SCHOOL OF PUBLIC HEALTH · UNIVERSITY of WASHINGTON
Department of Biostatistics

Current Issues in the Design and Analysis of Stepped Wedge Trials

Jim Hughes

NIH PRAGMATIC TRIALS
COLLABORATORY GRAND ROUNDS

DEC 1, 2023

Outline

1. **Background**
2. Key design considerations
3. Analysis recommendations

Stepped Wedge Design

		Time			
		1	2	3	4
cluster	1	0	1	1	1
	2	0	0	1	1
	3	0	0	0	1

0 = control
1 = treatment

- N clusters randomized to Q sequences
- Interest is in the "intervention effect"

Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? International Journal of Epidemiology. 49:1043-1052, 2020.

Stepped Wedge Design

		Time			
		1	2	3	4
cluster	1	0	1	1	1
	2	0	0	1	1
	3	0	0	0	1

- Clustered data
- Intervention effect is partly confounded with Time
- More information on earlier compared to later exposure times

Outline

1. Background
2. **Key design considerations**
3. Analysis recommendations

Key Design Considerations

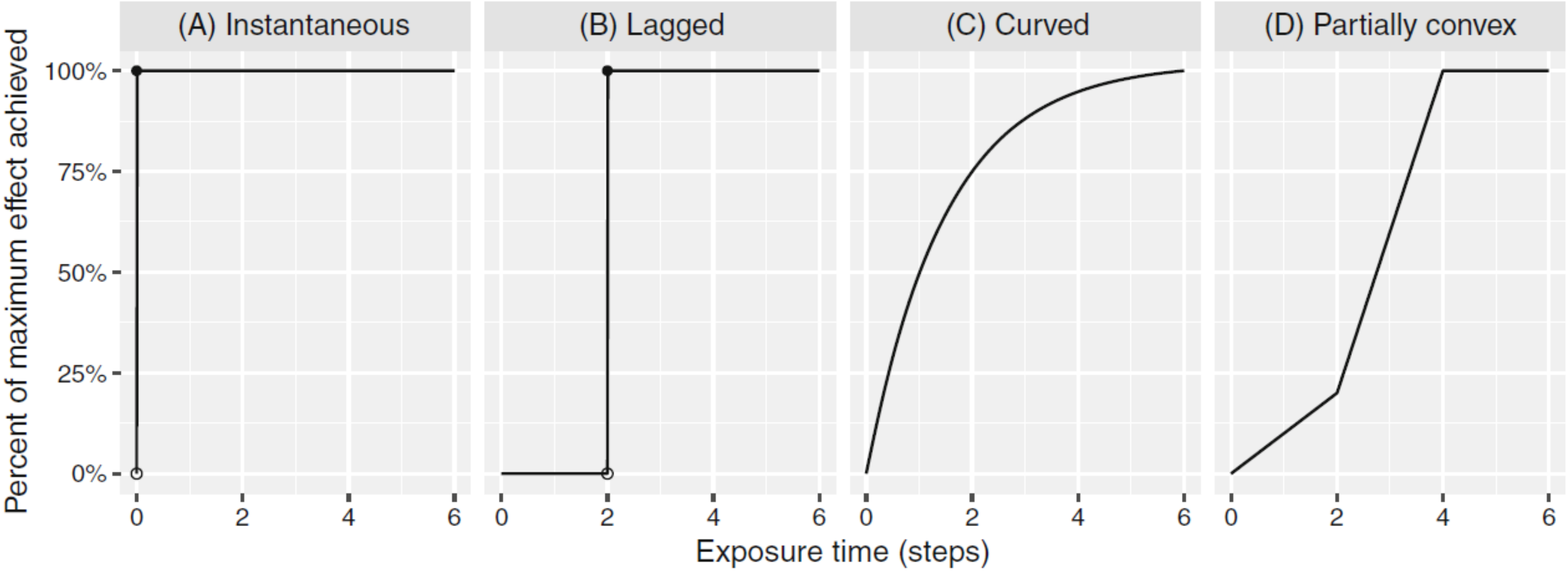
1. What is the estimand?
    a. Population, treatment, endpoint, summary measure, intercurrent events[1]
    b. Impact of informative cluster size on estimands[2]
    c. Impact of exposure time on treatment effect
2. What are the key sources of variation?
    a. Variation between cluster means
    b. Variation in temporal trend
    c. Variation in treatment effect
3. How will outcome data be collected?
    a. Cross-sectional design
    b. Cohort design

