Perelman
SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

SOCIETY FOR EPIDEMIOLOGIC RESEARCH ANNUAL MEETING

Ecological regression in health policy evaluation: A guilt-free dessert?

Nicholas J. Seewald, Ph.D.
Assistant Professor of Biostatistics
June 13, 2025

DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS

Scan for slides

Motivating Example: Medical Cannabis Laws and Opioid Prescribing in the U.S.

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

Question: What are the effects of state medical cannabis laws on receipt of opioid pain treatment among patients with chronic non-cancer pain?

McGinty EE, Tormohlen KN, Seewald NJ, et al. Effects of U.S. State Medical Cannabis Laws on Treatment of Chronic Noncancer Pain. *Ann Intern Med*. 2023;176(7):904-912.

Scan for slides

Data for Health Policy Evaluation

Many health policy evaluations start with "disaggregated" individual-level data (e.g., insurance claims, EHR, etc.)

Intuitively, we like this!

- Allows more choices about the population of interest
  - Continuous enrollment, samples with certain diagnoses, etc.
- Allows outcome / covariate construction

BUT! Data becomes large, computational constraints kick in, and aren't policies inherently cluster-level interventions?

Perelman School of Medicine University of Pennsylvania

Scan for slides

Motivating Example: Medical Cannabis Laws and Opioid Prescribing in the U.S.

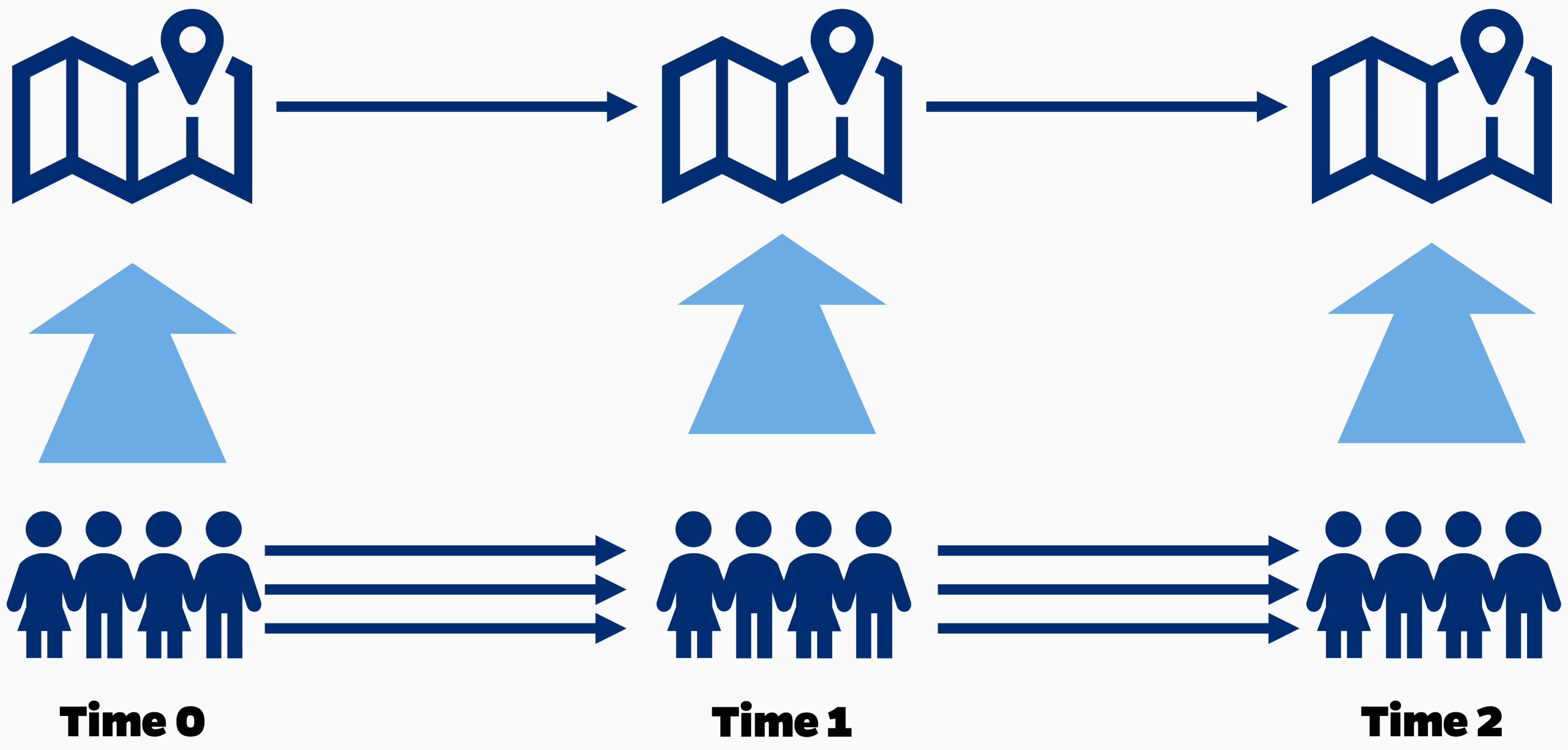
Data are individual-level commercial health insurance claims.

