Perelman SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

NCSU DEPARTMENT OF STATISTICS SEMINAR

Design Thinking for Policy Evaluation: The Policy Trial Emulation Framework

Nicholas J. Seewald, Ph.D.
Assistant Professor of Biostatistics
April 4, 2025

DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

Scan for slides

Some Framing

In statistics, we're often taught with statements that begin with

"Given data X…"

But *getting* data X is often very hard!

The goal of this talk is to get you thinking very deeply about **design**.

Scan for slides

The Goal of Policy Evaluation

In general:

> *"What is the effect of [a policy] on [outcome(s) of interest] over [a defined period of time], relative to what would have happened in the absence of the policy?"*

Design thinking is useful in operationalizing this question.

Scan for slides

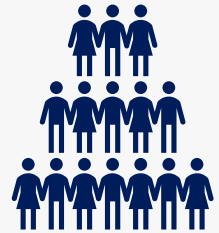
Challenges of Policy Evaluation

- Can be difficult to isolate policy of interest
  - "Current events are happening as we speak." - Matt Rogers, *Las Culturistas*
- Confounding by time
- Heterogeneous policies
- Small sample size

Some More Framing

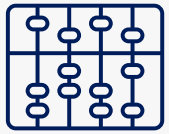
Researchers often start with a data structure and let everything flow out of that. That is… not great!

Data Structure



Population

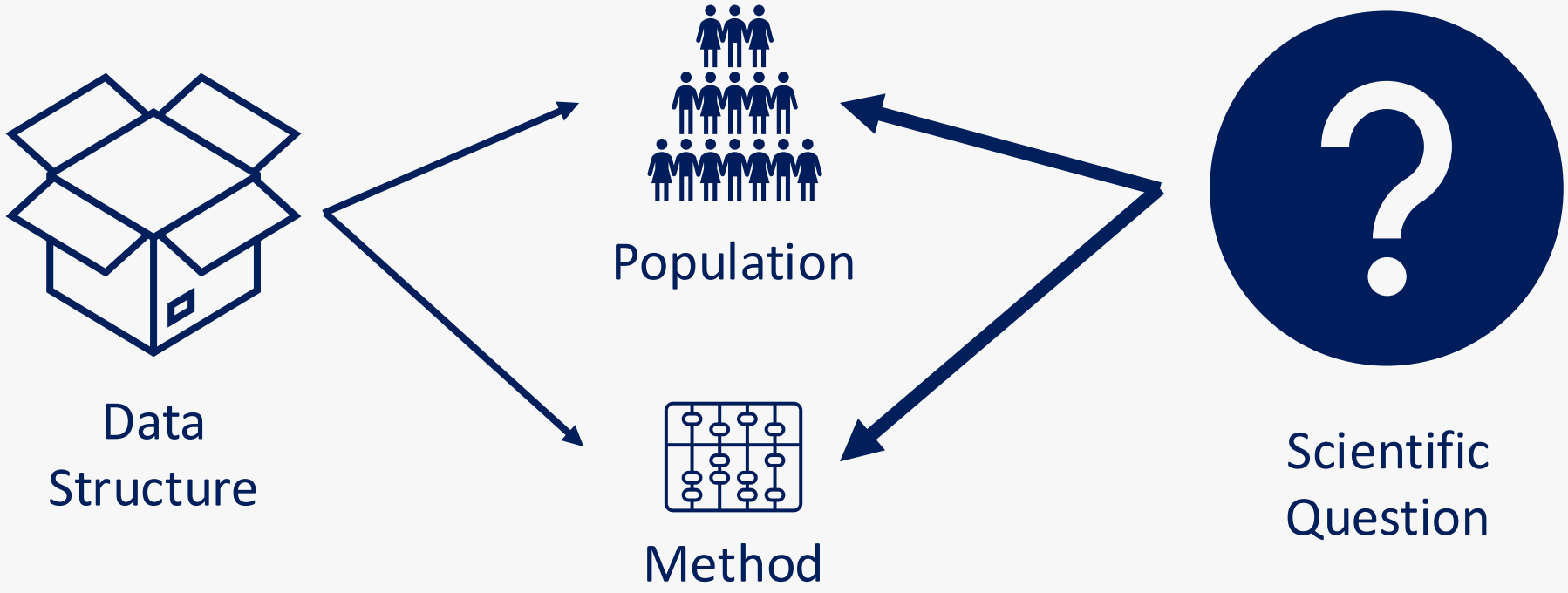
Scientific Question



Method

Some More Framing

We **should** be using the question to inform decisions, with the data structure used to guide practical realities.



