Practical Guidance on Whether and When to Aggregate Individual-Level Data for Causal Health Policy Evaluation

Nicholas J. Seewald, PhD

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

16 May 2024

American Causal Inference Conference

Perelman
SCHOOL OF MEDICINE
UNIVERSITY *of* PENNSYLVANIA

This is a weird talk.

I’m $\sim$80% sure I have some idea of what’s going on.

This is a pitch to get you to come talk to me so I can get to $\sim$90%.

Policy Evaluation is Hard

- Sample size is often quite limited
- “Policy-level” units are large and meaningful (e.g., states)

Do Medical Cannabis Laws Change 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 among adults with chronic non-cancer pain

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

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

Individual-Level Data in Health Policy Evaluation

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

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

But many methods use/require *aggregated* (i.e., policy-level unit-time) data. Is that okay?

Individual-Level Data is Better, Right?

Intuition suggests that individual-level data would be better than aggregated data:

- More data is more information
- Adjust for individual-level confounding
- Appropriately account for nuanced functional forms

But “treatment” is at the state level.

Medical Cannabis Study: Data

Data are individual-level commercial health insurance claims.

- Individuals included if they have a chronic non-cancer pain diagnosis in the pre-law period **and** are continuously enrolled in commercial health insurance for the full study period.
- Monthly data on diagnoses, opioid and non-opioid pain prescriptions, procedures, etc.
- No data on cannabis use, OTC pain treatments, etc. (things not covered by insurance)

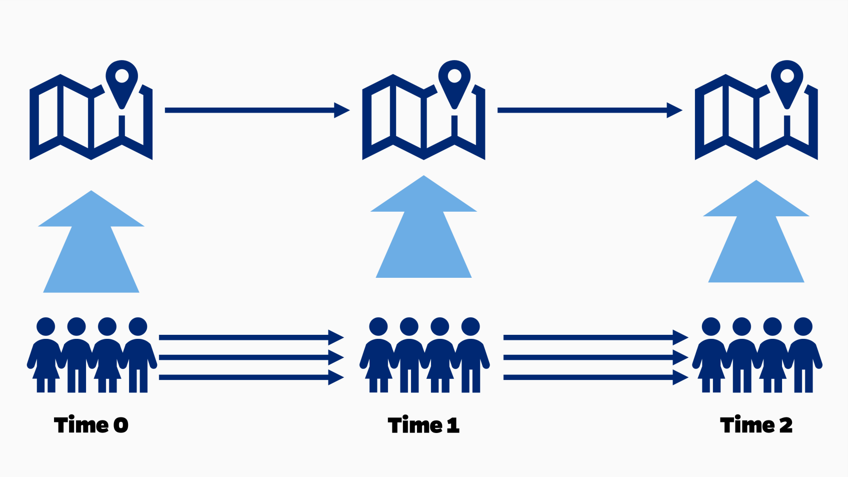
We have rich data on individual outcome trajectories, and think we should use it!

Computational Feasibility

Oftentimes, data has to be analyzed on remote servers.

Computational resources are often very constrained: if we can use smaller data without losing much, that'd be great.

Unit-Time Aggregation



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

Questions

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

Difference-in-Differences

Right now, let's think of diff-in-diff as the two-way fixed-effects model for a continuous outcome (and simultaneous treatment adoption in treated units).

With individual-level data,

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

where γ indexes policy-level unit, i individual, and t time, and $A_{\gamma t}$ is 1 iff unit γ implemented the policy at or before time t.

Difference-in-Differences

Right now, let's think of diff-in-diff as the two-way fixed-effects model for a continuous outcome (and simultaneous treatment adoption in treated units).

With individual-level data,

$$Y_{\gamma i t} = \beta_{0\gamma} + \beta_{1t} + \beta_2 A_{\gamma t} + \varepsilon_{\gamma i t},$$

where γ indexes policy-level unit, i individual, and t time, and $A_{\gamma t}$ is 1 iff unit γ implemented the policy at or before time t.

The individual-level index appears only in the error! Without covariates and assuming balanced cluster sizes, estimation & inference should be identical for individual-level and aggregated data.

