamond, A., and Hainmueller, J. (2010). *J Am Stat Assoc*.
Ben-Michael, E., Feller, A., and Rothstein, J. (2021). *J Am Stat Assoc*.

Medical Cannabis Laws Study: Results



Change in proportion of chronic noncancer pain patients receiving *any opioid prescription*, per month, attributable to state medical cannabis law in first 3 years of implementation

McGinty, E. E. et al. (2023). *Ann Intern Med*.

Texas Agency Integration Study: Results

- No observed changes in mental health service outcomes among people with cognitive disabiliies and mental health conditions after integration.
- Deep qualitative component of study explores possible explanations: see paper to appear!

Stone, E. M. et al. (To Appear). *Community Ment Hlt J*.

Discussion

Explicit head-to-head comparison of target trial and a non-experimental policy evaluation helps identify threats to causal inference.

- Poorly-defined exposure inappropriately grouping different policies $\rightarrow$ estimate effect of some average policy that doesn't exist, ignores heterogeneity
- Failure to account for confounding policies $\rightarrow$ could estimate effect of wrong thing

Discussion

Strong agreement between trial emulation and target trial allows for use of causal language.

- Use “estimated effect” to acknowledge statistical and causal uncertainty
- Emphasize confidence intervals

Anecdotally, explicit comparisons and transparency about design and analysis have greatly improved understanding of our non-experimental studies and their results.

Acknowledgements

NIDA R01DA049789 (PI: McGinty)

Draft paper under review at *Annals of Internal Medicine* available on request.

seewaldn@pennmedicine.upenn.edu
nickseewald.com