- Individuals included if they have a chronic non-cancer pain diagnosis pre-law and are continuously present in data for full study period
- Monthly data on diagnoses, opioid Rx, non-opioid Rx, pain procedures, etc.
- 7-year study periods → 84 measurement occasions per person

Computation is *extremely* expensive. **Can we aggregate to state-month without losing information?**

McGinty EE, Tormohlen KN, Seewald NJ, et al. Effects of U.S. State Medical Cannabis Laws on Treatment of Chronic Noncancer Pain. *Ann Intern Med*. 2023;176(7):904-912.

Unit-Time Aggregation




```
stats::aggregate(Y ~ state + time, data, mean)
```

The Ecological Fallacy

Data aggregation might introduce worries about ecological bias.

I argue it should not:

- Policies are inherently cluster-level
- Policy *scholars* think about cluster-level effects
- Policy*makers* think about cluster-level effects

So, can we just do ecological regression and be done with it?

Two Big Questions

1. Are difference-in-difference analyses using individual-level data **more statistically efficient** than those using aggregate-level data?
2. Does individual-level data allow for **better control of confounding**?

Difference-in-Differences

Consider a continuous outcome with all exposed units exposed simultaneously.

If exposure effect is constant, we can fit the **two-way fixed effects model:**

$$Y_{\gamma it} = \beta_{0\gamma} + \beta_{1t} + \beta_2 A_{\gamma t} + \epsilon_{\gamma it},$$

where

- γ indexes cluster (exposure units)
- i indexes individuals inside clusters
- t indexes time
- $A_{\gamma t} = 1$ iff unit γ is first exposed at or before time t

NOTE: i appears only in the error! With balanced clusters & no covariates, estimation & inference is identical for individual- and aggregate-level data.

Ecological Regression?

$$Y_{\gamma it} = \beta_{0\gamma} + \beta_{1t} + \beta_2 A_{\gamma t} + \epsilon_{\gamma it}$$

vs.

$$\bar{Y}_{\gamma it} = \beta_{0\gamma} + \beta_{1t} + \beta_2 A_{\gamma t} + \bar{\epsilon}_{\gamma it}$$

Differences in these models might arise from:

1. Covariate adjustment
2. Clustering standard errors

Perelman SCHOOL OF MEDICINE UNIVERSITY of PENNSYLVANIA

Simulation Study: Generative Model

Idea: Simulate data from a simple but flexible generative model and analyze it using various approaches.

$$Y_{\gamma it} = \beta_0 + \beta_1(t) + \beta_2 A_{\gamma t} + \beta_3\big((t - t_*)_+\big)A_{\gamma t} + \boldsymbol{\eta}_t^\top \boldsymbol{X}_{\gamma it} + \boldsymbol{\xi}_t^\top \boldsymbol{X}_{\gamma it} A_{\gamma t} + b_{\gamma i} + c_{\gamma t} + \epsilon_{\gamma it}$$

This allows for:

- Time-varying treatment effects
- Time-varying covariate effects
- Time-varying effect modification
- Complex dependency structures across observations

Simulation Study: Setting

Limited, but common settings:

- Continuously-enrolled sample (i.e., closed cohorts)
- Balanced panels
- Simultaneous exposure
- Similar number of treated and control states (Rokicki et al. 2018)

Analytic approaches are <u>extremely</u> mechanical:
fit two-way fixed effects model and cluster SEs

Rokicki S, Cohen J, Fink G, Salomon JA, Landrum MB. Inference With Difference-in-Differences With a Small Number of Groups: A Review, Simulation Study, and Empirical Application Using SHARE Data. *Medical Care*. 2018;56(1):97-105.