Difference-in-Differences

$$Y_{\gamma i t} = \beta_{0\gamma} + \beta_{1t} + \beta_2 A_{\gamma t} + \varepsilon_{\gamma i t}$$

vs.

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

Differences might come from:

1. Covariate adjustment
2. Clustering standard errors

Simulation Study: Generative Model

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

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

- $A_{\gamma t} = \mathbb{1}\{\text{unit } \gamma \text{ is treated at time } t\}$
- t^* is the first post-treatment timepoint
- $\mathbf{X}_{\gamma it}$ is a vector of covariates
- $b_{\gamma i}$ and $c_{\gamma t}$ are random intercepts for individual and policy-level unit-time.

Simulation Study: Generative Model

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

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

- Random effects induce three distinct correlations:
  - Within-person correlation
  - Within-period correlation
  - Between-period correlation
- Time-varying treatment effects and effect heterogeneity are allowed
- Necessarily simpler than real data!

Simulation Study: Setting

Current focus has been on limited but common settings

- Continuously-enrolled sample (i.e., no changing case mix)
- Balanced panels
- Simultaneous treatment adoption
- Similar number of treated and control states (Rokicki et al. 2018)

Rokicki, S. et al. (2018). *Medical Care*.

Simulation Study: Setting

Current focus has been on limited but common settings

- Continuously-enrolled sample (i.e., no changing case mix)
- Balanced panels
- Simultaneous treatment adoption
- Similar number of treated and control states (Rokicki et al. 2018)

Analytic strategies are, and I cannot emphasize this enough, **entirely mechanical**.

Rokicki, S. et al. (2018). *Medical Care*.

Clustered Standard Errors, No Covariates

Moderate within- and between-person correlation: $\text{ICC}_{\text{indiv}} = 0.5$, $\text{ICC}_{\text{policy}} = 0.4$. 2000 simulations, 500 individuals per state.

$$Y_{\gamma it} = \beta_0 + \beta_1 t + \beta_2 A_{\gamma t} + \beta_3 (t - t^*)_+ A_{\gamma t} + b_{\gamma i} + c_{\gamma t} + \varepsilon_{\gamma it}$$

	Bias	SE	95% Coverage
Individual data, OLS SE	0.000	0.014	0.971
Individual data, person-clustered SE	0.000	0.013	0.955
Individual data, state-clustered SE	0.000	0.012	0.928
Aggregate data, OLS SE	0.000	0.013	0.953
Aggregate data, state-clustered SE	0.000	0.013	0.954

Confounding in Difference-in-Differences

“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. and Hatfield, L. A. (2021). *Health Services Research*.

Time-Invariant Covariates

$$Y_{\gamma it} = \beta_0 + \beta_1 t + \beta_2 A_{\gamma t} + \beta_3 X_{\gamma i} + b_{\gamma i} + c_{\gamma t} + \epsilon_{\gamma it}$$

Aggregate analysis model can't adjust for X: $\bar{X}_s$ is collinear with state fixed effects.

Time-Invariant Covariate, Time-Invariant Effect

$$Y_{\gamma it} = \beta_0 + \beta_1 t + \beta_2 A_{\gamma t} + \beta_3 X_{\gamma i} + b_{\gamma i} + c_{\gamma t} + \epsilon_{\gamma it}$$

	Bias	SE	RMSE	95% Coverage
Individual, unadj., OLS SE	0.000	0.030	0.013	1.000
Individual, unadj., person-clustered SE	0.000	0.013	0.013	0.942
Individual, unadj., state-clustered SE	0.000	0.012	0.013	0.922
Individual, adj., OLS SE	0.000	0.014	0.013	0.965
Individual, adj., person-clustered SE	0.000	0.013	0.013	0.942
Individual, adj., state-clustered SE	0.000	0.012	0.013	0.922
Aggregated, unadj., OLS SE	0.000	0.013	0.013	0.942
Aggregated, unadj., state-clustered SE	0.000	0.013	0.013	0.945