Designing for Policy Evaluation

High-quality study design helps alleviate concerns about

- Isolating the policy of interest
- Confounding by time
- Heterogeneous policies

“Mixed-methods” approaches allow better understanding of

- “Treatment” definition
- Implementation time
- Effects (or lack thereof)

Target Trial Emulation (TTE)

A **design** 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 analogue that gets as close as possible.
- Common in epidemiology, but broadly applicable
- *Not magic!* TTE *per se* does not guarantee quality.

doi.org/nmmw

Controversy! Gasp!

Health policy applications often require different considerations than studies of individual-level interventions.

- Policies are cluster-level interventions
- Policy evaluations require natural experiments
- Sample sizes are often small
- Policy-level units are not exchangeable (e.g., states)

The practical reality of policy evaluation requires trade-offs from ideal trial emulation best practices.

Components of Policy Trial Emulation

1. Units and eligibility criteria
2. Definitions of exposure and comparison conditions
3. Assignment mechanism
4. Baseline / time zero and follow-up
5. Outcomes
6. Causal estimand
7. Statistical analysis and assumptions

This all happens **before** analysis!

1. Units & Eligibility

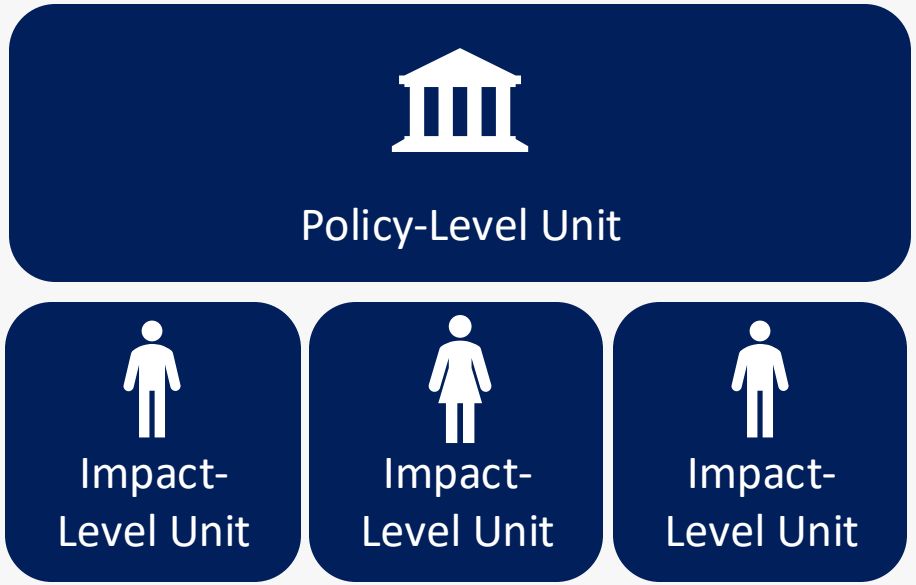
WHO ARE WE STUDYING?

Units and Eligibility Criteria

Policy evaluations must consider

1. "Policy-level" units that could implement the policy or comparison condition
2. "Impact-level" units that the policy is designed to affect and on which outcomes are measured.

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



Units and Eligibility Criteria

Policy evaluations must consider

1. "Policy-level" units that could implement the policy or comparison condition
2. "Impact-level" units that the policy is designed to affect and on which outcomes are measured.

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

Policy-Level Units

Hypothetical Policy Trial

- Units that *could* implement the policy (states, organizations, etc.)
- Monitored longitudinally

- Eligibility criteria *would* be based only on pre-policy information:
  - "has not implemented the policy before" or more complex (e.g., "has not previously implemented policies X, Y, Z")

Policy Trial Emulation

- Units that *did* implement the policy or *did* implement the comparison condition
- Monitored longitudinally

- Eligibility criteria *should* be based only on pre-policy information:
  - "has not implemented the policy before" or more complex (e.g., "has not previously implemented policies X, Y, Z")

Impact-Level Units

Hypothetical Policy Trial

- Units that the policy is designed to affect.
  - the policy-level units themselves, *or*
  - sub-units nested in policy-level units on which outcomes are measured, ideally from the population the policy is designed to affect.
- Eligibility based only on pre-policy information:
  - "Lives in state X" for policies that apply to everyone
  - "Lives in state X and was diagnosed with Y before the policy", etc.
- Retention efforts if followed longitudinally

Policy Trial Emulation

- Same deal!