Correlation Structures

We consider three types of dependency in the data:

- Within-individual correlation: $\mathrm{Cor}(Y_{\gamma it}, Y_{\gamma is}) =: \rho_{ts}$
- Within-period correlation: $\mathrm{Cor}(Y_{\gamma it}, Y_{\gamma jt}) =: \phi_t$
- Between-period correlation: $\mathrm{Cor}(Y_{\gamma it}, Y_{\gamma js}) =: \psi_{ts}$

Generally, $\psi \leq \phi < \rho$

"Block Exchangeable" Correlation, No Covariates

$$Y_{\gamma it} = \beta_0 + \beta_t t + \beta_2 A_{\gamma t} + \beta_3 (t - t_*)_+ A_{\gamma t} + b_{\gamma i} + \boldsymbol{c}_{\boldsymbol{\gamma}} + \epsilon_{\gamma it}$$

Within-person correlation $\rho = 0.3$

Within-period correlation $\phi = 0.2$

Between-period correlation $\psi = 0.2$

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	0.0	0.019	0.948
SE clustered by state	0.0	0.019	0.948
Individual-Level Data			
OLS SE	0.0	0.020	0.964
SE clustered by individual	0.0	0.019	0.942
SE clustered by state	0.0	0.019	0.940
SE clustered by individual and state	0.0	0.019	0.940
SE clustered by state and time	0.0	0.019	0.924
True mixed model	0.0	0.019	0.944

Just use the aggregated data!

"Nested Exchangeable" Correlation, No Covariates

$$Y_{\gamma it} = \beta_0 + \beta_t t + \beta_2 A_{\gamma t} + \beta_3 (t - t_*)_+ A_{\gamma t} + b_{\gamma i} + \boldsymbol{c}_{\boldsymbol{\gamma t}} + \epsilon_{\gamma it}$$

Within-person correlation $\rho = 0.3$

Within-period correlation $\phi = 0.2$

Between-period correlation $\psi = 0.1$

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	0.1	0.124	0.938
SE clustered by state	0.1	0.125	0.936
Individual-Level Data			
OLS SE	0.1	0.023	0.302
SE clustered by individual	0.1	0.020	0.266
SE clustered by state	0.1	0.122	0.926
SE clustered by individual and state	0.1	0.122	0.926
SE clustered by state and time	0.1	0.122	0.916
True mixed model	0.1	0.124	0.944

Individual-level analysis <u>must</u> correctly cluster SEs.

Perelman School of Medicine University *of* Pennsylvania

Confounding in Diff-in-Diff

"Only covariates that differ by treatment group and are associated with outcome *trends* are confounders in diff-in-diff."

- Time-invariant covariates are confounders if they have time-varying effects on the outcome
- Time-varying covariates are confounders if they have time-varying effects on the outcome or evolve differently in treated and control groups.

Zeldow B, Hatfield LA. Confounding and regression adjustment in difference-in-differences studies. *Health Services Research*. 2021;56(5):932-941.

Block Exchangeable Correlation, Unconfounded

$$Y_{\gamma it} = \beta_0 + \beta_t t + \beta_2 A_{\gamma t} + \beta_3 (t - t_*)_+ A_{\gamma t} + \boldsymbol{\eta_1 X_{\gamma i}} + b_{\gamma i} + \boldsymbol{c_\gamma} + \epsilon_{\gamma it}$$

Within-person correlation $\rho = 0.3$

Within-period correlation $\phi = 0.2$

Between-period correlation $\psi = 0.2$

Results shown for correctly adjusted models.

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	0.1	0.030	0.950
SE clustered by state	0.1	0.030	0.940
Individual-Level Data			
OLS SE	0.1	0.032	0.958
SE clustered by individual	0.1	0.030	0.946
SE clustered by state	0.1	0.029	0.928
SE clustered by individual and state	0.1	0.029	0.929
SE clustered by state and time	0.1	0.029	0.932
True mixed model	0.1	0.030	0.948

Just use the aggregated data!

Nested Exchangeable Correlation, Unconfounded

$$Y_{\gamma it} = \beta_0 + \beta_t t + \beta_2 A_{\gamma t} + \beta_3 (t - t_*)_+ A_{\gamma t} + \boldsymbol{\eta_1 X_{\gamma i}} + b_{\gamma i} + \boldsymbol{c_{\gamma t}} + \epsilon_{\gamma it}$$

Within-person correlation $\rho = 0.3$

Within-period correlation $\phi = 0.2$

Between-period correlation $\psi = 0.1$

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	−0.1	0.195	0.938
SE clustered by state	−0.1	0.195	0.936
Individual-Level Data			
OLS SE	−0.1	0.037	0.294
SE clustered by individual	−0.1	0.020	0.262
SE clustered by state	−0.1	0.187	0.924
SE clustered by individual and state	−0.1	0.187	0.924
SE clustered by state and time	−0.1	0.185	0.903
True mixed model	−0.1	0.195	0.946

Individual-level analysis must correctly cluster SEs and is still slightly inefficient. **Weird!**

Perelman School of Medicine
University of Pennsylvania

Block Exchangeable Correlation, Confounded

$$Y_{\gamma it}$$
$$= \beta_0 + \beta_t t + \beta_2 A_{\gamma t}$$
$$+ \beta_3 (t - t_*)_+ A_{\gamma t}$$
$$+ \boldsymbol{\eta_1}(\boldsymbol{t}) \boldsymbol{X_{\gamma i}}$$
$$+ b_{\gamma i} + \boldsymbol{c_\gamma} + \epsilon_{\gamma it}$$

$$E\left[X_{\gamma i} \mid A_{\gamma T} = 1\right] = 5$$
$$\mathrm{E}\left[X_{\gamma i} \mid A_{\gamma T} = 1\right] = 2$$

Results shown for correctly adjusted models.

