How to achieve model-robust inference in stepped wedge trials with model-based methods?

Bingkai Wang

Assistant Professor, Department of Biostatistics
University of Michigan

https://bingkaiwang.com/

Joint work with Xueqi Wang and Fan Li

SCT Annual Meeting
May 20, 2026

Disclosure

This research was supported by Patient-Centered Outcomes Research Institute Awards® (PCORI® Awards ME-2020C3-21072 and ME-2022C2-27676) and the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R00AI173395. The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health or PCORI®, its Board of Governors, or the Methodology Committee.

Cluster randomized trials (CRTs)

- Clusters (hospitals, nursing homes, clinics) of individuals are randomized.
- **Example**: schematic for two-arm study with 12 clusters

	Time
	1
Clusters 1-6	O
Clusters 7-12	A

O: control; A: treatment

- Methods for planning & analyzing parallel-arm CRTs have been relatively well documented.

Stepped-wedge cluster randomized trials (SW-CRTs)

- **SW-CRT:** increasingly popular where the intervention is rolled out
  - in the first time period, all clusters are untreated; in the last time period, all clusters are treated
  - the timing of crossover to intervention is randomized
- **Example**:

	Time				
	1	2	3	4	5
Clusters 1-3	O	A	A	A	A
Clusters 4-6	O	O	A	A	A
Clusters 7-9	O	O	O	A	A
Clusters 10-12	O	O	O	O	A

O: control; A: treatment

Analysis of SW-CRTs

- Essential task: estimate the average treatment effect (ATE)
- Considerations:
  - longitudinal data
  - open/closed cohort
  - inter-subject correlation

Model-based analysis

- (Generalized) linear mixed model **most commonly used**
- Common practice[1]:

$$Y_{ijk} = \underbrace{\beta_j}_{\text{secluar trend}} + \underbrace{\beta_A A_{ij}}_{\text{intervention effect}} + \underbrace{\beta_X X_{ik}}_{\text{covariate adjustment}} + \underbrace{\gamma_i}_{\text{random effect}} + \underbrace{\epsilon_{ijk}}_{\text{risidual error}}$$

- Non-continuous outcomes will involve link functions.

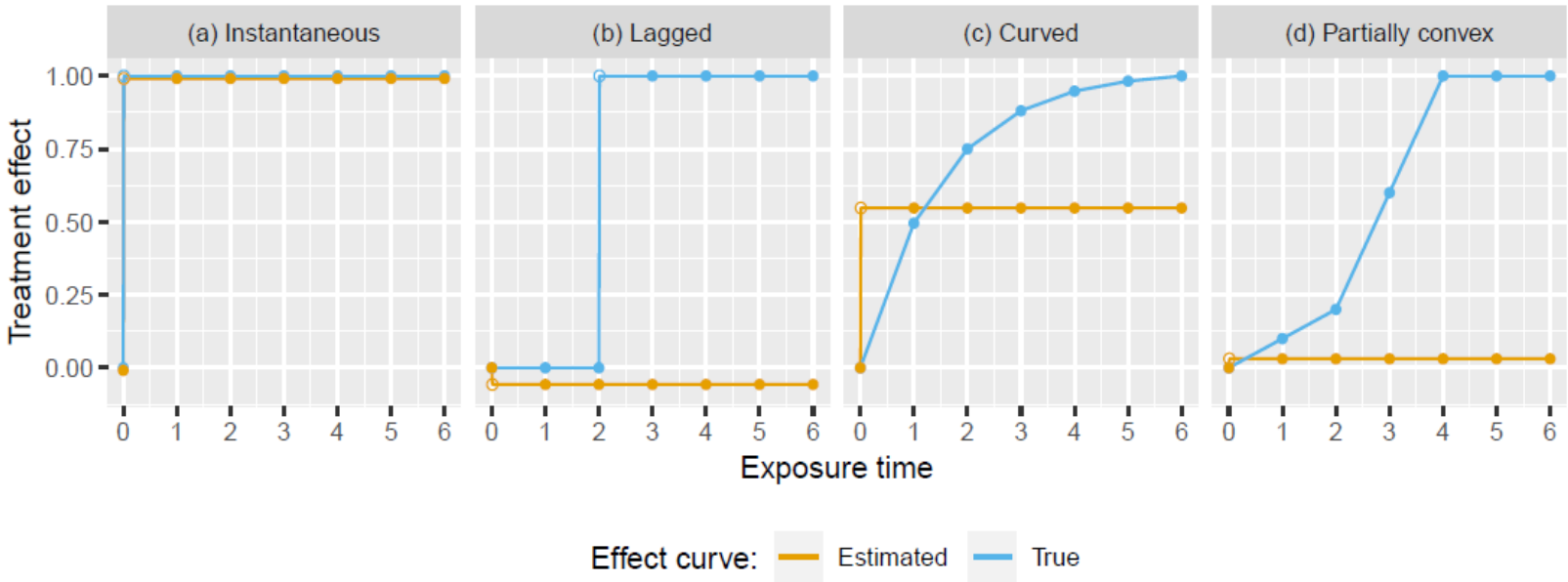
[1] Hussey MA, Hughes JP (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials.

Some challenges in model-based analysis

- **Unclear causal estimands**
  - Regression coefficients often target conditional treatment effects.
- **Risk of model misspecification**
  - When the assumed model DEVIATES from the true data-generating process, what does β_A represent?

What do we know?—Causal estimands

- Treatment effect may vary by the exposure time[2]



A constant treatment effect model may not represent the average of exposure-time-specific treatment effect.

[2] Kenny, A., Voldal, E., Xia, F., Heagerty, P.J., and Hughes, J.P. (2022) Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Statistics in Medicine*, 41(22), 4311-4339

What do we know?—Model misspecification

- Misspecified random effects leads to biased model-based variance (Kasza and Forbes, 2019; Bowden et al., 2021; Voldal et al., 2022)
- Robust sandwich variance estimator provides nominal coverage under random-effects structure misspecification (Ouyang et al., 2024)
- Both results obtained assuming (1) remaining model ingredients correctly specified (2) no covariates

Objectives

- **Gap**:
  - insufficient attention to estimands
  - limited knowledge about model-robustness
- **Objectives**:
  - articulate a potential outcomes framework to define estimands that allow for treatment effect heterogeneity across time axes
  - adapt linear mixed models and GEE to achieve model-robustness

Roadmap

- **The causal framework**
- Robustness of linear mixed models and GEE
- Demonstration

Notation: indices

- Cluster $i = 1, \ldots, I$
- Period $j = 1, \ldots, J$
- Individual $k = 1, \ldots, N_i$
- N_i: finite, source population size in cluster i

Notation: random variables

- Y_{ijk}: observed outcome of individual k in cluster i in period j
- $Z_i = j$: if cluster i starts receiving treatment at period j
- $\boldsymbol{X}_{ik}$: baseline covariates for individual k in cluster i

Notation: potential outcomes

- $Y_{ijk}(z)$: potential outcome of individual k in cluster i during period j, had the cluster been first treated in period z for $1 \leq z \leq j$
- $Y_{ijk}(0)$: the untreated potential outcome
- Causal consistency:
  $Y_{ijk} = \sum_{z=1}^{j} I\{Z_i = z\} Y_{ijk}(z) + I\{Z_i > j\} Y_{ijk}(0), \qquad j \in \{1, \ldots, J\}$

Assumptions

Formalize a sampling-based framework:

Assumption 1:
i.i.d. clusters, allowing arbitrary within-cluster correlation.

Assumption 2:
Treatment assignment is randomly rolled out.

Assumption 3:
Enrollment is non-informative.

- Covering cross-sectional, closed-cohort, and open-cohort designs.

Causal estimands

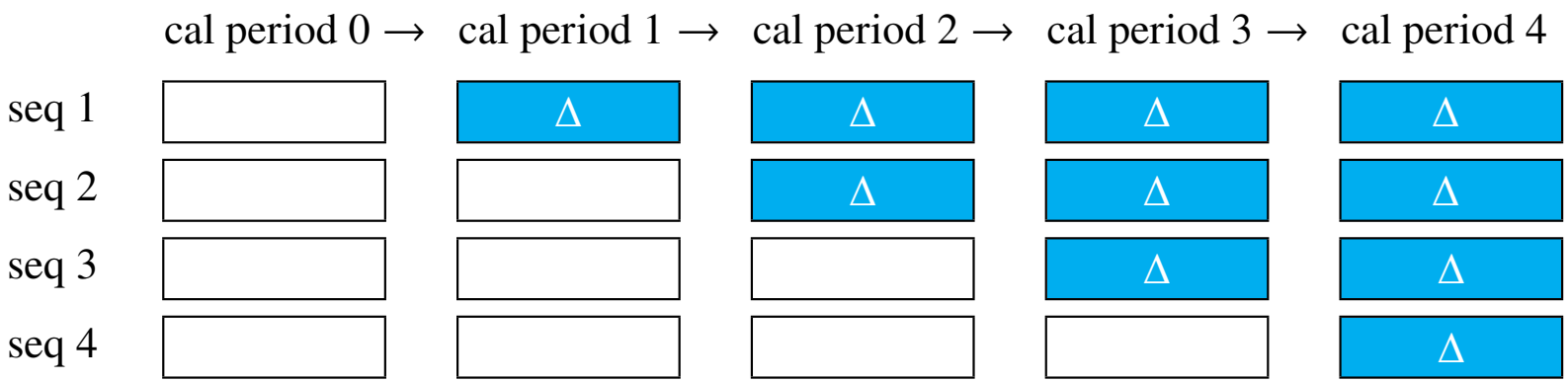
- Define the average causal effect given exposure time d and calendar time j

$$\Delta_j(d) = E\{Y_{ijk}(j-d+1)\} - E\{Y_{ijk}(0)\}, \quad 1 \le d \le j \le J$$

  - Model-free definition!
  - $J(J+1)/2$ estimands in total
  - Four interpretable examples to follow

Case 1: Constant treatment effect structure Δ

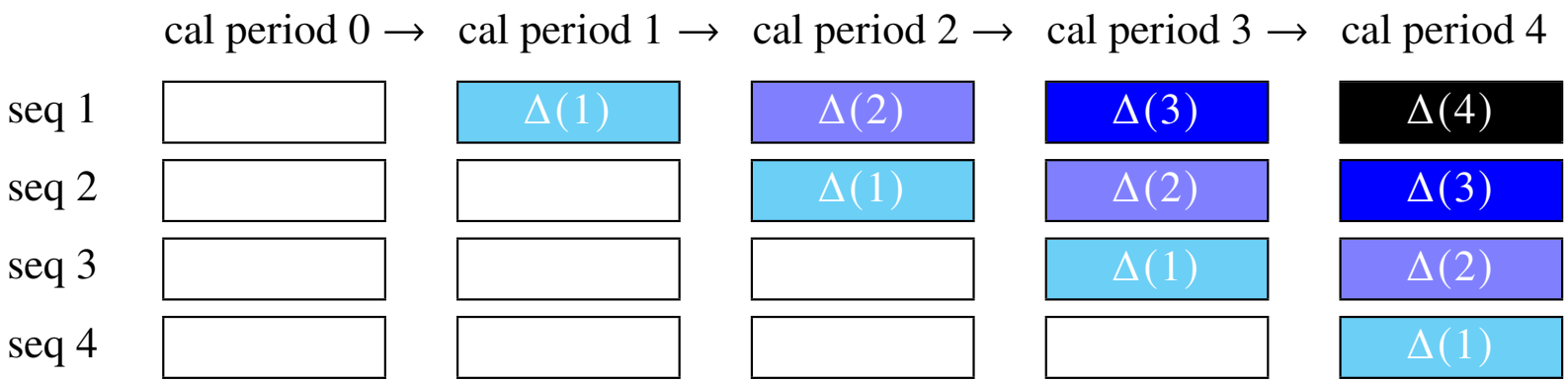
- Assume $\Delta_j(d)$ is constant for all j and d, i.e., $\Delta_j(d) \equiv \Delta$



  - Treatment effect is expected to be immediate and sustained
  - Example: a new emergency operation to improve the survival rate

Case 2: Duration-specific treatment effect structure $\boldsymbol{\Delta}^D = (\Delta(1), \ldots, \Delta(J))^\top$

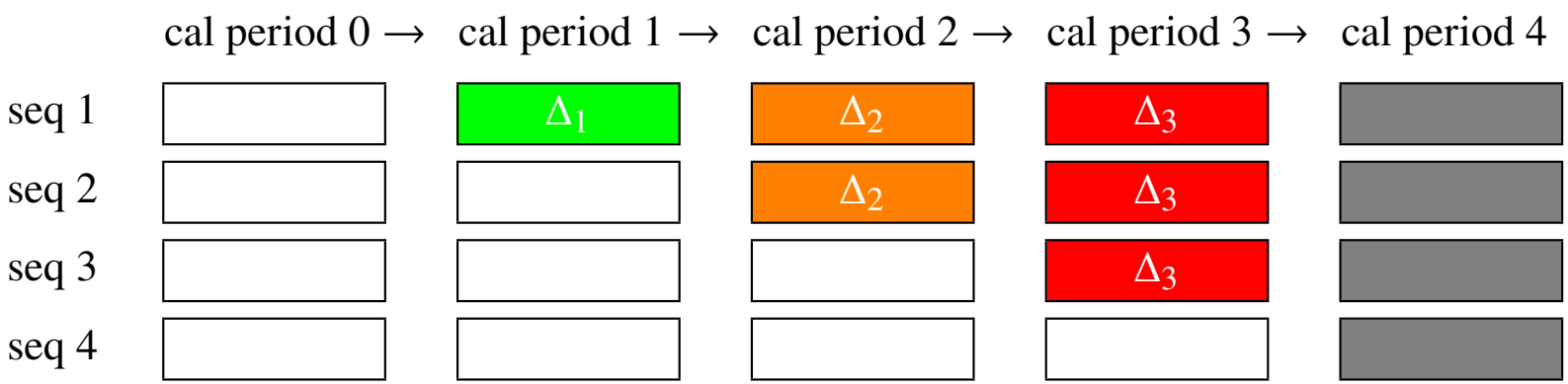
- Consider $\Delta_j(d)$ to be constant across j, but vary by d, i.e., $\Delta_j(d) = \Delta(d)$



  - Example: delayed effect or learning effect
  - ignoring duration effects (exposure-time heterogeneity) can lead to erroneous conclusions

Case 3: Period-specific treatment effect structure $\boldsymbol{\Delta}^P = (\Delta_1, \ldots, \Delta_{J-1})^\top$

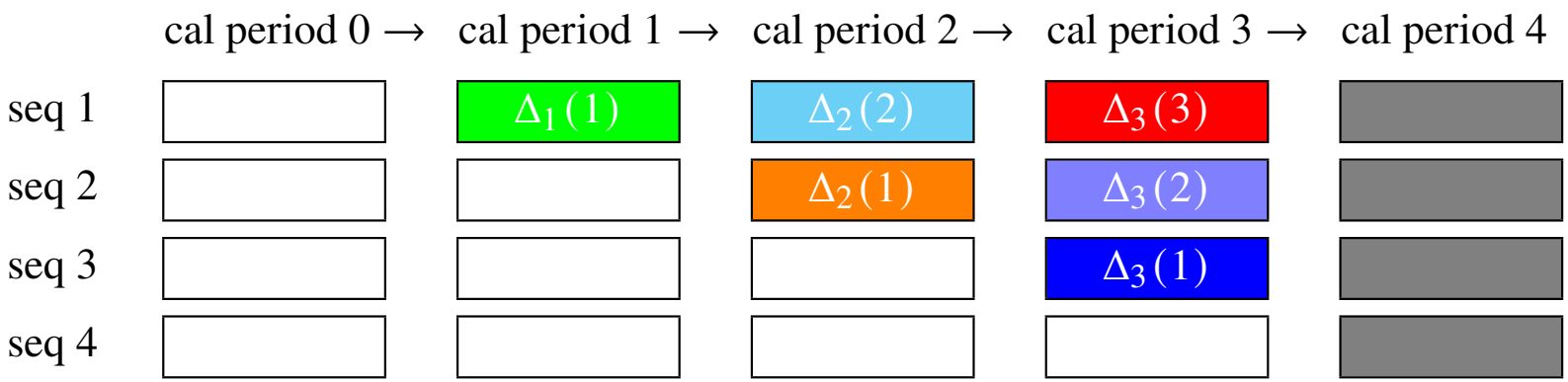
- Allow the treatment effect to vary by the period of measurement, but not the duration of treatment, i.e.,$\Delta_j(d) = \Delta_j$



  - treatment effect is seasonal, or is affected by external events (pandemic)
  - Δ_J not defined because $Y_{iJk}(0)$ is truncated by design

Case 4: Saturated treatment effect structure $\boldsymbol{\Delta}^S$ $= (\Delta_1(1), \Delta_2(1), \Delta_2(2), \ldots, \Delta_{J-1}(1), \ldots, \Delta_{J-1}(J-1))^\top$

- The finest set of estimands



  - make the least assumptions but include the most parameters to estimate, but not always easy to interpret

How to select among the treatment effect models?

- Trade-off: simpler model requires stronger assumption but fewer parameters to estimate.
- Domain knowledge.
- Likelihood ratio test in linear mixed models.
- typical summary estimands in Cases 2-4 as an overall effect measure:
  - $\Delta^{D\text{-avg}} = J^{-1} \sum_{d=1}^{J} \Delta(d)$
  - $\Delta^{P\text{-avg}} = (J-1)^{-1} \sum_{j=1}^{J-1} \Delta_j$
  - $\Delta^{S\text{-avg}} = \{(J-1)J\}^{-1} 2 \sum_{j=1}^{J-1} \sum_{d=1}^{j} \Delta_j(d)$

Roadmap

- The causal framework
- **Robustness of linear mixed models and GEE**
- Demonstration

Working linear mixed model

$$Y_{ijk} = \beta_{0j} + TE_{ij} + \boldsymbol{\beta}_{\boldsymbol{X}}^{\top} \boldsymbol{X}_{ik} + RE_{ij} + \varepsilon_{ijk},$$

- β_{0j}: intercept parameter for period j
- TE_{ij}: working treatment effect structure
- $\boldsymbol{\beta}_{\boldsymbol{X}}$: coefficient for baseline covariates
- RE_{ij}: working random-effects structure to account for intracluster correlation
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$: independent error

Working treatment effect structure TE_{ij}

- constant treatment effect Δ:

  $TE_{ij} = \beta_Z I\{Z_i \leq j\}$

- duration-specific treatment effect $\boldsymbol{\Delta}^D$:

  $TE_{ij} = \sum_{d=1}^{j} \beta_{Zd} I\{Z_i = j - d + 1\}$

- period-specific treatment effect $\boldsymbol{\Delta}^P$:

  $TE_{ij} = \beta_{jZ} I\{Z_i \leq j\}$

- saturated treatment effect $\boldsymbol{\Delta}^S$:

  $TE_{ij} = \sum_{d=1}^{j} \beta_{jZd} I\{Z_i = j - d + 1\}$

Working random-effects structure RE_{ij}

- $RE_{ij} = 0$: independence
- $RE_{ij} = \alpha_i$ with $\alpha_i \sim N(0, \tau^2)$: exchangeable
- $RE_{ij} = \alpha_i + \gamma_{ij}$ with $\alpha_i \sim N(0, \tau^2)$ and $\gamma_{ij} \sim N(0, \kappa^2)$: nested exchangeable

Variance is based on the robust sandwich variance estimator (e.g.[3])

[3] Ouyang Y, Taljaard M, Forbes AB, Li F. (2024) Maintaining the validity of inference from linear mixed models in stepped-wedge cluster randomized trials under misspecified random-effects structures. *Statistical Methods in Medical Research*.

A summary of working structures and variance estimators

Table 1: Summary of estimands, model specifications, and estimators when a working linear mixed model is considered.

Treatment effect structure	TE_{ij} in model (3)	Point estimator	Variance estimator
Constant	$\beta_Z I\{Z_i \leqslant j\}$	$\widehat{\beta}_Z$	$\widehat{V}$
Duration-specific	$\sum_{d=1}^{j-1} \beta_{Zd} I\{Z_i = j-d+1\}$	$\widehat{\boldsymbol{\beta}}_Z^D = (\widehat{\beta}_{Z1}, \ldots, \widehat{\beta}_{ZJ})^\top$	$\widehat{\mathbf{V}}^D$
Period-specific	$\beta_{jZ} I\{Z_i \leqslant j\}$	$\widehat{\boldsymbol{\beta}}_Z^P = (\widehat{\beta}_{1Z}, \ldots, \widehat{\beta}_{J-1,Z})^\top$	$\widehat{\mathbf{V}}^P$
Saturated	$\sum_{d=1}^{j-1} \beta_{jZd} I\{Z_i = j-d+1\}$	$\widehat{\boldsymbol{\beta}}_Z^S = (\widehat{\beta}_{1Z1}, \ldots, \widehat{\beta}_{J-1,Z,J-1})^\top$	$\widehat{\mathbf{V}}^S$

What does stepped wedge randomization offer?