Example: Medical Cannabis Laws

Consider a study designed to understand the effects of state medical cannabis laws on opioid prescribing among individuals with chronic non-cancer pain.

Policy-level units:

- States that did or did not implement a medical cannabis law, 2014-2019

Impact-level units:

- Individuals living in a policy-level unit with a chronic non-cancer pain diagnosis in the 3 years prior to “time zero”


Annals of Internal Medicine

ORIGINAL RESEARCH

Effects of U.S. State Medical Cannabis Laws on Treatment of Chronic Noncancer Pain

Emma E. McGinty, PhD; Kayla N. Tormohlen, PhD; Nicholas J. Seewald, PhD; Mark C. Bicket, MD, PhD; Alexander D. McCourt, JD, PhD; Lainie Rutkow, JD, PhD; Sarah A. White, MS; and Elizabeth A. Stuart, PhD

doi.org/khxp

Available Data Affects Emulation Quality

- "Group panel" data aggregated to policy level is common
  - Might not be possible to restrict to target population (→ weaker study)
  - Okay if aggregated from target population (e.g., all individuals with SMI) or if target population is very general

- Impact-level data enables additional eligibility criteria
  - Can restrict to target population (→ stronger study)
  - (Probably) doesn't improve efficiency for continuous outcomes

Longitudinal Follow-Up of Impact-Level Units

In policy trial emulation, following impact-level units longitudinally vs. in repeated cross-sections changes the *sampling frame*.

"Continuous presence" requirement can mimic high-quality retention efforts in an RCT

- Maybe inappropriate if exposure affects probability of continuous presence
- Not requiring this probably leads to missing service use and allows patient case-mix to change over time
  - Threatens internal validity but improves external validity (weighting can help!)
- Impacts generalizability

2. Exposure & Comparison Conditions

WHAT ARE WE STUDYING?

Definitions of Exposure & Comparison Conditions

Hypothetical Target Trial

- Exposure would be *one* policy that all implementing units are assigned to implement.
- Comparison could be a specific alternative policy, or “business as usual”

Policy Trial Emulation Analogue

- Specific details of each policy can be quite heterogeneous
  - E.g., specialty mental health clinics implement cardiovascular care management to different extents or in different ways.

Defining the Exposure

LEGAL EPIDEMIOLOGY

- Use **qualitative methods** to identify a class (or small number of classes) of similar policies that will be the exposure(s).
- Definition should be precise to help disentangle effects of interest & avoid confounding policies.
- Could emulate a multi-arm trial.

ANNUAL REVIEWS **Further**

Click here to view this article's online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

A Transdisciplinary Approach to Public Health Law: The Emerging Practice of Legal Epidemiology

Scott Burris,[1] Marice Ashe,[2] Donna Levin,[3] Matthew Penn,[4] and Michelle Larkin[5]

[1]National Program Office, Public Health Law Research Program, Beasley School of Law, Temple University, Philadelphia, Pennsylvania 19122; email: scott.burris@temple.edu

[2]ChangeLab Solutions, Oakland, California 94612; email: mashe@changelabsolutions.org

[3]Network for Public Health Law, St. Paul, Minnesota 55105; email: dlevin@networkforphl.org

[4]Centers for Disease Control and Prevention, Atlanta, Georgia 30333; email: itv1@cdc.gov

[5]Robert Wood Johnson Foundation, Princeton, New Jersey 08543; email: mlarkin@rwjf.org

Annu. Rev. Public Health 2016. 37:135–48

Keywords

Defining the Comparison Group

Best practices for trial emulation:

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

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

Choosing Comparators for Policy Evaluation

Unexposed at Baseline

- Avoids conditioning on post-treatment information
- 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

- Chosen using knowledge of future policy status – could lead to bias!
- Clearly not ideal in the target trial framework, but
- the comparison group remains unchanged over time.

Never-Exposed Comparators

Very commonly used in policy evaluations, but

- 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 ideal target trial emulation.*

Options for redesigning the study:

- 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.

Example: Medical Cannabis Laws

Our **exposure** was

- *Implementation of a medical cannabis law that allowed qualified patients to purchase cannabis at a dispensary*, and
- *Lack of implementation of a recreational cannabis law.*

The **comparison condition** was

- *Lack of implementation of a medical cannabis law, 2010-2022.*
  - (i.e., never-exposed comparators)

3. Assignment Mechanism

HOW DID UNITS DECIDE TO IMPLEMENT OR NOT IMPLEMENT THE POLICY?