[1] ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinicaltrials-guideline-statistical-principles_en.pdf

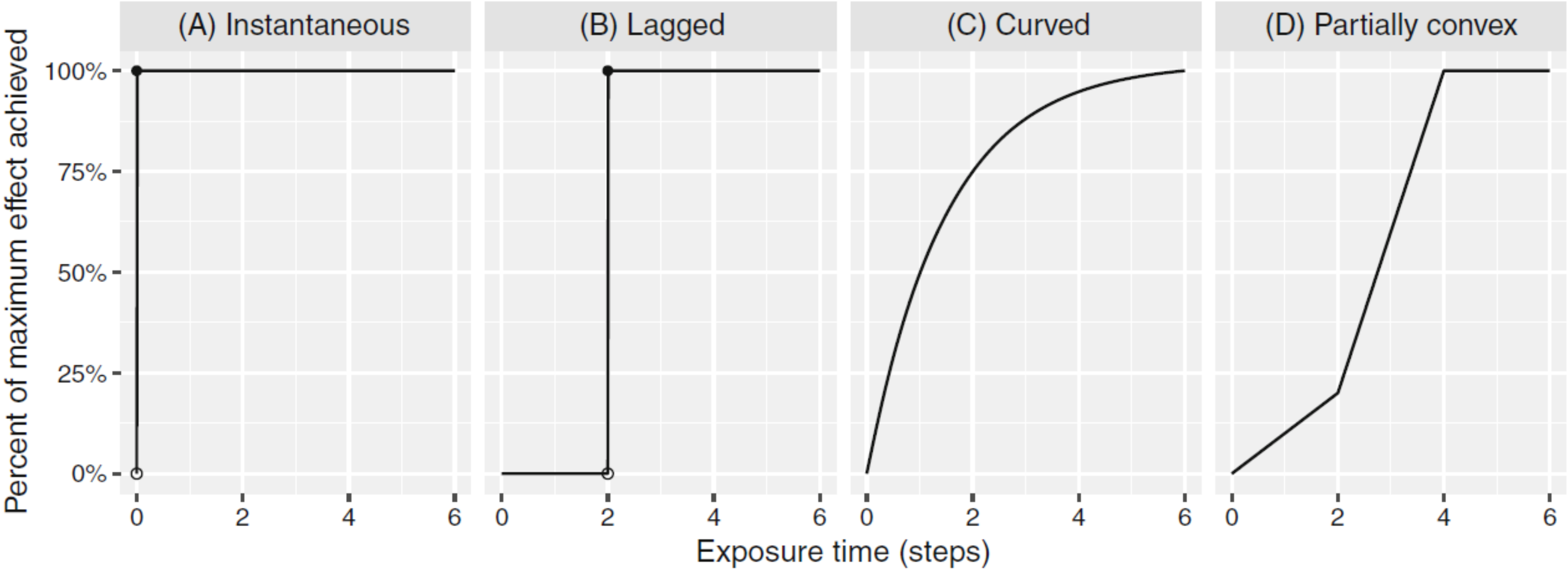
[2] Kahan et al. *International J Epidemiology* 52(1):107-118, 2023

How will the treatment effect the outcome?

- Magnitude of effect (key power consideration)
- Variation in effect over exposure time