Results: LLM yields consistent and valid confidence intervals as long as the working treatment effect structure is correct.

All other aspects of the working model can be misspecified.

Roadmap

- The causal framework
- Robustness of linear mixed models and GEE
- **Demonstration**

SMARThealth India

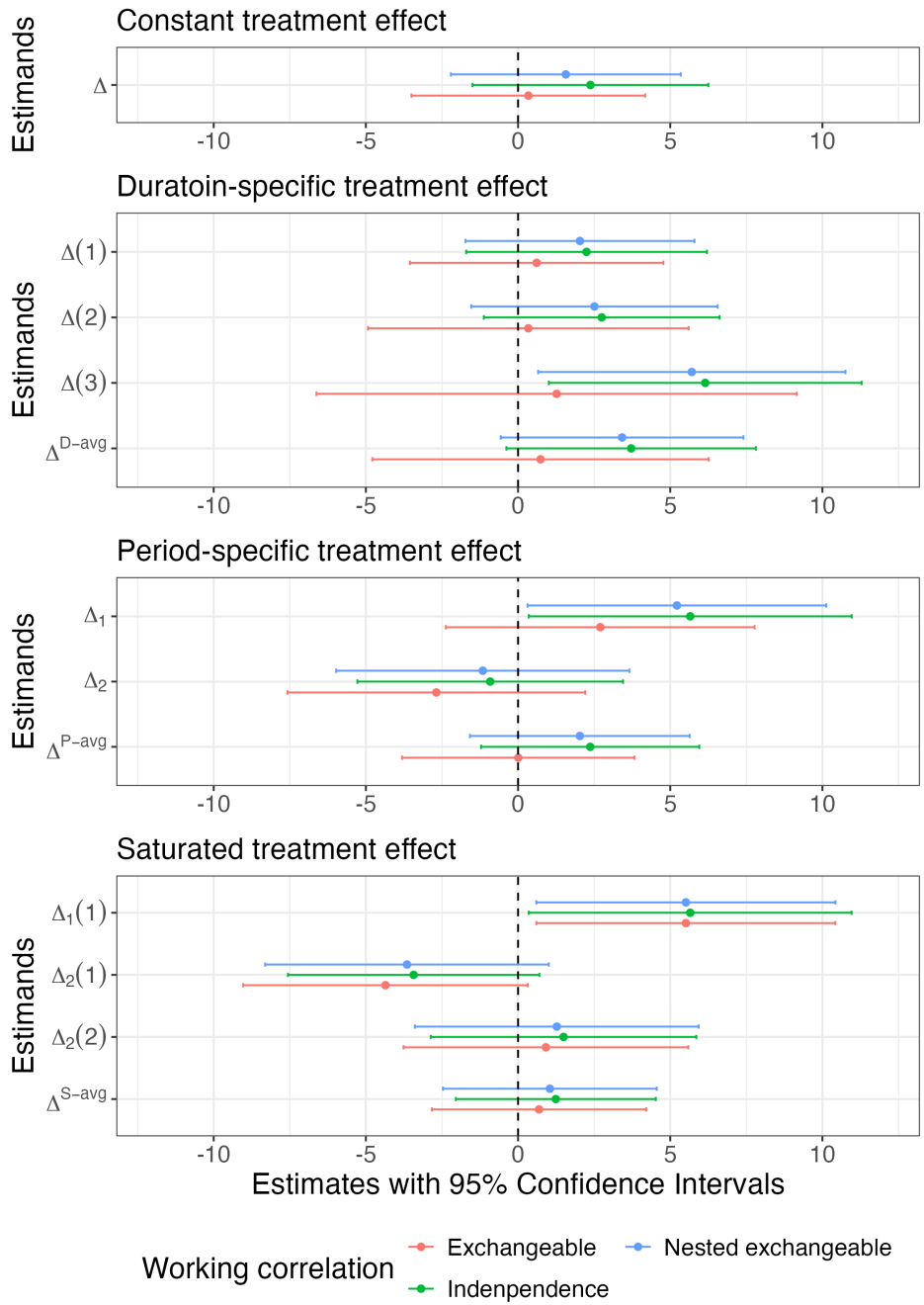
- **SMARThealth India**: SW-CRT studying a mobile health intervention on cardiovascular disease in rural India (Peiris et al. 2019)
  - 18 primary health centers
  - 3 randomization sequences
  - 6 months per period
  - 120 participants per cluster-period
  - Primary outcome: systolic blood pressure

Data analysis results

Finding: Most treatment effects are not statistically significant.

We detected strong treatment effect heterogeneity across **calendar time** ($p < 0.01$) but not across exposure time ($p>0.05$).

Possible reasons: a strong heatwave in Andhra Pradesh during the first rollout period and the possibility that standard care improved over calendar time.



Implications of results

- Select estimands carefully: rigor versus cost
- Model robustness in SW-CRT:
  - covariate effects, random-effects structure and error structure
  - but need to correctly acknowledge the treatment effect structure in your working model!

Acknowledgement

- More details available in
  - Wang B, Wang X, Li F. How to achieve model-robust inference in stepped wedge trials with model-based methods?. Biometrics. 2024 Dec;80(4):ujae123.
- Research in this article was supported by Patient-Centered Outcomes Research Institute Awards® (PCORI® Awards ME-2020C3-21072 and ME-2022C2-27676) & the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R00AI173395.
- The statements presented are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health or PCORI®, its Board of Governors, or the Methodology Committee.

References

Li, F., Hughes, J. P., Hemming, K., Taljaard, M., Melnick, E. R., Heagerty, P. J. (2021). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Statistical Methods in Medical Research*, 30(2), 612-639.

Nevins, P., Ryan, M., Davis-Plourde, K., Ouyang, Y., Pereira Macedo, J. A., Meng, C., ..., **Li, F.**, Taljaard, M. (2024). Adherence to key recommendations for design and analysis of stepped-wedge cluster randomized trials: A review of trials published 2016-2022. *Clinical Trials*, 21(2), 199-210.

Wang, B., Harhay, M. O., Tong, J., Small, D. S., Morris, T. P., **Li, F.** (2024). On the mixed-model analysis of covariance in cluster-randomized trials. *Statistical Science*. In Press.

Bowden, R., Forbes, A. B., Kasza, J. (2021). Inference for the treatment effect in longitudinal cluster randomized trials when treatment effect heterogeneity is ignored. *Statistical Methods in Medical Research*, 30(11), 2503-2525.

Voldal, E. C., Xia, F., Kenny, A., Heagerty, P. J., Hughes, J. P. (2022). Random effect misspecification in stepped wedge designs. *Clinical Trials*, 19(4), 380-383.

Ouyang, Y., Taljaard, M., Forbes, A. B., **Li, F.** (2024). Maintaining the validity of inference from linear mixed models in stepped-wedge cluster randomized trials under misspecified random-effects structures. *Statistical Methods in Medical Research*, 09622802241248382.

Peiris, D., Praveen, D., Mogulluru, K., Ameer, M. A., Raghu, A., Li, Q., ..., Patel, A. (2019). SMARThealth India: a stepped-wedge, cluster randomised controlled trial of a community health worker managed mobile health intervention for people assessed at high cardiovascular disease risk in rural India. *PLoS One*, 14(3), e0213708.

References

Kenny, A., Voldal, E. C., Xia, F., Heagerty, P. J., Hughes, J. P. (2022). Analysis of stepped wedge cluster randomized trials in the presence of a timeâvarying treatment effect. *Statistics in Medicine*, 41(22), 4311-4339.

Hussey, M. A., Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2), 182-191.

Sun, L., Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175-199.

Maleyeff, L., Li, F., Haneuse, S., Wang, R. (2022). Assessing exposure-time treatment effect heterogeneity in stepped wedge cluster randomized trials. *Biometrics*, 79(3), 2551-2564.

Chen, X., Li, F. (2024). Model-assisted analysis of covariance estimators for stepped wedge cluster randomized experiments. *Scandinavian Journal of Statistics*.

Wang, B., Wang, X., Wang, R., Li, F. (2024). How to achieve model-robust inference in stepped wedge trials with model-based methods?. *Biometrics*.

Tian, Z., Li, F. (2024). Information content of stepped wedge designs under the working independence assumption. *Journal of Statistical Planning and Inference*, 229, 106097.

Kasza, J., Hemming, K., Hooper, R., Matthews, J. N. S., Forbes, A. B. (2019). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research*, 28(3), 703-716.

Extra: Simulations with binary outcomes and odds ratio estimands

- $Y_{ijk}(z) \sim \mathcal{B}(\mu_{ijk}(d))$, where $d = I\{z > 0\}(j - z + 1)$ and

$$\text{logit}(\mu_{ijk}(d)) = \beta_{0j} + \beta_{Zd}(X) + \left(X_{ik1} + \frac{j}{J} X_{ik2}^2\right) + \alpha_i + \gamma_{ij},$$

where

$$\beta_{Zd}(X) = I\{d > 0\}\left\{0.72 + 0.18d + \frac{1}{4}(X_{ik1} - \overline{X}_{i\cdot 1}) + \frac{d}{10}(X_{ik2}^2 - \overline{X^2}_{i\cdot 2}) + \beta_i\right\}$$

- A duration-specific treatment effect on the logit scale $\Rightarrow$ saturated treatment effect structure marginally due to β_{0j}
- β_i is the random intervention effect on the logit scale
- Marginal odds ratio estimands

$$\Phi_j(d) = \frac{E[Y_{ijk}(j - d + 1)]}{1 - E[Y_{ijk}(j - d + 1)]} \Big/ \frac{E[Y_{ijk}(0)]}{1 - E[Y_{ijk}(0)]},$$

for $(j, d) \in \{(1, 1), (2, 1), (2, 2)\}$.

Extra: Simulation results with binary outcomes

- Linear mixed model & independence GEE with proper g-computation
- Logistic linear mixed model (GLMM) with nested exchangeable random-effects structure
  - all without covariate adjustment
  - all under the saturated treatment effect specification

I	Estimand	Linear Mixed Model			GEE			GLMM[b]		
		Bias	ASE/ESE[d]	ECP[e]	Bias	ASE/ESE[d]	ECP[e]	Bias	ASE/ESE[d]	ECP[e]
30	$\Phi_1(1)$	0.061	0.891	0.915	0.060	0.884	0.913	0.153	0.801	0.899
	$\Phi_2(1)$	0.075	0.901	0.915	0.069	0.897	0.921	0.133	0.775	0.899
	$\Phi_2(2)$	0.089	0.897	0.910	0.092	0.876	0.916	0.216	0.858	0.928
100	$\Phi_1(1)$	0.024	0.989	0.952	0.023	0.982	0.953	0.127	0.883	0.911
	$\Phi_2(1)$	0.008	0.951	0.935	0.004	0.984	0.937	0.079	0.821	0.887
	$\Phi_2(2)$	0.010	0.961	0.934	0.009	0.974	0.930	0.151	0.913	0.917
1000	$\Phi_1(1)$	0.002	1.007	0.952	0.002	1.012	0.950	0.107	0.905	0.674
	$\Phi_2(1)$	0.002	0.926	0.932	0.000	0.929	0.927	0.076	0.801	0.771
	$\Phi_2(2)$	0.005	1.004	0.952	0.005	1.007	0.952	0.151	0.958	0.681

[d] ASE/ESE: averaged standard error (SE) divided by empirical SE.

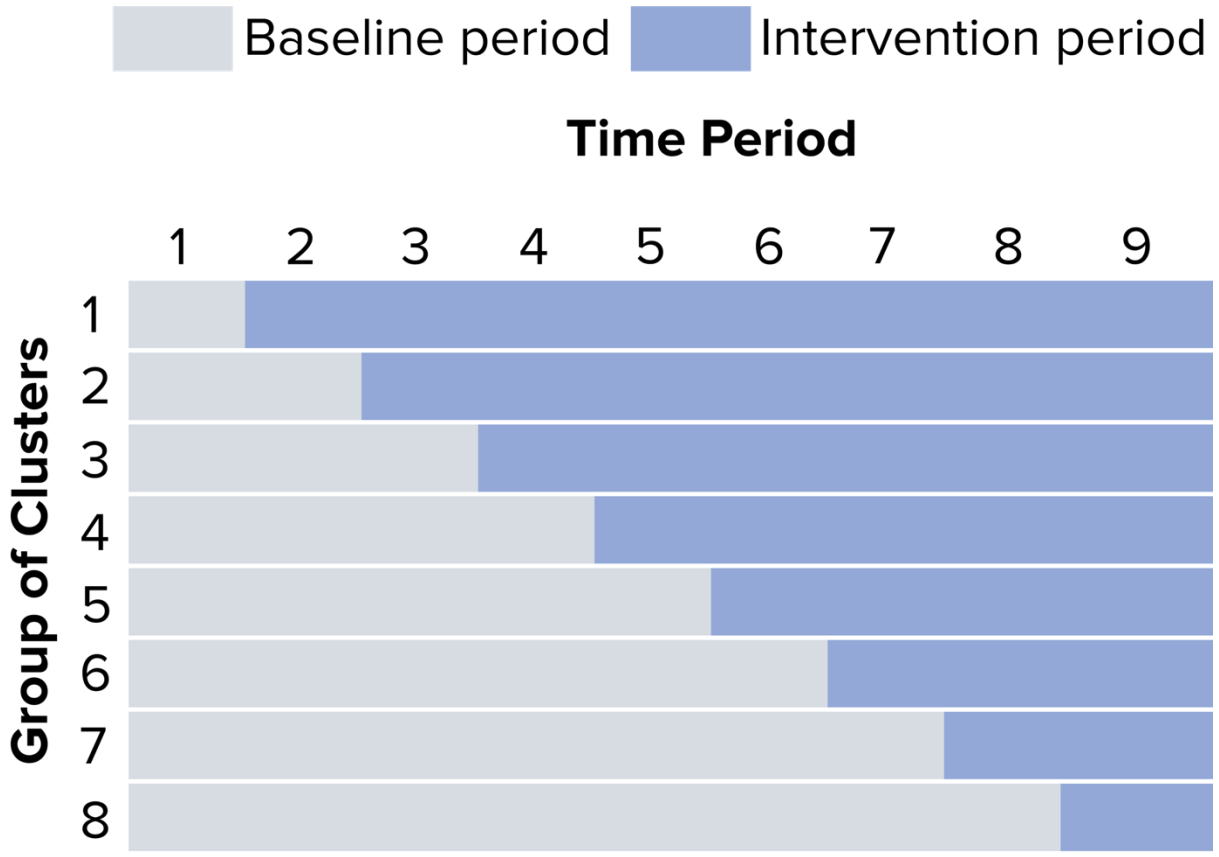
[e] ECP: empirical coverage probability.

Bayesian Hierarchical Penalized Spline Models for Stepped Wedge Cluster Randomized Trials

Danni Wu, Hyung Park, Corita Grudzen, Keith Goldfeld

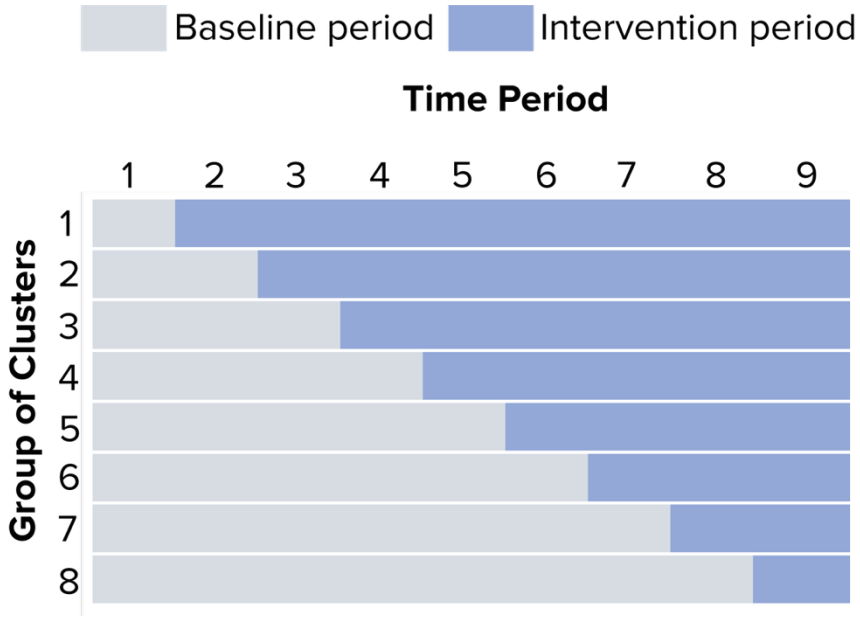
Stepped wedge cluster randomized trials (SWCRTs)

SWCRTs: clusters are assigned in a random sequence to transition from control to intervention status



freshspectrum.com & steppedwedgehog.blog
rdatagen.net

Motivation: Addressing confounding by temporal trends in SWCRTs with many more time periods and clusters



Traditional frequentist methods?
Tend to underestimate intervention effect's confidence interval coverage

Bayesian methods?
Limited models for SWCRTs

We propose

1. A Bayesian spline model for addressing confounding by temporal trends in SWCRTs
2. Bayesian spline modeling for estimating time-varying intervention effects

Traditional frequentist model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_{jt}$$

Variable/Parameter	Meaning
Y_{ijt}	Outcome for the i^{th} subject from the j^{th} cluster at time t

Rennert L et al. *BMC Medical Research Methodology.* 2021

Traditional frequentist model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_{jt}$$

Variable/Parameter	Meaning
Y_{ijt}	Outcome for the i^{th} subject from the j^{th} cluster at time t
A_{jt}	Binary treatment assignment (1/0)

Traditional frequentist model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_{jt}$$

Variable/Parameter	Meaning
Y_{ijt}	Outcome for the i^{th} subject from the j^{th} cluster at time t
A_{jt}	Treatment assignment for the j^{th} cluster at time t (1/0)
$g(.)$	Outcome-specific link

Traditional frequentist model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_{jt}$$

Variable/Parameter	Meaning
Y_{ijt}	Outcome for the i^{th} subject from the j^{th} cluster at time t
A_{jt}	Treatment assignment for the j^{th} cluster at time t (1/0)
$g(.)$	Outcome-specific link
α	Intercept

Traditional frequentist model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_{jt}$$

Variable/Parameter	Meaning
Y_{ijt}	Outcome for the i^{th} subject from the j^{th} cluster at time t
A_{jt}	Treatment assignment for the j^{th} cluster at time t (1/0)
$g(.)$	Outcome-specific link
α	Intercept
τ	Intervention effect

Traditional frequentist model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_{jt}$$

Variable/Parameter	Meaning
Y_{ijt}	Outcome for the i^{th} subject from the j^{th} cluster at time t
A_{jt}	Treatment assignment for the j^{th} cluster at time t (1/0)
$g(.)$	Outcome-specific link
α	Intercept
τ	Intervention effect
s_{jt}	Cluster j specific time effect at time t

Bayesian immediate effect model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_j(t)$$

Bayesian immediate effect model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_j(t)$$
$$= \alpha + \tau \cdot A_{jt} + \boldsymbol{\beta}_{bj} \cdot \boldsymbol{B}_t$$

$\boldsymbol{B}_t$: B-spline basis vector for time t (p-dimensional)

Bayesian immediate effect model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_j(t)$$
$$= \alpha + \tau \cdot A_{jt} + \boldsymbol{\beta}_{bj} \cdot \boldsymbol{B}_t$$

Parameter	Prior distribution
$\boldsymbol{\beta}_{bj}$ (Spline coefficients in cluster j)	$\text{Normal}\left(\mu = \boldsymbol{\beta}, \Sigma = \sigma_b^2 I_{p\times \text{p}}\right)$ $\sigma_b \sim \text{t}_{\text{student}}(\text{df} = 3, \mu = 0, \sigma = 2.5)$

Bayesian immediate effect model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_j(t)$$
$$= \alpha + \tau \cdot A_{jt} + \boldsymbol{\beta}_{bj} \cdot \boldsymbol{B}_t$$

Parameter	Prior distribution
$\boldsymbol{\beta}_{bj}$ (Spline coefficients in cluster j)	$\text{Normal}\left(\mu = \boldsymbol{\beta}, \Sigma = \sigma_b^2 I_{p\times p}\right)$ $\sigma_b \sim \text{t}_{\text{student}}(\text{df} = 3, \mu = 0, \sigma = 2.5)$
$\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)$ (Overall coefficients across all clusters)	$\beta_1 \sim \text{Normal}\left(0, 1\right)$ $\beta_m \sim \text{Normal}\left(\beta_{m-1}, \sigma_\beta^2\right), \text{m} = 2, \ldots, p$ $\sigma_\beta \sim \text{Normal}\left(0, 1\right)$

