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

Nicholas J. Seewald, PhD   Kayla Tormohlen, PhD   Emma E. (Beth) McGinty, PhD   Elizabeth A. Stuart, PhD

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Research Question

Health policy researchers often have questions about the effects of state policy on individual-level outcomes collected over multiple time periods.

Example: Limited evidence suggests that medical cannabis may be an effective substitute for opioids in pain management. This raises a question about the effect of medical cannabis laws on receipt of opioid treatment among individuals with chronic non-cancer pain This might be addressed using a large health insurance claims database which would track individuals' receipt of such treatment.

Individual-level longitudinal insurance claims data is very large, and can be difficult to work with. "Rolling up" (aggregating) data to, e.g., state-months can make it much easier to work with.

When, if ever, can a researcher roll up individual-level data to answer a question about the effects of a health policy?

Individual vs. Aggregate Data: Pros & Cons

Individual-level longitudinal data:

- Lots of data! Rich data on individual trajectories over time seems like it would be useful to use when assessing effects of policy.
- Likely requires big data techniques to analyze.

Aggregate-level longitudinal data:

- Significantly easier to work with: doesn't require big data techniques.
- Intuition suggests that rolling up individual data might lead to loss of statistical efficiency.

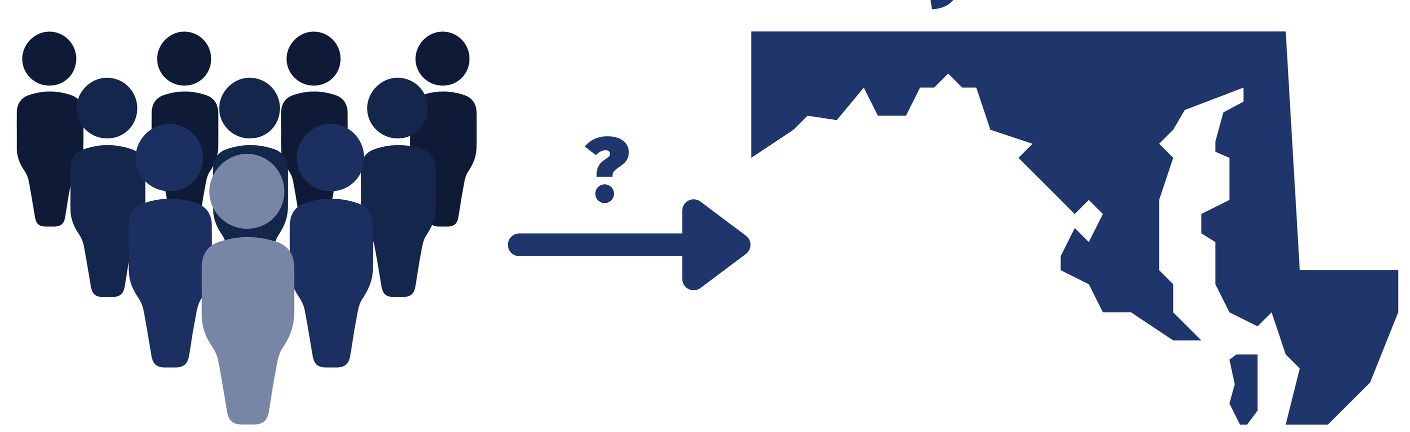
Simulation Study Design

We designed a simulation study to mimic individual-level data from a large-scale longitudinal administrative database (e.g., health insurance claims data). The data are generated from the model

$$Y_{sit} = \beta_0 + \beta_1 t + \beta_2 A_{st} \mathbb{1}\{t \geq 5\} + \beta_3(t) X_{sit} + b_{0si} + b_{0s} + \epsilon_{sit}$$

where s indicates state, i an individual, and t the measurement occasion. We use random intercepts (b_0's) for states and individuals to induce both within- and between-person correlation in states: at any given time, individuals' observations are related to their own past and future observations, as well as observations from other individuals in their state. The covariate X_{sit} is allowed to vary over time and have a time-varying effect on the outcome. We simulate 10 simultaneously-treated states and 10 control states, each with 10 measurements and 500 individuals per state.

Intuition says the more data the better. Is that always true?



We consider the following scenarios:

- No effect of X_{sit}
- Constant covariate with constant effect (i.e., $X_{sit} = X_{si}$ and $\beta_3(t) = \beta_3$)
- Constant covariate with time-varying effect (i.e., $X_{sit} = X_{si}$)
- Time-varying covariate with constant effect (i.e., $\beta_3(t) = \beta_3$)
- We vary the within- and between-person correlation between between 0.1 and 0.5.

Modeling Approaches

Individual-level models:

- State and time fixed effects with (1) no cluster standard error (SE) adjustment, (2) cluster SE adjustment at individual level, (3) cluster SE adjustment at state level

Aggregate-level models:

- State and time (two-way) fixed effects with and without state cluster SE adjustment
- Generalized estimating equations (GEE) with an exchangeable working correlation structure

In all models:

- When we include a time-varying covariate, we estimate separate effects at each timepoint (i.e., we interact with the time fixed effects).
- We fit the (non-GEE) models with ordinary least-squares (OLS) regression and optionally cluster adjust SEs.

Selected Preliminary Results

Below, we show standard errors, root mean squared errors, and 95% confidence interval coverage for 2000 simulations in a setting in which there is a **time-varying covariate** that evolves in the same way in both treatment and control groups and has a **constant effect** on the outcome. *Results from other scenarios are available online: scan the QR code!*

Model	SE	RMSE	95% CI Coverage
Individual w/ OLS SE	0.017	0.016	96.8%
Individual w/ Indiv. SE	0.016	0.016	95.1%
Individual w/ State SE	0.015	0.016	91.4%
Aggregate w/ OLS SE	0.018	0.019	95.0%
Aggregate w/ State SE	0.018	0.019	91.7%
Aggregate GEE	0.011	0.013	88.7%

Conclusions & Takeaways

- True treatment effect was recovered in all scenarios in which we expected unconfoundedness.
- State-level cluster adjustment of standard errors in individual-level models resulted in inappropriate deflation and undercoverage of 95% conf. intervals
- Because polices are implemented at the state level, individual-level information does not appear to be useful in estimating policy effects, based on limited simulations.

Based on (so far) limited simulations, we have initially found that using individual-level data offers no meaningful gain in statistical efficiency versus aggregate-level data in evaluating state health policy.

Future Work

- Expand the variety of scenarios under which we simulate to better identify times when individual-level data might be useful.
- More realistic within-person correlation structures, including AR(1)
- Expansion to additional outcome types, including binary and count outcomes. This poses an additional challenge: individual-level and aggregate models may be on different scales.

Acknowledgements

Research reported in this publication was supported by the National Institute on Drug Abuse of the National Institutes of Health under award number **R01DA049789**. The content of this poster is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Get in touch!

Email: nseewal1@jhu.edu
Twitter: @nickseewald
Web: nickseewald.com

Scan Me! Poster and additional materials