Novel Methods to Address Treatment Effect Heterogeneity in Cluster Randomized Trials

Invited Session 7 | Mon May 20, 2024 | 2:00pm - 3:30pm

2024 BOSTON

SCT | 45TH ANNUAL MEETING

Disclosures

- No relevant disclosures (Dustin Rabideau)

Session Agenda

2:00 - 2:05 (5 min)	Introduction	**Dustin Rabideau**
2:05 - 2:30 (25 min)	Unified methods for designing parallel and longitudinal cluster randomized trials to detect treatment effect heterogeneity	**Fan Li**
2:30 - 2:55 (25 min)	Detecting treatment effect heterogeneity in cluster randomized trials	**Rui Wang**
2:55 - 3:20 (25 min)	Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect	**Avi Kenny**
3:20 - 3:30 (10 min)	Discussion, Q&A	**Dustin Rabideau & Speakers**

Dustin Rabideau, PhD

MASSACHUSETTS GENERAL HOSPITAL

Associate Director, MGH Biostatistics

HARVARD MEDICAL SCHOOL

Assistant Professor of Medicine

- Robust statistical methods for cluster randomized trials

Fan Li, PhD

Session Organizer

YALE SCHOOL OF PUBLIC HEALTH

Assistant Professor of Biostatistics

- Methods for designing & analyzing pragmatic trials
- Causal inference for randomized trials & observational studies

Unified methods for designing parallel and longitudinal cluster randomized trials to detect treatment effect heterogeneity

Fan Li

Department of Biostatistics
Center for Methods in Implementation and Prevention Science (CMIPS)
Yale School of Public Health

https://lifan90.com/

The Society for Clinical Trials (SCT) 45th Annual Meeting
Boston, May 20, 2024

Acknowledgement

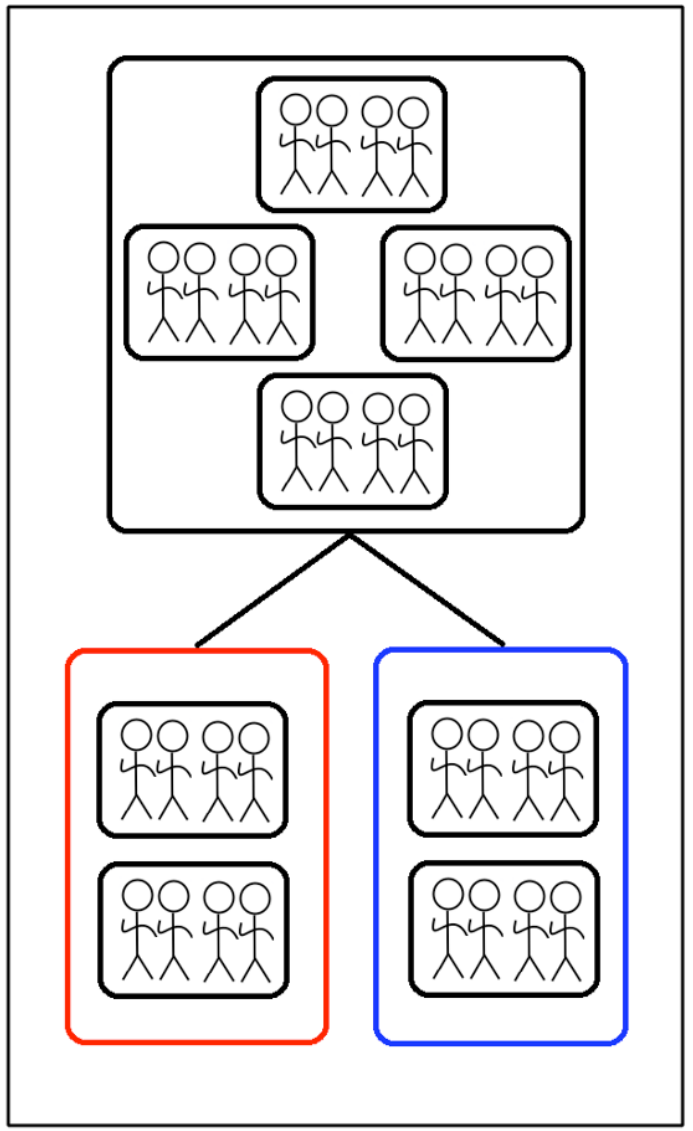
- This work is supported by a Patient-Centered Outcomes Research Institute Award **ME-2020C3-21072**. The statements presented are solely the responsibility of the presenter/authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.
    - Co-Investigators and collaborators: Patrick Heagerty, Rui Wang, Denise Esserman, Mary Ryan, Kendra Plourde, Monica Taljaard, Michael Harhay, Guangyu Tong, Xinyuan Chen, Xueqi Wang, Jiaqi Tong and many other stakeholders
- Support from **NIA IMPACT Collaboratory**
    - feedback from the Design & Statistics Core
- Support from **NIH Pragmatic Clinical Trials Collaboratory**
    - feeback from the Biostatistics & Study Design Core