Random walk prior

Brezger A, Lang S. ***Computational Statistics & Data Analysis.***

Bayesian immediate effect model

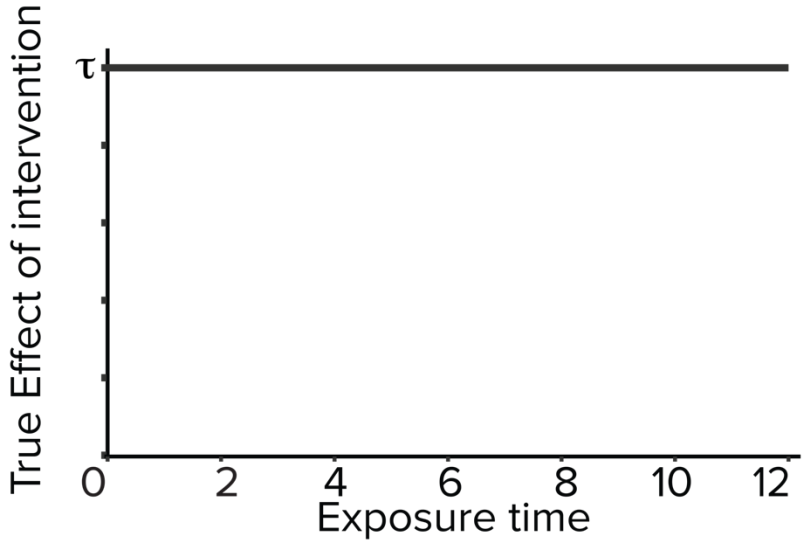
$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_j(t)$$

Parameter	Prior distribution
α (overall intercept)	$\text{Normal}\,(0, 1)$

Bayesian immediate effect model

$$g\left(E\left(Y_{ijt}\middle|t, A_{jt}\right)\right) = \alpha + \tau \cdot A_{jt} + s_j(t)$$

Parameter	Prior distribution
α (overall intercept)	$\text{Normal}\ (0, 1)$
τ (Intervention effect)	$\text{Normal}\ (\mu = 0, \sigma = 5)$



Immediate effect

(duration since the initiation of the intervention in a cluster)

Simulation setting

Assuming $T = 12$, $J = 10$, $I = 10$, and a continuous outcome

$$Y_{ijt} = \alpha_j + \tau \cdot A_{jt} + s_j(t) + \epsilon_{ijt}$$

Variable/Parameter	**Generated from Distribution**
α_j (Cluster-specific random intercept)	$\text{Normal}\,(\mu = 0, \sigma = 0.5)$
τ (Intervention effect)	$0,1,2,3,4,5$
$s_j(t)$ (Cluster-specific temporal effect)	$s_j(t) = -0.01\text{t}^2 + \gamma_{jt}$
γ_{jt} (Cluster-specific random time effect)	$\Gamma_j = (\gamma_{j1}, \ldots, \gamma_{jT}) \sim MVN(\mathbf{0}, \Sigma_{T\times T})$; $\Sigma_{T\times T} = DRD$ $D = 0.6 \cdot I_{T\times T}$ $R = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix}$, ρ=0.95
ϵ_{ijt} (Individual and time-specific noise)	$\text{Normal}\,(\mu = 0, \sigma = 1)$

Conventional frequentist models for comparison

1. Linear mixed effect model treating <u>time as factor</u>
   - $Y_{ijt} = \alpha + \gamma_t + b_{1jt} + \tau \cdot A_{jt} + b_{0j} + \epsilon_{ijt}$

Fixed effect for time

Random effect for time t within cluster j

Conventional frequentist models for comparison

1. Linear mixed effect model treating time as factor

$$Y_{ijt} = \alpha + \gamma_t + b_{1jt} + \tau \cdot A_{jt} + b_{0j} + \epsilon_{ijt}$$

2. Linear mixed effect model treating <u>time as continuous variable</u>

$$Y_{ijt} = \alpha + (\beta_2 + b_{1j})t + \tau \cdot A_{jt} + b_{0j} + \epsilon_{ijt}$$

Slope for time

Random slope for time

Conventional frequentist models for comparison

1. Linear mixed effect model treating time as factor

   $Y_{ijt} = \alpha + \tau \cdot A_{jt} + \gamma_t + b_{0j} + b_{1jt} + \epsilon_{ijt}$

2. Linear mixed effect model treating time as continuous variable

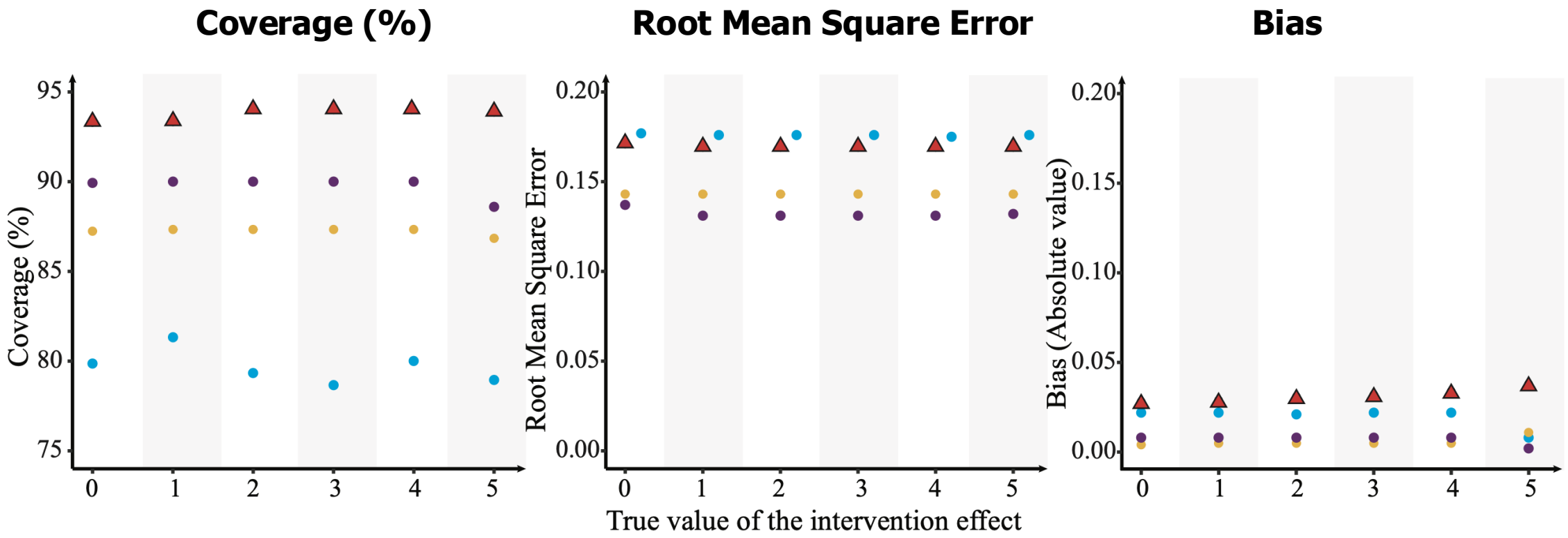
   $Y_{ijt} = \alpha + (\beta_2 + b_{1j})t + \tau \cdot A_{jt} + b_{0j} + \epsilon_{ijt}$

3. Generalized additive models smoothing time trend

   $Y_{ijt} = \alpha + s(t) + s_j(t) + \tau \cdot A_{jt} + \epsilon_{ijt}$

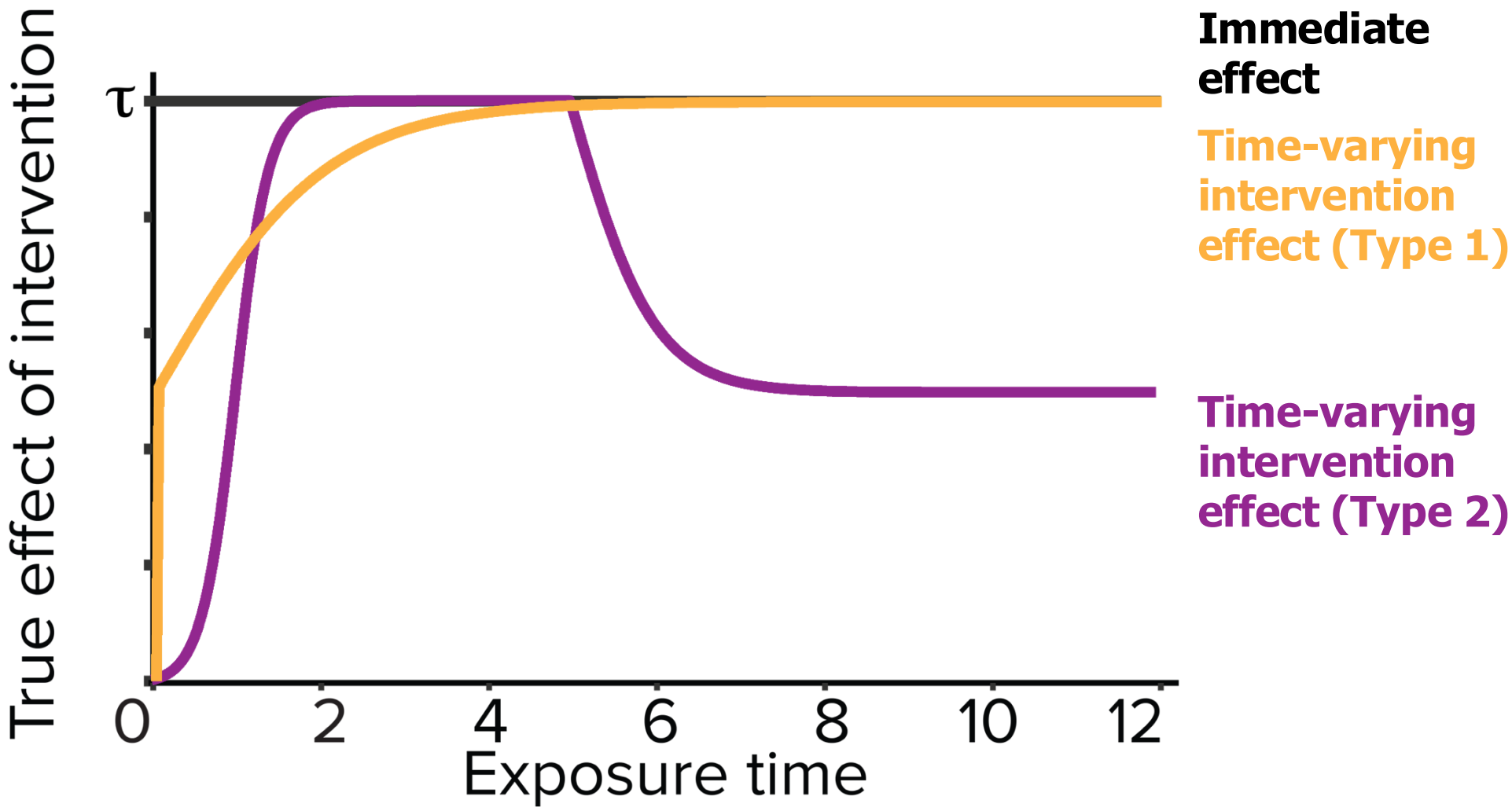
Fixed smooth function for time Cluster-specific random smooth function for time

Simulation results – comparative evaluation



Bayesian spline modeling for estimating time-varying intervention effects

Time-varying intervention effect



Bayesian time-varying effect model with cluster-specific random effect

$t_j{}^*$: exposure time; the duration since the initiation of the intervention in the j^{th} cluster
t: study time; the elapsed time since the onset of the entire study

$$g\left(E\left(Y_{ijt}\middle|t, t_j{}^*, A_{jt}\right)\right) = \alpha + \tau_j(t_j{}^*) \cdot A_{jt} + s_j(t)$$

$$= \alpha + e^{u_j} \cdot \tau\left(t_j{}^*\right) \cdot A_{jt} + s_j(t)$$

$\mu_j \sim \text{Normal}(0, \sigma_u^2)$
$\sigma_u \sim \text{Normal}(0, 0.2^2)$

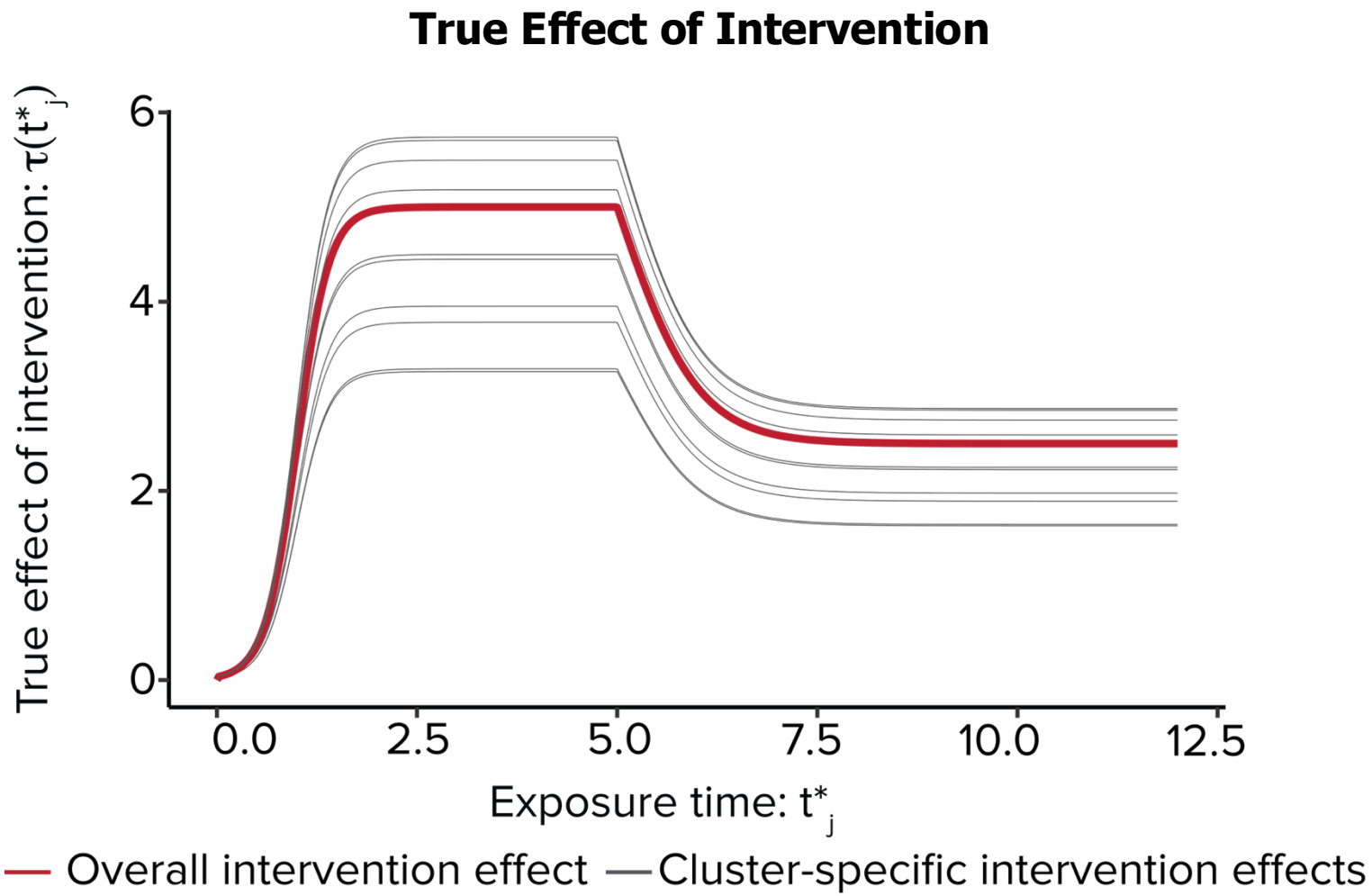
$$= \alpha + e^{u_j} \cdot \boldsymbol{\beta}^* \cdot \boldsymbol{B}_{t_j{}^*} \cdot A_{jt} + \boldsymbol{\beta}_{bj} \cdot \boldsymbol{B}_t$$

Random walk prior

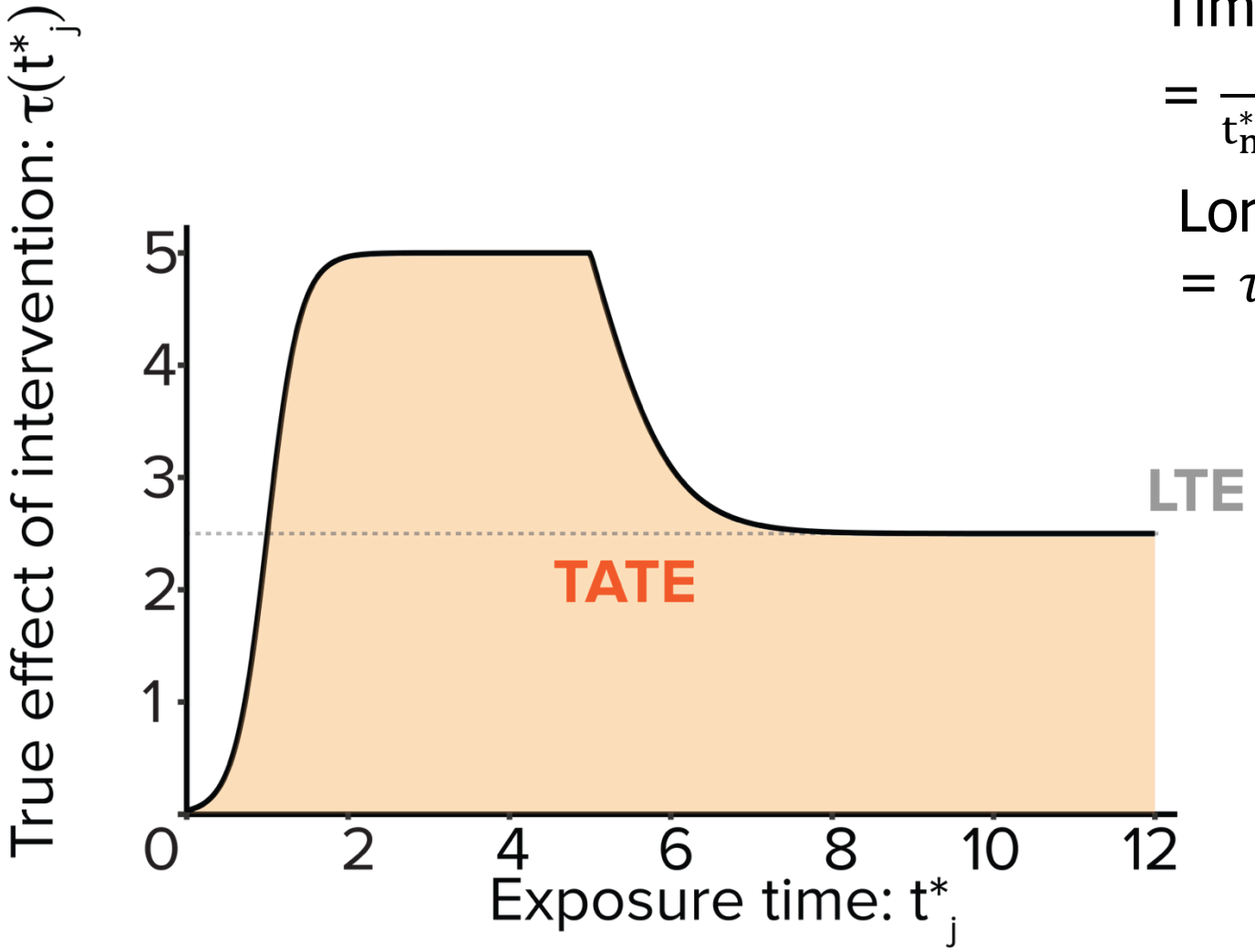
e: exponential function
e^{u_j}: cluster-specific effect modifier; account for the cluster-specific variations in the intervention effect ($\tau\left(t_j{}^*\right)$)

Simulation setting

Assuming $T = 22$, $J = 10$, $I = 10$, and a continuous outcome



Two summary measure for the magnitude of the intervention effect curve



Time-averaged treatment effect (TATE)

$$= \frac{1}{t^*_{max}} \int_0^{t^*_{max}} \tau(t^*)\, dt^*$$

Long-term treatment effect (LTE)

$$= \tau(t^*_{max})$$

Model for comparison: Bayesian time-varying effect model without cluster-specific random effect

$$\tau_j(t_j^*)$$

$$g\left(E\left(Y_{ijt}\middle|t, t_j^*, A_{jt}\right)\right) = \alpha + \tau\left(t_j^*\right) \cdot A_{jt} + s_j(t)$$

Existing model: Bayesian monotone effect curve model

$$g\left(E\left(Y_{ijt}\middle|t, t_j^*, A_{jt}\right)\right) = \alpha + \tau\left(t_j^*\right) \cdot A_{jt} + s_j(t) + \epsilon_{ijt}$$