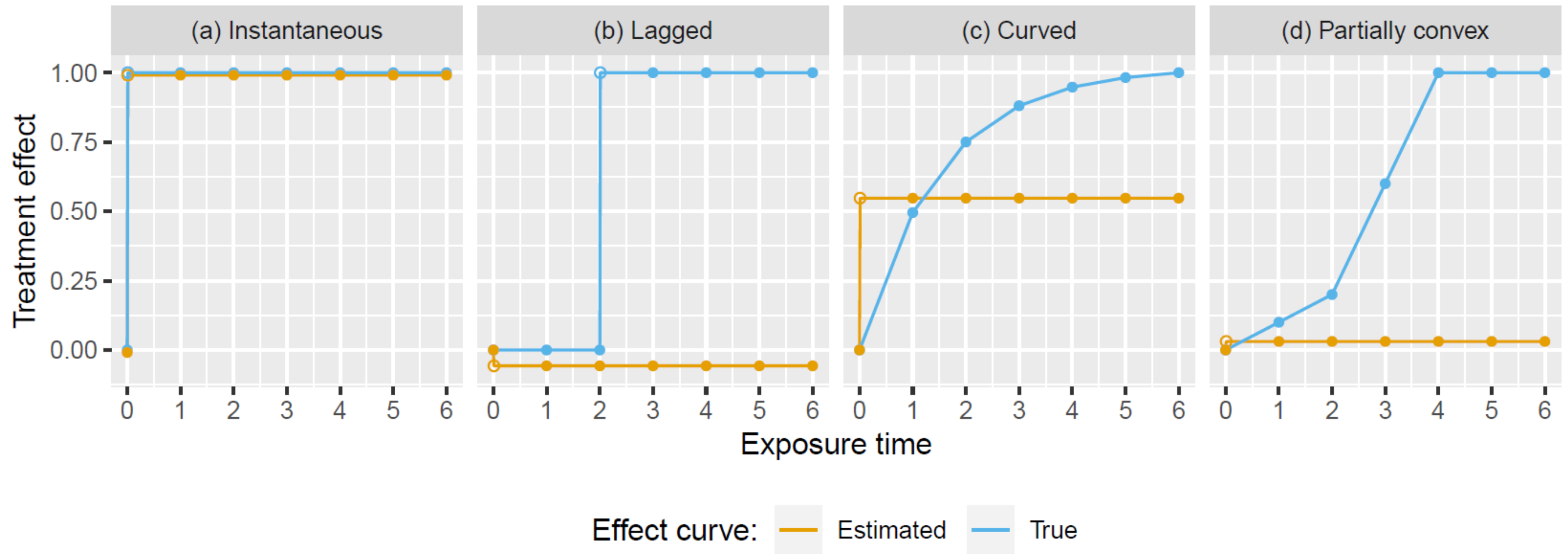
What is the estimand?



1. Average treatment effect over exposure time?
2. Treatment effect at a point in time?
3. Average treatment effect after a (predefined) transition period?

Treatment effect not constant

- What happens if you assume the treatment effect is immediate and constant (IT model), but it's not?[1]



[1] Kenny et al. *Stat. Medicine* 41:4311-4339, 2022

What is the estimand?

A general approach (ETI model):

$$\theta = \sum_{s} w(s) * \delta(s)$$

w(s) = weight at exposure time s
δ*(s)* = treatment effect at exposure time s

		Time Period						
		1	2	3	4	5	6	7
Sequence	1	0	1	2	3	4	5	6
	2	0	0	1	2	3	4	5
	3	0	0	0	1	2	3	4
	4	0	0	0	0	1	2	3
	5	0	0	0	0	0	1	2
	6	0	0	0	0	0	0	1

What is the estimand?

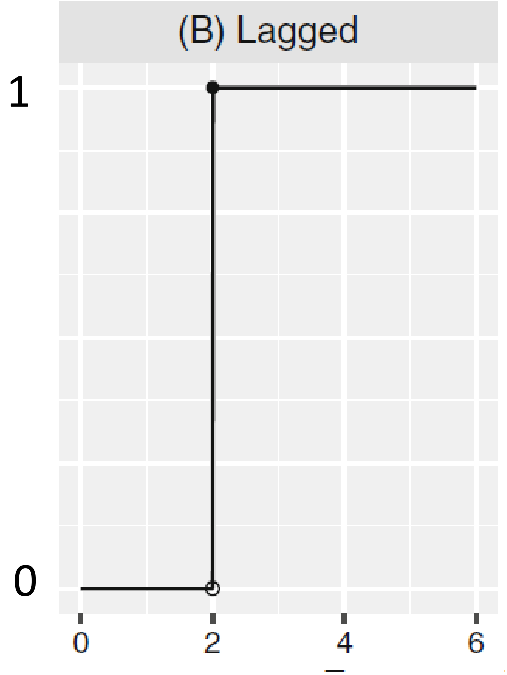
$$\theta = \sum_{s} w(s) * \delta(s)$$

Three possible estimands:

ATE: w = (1,1,1,1,1,1)/6

LTE: w = (0,0,0,0,0,1)

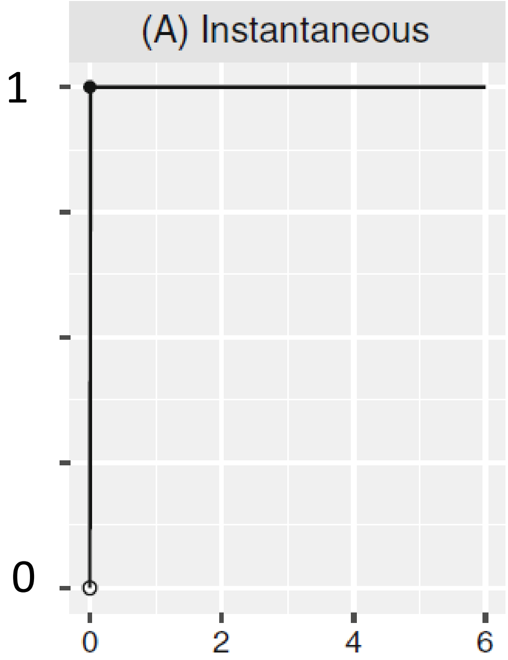
ATE w transition: w = (0,0,1,1,1,1)/4



What is the estimand?

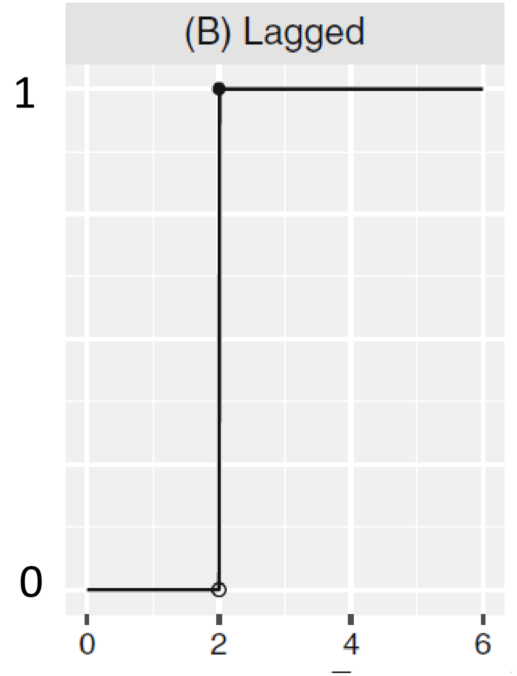
$$\theta = \sum_s w(s) * \delta(s)$$

- The standard IT estimator also has the above form e.g. if one assumes working independence (GEE), the weights end up being w = (6,5,4,3,2,1)/21
- This leads to a more efficient estimate than the ATE **if the IT assumption holds**.
- *BUT likely does not correspond to any estimand of interest if the IT assumption is violated*



What is the estimand?

- Estimating a separate treatment effect for each exposure lag is robust, but inefficient.
- Additional assumptions can gain efficiency (at the cost of possible loss of robustness)
- Example:
  - δ(1) = treatment effect at exposure times 1 and 2
  - δ(2) = treatment effect at exposure times 3 – 6
  - estimand is $\theta = \delta(2)$



		Time Period						
		1	2	3	4	5	6	7
Sequence	1	0	1	1	2	2	2	2
	2	0	0	1	1	2	2	2
	3	0	0	0	1	1	2	2
	4	0	0	0	0	1	1	2
	5	0	0	0	0	0	1	1
	6	0	0	0	0	0	0	1