Assignment Mechanism

Hypothetical Target Trial	Policy Trial Emulation Analogue
Cluster-randomized	Not randomized
Possibly stratified	(Usually) emulates cluster randomization
Almost certainly unblinded	Almost certainly unblinded
Unconfounded	Affected by known and unknown characteristics of policy-level units

4. Baseline / Time Zero

WHEN DID UNITS DECIDE TO IMPLEMENT OR NOT IMPLEMENT THE POLICY?

Baseline / Time Zero

Hypothetical Target Trial

Time of randomization

- Recruitment & prep done prior, so policy can be implemented immediately

Policy Trial Emulation Analogue

When the policy could start impacting outcomes

Complicated for comparison units. When could they have implemented the policy but did not?

Baseline / Time zero

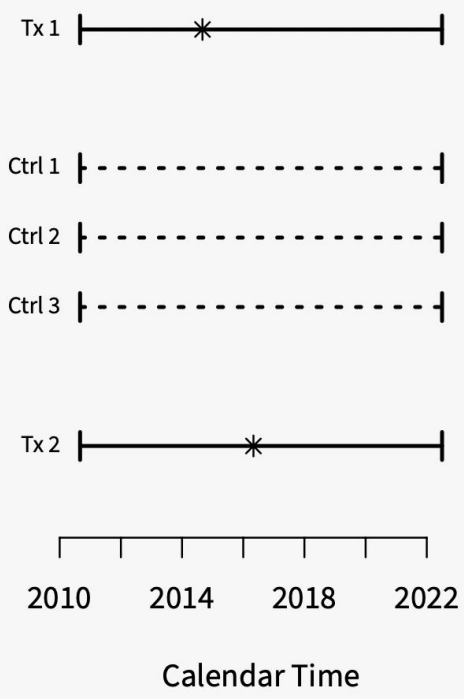
A bad definition can lead to bias (conditioning on post-treatment information)

"Staggered adoption" yields even more complexity. One solution is **serial trial emulation:**

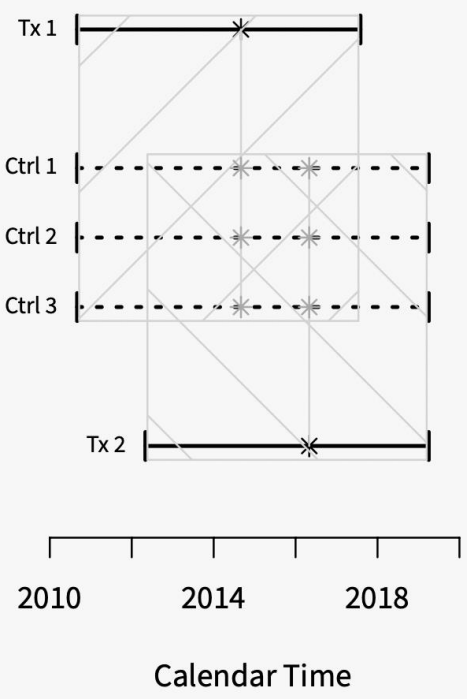
- Define baseline for each treated unit, then use those calendar times to define a series of baselines for comparators
- Creates multiple trial emulations, one per unique policy implementation date

Serial Trial Emulation

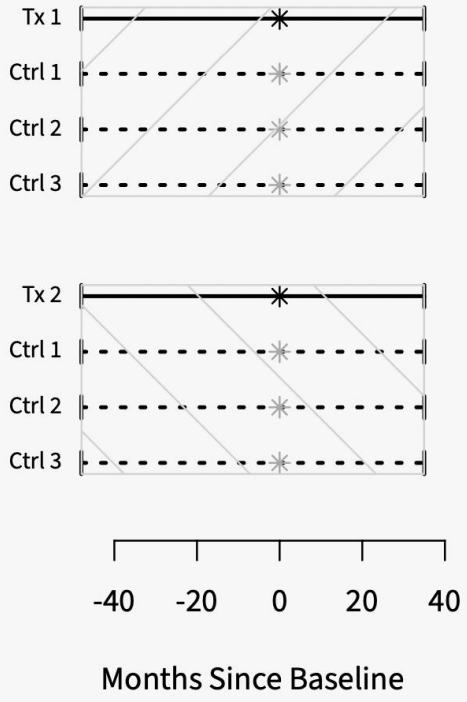
1. Identify Implementation Dates



2. Map Implementation Dates and Study Periods onto Controls



3. Create Unique Trials Aligned in Relative Time



✳ Policy implemented

5. Outcomes and Follow-Up

WHAT ARE WE MEASURING AND WHEN?