⬇

Monotonic step function

Kenny A et al. ***Statistics in Medicine***. 2022

Existing model: frequentist natural cubic spline model with cluster-specific random effect

$$
\begin{aligned}
Y_{ijt} &= \alpha + s_j(t) + \tau_j(t_j^*) \cdot A_{jt} + \epsilon_{ijt} \\
&= \alpha + s(t) + b_{0j} + \tau_j(t_j^*) \cdot A_{jt} + \epsilon_{ijt} \\
&= \alpha + \boldsymbol{\beta} \cdot \boldsymbol{B}_{ns} + b_{0j} + (\boldsymbol{\beta}^* \cdot \boldsymbol{B}_{ns}^* + b_{1j}) \cdot A_{jt} + \epsilon_{ijt}
\end{aligned}
$$

Cluster-specific random intercept

Cluster-specific random intervention effect

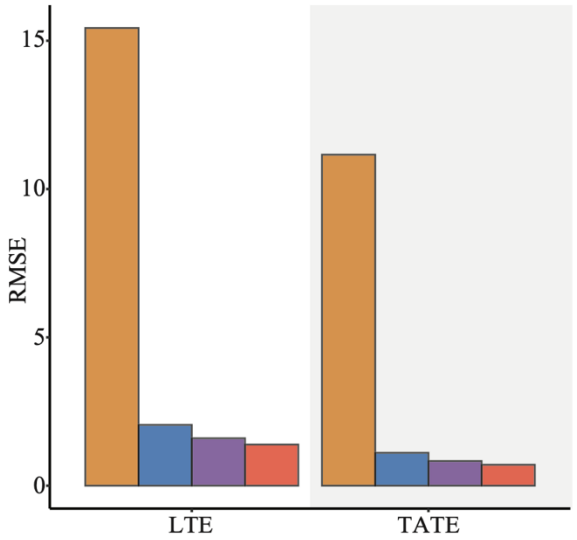
$$
\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2 & \rho\sigma\upsilon \\ \rho\sigma\upsilon & \upsilon^2 \end{pmatrix} \right)
$$

$\boldsymbol{B}_{ns}$: natural cubic spline basis function

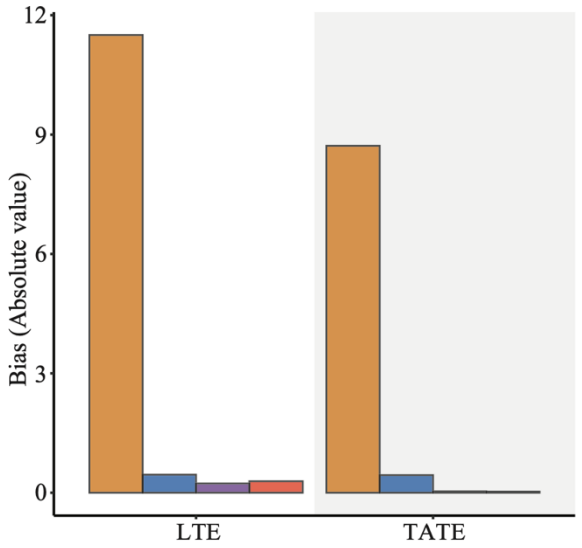
Kenny A et al. ***Statistics in Medicine***. 2022

Simulation results

Root Mean Square Error



Bias



- Bayesian monotone effect curve model
- Bayesian time-varying effect model
- Frequentist natural cubic spline model
- **Bayesian time-varying model with cluster-specific random effect**

The proposed Bayesian model has higher coverage rates (%) than the Frequentist model

Exposure time point	Bayesian cluster-specific time varying effect model	Frequentist natural cubic spline model
2	72	56
4	95	71
6	86	53
8	90	75
10	90	74
12	87	72
14	85	73
16	83	73
18	85	67
20	88	69

Application: Primary palliative care for emergency medicine (PRIM-ER study)

Goal:
Reduce disposition to an acute care setting

Intervention:
Primary palliative care training and computer decision support

Data source:
29 emergency departments in the U.S. (56 months)
~100,000 older patients with serious, life-limiting illnesses

Binary (yes/no) outcome:

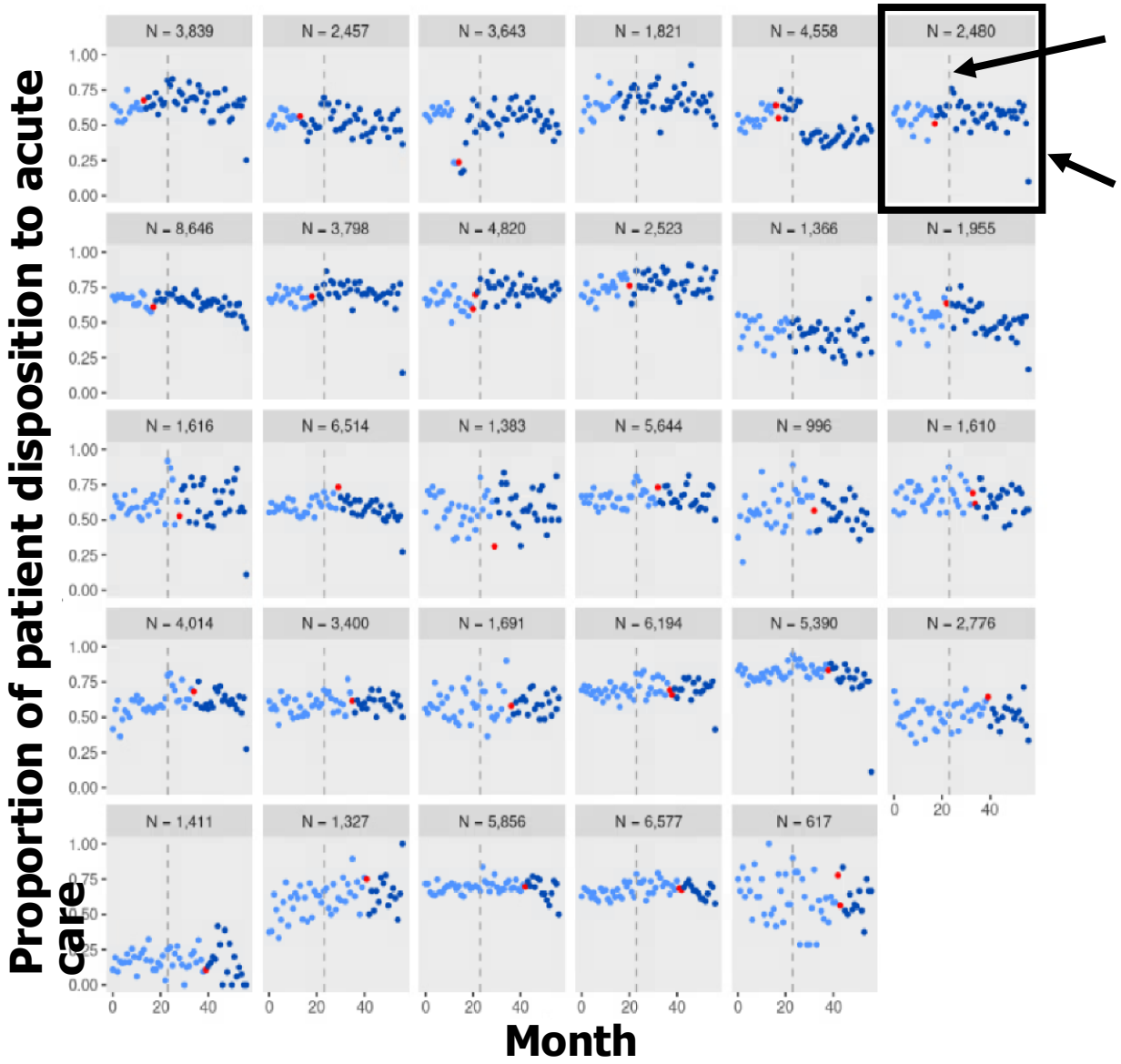
patient is disposed to an acute care setting

Covariates:

- Intervention status (control/intervention)
- COVID-19 pandemic period (yes/no)
- Interaction between intervention status and COVID-19 pandemic period

Grudzen C et al. *JAMA*. 2025

The observed proportions of patient disposition to an acute care setting

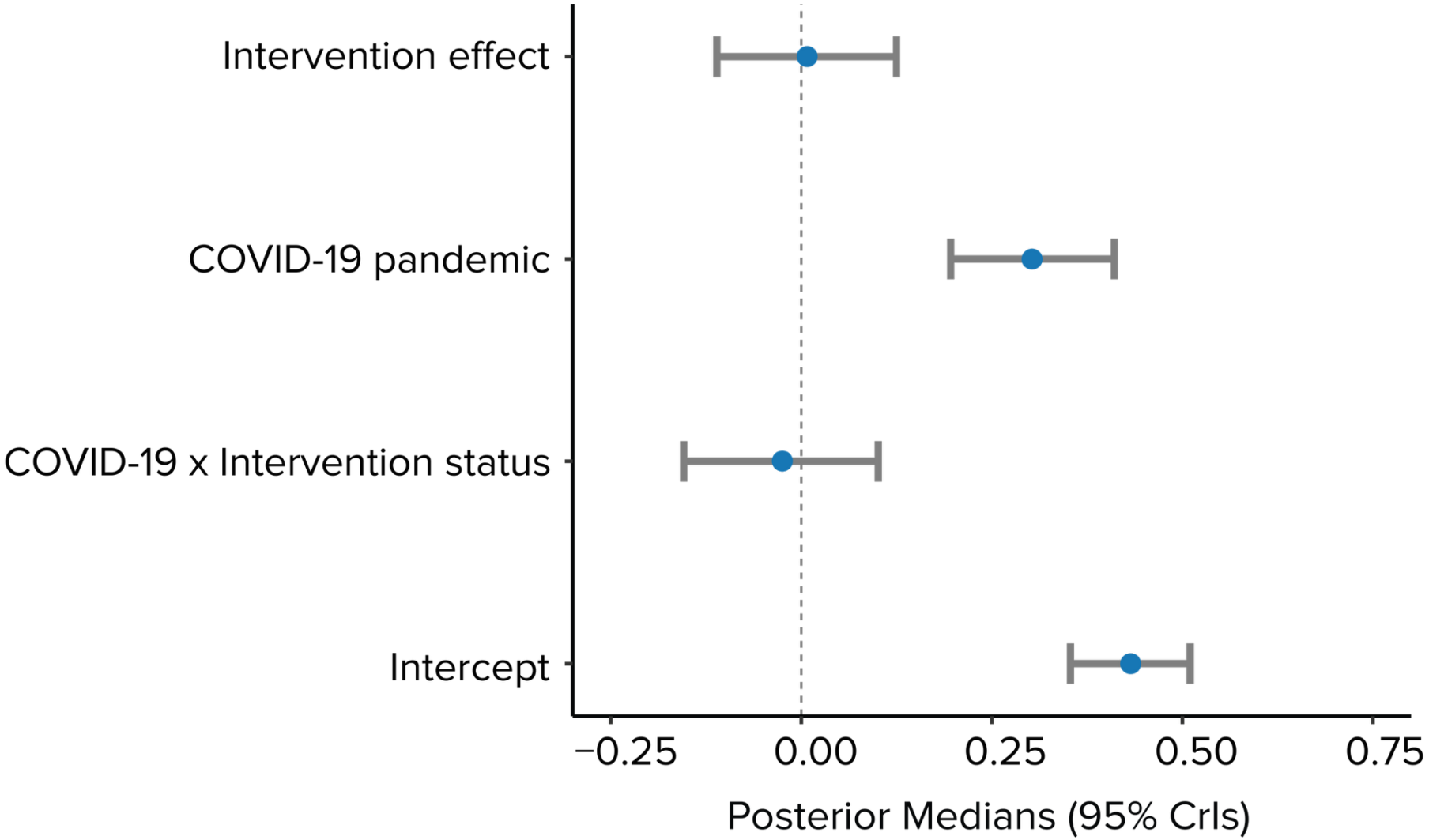


COVID-19 pandemic start

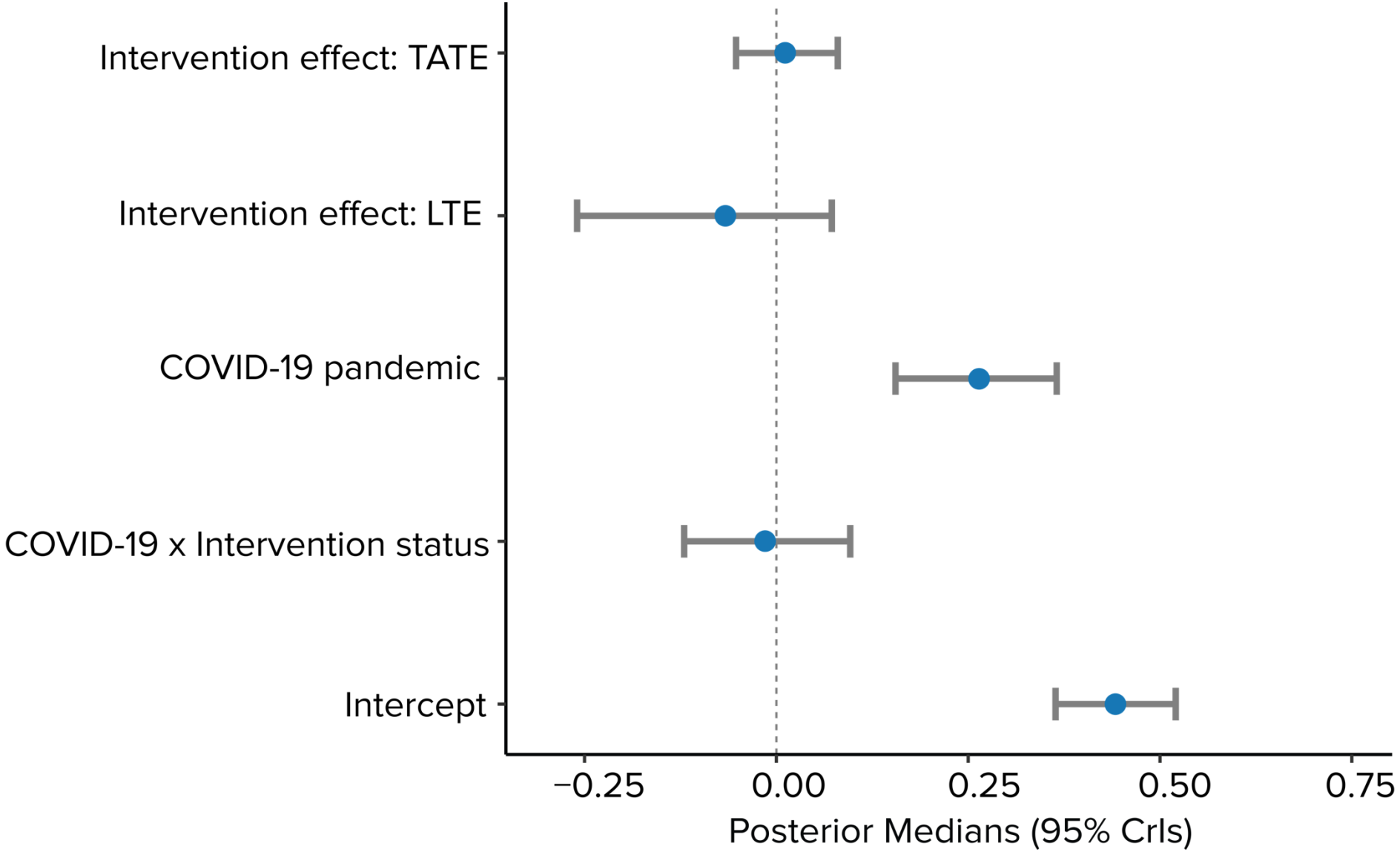
One cluster

- Control status
- Intervention status
- Training happens

Bayesian immediate effect model results



Bayesian time varying effect model results



Conclusion

Bayesian hierarchical penalized spline models for advancing SWCRT Analysis

- Provide more reliable interval estimations while maintaining high estimation accuracy
- Solve low coverage issue of conventional frequentist methods
- Offer a robust statistical alternative

Implications:

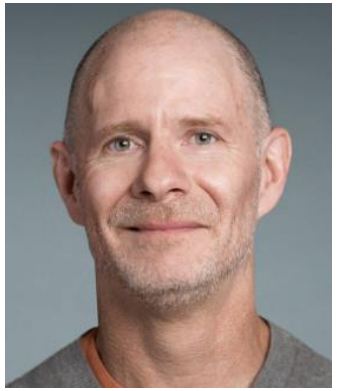
- Helpful for understanding treatment effect dynamics
- Inform policy-making and improves clinical practice

RESEARCH ARTICLE | Open Access

Bayesian Hierarchical Penalized Spline Models for Immediate and Time-Varying Intervention Effects in Stepped Wedge Cluster Randomized Trials

Danni Wu, Hyung G. Park, Corita R. Grudzen, Keith S. Goldfeld

Acknowledgements



Dr. Keith Goldfeld

Dr. Hyung Park

Dr. Corita Grudzen

Which Small-Sample Correction Should Be Used When Analyzing Stepped-Wedge Designs with Time-Varying Treatment Effects?

A simulation study of robust variance estimators for ETI models

Yongdong Ouyang, PhD

Department of Biostatistics and Bioinformatics

Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA

University at Buffalo, Buffalo, New York, USA

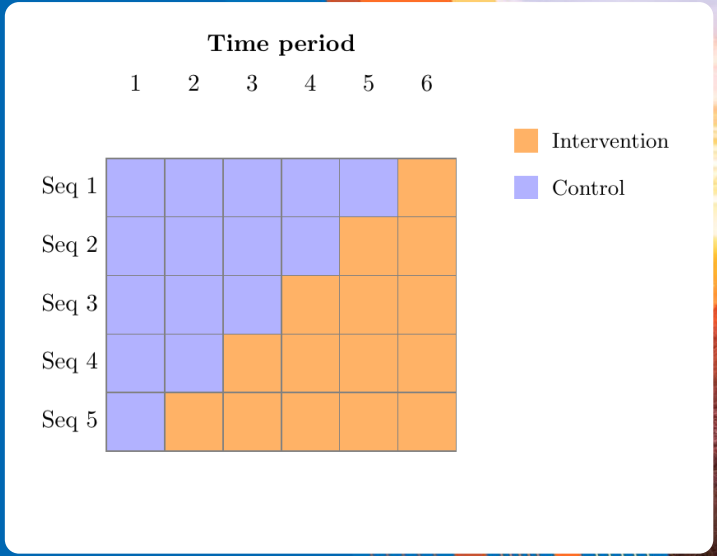
SCT | 47TH ANNUAL MEETING

Acknowledgment

This is joint work with: **Dr. Fan Li** (Yale University), **Dr. James Hughes** (University of Washington), **Dr. Monica Taljaard** (Ottawa Hospital Research Institute)

What Is a Stepped-Wedge Cluster Randomized Trial?

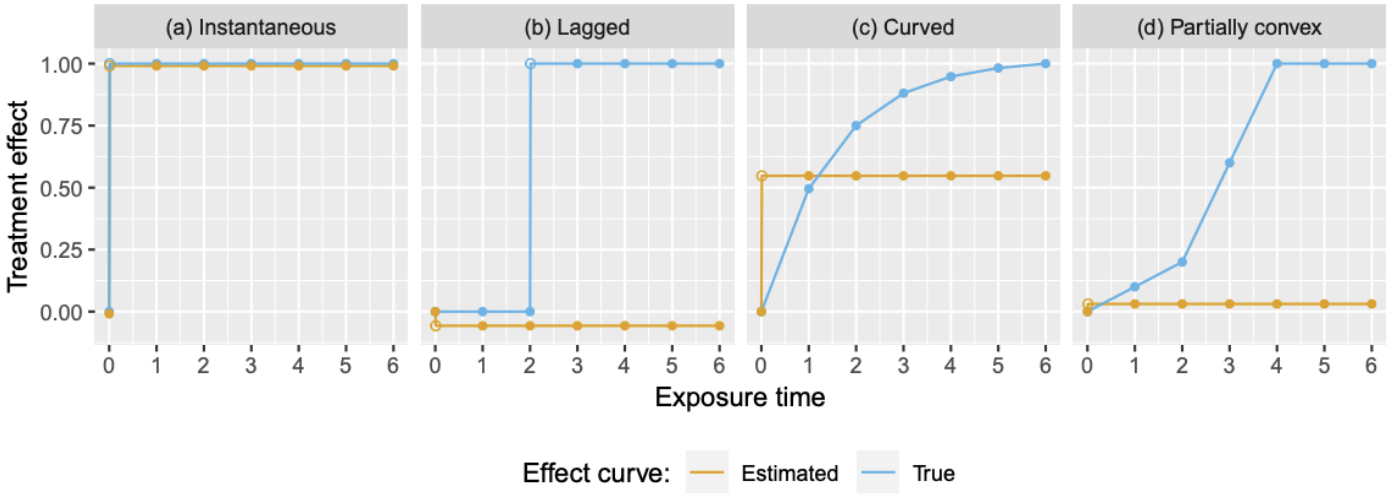
- Clusters (clinics, schools, villages) are the unit of randomization.
- All clusters start in the control condition.
- Clusters cross over to the intervention at staggered, randomized times.
- By the final period, all clusters are exposed.