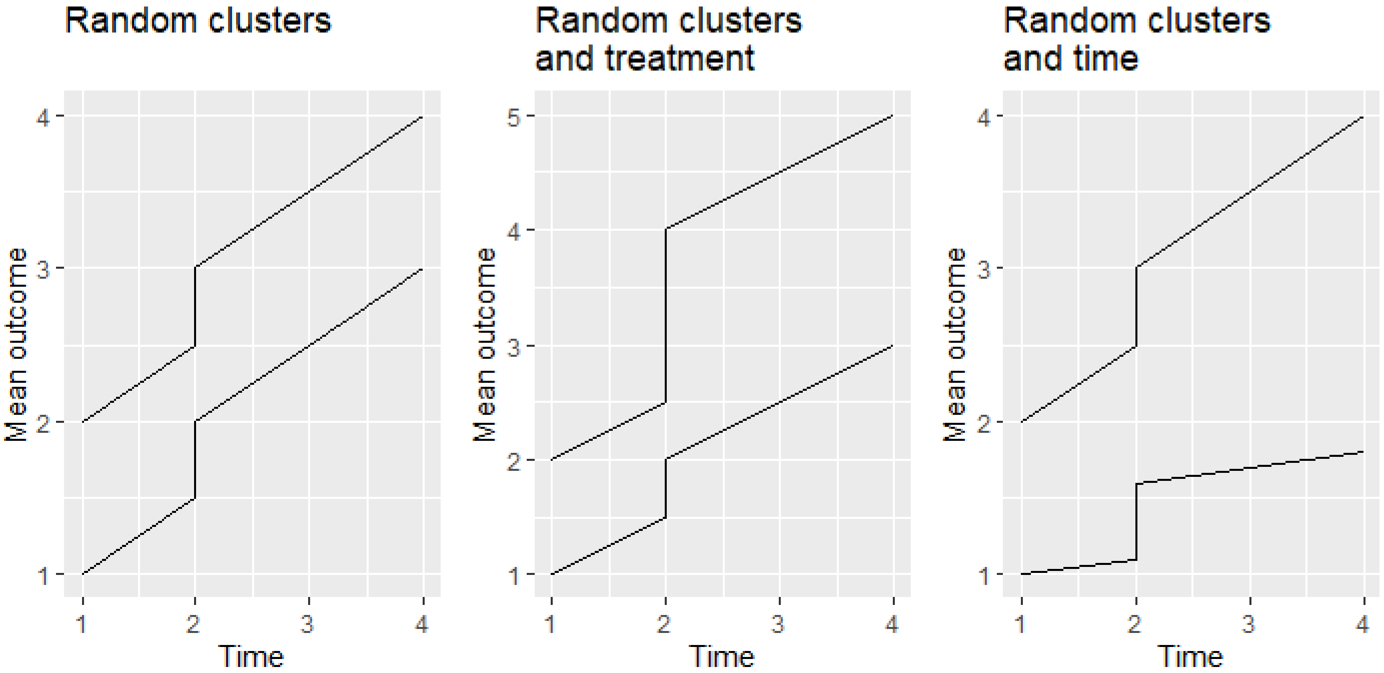
Considerations in choosing estimand

1) What is scientifically meaningful?
    - Is there prior information on form of exposure time curve?
2) Robustness
3) Efficiency

SCHOOL OF PUBLIC HEALTH · UNIVERSITY of WASHINGTON

Department of Biostatistics

Key Sources of Variation



Cluster means: How much variation do you expect in the cluster means (in the absence of treatment)?

Treatment effect: How much variation do you expect in the treatment effect from cluster to cluster?

Temporal trend: In the absence of treatment, how much variation do you expect in the temporal trend from cluster to cluster?

Key Sources of Variation

For **power calculations** …

- Always include a random cluster effect
  - Always include a random individual effect for cohort designs
- “Better” to include too many random effects in power calculation than too few
- Get random effect variances from
  - Prior data
  - Expert opinion[1]

[1] Hughes et al. *Contemporary Clinical Trials* 45(Pt A):55-60, 2015

Example

ADDRESS – BP trial

	Study period (year/month)													
	Year 1				Year 2				Year 3				Year 4	
Sequence	3	6	9	12	3	6	9	12	3	6	9	12	3	6
1	TAU	TAU	TAU	TAU	P1	P2	P3	P4	F5	F6	F7	F8	F9	F10
2		TAU	TAU	TAU	TAU	P1	P2	P3	P4	F5	F6	F7	F8	F9
3			TAU	TAU	TAU	TAU	P1	P2	P3	P4	F5	F6	F7	F8
4				TAU	TAU	TAU	TAU	P1	P2	P3	P4	F5	F6	F7
5					TAU	TAU	TAU	TAU	P1	P2	P3	P4	F5	F6

