Supplementary Material for “Ready to Roll?” Poster at AcademyHealth ARM 2022

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

2022-06-06

Setting

We will consider the setting of a state policy evaluation in which individual-level data is available through, say, a health insurance claims database or other large-scale longitudinal administrative database. The outcome of interest is collected on all individuals in the sample a total of T times. Each individual is continuously enrolled in their health insurance plan, and so is available for outcome measurement at each occasion. Each individual lives in exactly one state for the entire study period, and, at least initially, we consider only one treated state, treated at time $t^* \in \{t_1, \dots, t_T\}$.

The outcome collected on the ith individual in the sth state at the tth measurement occasion is denoted Y_{sit}, where $t = 1, \dots, T$, $i = 1, \dots, n_s$ and $s = 1, \dots S$. We assume measurement occasions are common to all individuals in the study, and that T_{post} occasions were at or after the time of treatment.

We will generate data at the individual level; therefore, we can imagine simulating from a three-level hierarchical model: time is nested in individuals, which are nested in states. Denote by Y_{sit} the outcome at time k for individual j in state i. In general, we generate data for S states, S_{tx} of which are treated. There are n_s individuals included in state s, each of which are observed at T_{obs} total measurement occasions. T_{post} measurement occasions are at or after the time of treatment.

There seems to be a serious issue with using state-level cluster adjustments to the standard errors when the number of treated states is small. This is backed up by Rokicki et al. (2018), who found that confidence interval coverage was severely compromised when the ratio of the number of treated units to the number of control units was far from 0.5. **Therefore, we proceed using 10 treated and 10 control states to avoid “sample size” issues.**

No Covariates, Nested Dependency Structure

We’ll start by simulating a continuous outcome with simple linear time trends and no covariates. The generative model for the outcome is

$$Y_{sit} = \beta_0 + \beta_1 t + \beta_2 A_{st} + b_{0si} + b_{0s} + \epsilon_{ijk}, \tag{1}$$

where $b_{0ij} \sim \mathcal{N}(0, \sigma^2_{\text{person}})$, $b_{0i} \sim \mathcal{N}(0, \sigma^2_{\text{state}})$, and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_{\text{error}})$. Note that eq. (1) involves only an intercept shift at the time of treatment: there is a constant treatment effect over time. Similarly, the generative model induces within-state correlation through a random intercept b_{0i} at the state level; therefore, the true state fixed effects are $\beta_0 + b_{0i}$.

We induce within-person and within-state correlation via the random effects. In particular, we choose σ^2_{person} and σ^2_{state} to achieve desired intra-cluster correlations (ICCs). Using the definitions of ICC at each level, we have

$$\sigma^2_{\text{person}} = \frac{(\text{ICC}_{\text{person}} - \text{ICC}_{\text{state}})\sigma^2_{\text{error}}}{1 - \text{ICC}_{\text{person}}} \tag{2}$$

and

$$\sigma^2_{\text{state}} = \frac{\text{ICC}_{\text{state}} \cdot \sigma^2_{\text{error}}}{1 - \text{ICC}_{\text{person}}}. \tag{3}$$

Low within- and between-person correlation

We choose parameters as follows. Note the low within-person and within-state ICCs.

	Value
β_0	15.00
β_1	0.10
β_2	0.20
σ^2_{error}	1.00
$\text{ICC}_{\text{person}}$	0.10
$\text{ICC}_{\text{state}}$	0.09

2000 simulations with 10 treated and 10 control states, 500 people per state, measurements per person:

	Estimate	SE	RMSE	Type-II Error Rate	95% CI Coverage
Individual data, OLS SE	0.199	0.013	0.013	0.000	0.953
Individual data, person-clustered SE	0.199	0.013	0.013	0.000	0.953
Individual data, state-clustered SE	0.199	0.012	0.013	0.000	0.925
Aggregate data, OLS SE	0.199	0.013	0.013	0.000	0.950
Aggregate data, state-clustered SE	0.199	0.013	0.013	0.000	0.950
Aggregate GEE, Exch. State Corr.	0.199	0.011	0.013	0.000	0.890

Table 1: Results of 2000 simulations assuming small dependency within people and within states. No covariates.