Why use it?

- Logistically attractive: phased rollout.
- Every cluster eventually receives the intervention.
- But: treatment is confounded with calendar time.

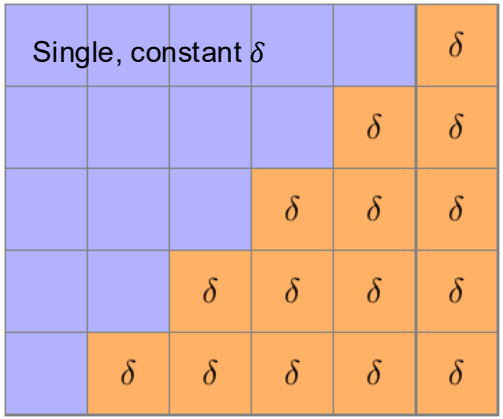
Constant vs. Time-vary treatment



Kenny et al. (2022): when effects truly vary with e, the IT model may converge to a weighted average with negative weights.

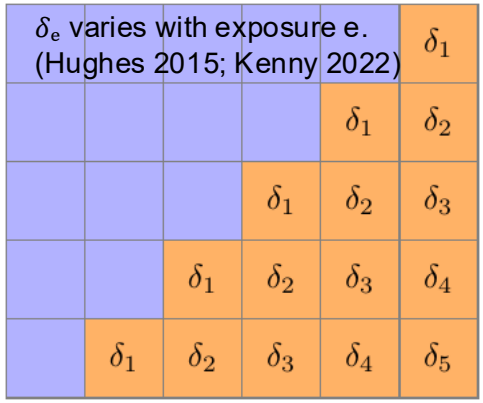
IT vs. ETI: Designs, Equations, and Estimands

(a) Immediate Treatment (IT)



$$\eta_{ijk} = \mu + \beta_j + \delta\, X_{ij} + \alpha_{ij}$$

(b) Exposure Time Indicator (ETI)



$$\eta_{ijk} = \mu + \beta_j + \sum_{e=1}^{J-1} \delta_e\, I_{ij}^{(e)} + \alpha_{ij}$$

Exposure time indicator (ETI) summary estimands

$$\text{TATE} = \frac{1}{J-1} \sum_{e=1}^{J-1} \delta_e \qquad \text{LTE} = \delta_{J-1}$$

Two Layers of Model Specification

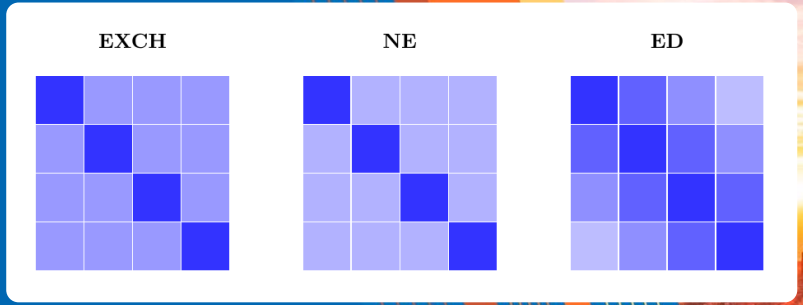
Fixed effects (treatment structure)

- IT vs. ETI
- Should be prespecified from mechanism, not chosen post hoc.

Random effects (correlation structure)

- Exchangeable (EXCH): single cluster RE
- Nested Exchangeable (NE): + cluster × period RE
- Exponential Decay (ED): between-period corr. decays
- + Random Intervention (RI): heterogeneous effects

Within-cluster correlation: period j vs. j'



EXCH: uniform. NE: WP-ICC > BP-ICC (constant). ED: BP-ICC decays with |j−j'|.

The problem: True correlation is unknown ⇒ misspecified random effects ⇒ biased model-based SEs and undercoverage. Worse with few clusters.

What Is Already Known, and What Is Not

Earlier work has shown: Robust variance estimator (RVEs)

- For IT models (continuous): Mancl-Derouen (MD) with t-based inference recover nominal coverage under misspecification (Ouyang et al. 2024).
- For IT models (binary, GEE): Kauermann and Carroll (KC) and Fay-Graubard (FG) corrections perform well (Ford & Westgate 2020; Thompson et al. 2021).

Open questions this paper addresses

- 1. Continuous outcomes: do RVEs give valid inference for TATE and LTE under misspecification in ETI models?
- 2. Binary outcomes: how do RVEs perform under correctly specified and misspecified random-effects structures?
- 3. Which RVE + reference distribution should be the default when clusters are few?

Setting Up the Sandwich: GLMM Framework

The cluster-level GLMM

$$g\big(\mathbb{E}[\mathbf{y}_i \mid \mathbf{b}_i]\big) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}))$$

- X, Z: fixed and random-effect design matrices.
- β contains the IT or ETI treatment effects.
- $R(\theta)$ encodes the random-effects structure.

Linearization (non-Gaussian outcomes)

Working pseudo-response:

$$\mathbf{P}_i = \boldsymbol{\Delta}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) + \boldsymbol{\eta}_i$$

Working marginal covariance:

$$\widehat{\mathbf{V}}_i = \mathbf{Z}_i\mathbf{R}(\hat{\boldsymbol{\theta}})\mathbf{Z}_i^\top + \hat{\boldsymbol{\Delta}}_i^{-1}\boldsymbol{\Sigma}_i\hat{\boldsymbol{\Delta}}_i^{-1}$$

The residual everyone uses:

$$\mathbf{e}_i = \hat{\mathbf{P}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$$

All four sandwich estimators share these residuals. They differ only in how they adjust the outer product before applying the bread.

General Sandwich Form

The cluster-robust variance estimator

$$\widehat{\mathbf{V}}_{\text{sand}} = \mathbf{M}^{-1}\left(\sum_{i=1}^{I} \mathbf{X}_i^\top \widehat{\mathbf{V}}_i^{-1}\, \mathbf{F}_i\, \mathbf{e}_i \mathbf{e}_i^\top\, \mathbf{F}_i^\top\, \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i\right) \mathbf{M}^{-1}$$

Bread (the same for all four)

$$\mathbf{M} = \sum_{i=1}^{I} \mathbf{X}_i^\top \widehat{\mathbf{V}}_i^{-1} \mathbf{X}_i$$

The sandwich estimators generally differ only in the residual-adjustment matrix F_i.

The Four Adjustments at a Glance

Estimator	$\mathbf{F}_i$	Key idea
Classic (Liang & Zeger 1986)	$\mathbf{I}$	No adjustment; biased when I smal
KC (Kauermann & Carroll; Bell & McCaffrey 2002)	$(\mathbf{I} - \mathbf{H}_i^\top)^{-1/2}$	Inflates residuals; moderate bias correction
MD (Mancl & DeRouen 2001)	$(\mathbf{I} - \mathbf{H}_i^\top)^{-1}$	Squared KC; approximate jackknife, more conservative
MBN (Morel, Bokossa, Neerchal 2003)	(different form)	$\widehat{\mathbf{V}}_{\text{MBN}} = c\,\widehat{\mathbf{V}}_{\text{classic}} + \delta I \phi \mathbf{M}^{-1}$ scaling + inflation

$$\mathbf{H}_i = \mathbf{X}_i \mathbf{M}^{-1} \mathbf{X}_i^\top \widehat{\mathbf{V}}_i^{-1}$$

Captures how much cluster i "pulls" its own fitted values toward itself.

Inference companion: Each RVE is paired with Normal N(0,1) or t with I−2 d.f. The choice substantially affects coverage.

Code: github.com/jphughes9/vcovCRglmer

Simulation Design

Common design grid: $I \in \{8, 16, 32\}$, $J \in \{5, 9\}, K \in \{10, 50\}$

Continuous outcomes

- Generated from ED-RI (most complex)
- Fitted: misspecified EXCH
- True effects: linear ramp 0 → 1 across exposure

Binary outcomes

- Generated from EXCH, NE, NE-RI
- Fitted: correct and misspecified (EXCH)
- Constant log-OR = 0.25
- Baseline event probability: 0.2, 0.5

Performance metrics

- Bias of the point estimate.
- **Empirical 95% CI coverage for TATE (primary metric).**
- All combinations of {Classic, KC, MD, MBN, model-based} × {Normal, t_{I-2}}.

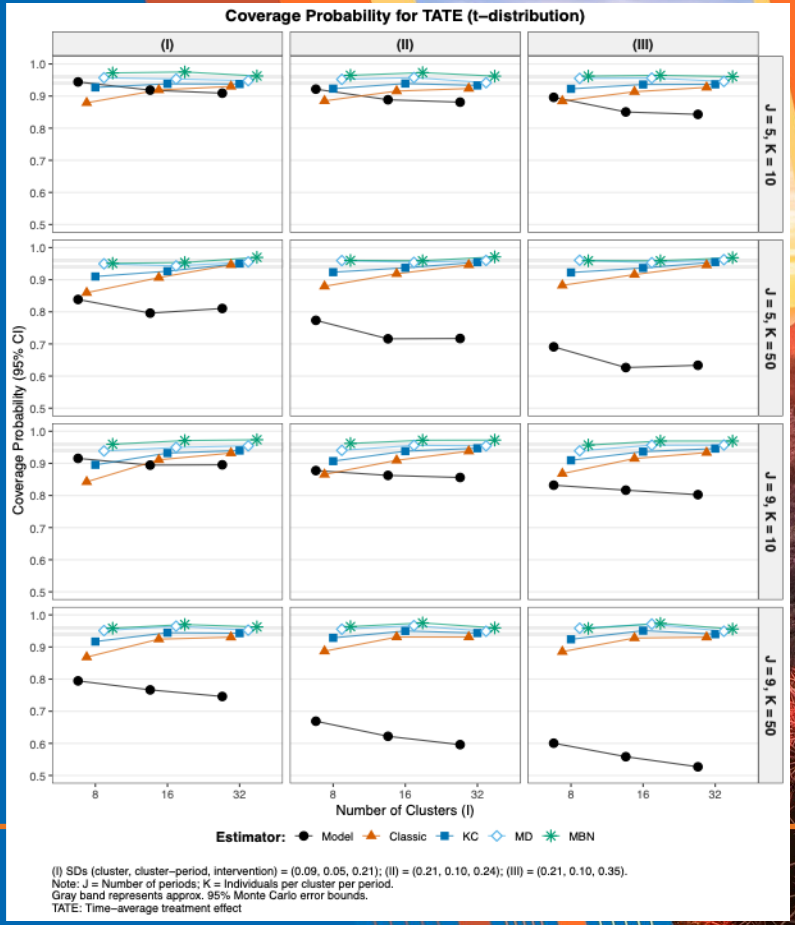
Continuous Outcomes: TATE Coverage

Setup: EXCH fitted, ED-RI true (misspecified).

Each panel: coverage vs. I for one (J, K) combination.

KEY Takeaway:

MD with t_{I-2} is the only RVE that hits the 0.95 line in every (J, K) panel. Model-based SEs undercover progressively as K grows.

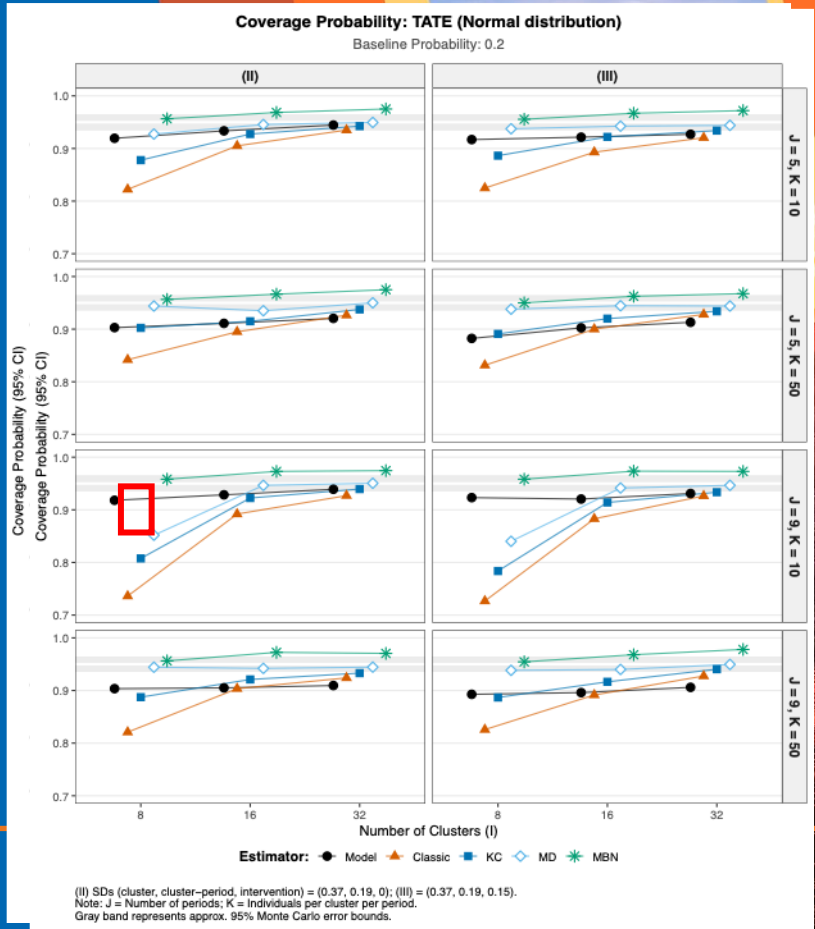


Binary Outcomes: TATE Coverage

Setup: misspecified RE structure, p_0 = 0.2.

KEY Takeaway:

- MBN (star) is uniformly stable, but conservative under t-distribution
- MD breaks down in 1CPS (J = 9, K = 10, I = 8); drops below 90% coverage.
- MBN under normal approximation maintained nominal 95% coverage across all scenarios

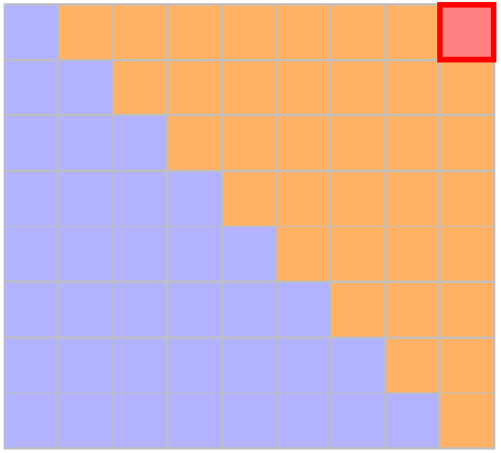


The 1CPS Vulnerability: Why MD Breaks Down

1-cluster-per-sequence (1CPS) design:
$I = 8, J = 9, K = 10.$

- The LTE is identified by a single cluster-period.
- With low baseline probability (p_0 = 0.2): zero-event cells become common ⇒ unstable LTE estimates ⇒ propagates into TATE bias.
- MD jackknife is highly sensitive to this high-leverage cell. MBN is much more stable.

$$\text{TATE} = \frac{1}{J-1}\sum_{e=1}^{J-1}\delta_e \qquad \text{LTE} = \delta_{J-1}$$



Only one cluster-period informs the LTE

Practical warning: Do not make inferences about exposure periods with zero or near-zero events. If sensible, combine late exposure periods (Hughes et al. 2024).

Practical Recommendations

For the TATE

- Continuous outcomes: MD with t_{I−2} as default.
- Binary outcomes: MBN with Normal reference as default.
- Avoid relying on model-based SEs when correlation may be misspecified.

For the LTE

- All methods are less reliable than for TATE.
- Treat LTE as a secondary or exploratory estimand whenever possible.
- Avoid MD for binary outcomes in sparse settings (especially 1CPS).
- If sparsity threatens: combine late exposure periods, or redesign.

Design-stage advice

- Prespecify IT vs. ETI from mechanism (uptake curve, biological lag), not data.
- Ensure adequate event support for the longest-exposure cells.

Limitations of This Study

- Limited set of random-effects structures; GEE-based ETI not evaluated.
- **While coverage probably may be maintained, the impact on power has not been evaluated.**
- **For binary outcomes, non-collapsibility of the OR can drive part of the IT vs. ETI divergence, independent of misspecification.**
- Cross-sectional, balanced, complete SW-CRTs only; not cohort, staircase, or hybrid designs.
- Continuous (LMM) and binary (logistic GLMM) outcomes only; no Poisson, negative binomial, log links.
- For binary outcomes, only a flat exposure-time profile was simulated; truly time-varying binary effects untested.

Take-Home Message

Valid inference for time-varying effects in SW-CRTs depends jointly on:

- 1. A prespecified, mechanism-informed estimand (IT, TATE, or LTE).
- 2. An appropriate treatment-effect time scale (exposure vs. calendar time).
- 3. A robust variance estimator that survives small-sample regimes.

Default recipes (this paper)

Continuous TATE → **MD + I_{t-2}**

Binary TATE → **MBN + Normal**

Code: github.com/jphughes9/vcovCRglmer

Fred Hutch Cancer Center

Does it matter? Evidence of time-varying treatment effects in published stepped wedge trials

Emily Voldal
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center
SCT Annual Meeting, Phoenix, Arizona
May 20, 2026

In collaboration with:
Yongdong Ouyang, Hao Wang, Avi Kenny, Julia Shaw, Fan Xia, Patrick Heagerty, Karla Hemming, Fan Li, Monica Taljaard, James P Hughes

Disclosures

- No relevant disclosures

Acknowledgements

- Special thanks to:
  - Authors who provided datasets to this repository
  - Students from Yale and the University of Washington
  - Staff at the University of Washington and Ottawa Hospital Research Institute
- Funded in part by NIH Grant AI029168.
  - The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Time-varying treatment effect

- Instead of an immediate and constant treatment effect, what if the treatment effect varies with exposure time?
- Treatment effect $\delta(t)$ is a function of exposure time $t = 1,2, \ldots$

Immediate (and constant) treatment effect

	Time 1	Time 2	Time 3	Time 4
Sequence 1	0	δ	δ	δ
Sequence 2	0	0	δ	δ
Sequence 3	0	0	0	δ

Time-varying treatment effect (over exposure time)

	Time 1	Time 2	Time 3	Time 4
Sequence 1	0	$\delta(1)$	$\delta(2)$	$\delta(3)$
Sequence 2	0	0	$\delta(1)$	$\delta(2)$
Sequence 3	0	0	0	$\delta(1)$

Estimand of interest

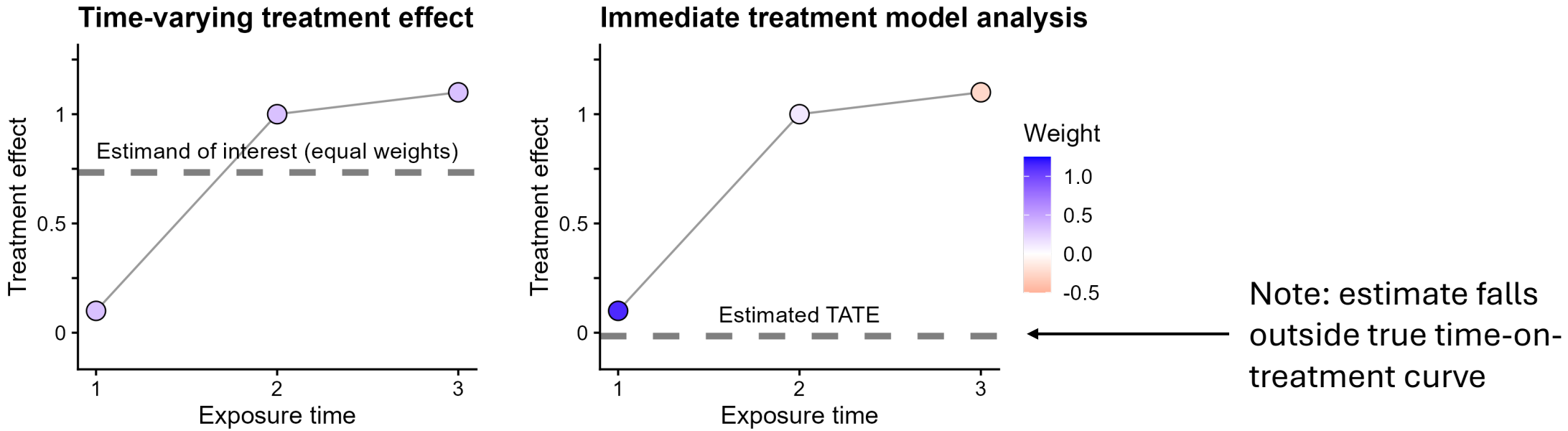
- Time-averaged treatment effect (TATE): average $\delta(t)$ over all observed t
  - E.g. average effect of intervention over first 3 years after implementation
- For immediate and constant treatment effect ($\delta(t) = \delta$), any average with weights summing to one estimates TATE