- Treatment is an implementation strategy designed to promote adoption of an evidence-based intervention for the treatment of uncontrolled hypertension.
- 5 sequences, 5 health care facilities per sequence
- 14 periods, categorized as Treatment as Usual (TAU), Treatment (P1 – P4), Followup (F5 – F10)
- Cohort of 20 individuals/cluster with uncontrolled hypertension; outcome is blood pressure control (yes/no)

Example

Parameter description	**Value**	
Number of observations per facility per time period	n = 20	
Outcome percent during TAU period	40%	
Outcome percent during exposure times three (P3) and four (P4)	60%	
Time trend (on logit scale)	.08/period	
	Variance (on logit scale)	Intra-cluster correlation
Cluster variance/Between-period ICC	0.1316	0.022
Cluster*time variance/Within-period ICC	0.1974	0.054
Individual variance/Within-individual ICC	2.5	0.430

- Variance component estimates from prior data
- Study time trend must be specified for non-linear models (ie logit)

Example

- R package `swCRTdesign` used for power calculations (also can use NIH RMR `SWGRT` sample size calculator)

Comparison	Comment	Power
TAU vs (P3 + P4)/2	Primary comparison	0.82
TAU vs (P3 + P4)/2	No data at P1, P2	0.73
TAU vs (F5+F6+F7+F8+F9+F10)/6		0.39
TAU vs (P3,P4)	Piece-wise constant treatment effect for (P1 – P2), (P3 – P4), (F5 – F10)	.94
TAU vs (P3,P4)	Piece-wise constant treatment effect for (P1 – P2), (P3 – P4), (F5), (F6), (F7), F(8), F(9), (F10)	.87

Design recommendations

1) Don’t make IT assumption unless well-justified
   - Consider both exposure time variation and estimand of interest
   - Additional assumptions (ie piece-wise constant effect) can increase power at cost of robustness
2) If a transition period is planned, include data from the transition period
3) If estimand is the effect at a point in time, maximize the number of observations at that exposure time
4) Including more variance components in power calculation reduces possibility of an underpowered trial
5) Power calculations in SW trials can sometimes seem counterintuitive!

Outline

1. Background
2. Key design considerations
3. **Analysis recommendations**

Common Analysis Approaches

- Most analyses rely on a parametric model of the form

$$\boxed{\begin{array}{l} g(E(Y_{ijk})) = \beta\left(t_{ij}\right) + \delta(s_{ij})x_{ij} \\ Var(Y_{ijk}) \text{ specification} \end{array}}$$

cluster i
time j
individual k

  - Link function (eg identity, logit) – *g(·)*
  - Model for changes over study time – $\beta(t_{ij})$
  - Model for changes over exposure time – $\delta(s_{ij})$
  - Model for Var(Y_{ijk})
- GLMM, GEE

Key questions/issues

- What model for study time?
- What model for exposure time?
- What variance structure?
- Other considerations

What model for study time?

- If $\beta(t_{ij})$ misspecified → δ likely biased
- Number of study time intervals is …
  - Small – maintain maximum flexibility by using indicators for each time period
  - Large (or continuous) – little research; maintain flexibility e.g. spline

What model for exposure time?

- If $\delta(s)$ misspecified → misleading estimates of the treatment effect
- Fitting a separate δ for each *s* is most robust

$$\hat{\theta} = \sum w(s)\delta(s) = w\hat{\delta}$$

$$Var(\hat{\theta}) = wVar(\hat{\delta})w^T$$

- Fitting piece-wise constant (or spline) for $\delta(s)$ can improve efficiency
- Estimates are straightforward to obtain in R or other packages (see extra slide)

