Perelman SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

ENAR 2025

How to Ask the Right Question: Choosing Estimands for Health Policy Research

Nicholas J. Seewald, PhD
Assistant Professor of Biostatistics

24 March 2025

DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS

Scan for slides

Context for this talk

This talk is a (very) early attempt to build a framework for translating policy evaluation questions to estimands.

Feedback & ideas are encouraged!

Some of the ideas come out of this paper:

Perelman SCHOOL OF MEDICINE UNIVERSITY of PENNSYLVANIA

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?"*

Policy Evaluation is Hard

- Policies are heterogeneous
- Policies aren’t implemented in a vacuum
- Small sample sizes
- Confounding by time
- Policymakers need to understand results

Formalizing an Estimand

"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?"

Answering this, and translating it to an estimand, requires operationalizing

- Who
- What
- When
- Where
- (Why)

Good Design Helps

Careful thinking about study design helps in defining a causal contrast

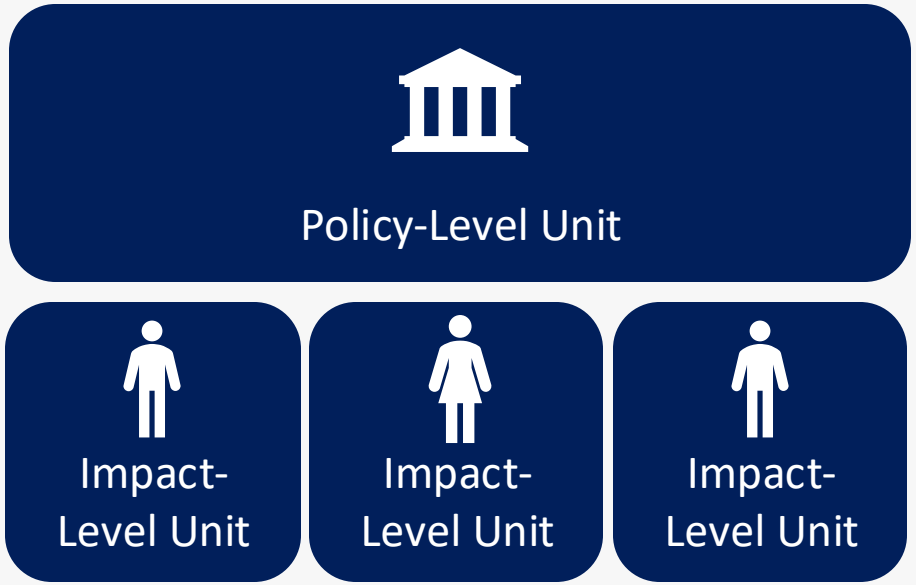
- Clear definitions of exposure & comparison conditions
- Clear thinking about “time zero”
- Clear identification of eligible units under study

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.



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.

Policy-Level Units

Policy-level units are

- units that *did* implement the policy 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 (e.g., “has not previously implemented policies X, Y, Z”)

Impact-Level Units

Impact-level units are those that the policy is designed to affect. Possibly

- 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 would be 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 impact-level units followed longitudinally

What are we studying?

WHAT DOES IT MEAN TO SAY $A_i = a$?

Defining the Exposure

"LEGAL EPIDEMIOLOGY"

- Rigorous & transparent framework for analysis of laws / policies
- 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.

JAMA Health Forum™

Viewpoint

Improving the Transparency of Legal Measurement in Health Policy Evaluation —A Guide for Researchers, Reviewers, and Editors

Benjamin A. Barsky, JD, PhD; Alina Schnake-Mahl, ScD, MPH; Cason D. Schmit, JD; Scott Burris, JD

Advances in quantitative research methods have led to acceptance that well-designed observational studies can enable causal inferences about the effects of policy interventions on health.[1] Indeed, *JAMA* recently announced exploring a range of study designs for which causal interpretations may be made in JAMA Network journals.[2] This development is welcome for research on the potential health effects of laws—including statutes, regulations, executive orders, formal agency guidance, and other documents that purport to require compliance with rules or standards—given that they are rarely amenable to evaluation using experimental methods. As researchers, reviewers, and editors refine criteria for assessing the appropriateness of causal claims in health policy evaluation research,[3] this Viewpoint emphasizes a critical methodological concern: the proper measurement of law.