Cluster randomized trials

- Cluster randomized trials (CRTs) randomize entire clusters/groups of individuals to treatment conditions
  - only feasible scheme
  - administrative and logistical considerations
- Increasingly seen in pragmatic clinical trials
- An essential task in planning studies is to ensure adequate power for detecting a clinically relevant effect size
- The average/overall treatment effect has been the primary pursuit
  - extensive literature on CRT study planning, with a focus on sample size and power calculation

A hypothetical example 1

- Plan for a CRT with 2 arms randomized in a 1 : 1 ratio
- Each nursing home is a cluster, and can include approximately 50 individuals (cluster size, m)
- For a given effect size (e.g., 0.2 standardized by outcome SD), how many nursing homes do we need to ensure 80% statistical power?
- What else goes into the equation?
  - intracluster correlation coefficient (ICC) [for the outcome of interest]



Intracluster correlation coefficient

- ICC often defined as

$$\rho_y = \frac{\text{between-cluster variance}}{\text{total variance}}$$

  - Characterizes the similarity of outcome values for pairs of individuals in the same cluster
  - Typically ranges from $0 \sim 0.2$, but rarely above
  - Plays an important role in determining the sample size for CRTs

$$\text{design effect} = 1 + (m - 1) \times \rho_y$$

- Increasingly more available from published literature, existing database, or pilot data

- R Shiny CRT Calculator[1]

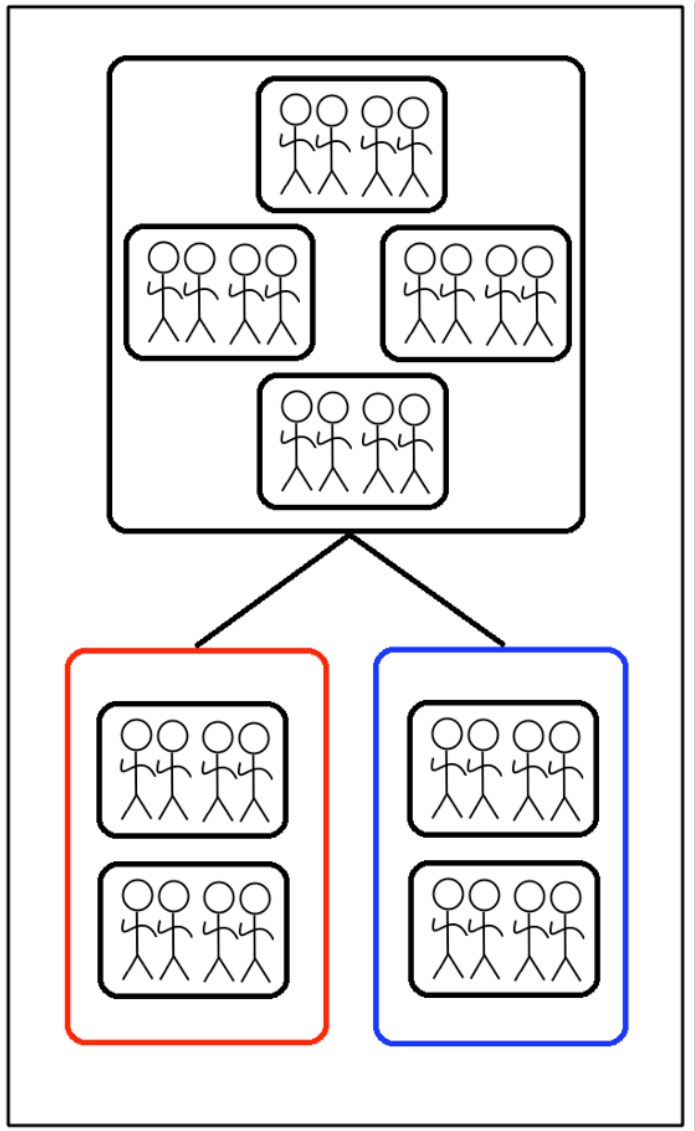
[1] Hemming K et al (2020). A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. *International Journal of Epidemiology*.