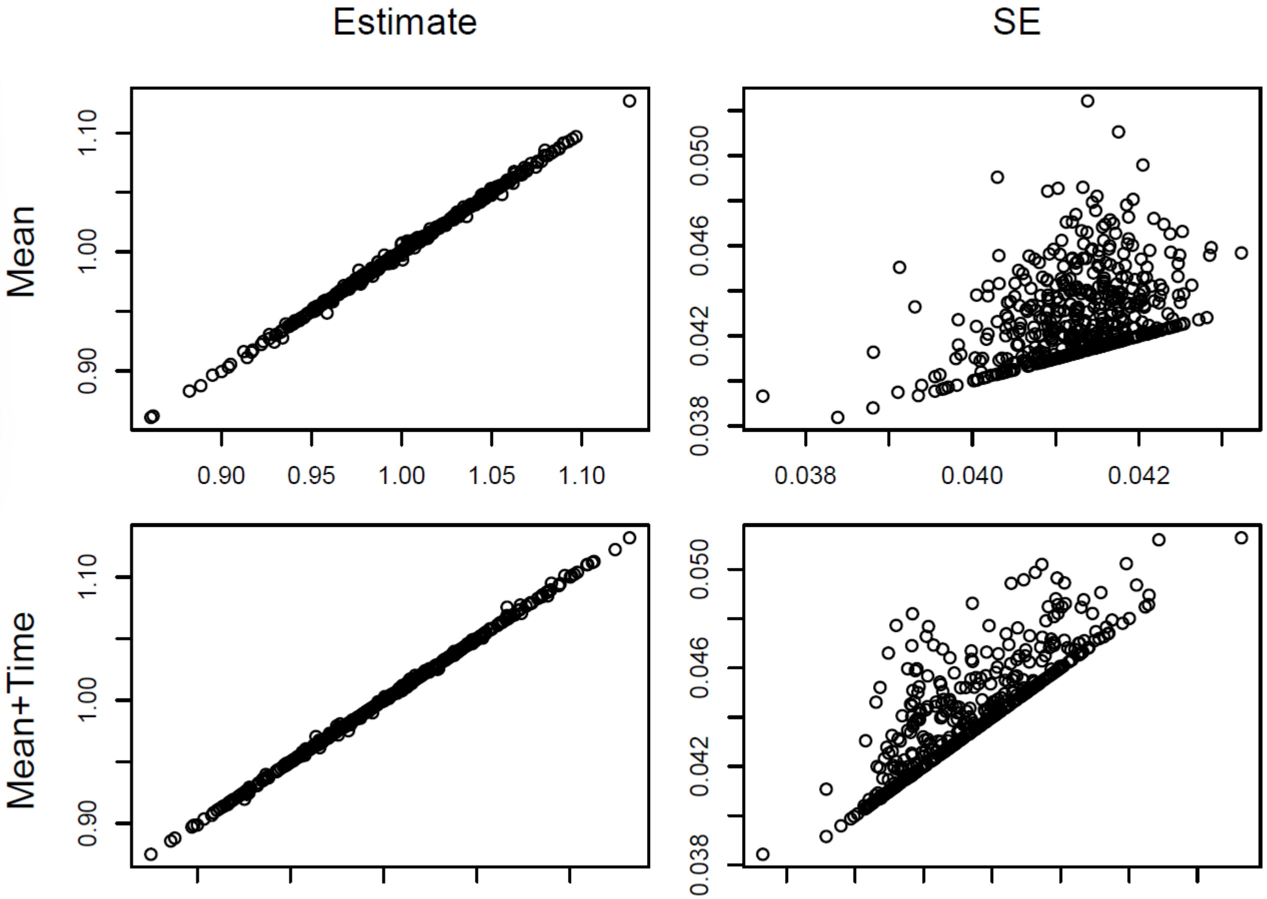
What variance structure?

- GLMM – variance components specified
  - Misspecification does not create bias
  - But can result in over or under estimation of $Var(\hat{\delta})$[1]
  - Too many variance components "better" than too few (conservative)
- GEE – "working" variance specified
  - Robust to misspecification of variance
  - May be inefficient

[1] Voldal et al *Stat Medicine 41:1751-1766,* 2022

What variance structure?

- Including all variance components in GLMM can lead to conservative SE



Other considerations

- Nonparametric & permutation-based methods
- Small number of clusters
  - With “small” numbers of clusters hypothesis tests may have inflated type I error rates[1]
  - Use small sample correction e.g. Kenward-Rodger
- Informative cluster size
  - Need to be careful about weighting and variance-specification to ensure correct estimand[2]

[1] Thompson et al. *Stat in Medicine* 30: 425–439, 2021

[2] Kahan et al. *International J Epidemiology* 52:107-118, 2023

SW Analysis Recommendations

1) Fit flexible study time effect (e.g. categorical time or spline)
2) Avoid fitting IT model unless very confident that treatment effect is immediate and constant
   - Exposure time indicator model most robust
   - Piecewise constant or spline model may increase power
3) Better to overfit than underfit random effects
   - Overfitting gives conservative SE
4) Use small sample correction if necessary

Questions?