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.

Validity of causal estimate in non-experimental study often relies on reasonably large number of pre-treatment measurement occasions.

Post-exposure follow-up should capture meaningful effects & changes therein.

6. Causal Estimand

WHAT POPULATION-LEVEL QUANTITY DESCRIBES THE QUESTION OF INTEREST?

Causal Estimand

Often, 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.

Expressed in **potential outcomes notation**:

- $Y(1)$ is the outcome that would be observed under exposure
- $Y(0)$ is the outcome that would be observed under no exposure

Categories of Causal Estimand

Typically the target (by convention)

Average treatment effect (ATE) compares expected counterfactual outcomes under exposure to those under the comparison condition on average over the entire population

- $E[Y(1) - Y(0)]$

Average treatment effect among the treated (ATT) compares observed outcomes in the exposed group to what would have happened had they been unexposed:

- $E[Y(1) - Y(0) \mid A = 1]$

Average treatment effect among comparators (ATC) compares observed outcomes in the unexposed group to what would have happened had they been exposed:

- $E[Y(1) - Y(0) \mid A = 0]$

Target Estimands & Target Audiences

Policymakers (hopefully) want to know what will happen if they implement something

- Question of interest for PA probably isn't "What happened in MD", but "What will happen in PA?"
  - ( That's not an ATT!)

We all love the ATT

The **average treatment effect among the treated (ATT)** compares what actually happened to the policy-implementing units to what would have happened in the absence of the policy.

$$ATT = E[Y(1) - Y(0)|A = 1]$$

A **ton** of methods estimate this.

But there are inherent limitations here!

Pros and Cons of the ATT

$$ATT = E[Y(1) - Y(0)|A = 1]$$

The ATT is nice because it

- is common, and so easy to communicate
- only requires imputing one counterfactual
- neatly describes *what happened*

One big problem, though:

- The ATT doesn't necessarily give actionable information to policymakers: it's inherently *post hoc*

Why do we prefer the ATT?

The ATE and ATC both require estimating $E[Y(1) \mid A = 0]$

This… feels weird! Usual identification assumptions often feel too strong.

- But: mismatch between the real question and what we can confidently do

We should try to get creative here!

7. Statistical Analysis

HOW DO WE ANALYZE DATA TO ANSWER THE QUESTION, AND UNDER WHAT ASSUMPTIONS?

Analytic Considerations

The hypothetical cluster-randomized target trial can use “standard” tools

Our non-experimental trial analogue probably can’t, because assignment is confounded.

- **Goal:** Estimate the estimand with reasonable assumptions.
- Methods usually use pre-baseline information from exposed & comparison units to extrapolate an estimate of the exposed group’s counterfactual outcomes under no policy.

Methods Explosion!

There's an increasingly large class of methods designed for this setting!

- **Difference-in-differences**
  - Two-way fixed effects
- **Synthetic controls**
  - Augmented synthetic controls
- Event studies