Time-Invariant Covariate, Time-Varying Effect

$$Y_{\gamma it} = \beta_0 + \beta_1 t + \beta_2 A_{\gamma t} + \beta_3 X_{\gamma i} + \beta_4 t X_{\gamma i} + b_{0,s} + b_{0,\gamma i} + \epsilon_{\gamma it}$$

	Bias	SE	RMSE	95% Coverage
Individual, unadj., OLS SE	5.182	0.043	5.182	0.000
Individual, unadj., person-clustered SE	5.182	0.075	5.182	0.000
Individual, unadj., state-clustered SE	5.182	1.410	5.182	0.000
Individual, adj., OLS SE	0.000	0.027	0.015	0.999
Individual, adj., person-clustered SE	0.000	0.015	0.015	0.959
Individual, adj., state-clustered SE	0.000	0.015	0.015	0.917
Aggregated, unadj., OLS SE	0.000	0.017	0.016	0.954
Aggregated, unadj., state-clustered SE	0.000	0.017	0.016	0.930

Time-Varying Covariate, Time-Invariant Effect

$$Y_{\gamma i t} = \beta_0 + \beta_1 t + \beta_2 A_{\gamma t} + \beta_3 X_{\gamma i} + \beta_4 X_{\gamma i t} + b_{0,s} + b_{0,\gamma i} + \epsilon_{\gamma i t} \qquad X_{\gamma i} \sim \mathcal{N}(\mu, \Sigma)$$

	Bias	SE	RMSE	95% Coverage
Individual, unadj., OLS SE	0.000	0.025	0.024	0.963
Individual, unadj., person-clustered SE	0.000	0.018	0.024	0.833
Individual, unadj., state-clustered SE	0.000	0.024	0.024	0.934
Individual, adj., OLS SE	0.000	0.022	0.013	0.999
Individual, adj., person-clustered SE	0.000	0.013	0.013	0.958
Individual, adj., state-clustered SE	0.000	0.012	0.013	0.934
Aggregated, unadj., OLS SE	0.000	0.025	0.024	0.962
Aggregated, unadj., state-clustered SE	0.000	0.026	0.024	0.960
Aggregated, adj., OLS SE	0.000	0.013	0.013	0.956
Aggregated, adj., state-clustered SE	0.000	0.013	0.013	0.962

Time-Varying Covariate, Time-Varying Effect

$$Y_{\gamma it} = \beta_0 + \beta_1 t + \beta_2 A_{\gamma t} + \beta_3 X_{\gamma i} + \beta_4 t X_{\gamma it} + b_{0,s} + b_{0,\gamma i} + \epsilon_{\gamma it}$$

$X_{\gamma it}$ is linear in time

	Bias	SE	RMSE	95% Coverage
Individual, unadj., OLS SE	9.949	0.037	9.949	0.000
Individual, unadj., person-clustered SE	9.949	0.018	9.949	0.000
Individual, unadj., state-clustered SE	9.949	0.024	9.949	0.000
Individual, adj., OLS SE	-0.001	0.059	0.082	0.845
Individual, adj., person-clustered SE	-0.001	0.081	0.082	0.940
Individual, adj., state-clustered SE	-0.001	0.079	0.082	0.935
Aggregated, unadj., OLS SE	9.949	0.071	9.949	0.000
Aggregated, unadj., state-clustered SE	9.949	0.215	9.949	0.000
Aggregated, adj., OLS SE	0.005	0.146	0.145	0.956
Aggregated, adj., state-clustered SE	0.005	0.133	0.145	0.895

Results

What we've seen so far:

- Differences in efficiency, if they exist, are small
- Seemingly quite similar bias control
- Individual-level data is harder to work with than aggregated data
- Individual-level data might be better if you're adjusting for complicated time-varying confounders

My current thinking

We think this is a question of **design** vs. **analysis**.

- Individual-level data is incredibly useful in *the design stage* of a policy evaluation!
    - Better sample identification, feature construction, outcome construction, etc.
- It's hard to distinguish between what's actually an issue with aggregation and what's model misspecfication.
- In *the analysis stage* (with diff-in-diff), aggregate-level data is more ergonomic and seems more or less the same.