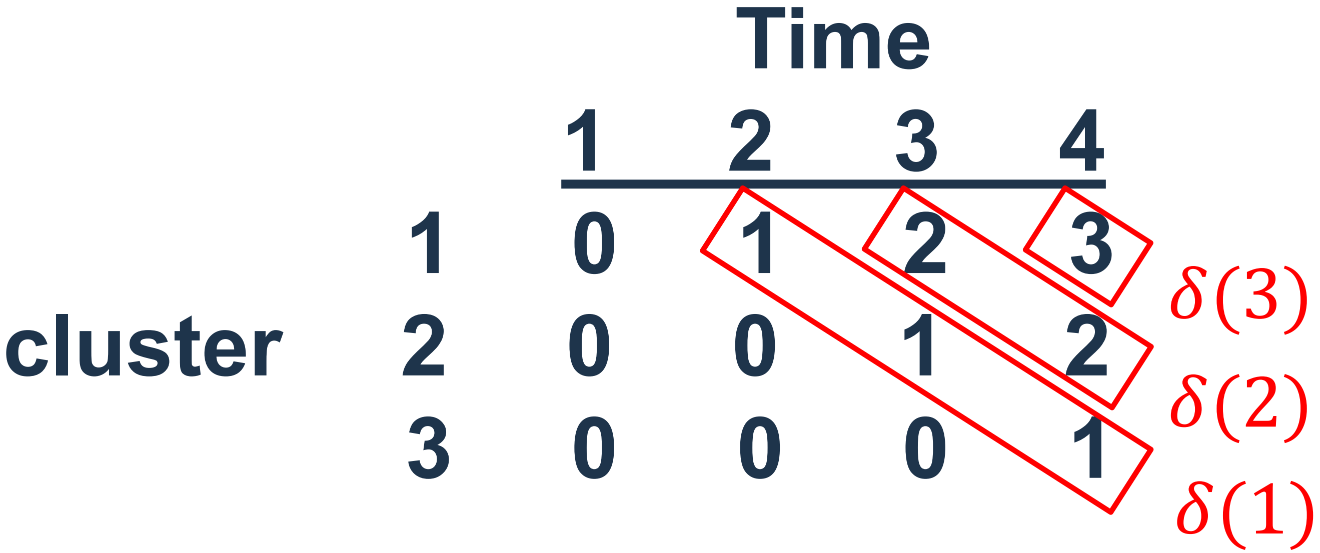
Estimate treatment effect ADDRESS-BP

```
## Assume ADDRESS-BP trial design and interesting in estimating effect at P3-P4
## Assume dataset with variables response, time, timeontx, cluster
ftime = factor(time)
ftimeontx = factor(timeontx)
ftx3 = factor(ifelse(timeontx==0,0,
              ifelse(timeontx<=2,1,ifelse(timeontx<=4,2,3))))

## Fitting a separate indicator for each exposure time s
##
rslt = lmer(response ~ ftime + ftimeontx + (1|fcluster))
# First 14 fixed effects correspond to grand mean and time
# Compare P3 + P4 to TAU
w=c(0,0,1,1,0,0,0,0,0)
est = sum(fixef(rslt)[15:24] * w)
se = sqrt(t(w) %*% vcov(rslt)[15:24,15:24] %*% w))

##Fitting a piecewise constant (P1-P2) (P3-P4) (F5-F10)
##
rslt = lmer(response ~ ftime + ftx3 + (1|fcluster))
# First 14 fixed effects correspond to grand mean and time
# Assume constant tx effect for P3, P4 and compare to TAU
est = fixef(rslt)[16]
se = sqrt(vcov(rslt)[16,16])
```

What’s happening?



		Time			
		1	2	3	4
cluster	1	0	1	2	3
	2	0	0	1	2
	3	0	0	0	1

$\delta(3)$

$\delta(2)$

$\delta(1)$

$$E(\hat{\delta}) = \sum_{s} w_s \delta_s$$

- Weights sum to 1 (as expected) but …weights can be > 1 and/or negative!
- Also, study time effect is biased, so treatment effect is compared to the wrong baseline