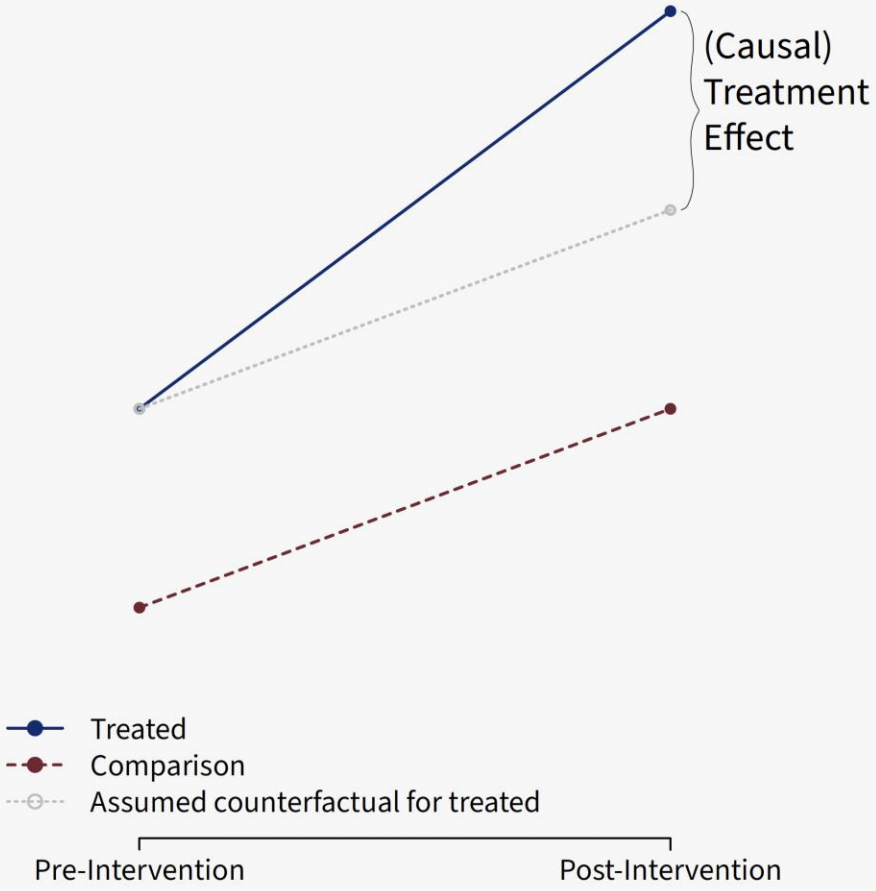
Different methods rely on different assumptions: be careful to be reasonable!

Difference in Differences (DiD)

Big Idea: Compare change in outcome over time between exposed and comparison groups.

Key Assumption: ”Parallel counterfactual trends”

- The exposed group’s outcome evolution would have looked like the comparison group’s outcome evolution had the exposed group not been exposed.



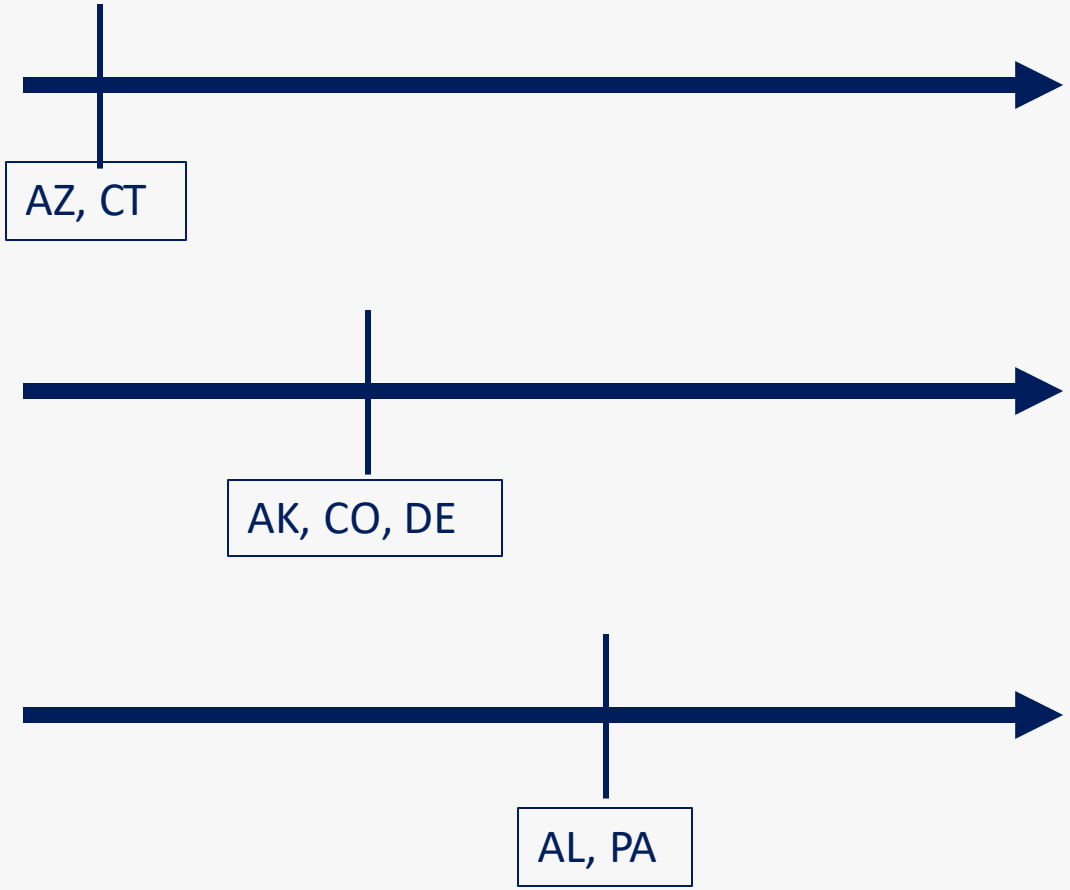
Staggered Adoption

Not every exposed unit is exposed at the same time!