	Time 1	Time 2	Time 3	Time 4
Sequence 1	0	$\delta(1)$	$\delta(2)$	$\delta(3)$
Sequence 2	0	0	$\delta(1)$	$\delta(2)$
Sequence 3	0	0	0	$\delta(1)$

$$\text{TATE} = \frac{1}{3}(\delta(1) + \delta(2) + \delta(3))$$

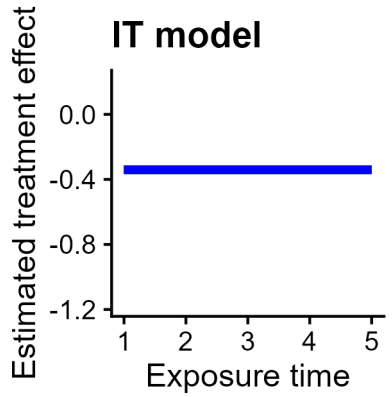
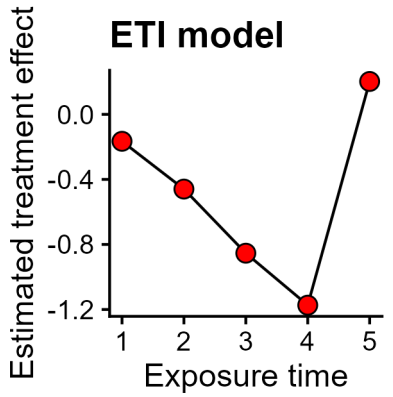
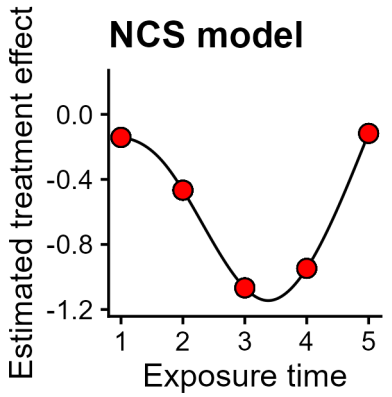
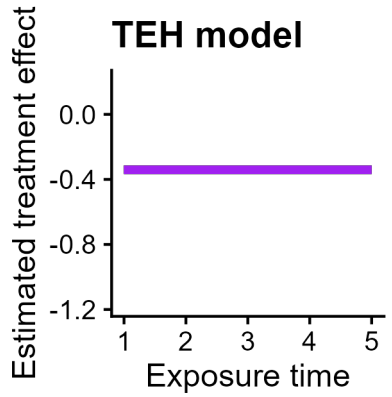
Bias from misspecification

- Fitting a simple (immediate treatment) model to data with a time-varying treatment effect can bias treatment effect estimate
  - Weights based on higher precision at earlier exposure times
- Solution: more flexible analysis models



Kenny et al. 2022.

Model choices for intervention effect

	Immediate treatment (IT)	**Exposure Time Indicator (ETI)**	**Natural Cubic Spline (NCS)**	**Treatment Effect Heterogeneity (TEH)**
Function of exposure time $t = 1, \ldots, T$	$\delta(t) = \delta$	$\delta(t) = \delta_t$	$\delta(t) = NCS\ of\ t$	$\delta(t) = \delta + \alpha_t$
Time-averaged treatment effect (TATE)	$TATE = \hat{\delta}$	$TATE = \frac{1}{T}\sum_{t=1}^{T} \hat{\delta}_t$	$TATE = \frac{1}{T}\sum_{t=1}^{T} \hat{\delta}_t$ where $\hat{\delta}_t$ are predictions from NCS	$TATE = \hat{\delta}$
Number of model coefficients	1	T	<T (depending on number of knots)	1 (plus one additional random effect)
Example of treatment effect curve ($\delta(t)$) shape				

Problem

- Typical immediate treatment analyses can be biased when treatment effects vary by exposure time
- Can be addressed by anticipating varying treatment effect during trial design and pre-specifying appropriate flexible models (at a cost)
- Challenge: little is known about time-varying effects in stepped wedge trials (SWTs), so difficult to make a priori decisions

Aims of repository analysis

- Inform best practice for designing and analyzing SWTs by:
  - Quantifying impact of model assumptions and differences between models
  - Summarizing prevalence of possible time-varying treatment effects
  - Examining shape and magnitude of time-varying treatment effects
- Repository of stepped wedge datasets from published trials

Inclusion/exclusion criteria

- Identified stepped wedge trials from existing literature review[1]; updated searches for recent trials
- Eligibility criteria:
  - Stepped-wedge, cluster-randomized trials
  - No completely observational trials
  - ≥5 clusters and ≥3 time periods
  - Exclude some especially unusual designs (e.g. multi-intervention trials)

[1]Nevins et al. 2024.

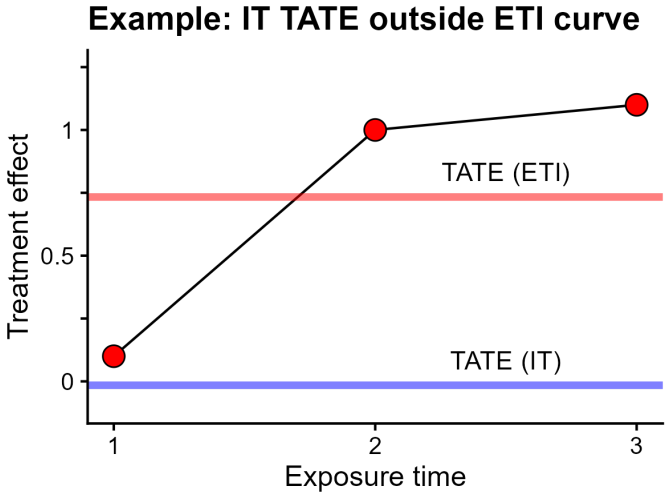
Analysis models

- Estimate time-averaged treatment effect from different treatment models (IT, ETI, NCS, and TEH)
- Random effects models (linear, logistic, and Poisson)
- Simple exchangeable random effects
  - Cross-sectional trial: random cluster intercept
  - Cohort trial: random cluster intercept and random individual intercept
- Robust standard errors appropriate for small samples
  - Mancl and DeRouen with t distribution and df correction[1]
  - Morel, Bokossa, and Neerchal[2] with Normal distribution
  - Not calculated for treatment effect heterogeneity (TEH) model (non-nested random effects)

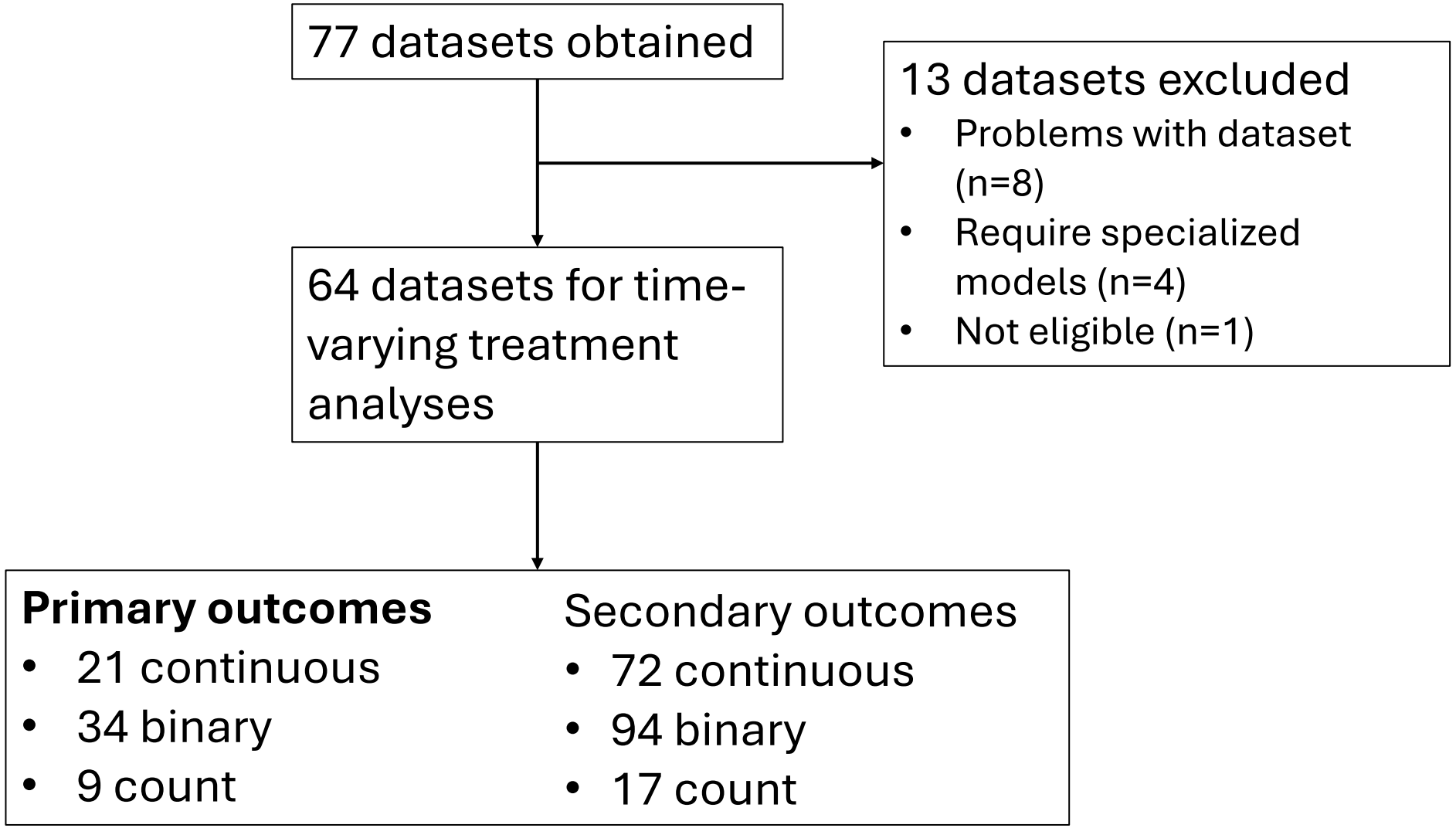
[1]Ouyang et al. 2024; [2]Morel, Bokossa, Neerchal 2003.

Comparison of models

- When is time-averaged treatment effect different?
  - Focus: immediate treatment (IT) vs saturated (ETI) models ($TATE_{IT}$ vs $TATE_{ETI}$, on linear/logit/log scale)
- Primary measure of a large impact of model choice: $TATE_{ETI}$ falls outside the 95% CI for $TATE_{IT}$
- Other metrics:
  - Percent change: $|TATE_{IT} - TATE_{ETI}|$, scaled by $TATE_{IT}$
  - Signs: whether $TATE_{IT}$ vs $TATE_{ETI}$ have different signs
  - TATE outside curve: whether $TATE_{IT}$ falls outside the range of time-specific treatment estimates from the ETI model $[\min(\widehat{\delta_t}), \max(\widehat{\delta_t})]$



Assembly of repository

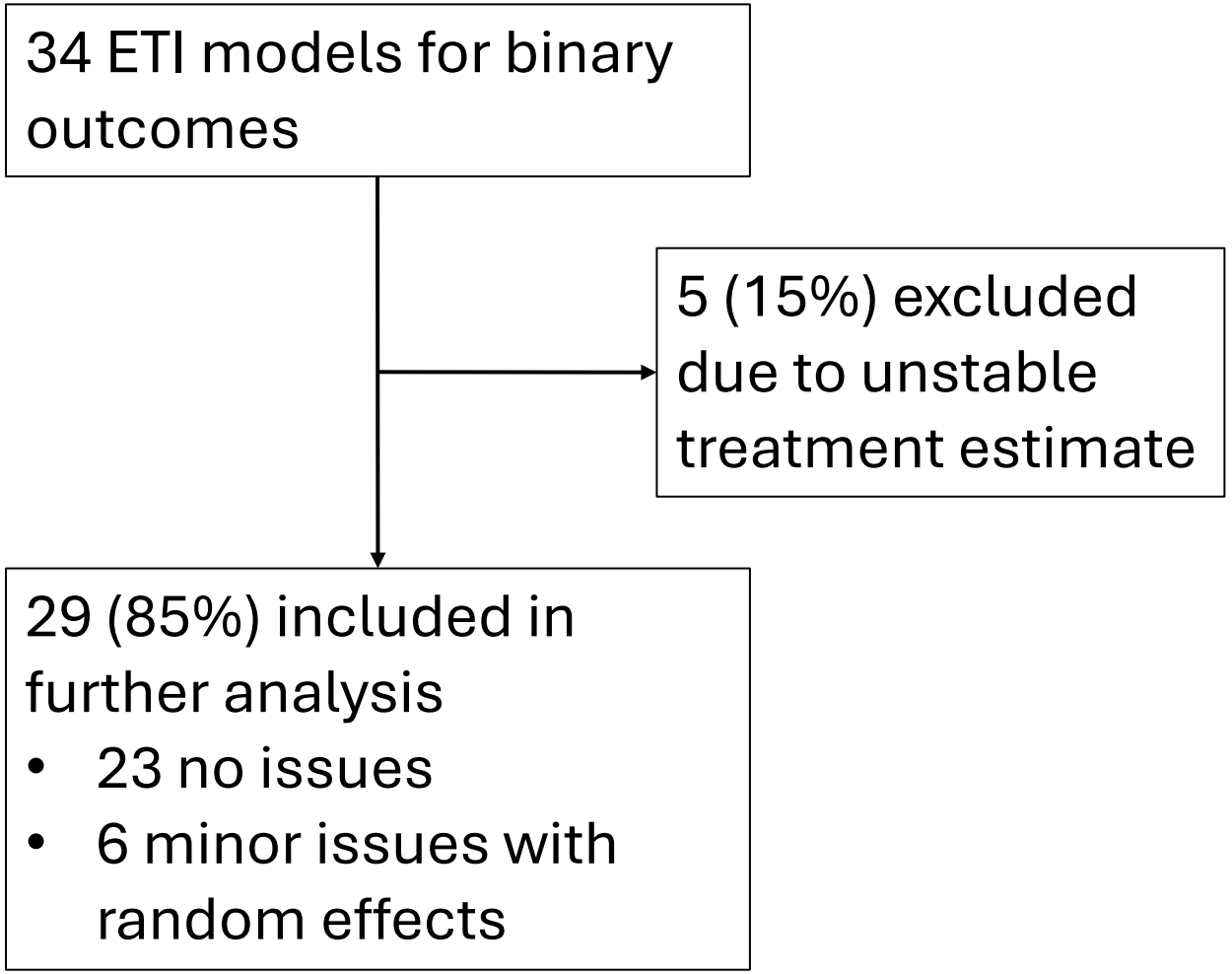


Characteristics of trials

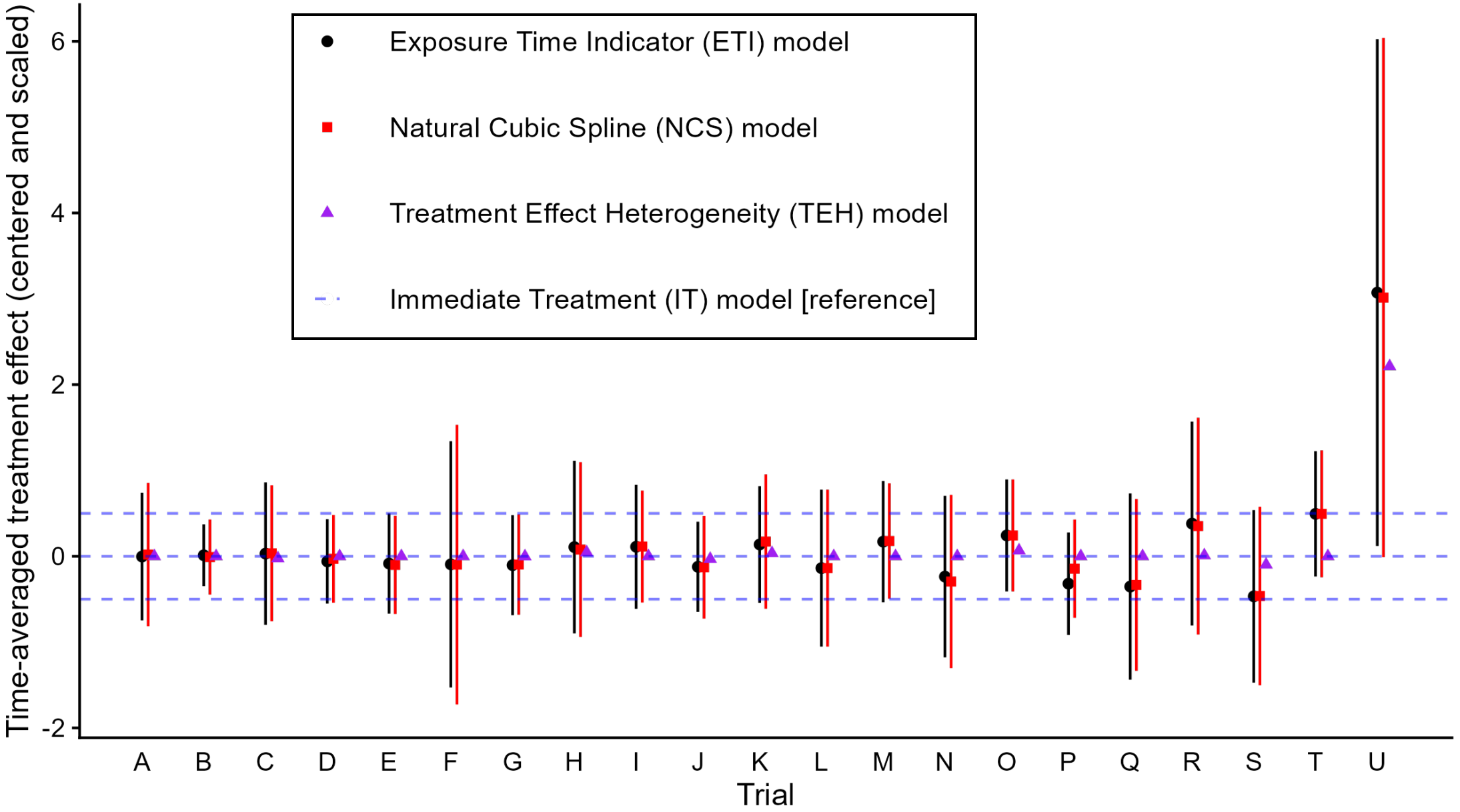
		Overall (n=64)
Study type, n (%)	Cross-sectional	48 (75%)
	Cohort	16 (25%)
Number of clusters randomized, median (IQR)		12 (8,18)
Number of sequences, median (IQR)		5 (4,7)
Type of cluster, n (%)	Hospitals/wards/ clinics	42 (66%)
	Geographical areas	8 (13%)
	Other	14 (22%)