Moderate within- and between-person correlation

We choose parameters as follows. Note the low within-person and within-state ICCs.

	Value
β_0	15.00
β_1	0.10
β_2	0.20
σ^2_{error}	1.00
$\text{ICC}_{\text{person}}$	0.50
$\text{ICC}_{\text{state}}$	0.40

2000 simulations with 10 treated and 10 control states, 500 people per state, measurements per person:

	Estimate	SE	RMSE	Type-II Error Rate	95% CI Coverage
Individual data, OLS SE	0.200	0.014	0.012	0.000	0.971
Individual data, person-clustered SE	0.200	0.013	0.012	0.000	0.952
Individual data, state-clustered SE	0.200	0.012	0.012	0.000	0.929
Aggregate data, OLS SE	0.200	0.013	0.012	0.000	0.954
Aggregate data, state-clustered SE	0.200	0.013	0.012	0.000	0.947
Aggregate GEE, Exch. State Corr.	0.200	0.011	0.012	0.000	0.899

Table 2: Results of 2000 simulations assuming small dependency within people and within states. No covariates.

Time-Invariant Individual Covariate

Next, we introduce a single individual-level, time-invariant, normally-distributed covariate $X_{si} \sim \mathcal{N}(\mu_{X,s}, \sigma^2_{X,s})$. The means and variances of the X_{si} are allowed to vary by state, but are not required to. We introduce this covariate into the model in a variety of ways, following Zeldow and Hatfield (2021).

Here, it is important to note that we are now *actually* using individual-level data in the individual-level model, so we expect to see improvement in standard errors in the individual-level analysis versus the aggregate-level analysis.

Constant Effect

Let X_{si} have a distribution dependent on state i. When X_{si} is time-invariant and does not have a time-varying effect on the outcome Y_{sit}, X_{si} does not confound the relationship between treatment A and outcome Y (Zeldow and Hatfield 2021).

$$Y_{sit} = \beta_0 + \beta_1 t + \beta_2 A_{st} \mathbb{1}\{t > t^*\} + \beta_3 X_{si} + b_{0si} + b_{0s} + \epsilon_{sit}. \tag{4}$$

Note that state-level fixed effects are not explicitly included in the model; they are generated from the random intercept b_{0i}. That is, for state s, the value of the state fixed effect is $\beta_0 + b_{0s}$.

The individual-level simple (unadjusted) analysis model is

$$Y_{sit} = \sum_{j=1}^{S} \beta_{0,j} \mathbb{1}\{j = s\} + \sum_{k=1}^{T} \beta_{1,k} \mathbb{1}\{k = t\} + \beta_2 A_{st} + \epsilon_{sit}.$$

The individual-level covariate-adjusted analysis model is

$$Y_{sit} = \sum_{j=1}^{S} \beta_{0,j} \mathbb{1}\{j = s\} + \sum_{k=1}^{T} \beta_{1,k} \mathbb{1}\{k = t\} + \beta_2 A_{st} + \beta_3 X_{si} + \epsilon_{sit}.$$

Similarly, the (desired) covariate-adjusted aggregate-level analysis model is

$$\bar{Y}_{s\cdot t} = \sum_{j=1}^{S} \beta_{0,j} \mathbb{1}\{j = s\} + \sum_{k=1}^{T} \beta_{1,k} \mathbb{1}\{k = t\} + \beta_2 A_{st} + \beta_3 \bar{X}_s + \epsilon_{st}.$$