- Staggered program rollout
- Policies adopted at different times

This can create **big** problems with traditional estimation.

- Traditional approach can be extremely biased if there are time-varying treatment effects under staggered adoption. (Goodman-Bacon 2021)

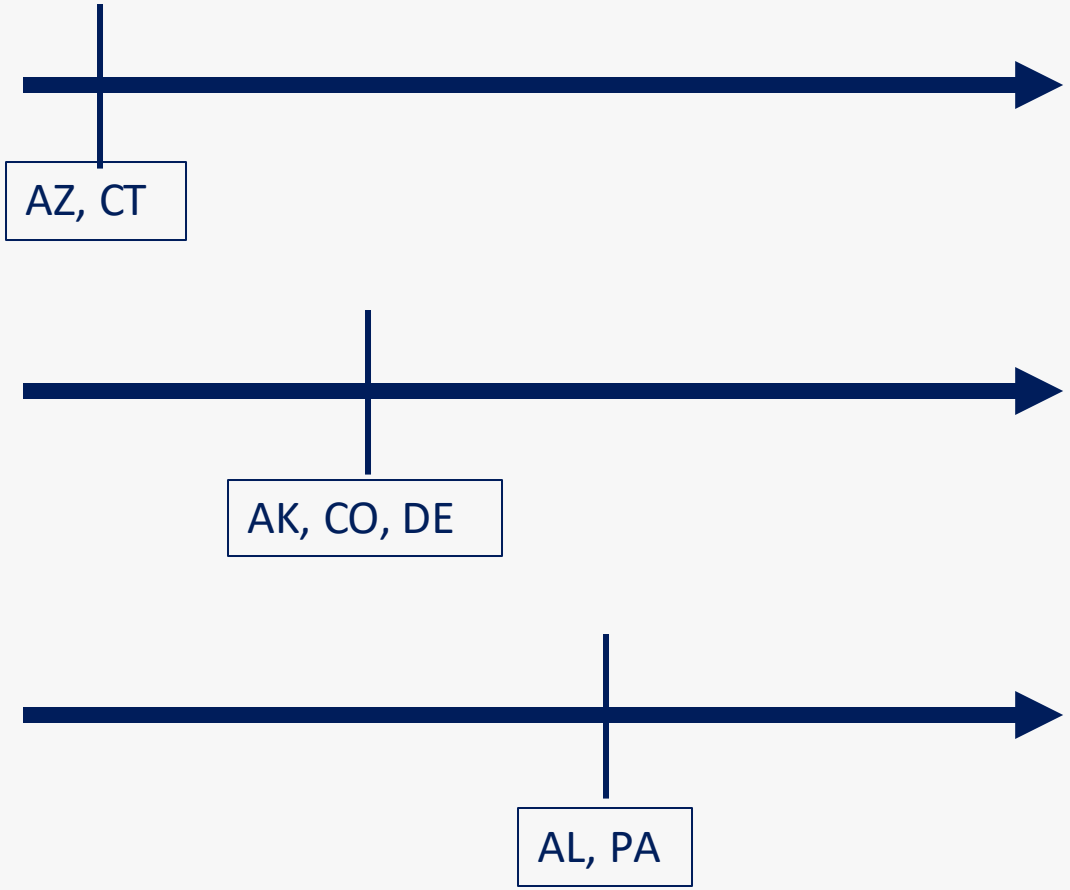


Goodman-Bacon A. Difference-in-differences with variation in treatment timing. J Econometrics. 2021 Dec;225(2):254–77.