Issues with fitting models

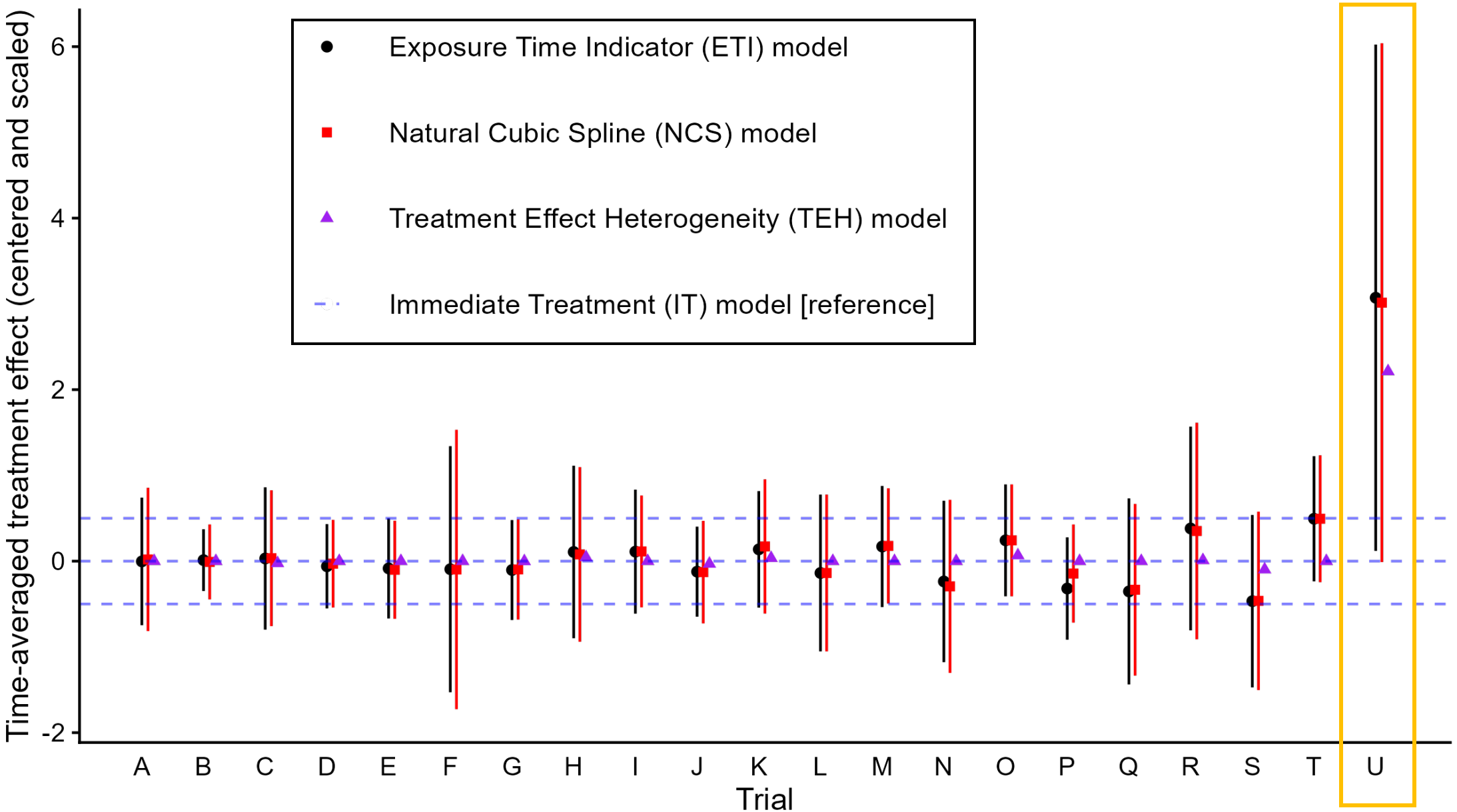
- Continuous and count outcomes: no major issues
- Binary outcomes: frequent issues with flexible ETI/NCS models
  - Often associated with rare outcomes, small sample size

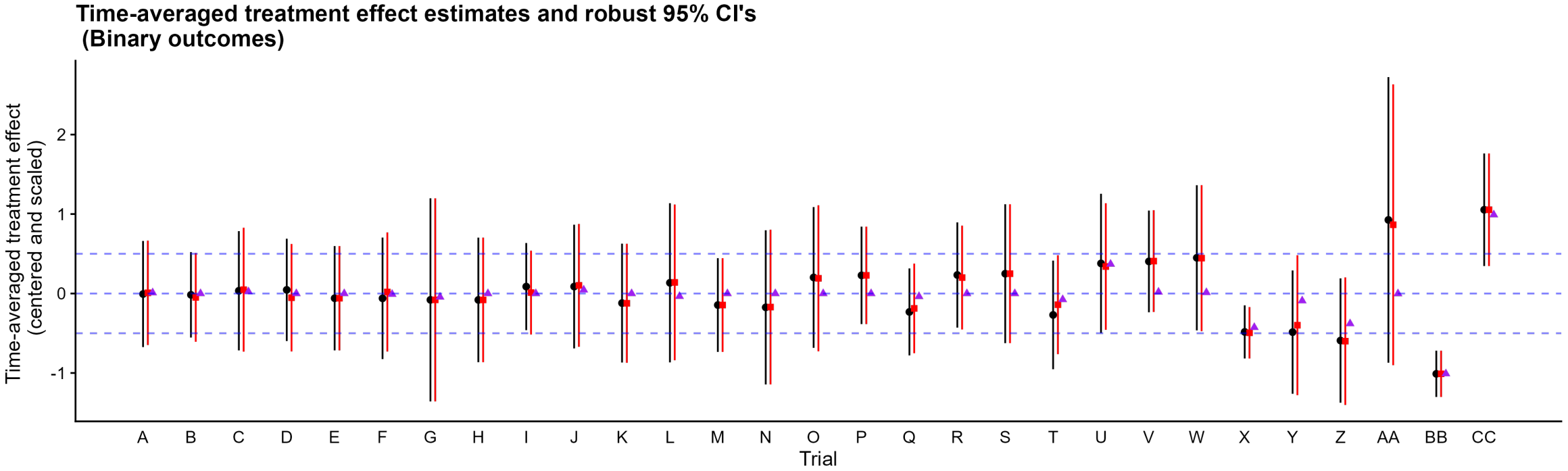


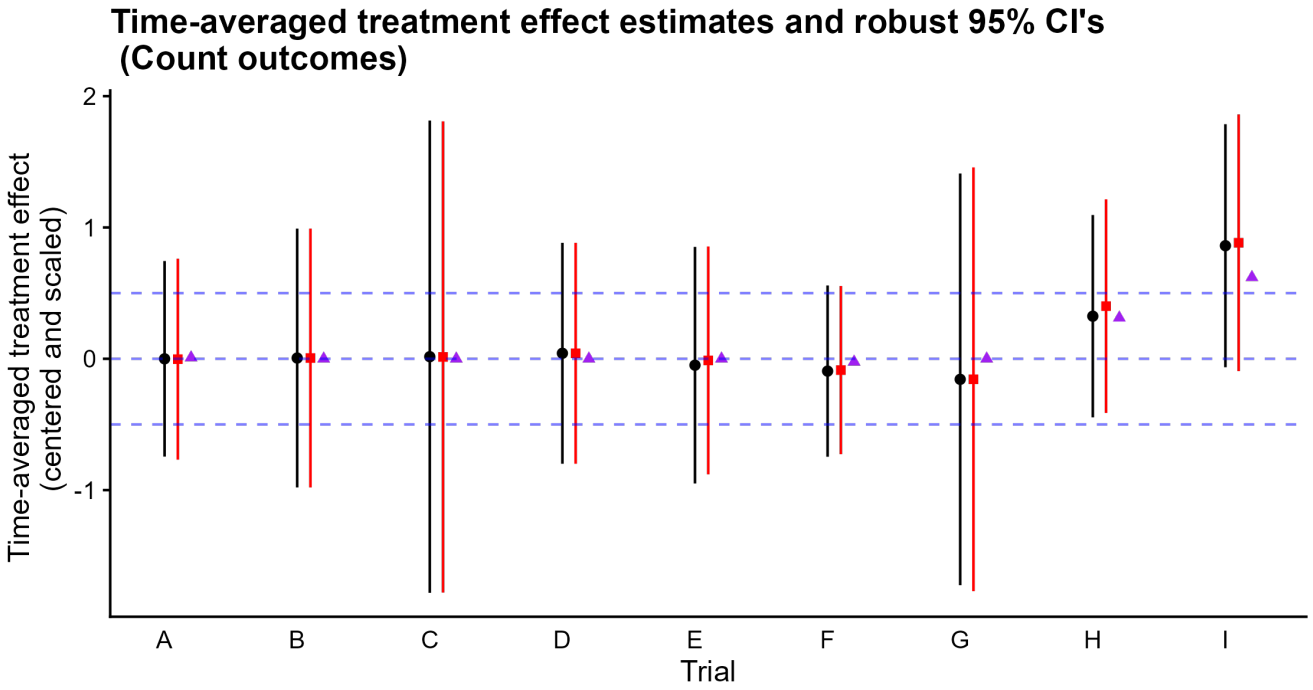
Time-averaged treatment effect estimates and robust 95% CI's (Continuous outcomes)

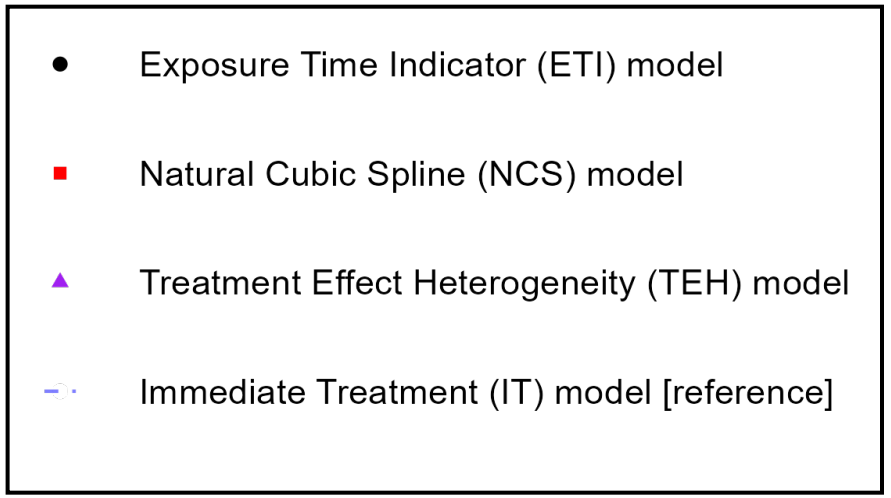


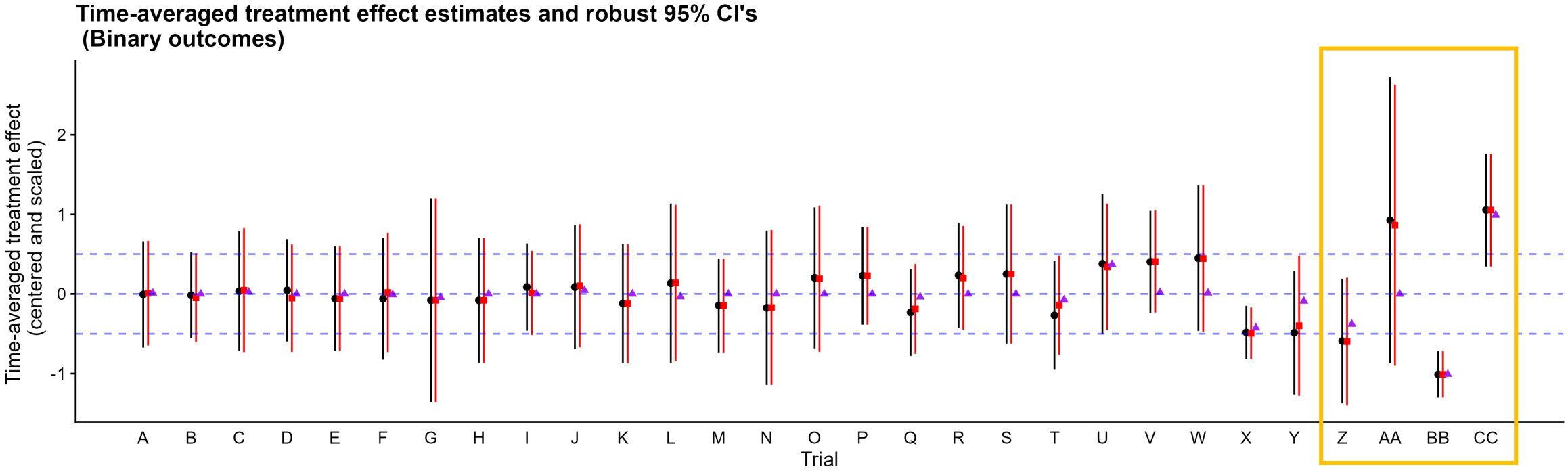
Time-averaged treatment effect estimates and robust 95% CI's (Continuous outcomes)

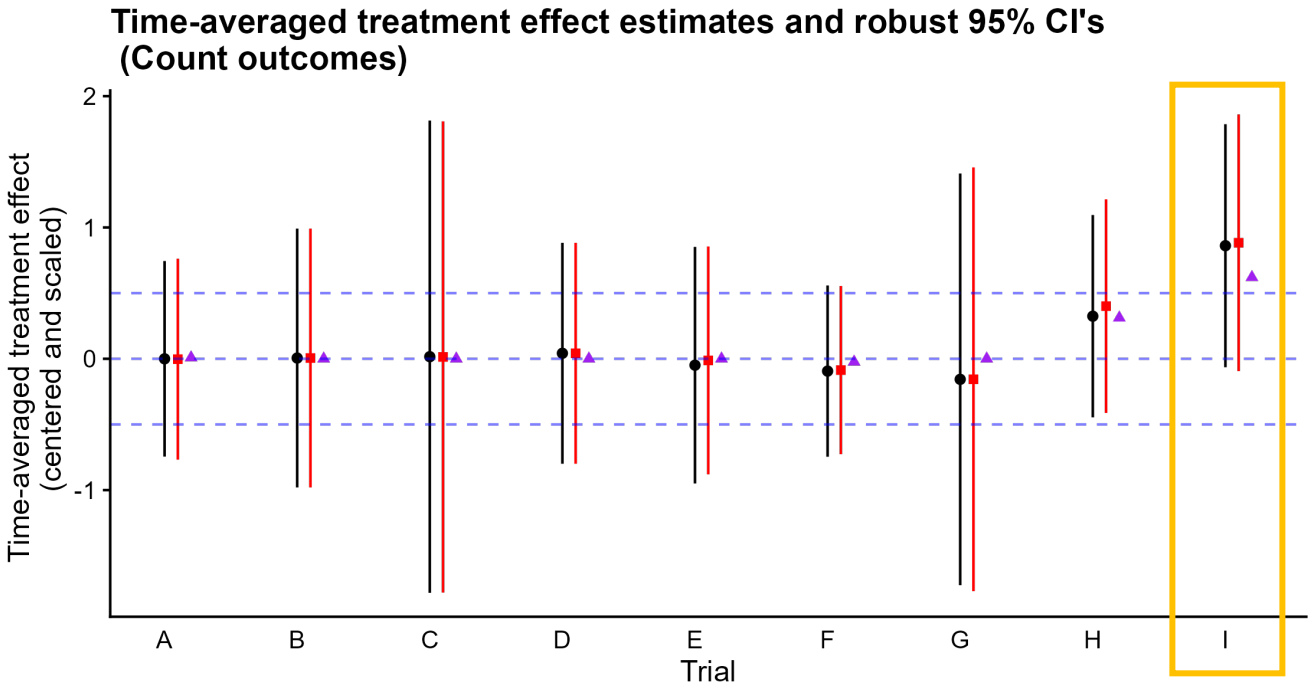


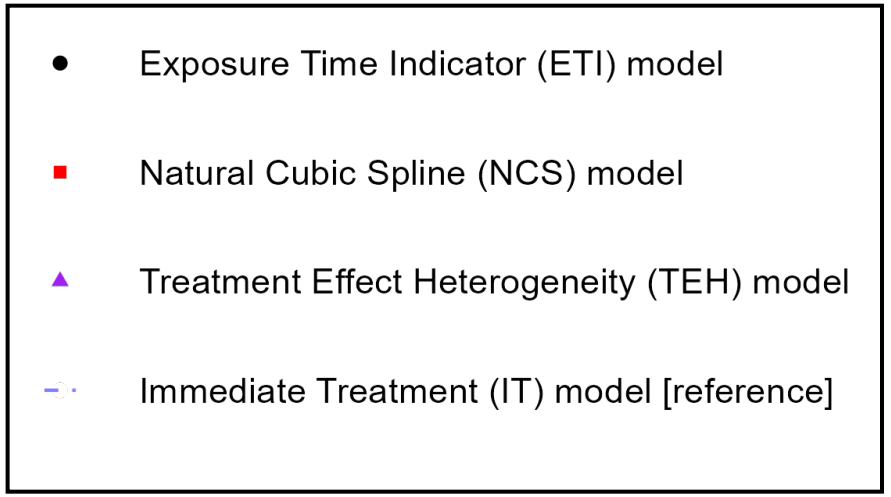

Time-averaged treatment effect estimates and robust 95% CI's
(Binary outcomes)
Time-averaged treatment effect
(centered and scaled)
2
1
0
-1
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA BB CC
Trial


Time-averaged treatment effect estimates and robust 95% CI's
(Count outcomes)
Time-averaged treatment effect
(centered and scaled)
2
1
0
-1
A B C D E F G H I
Trial


Exposure Time Indicator (ETI) model
Natural Cubic Spline (NCS) model
Treatment Effect Heterogeneity (TEH) model
Immediate Treatment (IT) model [reference]


Time-averaged treatment effect estimates and robust 95% CI's
(Binary outcomes)
Time-averaged treatment effect
(centered and scaled)
2
1
0
-1
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA BB CC
Trial


Time-averaged treatment effect estimates and robust 95% CI's
(Count outcomes)
Time-averaged treatment effect
(centered and scaled)
2
1
0
-1
A B C D E F G H I
Trial


Exposure Time Indicator (ETI) model
Natural Cubic Spline (NCS) model
Treatment Effect Heterogeneity (TEH) model
Immediate Treatment (IT) model [reference]

Exposure Time Indicator (ETI) vs. Immediate Treatment (IT)

	Continuous outcomes (n=21)	Binary outcomes (n=29)	Count outcomes (n=9)
ETI estimate falls outside 95% CI for IT, n (%)	1 (5%)	4 (14%)	1 (11%)
Estimates have different signs, n (%)	4 (19%)	5 (17%)	0 (0%)
IT estimate outside ETI curve, n (%)	6 (29%)	13 (45%)	3 (33%)
Percent change of estimate, median (IQR)	79% (42%, 228%)	63% (38%, 147%)	44% (41%, 772%)
Relative robust SE, median (IQR)	1.46 (1.19, 2.01)	1.50 (1.28, 1.75)	1.80 (1.54, 1.97)

- Some indicators of time-varying treatment effects are more common in binary outcomes than continuous outcomes - but sample size is limited
- Many trials have some interesting differences between models, but not broad agreement across all indicators of IT vs. ETI differences

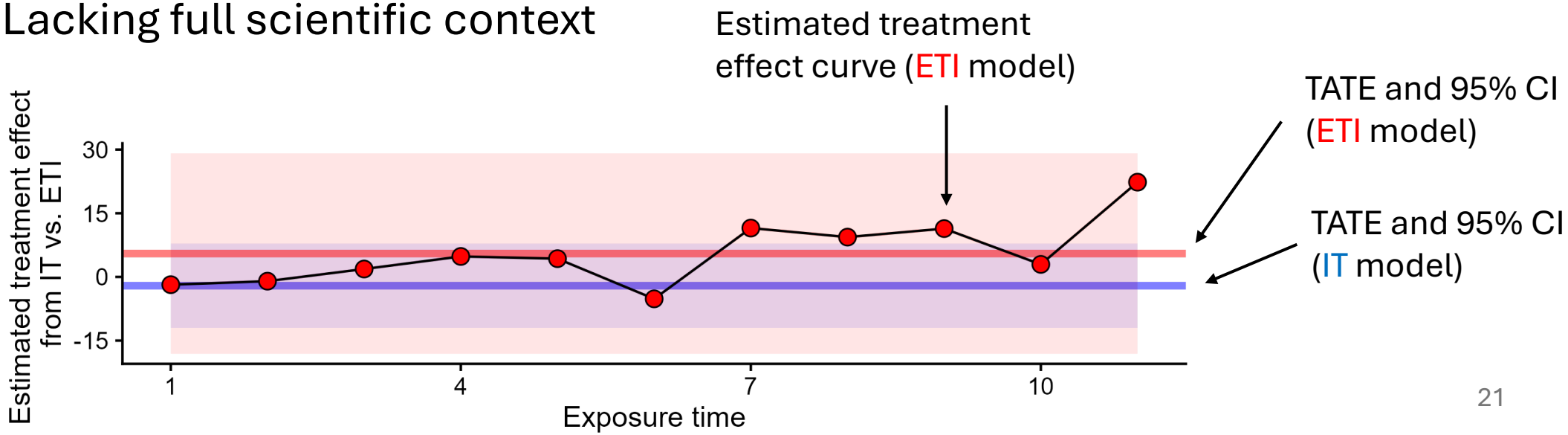
Characteristics of trials with large impact

- Large impact of model choice: ETI estimate of TATE falls outside of the robust 95% CI for the IT estimate
- Overall: 6/59 (10%)

		Overall (n=59)	Large impact (n=6)	Not a large impact (n=53)
Study type, n (%)	Cross-sectional	43 (73%)	2 (33%)	41 (77%)
	Cohort	16 (27%)	4 (67%)	12 (23%)
Number of clusters analyzed, median (IQR)		12 (8.5,24)	**26 (14,61)**	12 (8,22)
Number of sequences analyzed, median (IQR)		5 (3.5,6.5)	4 (3,5.5)	5 (4,7)
Number of time periods analyzed, median (IQR)		8 (5,11)	7.5 (5.5,9.5)	8 (5,11)