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	0.3	0.793	0.968
SE clustered by state	0.3	0.732	0.910
Individual-Level Data			
OLS SE	0.0	0.058	0.972
SE clustered by individual	0.0	0.054	0.958
SE clustered by state	0.0	0.053	0.942
SE clustered by individual and state	0.0	0.053	0.942
SE clustered by state and time	0.0	0.050	0.906
True mixed model	0.0	0.054	0.960

Block Exchangeable Correlation, Confounded

$$
\begin{aligned}
Y_{\gamma it} &= \beta_0 + \beta_t t + \beta_2 A_{\gamma t} \\
&+ \beta_3 (t - t_*)_+ A_{\gamma t} \\
&+ \boldsymbol{\eta_1(t) X_{\gamma i}} \\
&+ b_{\gamma i} + \boldsymbol{c_\gamma} + \epsilon_{\gamma it}
\end{aligned}
$$

$$E\left[X_{\gamma i} \mid A_{\gamma T} = 1\right] = 5$$
$$\mathrm{E}\left[X_{\gamma i} \mid A_{\gamma T} = 1\right] = 2$$

Results shown for correctly adjusted models.

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	0.3	0.793	0.968
SE clustered by state	0.3	0.732	0.910
Individual-Level Data			
OLS SE	0.0	0.058	0.972
SE clustered by individual	0.0	0.054	0.958
SE clustered by state	0.0	0.053	0.942
SE clustered by individual and state	0.0	0.053	0.942
[illegible]	0.0	0.050	0.906
[illegible]	.0	0.054	0.960

When time-invariant confounder is imbalanced at baseline, aggregation leads to efficiency loss

Perelman School of Medicine University of Pennsylvania

Nested Exchangeable Correlation, Confounded

$$
\begin{aligned}
&Y_{\gamma it} \\
&= \beta_0 + \beta_t t + \beta_2 A_{\gamma t} \\
&+ \beta_3 (t - t_*)_+ A_{\gamma t} \\
&+ \boldsymbol{\eta_1}(\boldsymbol{t}) \boldsymbol{X_{\gamma i}} \\
&+ b_{\gamma i} + \boldsymbol{c_{\gamma t}} + \epsilon_{\gamma it}
\end{aligned}
$$

$$E\left[X_{\gamma i} \mid A_{\gamma T} = 1\right] = 5$$
$$\mathrm{E}\left[X_{\gamma i} \mid A_{\gamma T} = 1\right] = 2$$

Results shown for correctly adjusted models.

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	−3.7	5.124	0.958
SE clustered by state	−3.7	4.766	0.910
Individual-Level Data			
OLS SE	0.0	0.066	0.516
SE clustered by individual	0.0	0.058	0.448
SE clustered by state	0.0	0.195	0.936
SE clustered by individual and state	0.0	0.195	0.936
SE clustered by state and time	0.0	0.192	0.928
True mixed model	0.0	0.201	0.964

Nested Exchangeable Correlation, Confounded

Individual-level CIs slightly under-cover, but are orders of magnitude more efficient unless you also adjust for state-level covariate means

NOTE: Without adjusting for cluster-level means, individual-level analysis answers an individual-level question. (not what we want!)

	% Bias	Std. Err.	95% CI Covg.
Aggregated Data (ecological models)			
OLS SE	-3.7	5.124	0.958
SE clustered by state	-3.7	4.766	0.910
Individual-Level Data			
OLS SE	0.0	0.066	0.516
SE clustered by individual	0.0	0.058	0.448
SE clustered by state	0.0	0.195	0.936
SE clustered by individual and state	0.0	0.195	0.936
SE clustered by state and time	0.0	0.192	0.928
True mixed model	0.0	0.201	0.964

Takeaways

This is a question of **design vs. analysis.**

- Individual-level data is very useful in the *design stage* of policy evaluation
  - Better sample identification, feature construction, outcome construction, etc.
- In the analysis stage (with DiD), aggregate-level data is more *ergonomic* and usually yields CIs with nominal coverage.
  - **Analyses using individual-level might struggle to achieve nominal coverage and can suffer when complex correlations are modeled wrong.**

It's hard to distinguish what's an issue with aggregation and what's an issue with model misspecification.

Perelman
SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

Perelman
SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

seewaldn@pennmedicine.upenn.edu

www.nickseewald.com