New Methods Handle Staggered Adoption

One common solution:

1. Group units treated at the same time
2. Estimate “group-time” effects for each such group
3. Aggregate group-time effects to estimate quantities of interest



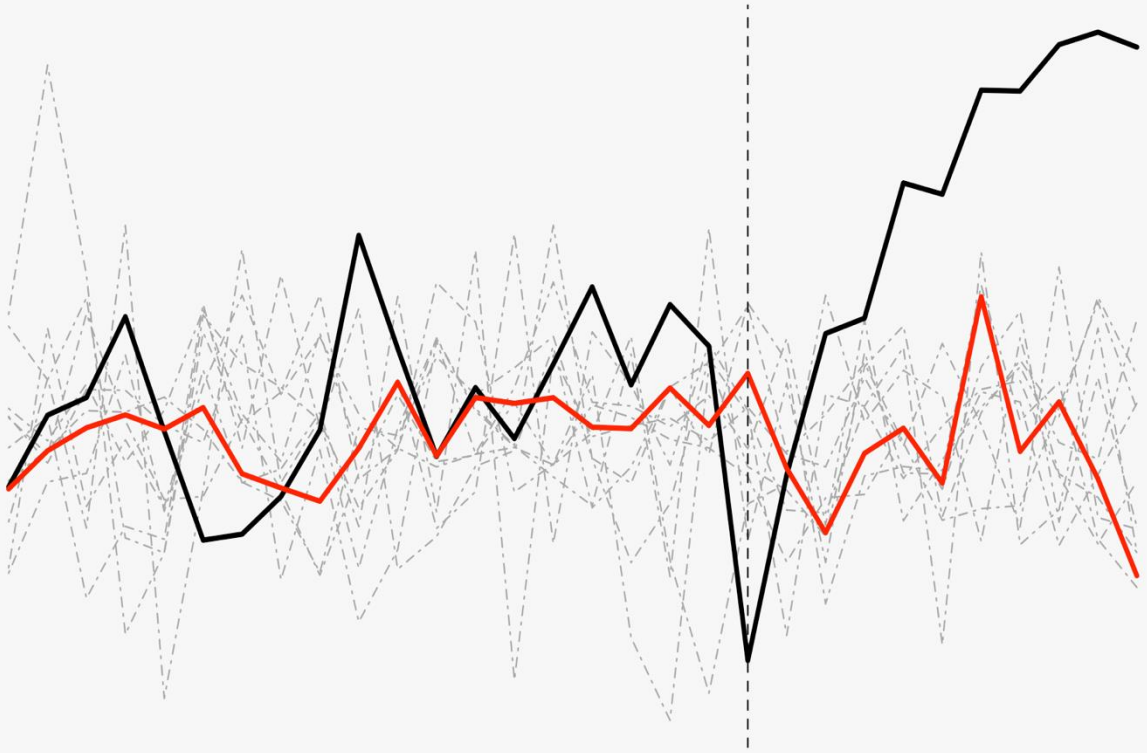
Callaway B, Sant’Anna PHC. Difference-in-Differences with multiple time periods. J Econometrics. 2021;225(2):200–30.

Synthetic Controls

Big Idea: Construct a weighted combination of non-implementing units that mimics the outcome trajectory of each implementing unit in the pre-policy period.

Then, extrapolate that combination forward to estimate the counterfactual for the exposed unit under no policy.

A useful variant is **augmented synthetic controls**, which incorporates covariates to get better pre-treatment fit.



Ben-Michael E, Feller A, Rothstein J. The Augmented Synthetic Control Method. *J Am Stat Assoc* 2021;**116**:1789–803.

Discussion

Scan for slides

Good study design is critical

Policy trial emulation provides a framework for thinking about good policy evaluation study design

- Think about the trial you would run if you could, then try to get as close as possible.

Closer alignment between hypothetical target trial and non-experimental analogue improves communication

- Clearly articulate similarities & differences across all 7 components
  - Use a table to compare target trial & emulation (Seewald, et al. 2024)
- Helps readers understand design better & calibrate confidence in results

Seewald NJ, McGinty EE, Stuart EA. Target Trial Emulation for Evaluating Health Policy. *Ann Intern Med* 2024.

Scan for slides

Good study design is not magic

Policy trial emulation does not guarantee quality.

- An emulated trial is not a trial.
- Calling something “trial emulation” doesn’t mean the trial was emulated well.

There will always be trade-offs.

Statistical tools for high-quality policy evaluation are available and accessible

Lots of methods innovation across disciplines – join us!

The key goal is often to estimate a good proxy for what would have happened in the absence of the policy.

Multi-disciplinary work is key

Rigorous policy research requires collaboration across disciplines

- Need both quantitative and qualitative approaches
- Working across fields improves communication and impact
- Challenging, but fun!

seewaldn@pennmedicine.upenn.edu

www.nickseewald.com