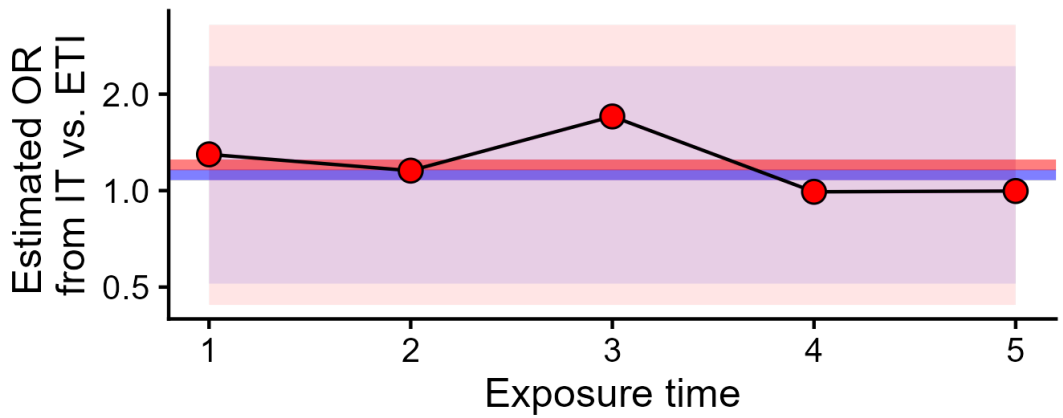
Evidence of trends in treatment effects

- Are differences in TATE caused by a trend in the treatment effect?
  - ETI model may also be misspecified (e.g. excluding other interactions)
  - Bias in IT depends on shape of curve
- Difficult to identify with certainty
  - Possible confounding of exposure time and other factors (e.g. calendar time interaction)
  - Few clusters, poor precision
  - Lacking full scientific context

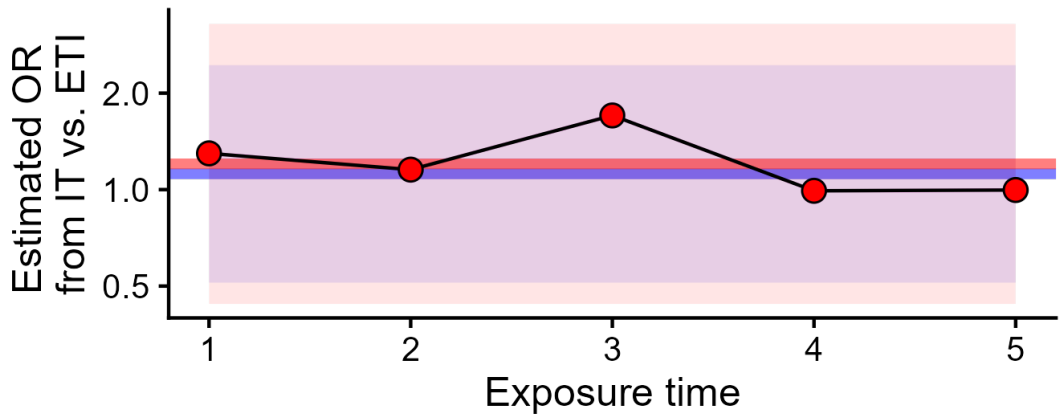
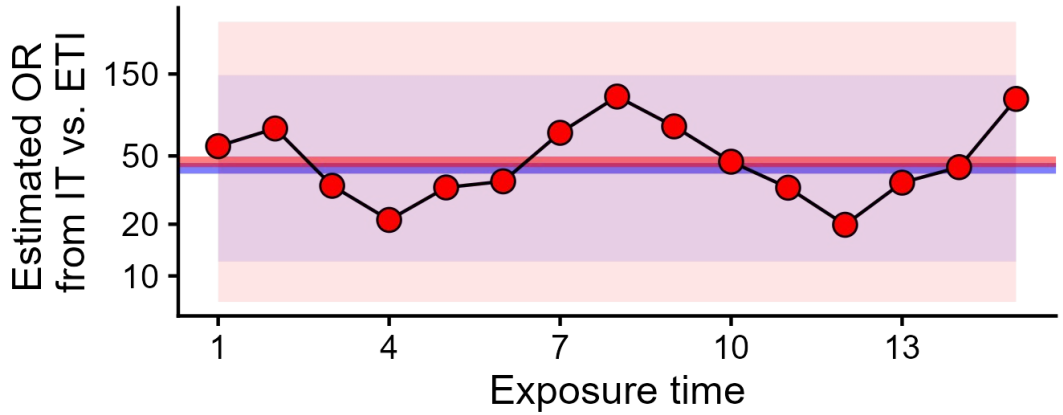


Evidence of systematic trend in the treatment effect (ETI)	Impact of model assumptions (ETI vs IT) on inference about treatment effect	
	Small	Large
Weak		
Strong		

Impact of model assumptions (ETI vs IT) on inference about treatment effect

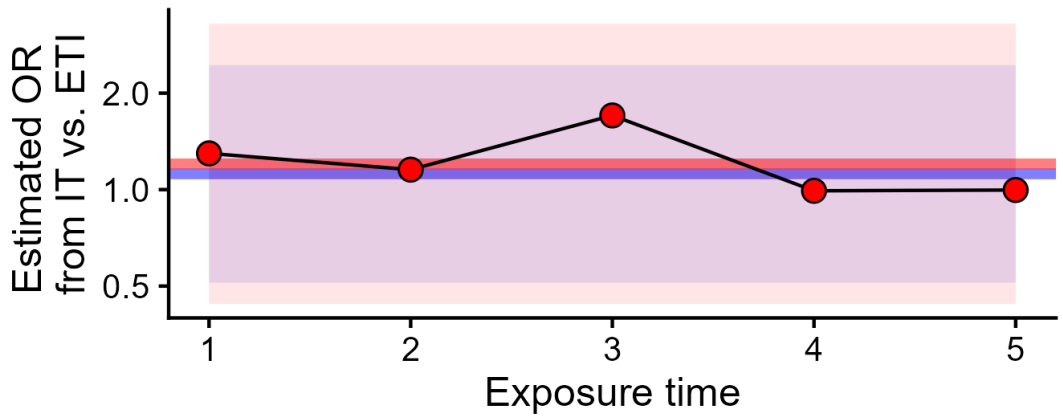
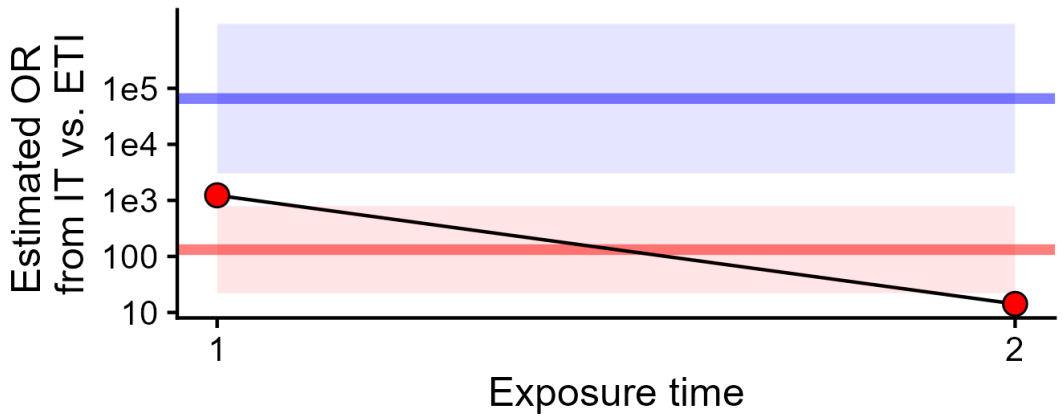
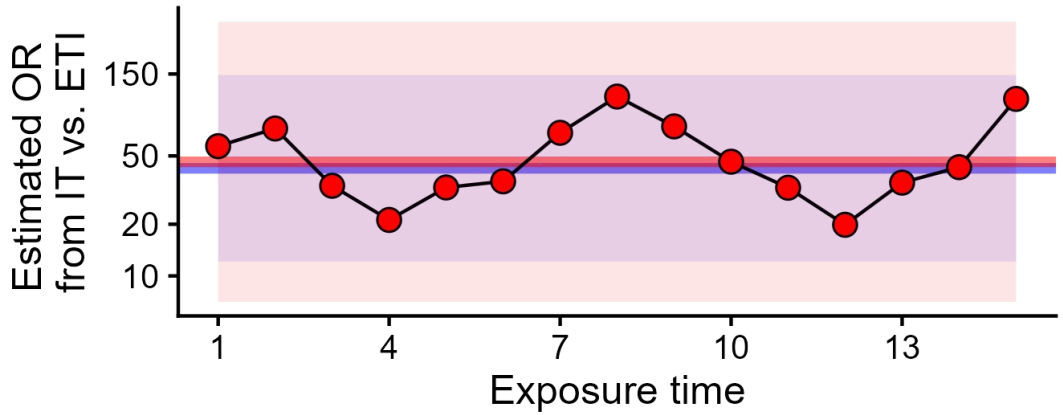
Evidence of systematic trend in the treatment effect (ETI)	Small	Large
Weak	 Short study duration, simple intervention	
Strong		

Impact of model assumptions (ETI vs IT) on inference about treatment effect

Evidence of systematic trend in the treatment effect (ETI)	Small	Large
Weak	 Short study duration, simple intervention	
Strong	 Intervention included ongoing monitoring and support to ensure quality and delivery of intervention	

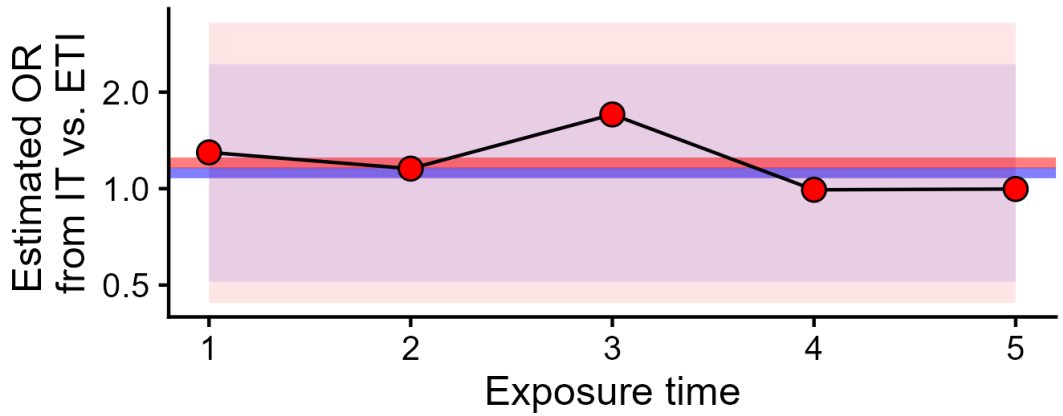
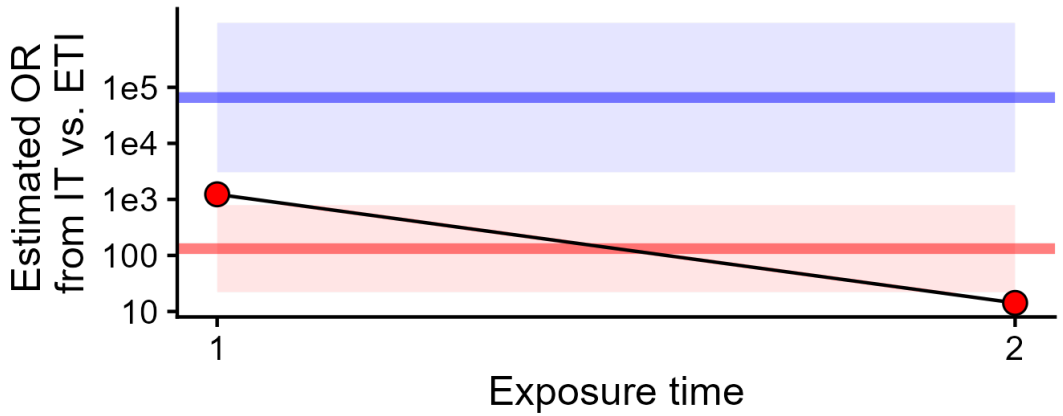
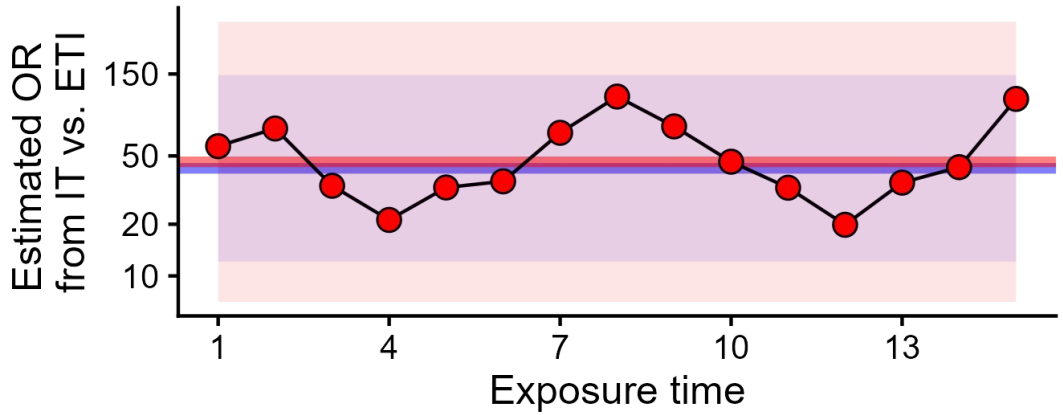
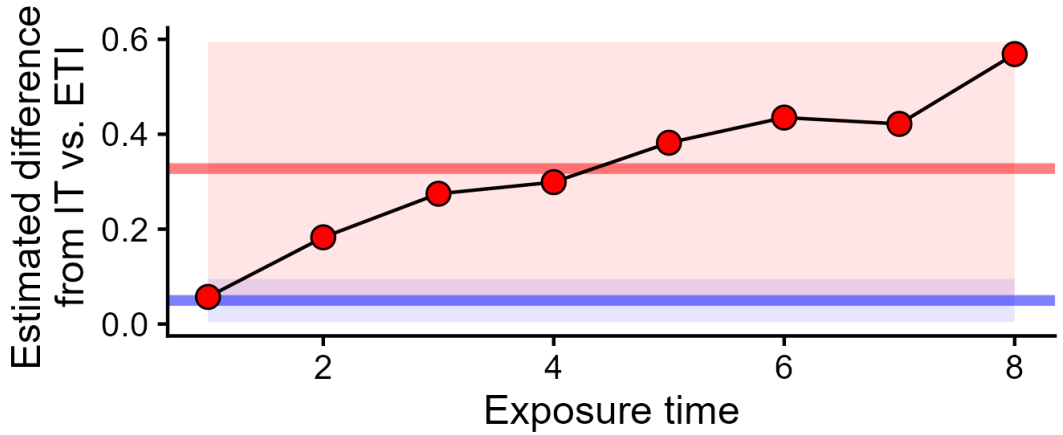
Impact of model assumptions (ETI vs IT) on inference about treatment effect

Evidence of systematic trend in the treatment effect (ETI)

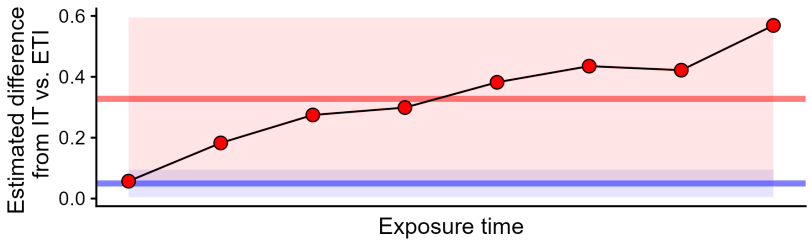
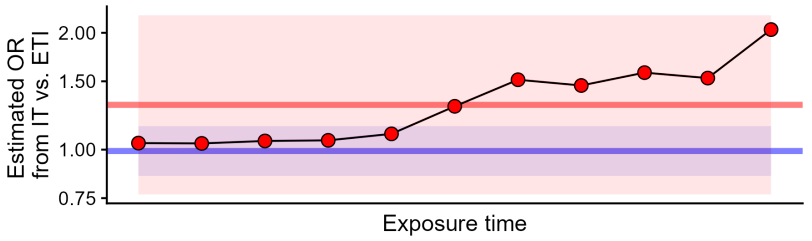
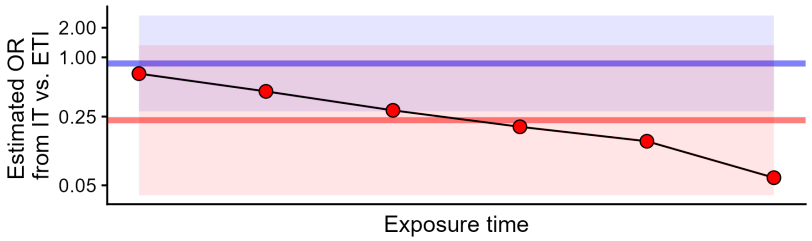
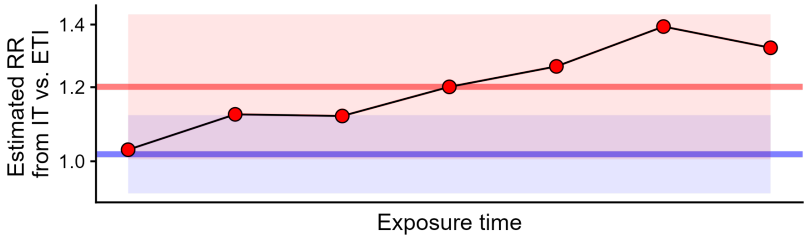
	Small	Large
Weak	 Short study duration, simple intervention	 Intervention is delivered once immediately after crossover; original analysis used flexible model
Strong	 Intervention included ongoing monitoring and support to ensure quality and delivery of intervention	

Impact of model assumptions (ETI vs IT) on inference about treatment effect

Evidence of systematic trend in the treatment effect (ETI)

	Small	Large
Weak	 Short study duration, simple intervention	 Intervention is delivered once immediately after crossover; original analysis used flexible model
Strong	 Intervention included ongoing monitoring and support to ensure quality and delivery of intervention	 Included tailoring of implementation strategies; original analysis assumed full effect after 2 periods

Trends in treatment effects with large impact of model choice

Time-on-treatment curve with ETI and IT TATE	Outcome	ETI vs. IT signs	IT is outside ETI curve	Percent change in TATE
	Continuous	Same sign	Yes	559%
	Binary	Different signs	Yes	-3390%
	Binary	Same sign	Yes	-929%
	Count	Same sign	Yes	959%

Insights on time-on-treatment curve shapes

- Trends causing large bias tend to be smooth and monotonic
  - Some don’t level off (e.g. linear)
  - Impact of treatment often increases over time, with biased immediate treatment estimate indicating little effect
- Cases with variation but no consistent trend (TEH) do occur
  - Often not associated with large bias
- Apparent trends in the treatment effect associated with moderate impact on the overall TATE are common
  - Large impact and compelling trend: 4/59 (7%)
  - Large impact and too short to understand trend: 2/59 (3%)
  - Moderate impact and compelling trend: 10/59 (17%)

Conclusions

- Time-varying treatment effects are common
  - In 10% of trials, treatment model choice had a large impact on results
  - Other differences common: 37% had simple IT TATE outside of curve from saturated ETI
  - Need to consider this possibility in design (including choice of estimands) and analysis (either primary or sensitivity)
- Roughly linear trends are especially concerning
  - Large bias in simple IT model
  - Length of trial and analysis plan need to be aligned with the estimand of interest, which may not be the TATE
- Flexible models for time-varying treatment effects are promising, but have some limitations
  - Require larger sample size[1]
  - Model instability for binary outcomes: saturated ETI models may not be a reliable option for rare outcomes

[1]Kenny et al. 2026.

Fred Hutch Cancer Center

Thank you

Questions:
Emily Voldal
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center
evoldal@fredhutch.org

References

- Kenny A, Voldal EC, Xia F, Heagerty PJ, Hughes JP. Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Statistics in Medicine*. 2022;41(22):4311-4339.
- Nevins P, Ryan M, Davis-Plourde K, et al. Adherence to key recommendations for design and analysis of stepped-wedge cluster randomized trials: A review of trials published 2016–2022. *Clinical Trials*. 2024;21(2):199-210.
- Ouyang Y, Taljaard M, Forbes AB, Li F. Maintaining the validity of inference from linear mixed models in stepped-wedge cluster randomized trials under misspecified random-effects structures. *Statistical Methods in Medical Research*. 2024;33(9):1497-1516.
- Morel JG, Bokossa MC, Neerchal NK. Small Sample Correction for the Variance of GEE Estimators. *Biometrical Journal*. 2003;45:395-409.
- Kenny A, Voldal EC, Xia F, et al. Factors affecting power in stepped wedge trials when the treatment effect varies with time. *Trials* 2026;27:241.