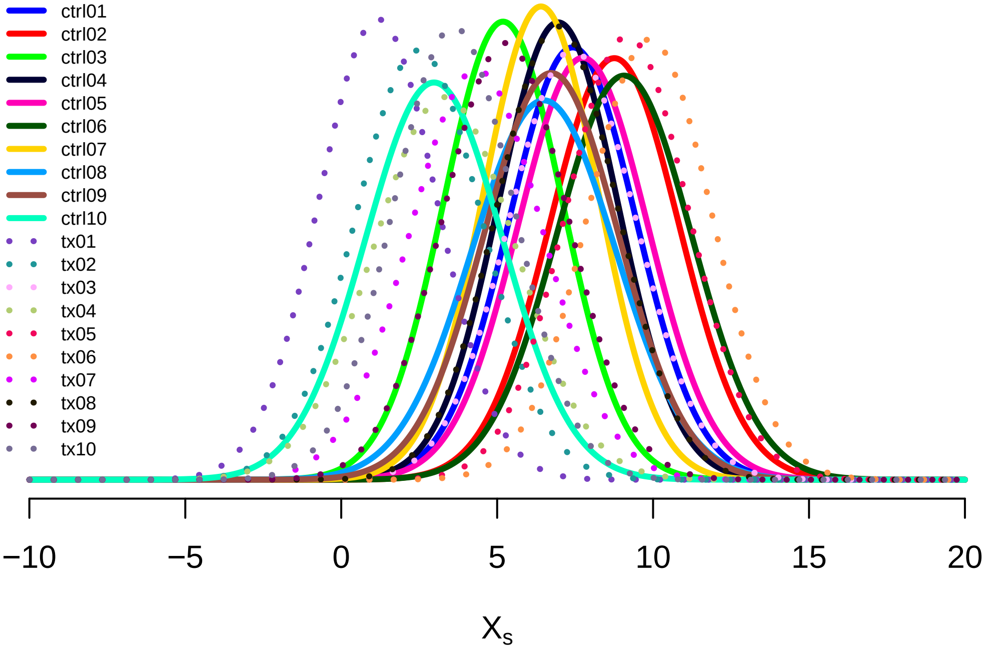
In reality, β_3 in the above aggregate model is unidentifiable: because $\bar{X}_s$ is a baseline covariate, it is collinear with the state fixed effect $\beta_{0,s}$. Therefore, it is omitted from the model, and we fit only the simple model:

$$\bar{Y}_{s\cdot t} = \sum_{j=1}^{S} \beta_{0,j} \mathbb{1}\{j = s\} + \sum_{k=1}^{T} \beta_{1,k} \mathbb{1}\{k = t\} + \beta_2 A_{st} + \epsilon_{st}.$$

	Value
β_1	0.10
β_2	0.20
β_3	1.00
σ^2_{error}	1.00

As above, intracluster correlation is induced using random effects. We choose parameters as follows (omitting fixed effects).

Per-state covariate distributions are depicted below.



	Estimate	SE	RMSE	Type-II Error Rate	95% CI Coverage
Individual Unadj. w/ OLS SE	0.200	0.029	0.013	0.000	1.000
Individual Unadj. w/ person-clustered SE	0.200	0.013	0.013	0.000	0.950
Individual Unadj. w/ state-clustered SE	0.200	0.012	0.013	0.000	0.926
Individual CA w/ OLS SE	0.200	0.014	0.013	0.000	0.964
Individual CA w/ person-clustered SE	0.200	0.013	0.013	0.000	0.950
Individual CA w/ state-clustered SE	0.200	0.012	0.013	0.000	0.926
Aggregate data w/ OLS SE	0.200	0.013	0.013	0.000	0.947
Aggregate data w/ state-clustered SE	0.200	0.013	0.013	0.000	0.938
Aggregate GEE w/ Exch. State Corr.	0.200	0.011	0.013	0.000	0.891

Table 3: Results of 2000 simulations assuming small dependency within people and within states. One time invariant covariate. CA = covariate adjusted.

References

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. doi:10.1097/MLR.0000000000000830