Author affiliations and article information are listed at the end of this article.

Defining the Comparison Group

Best practices:

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, 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.

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.

Impact on the Estimand

Definition of exposure & comparison conditions necessarily changes the estimand

- What's $A = 1$? What's $A = 0$?

If using never-exposed comparators, $A = 0$ becomes "non-implementation for the entire post period"

- Now we're asking a different question!

Measurement & Outcomes

What are we measuring? What does the estimand actually mean?

All sorts of messiness here:

- Retrospective analysis limits us to already-collected data
- Questions about measurement
- Data might not be available for the target population

Even more complexity:

- Degree of implementation can be highly variable, but this can be very difficult to measure. (Work is ongoing!)

When are we studying the policy?

Policy evaluation is naturally longitudinal

We **must** have longitudinal data before and after policy implementation to evaluate the policy!

Policy effects are likely time-varying

- Ramp-ups, delayed effects, waning effects, etc.

Policies aren't implemented in a vacuum

- Policy-level units are doing other stuff around the same time!
- Legislative packages, phased rule changes, etc.
- "Current events are happening as we speak" -Matt Rogers, *Las Culturistas*

Griffin BA, *et al.* Methodological considerations for estimating policy effects in the context of co-occurring policies. *Health Serv Outcomes Res Method* 2023;**23**:149–65.

Study Periods

Defining a study period can be difficult!

- Estimation methods typically rely on “long-enough” pre-periods (e.g., synthetic controls)
  - But, if you go too far back, you could model “old” dynamics
- Longer follow-up periods are interesting: what’s the effect of the policy X years after implementation?
  - But, if other things happen, we could accidentally attribute “new” dynamics to the “old” policy.

Translation to an Estimand

Do we care about estimating a policy effect

- at a single point in time?
- over some specific follow-up period?
- over the entire follow-up period?

Current best practice is probably to estimate time-specific effects on as granular a level as possible, then think carefully about aggregation to a desired interval

- See, e.g., Callaway & Sant'Anna's "group-time ATTs"

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

Where are we studying the policy?

Place is (very likely) important

Policy-level units are commonly non-exchangeable

- States, e.g., can be quite different from each other

Geographically proximal controls can help improve face validity

- But, that could introduce interference or spillover.

Impact of Geography on the Estimand

Usual interference approaches bundle together units

- Estimand becomes effect on outcomes under “bundled” treatments across interfering units

Might not be ideal in settings with small N (e.g., state policy)

- Bundling units can lead to even less power

Who is this for?

"WHO" PART DEUX

Who is the question for?

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?"
- Requires a different counterfactual than what we usually estimate

Cairney P. The myth of 'evidence-based policymaking' in a decentred state. *Public Policy and Administration* 2022;**37**:46–66.

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*

Other Common Simple Estimands

Average treatment effect (ATE):

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

"What would be the effect of the policy if all units implemented it?"

Average treatment effect among comparators (ATC):

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

"What would be the effect of the policy if all non-implementing units implemented it?"

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.

We should try to get creative here!

Where can we go from here?

Scan for slides

Policy evaluation is hard.

Understanding the effects of a policy is really challenging by nature.

Design thinking can help, but engagement with substantive expertise and different ways of thinking (including qualitative research!) is critical.

We need to think creatively when building methods.

Scan for slides

Where can we go from here?

Throughout the session, we'll see creative thinking being used to address complex policy questions:

- How do we handle heterogeneous exposures? (Gary)
- How do we handle recurrent events? (Arman)
- How do we handle spillover & heterogeneity thereof? (Fei)