Ready to Roll? Practical Guidance on Whether and When to Aggregate Data in Health Policy Evaluation

Nicholas J. Seewald, PhD

Department of Health Policy and Management
Johns Hopkins Bloomberg School of Public Health

Joint with K. Tormohlen, E.E. McGinty, and E.A. Stuart

ICHPS 2023, 10 January 2023

JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Slides are online!

slides.nickseewald.com/ichps2023.pdf

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* 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.

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*.

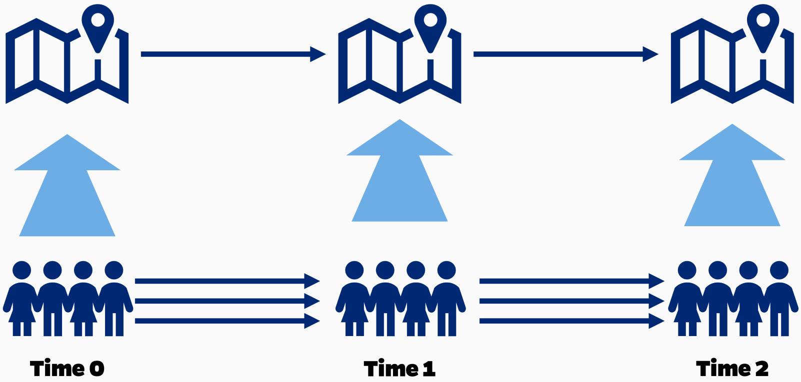
Medical Cannabis Study: Data

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.

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

State-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**?

Simulation Study: Generative Model

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

$$Y_{sit} = \beta_0 + \beta_1 t \beta_2 A_{st} + \beta_3 (t_k - t^*)_+ A_{st} + \eta^\top \mathbf{X}_{sit} + \gamma^\top \mathbf{X}_{sit} A_{st} \\ + b_{0,s} + b_{0,si} + b_{0,st} + \varepsilon_{sit}$$

- $A_{st} = \mathbb{1}\{\text{state } s \text{ is treated at time } t\}$
- t^* is the first post-treatment timepoint
- $\mathbf{X}_{sit}$ is a vector of covariates
- $b_{0,s}, b_{0,si}, b_{0,st}$ are state-, person-, and time-level random intercepts

Simulation Study: Generative Model

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

$$Y_{sit} = \beta_0 + \beta_1 t \beta_2 A_{st} + \beta_3 (t_k - t^*)_+ A_{st} + \eta^\top \mathbf{X}_{sit} + \gamma^\top \mathbf{X}_{sit} A_{st} + b_{0,s} + b_{0,si} + b_{0,st} + \varepsilon_{sit}$$

- 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)

Analytic approaches considered are "naive"

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

Question 1: Do we lose information in aggregated analyses?

A preview: OLS

OLS estimators are identical for individual- and aggregate-level data in a two-way fixed effects model

Individual-level model:

$$Y_{sit} = \beta_{0,s} + \beta_{1,t} + \beta_2 A_{st} + \varepsilon_{sit}$$

Aggregate-level model:

$$Y_{s\cdot t} = \beta_{0,s} + \beta_{1,t} + \beta_2 \bar{A}_{st} + \varepsilon_{st}$$

Differences might arise from clustering standard errors or introducing covariates.

Clustered Standard Errors, No Covariates

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

	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

Question 2: Do individual-level models allow better control of confounding?

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 Covariate, Time-Invariant Effect

$$Y_{sit} = \beta_0 + \beta_1 t + \beta_2 A_{st} + \beta_3 X_{si} + b_{0,s} + b_{0,si} + \epsilon_{sit}$$

	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

$$Y_{sit} = \beta_0 + \beta_1 t + \beta_2 A_{st} + \beta_3 X_{si} + \beta_4 t X_{si} + b_{0,s} + b_{0,si} + \epsilon_{sit}$$

	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

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

	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

$$Y_{sit} = \beta_0 + \beta_1 t + \beta_2 A_{st} + \beta_3 X_{si} + \beta_4 t X_{sit} + b_{0,s} + b_{0,si} + \epsilon_{sit}$$

X_{sit} 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

Conclusions

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.
- In *the analysis stage* (with diff-in-diff), aggregate-level data is more ergonomic and seems more or less the same.

Acknowledgements

- NIDA R01DA049789 (PI: McGinty)
- Elizabeth Stuart, Beth McGinty, Kayla Tormohlen
- Beth Ann Griffin, Laura Hatfield, Carrie Fry, Avi Feller, Eli Ben-Michael, Mariel Finucane, Dan Thal, Colleen Barry
- **Maybe you??**

nseewal1@jhu.edu

@nickseewald@mathstodon.xyz