Beyond the overall effect

- What if we wish to test the difference in treatment effect between different subgroups in CRTs?
- Interest is growing in understanding whether the treatment effect varies among pre-specified patient subgroups
  - defined by baseline demographics: sex, racial groups etc
  - health-equity variables that are relevant for health disparity studies
- An important form of treatment effect heterogeneity, but how to plan such a CRT?
- What are methods or simple tools like the Shiny CRT that enables convenient sample size & power calculation for such confirmatory heterogeneity of treatment effect (HTE) analysis in a CRT?

A hypothetical example 2

- Plan for a CRT with 2 arms randomized in a 1 : 1 ratio
- Each nursing home is a cluster, and can include approximately 50 individuals (cluster size, m)
- For a given effect size (e.g., treatment effect difference between white and minority), how many nursing homes do we need to ensure 80% statistical power?
- What goes into the equation?
  - ICC of the outcome
  - anything else?



Objectives

- We focus on demystifying sample size requirements for assessing confirmatory treatment effect heterogeneity with measured baseline cluster-level or individual-level covariates
  - what are the key ingredients that drive the statistical power of an interaction test?
  - **example 1**: parallel-arm design
  - **example 2**: stepped-wedge design
  - **other designs**
- Briefly introduce a new free R Shiny app being developed for this objective

Testing an overall effect: Recap

- Consider a **parallel two-arm CRT** with n clusters
- Let Y_{ij} be a continuous outcome for the jth individual ($j = 1, \ldots, m$) in the ith cluster ($i = 1, \ldots, n$)
- Let W_i be the cluster-level treatment indicator (= 1 if treated)
- Unadjusted linear mixed model for average treatment effect is given by

$$Y_{ij} = \alpha_1 + \alpha_2 W_i + \lambda_i + \xi_{ij},$$

  where $\lambda_i \sim \mathcal{N}(0, \sigma^2_\lambda)$ and $\xi_{ij} \sim \mathcal{N}(0, \sigma^2_\xi)$

- Treatment effect quantified by α_2, the classical design effect (DE $= 1 + (m-1)\rho_y$, $\rho_y = \sigma^2_\lambda / (\sigma^2_\lambda + \sigma^2_\xi)$) is derived based on this unadjusted model for study planning

Testing treatment effect difference

- A particular baseline covariate is considered as potential effect modifier of scientific interest
- For testing possible treatment effect heterogeneity with respect to covariate X_{ij} (e.g., age, gender and race), can modify the above model

$$Y_{ij} = \beta_1 + \beta_2 W_i + \beta_3 X_{ij} + \beta_4 X_{ij} W_i + \gamma_i + \epsilon_{ij}$$

  where $\gamma_i \sim \mathcal{N}(0, \sigma^2_\gamma)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\epsilon)$
- For binary X_{ij} (race), β_4 encodes difference in treatment effect among white and non-white subpopulations – HTE parameter ($\mathcal{H}_0 : \beta_4 = 0$) – target hypothesis of an interaction test
- Essentially a linear mixed analysis of covariance (ANCOVA) model

What are the design parameters?

Assume a univariate individual-level effect modifier X_{ij}, recall the ANCOVA model

$$Y_{ij} = \beta_1 + \beta_2 W_i + \beta_3 X_{ij} + \beta_4 X_{ij} W_i + \gamma_i + \epsilon_{ij}$$

- Assume equal cluster size m (can be relaxed[2])
- Assume $1:1$ allocation
- Total outcome variance (adjusted): $\sigma^2_{y|x} = \sigma^2_\gamma + \sigma^2_\epsilon$
- Outcome-ICC (adjusted): $\rho_{y|x} = \sigma^2_\gamma / \sigma^2_{y|x}$
- Covariate-ICC: ρ_x measures the degree of similarity between effect modifiers in the same cluster
  - **example**: if $X_{ij} = \mu_1 + b_i + c_{ij}$, $b_i \sim \mathcal{N}(0, \sigma^2_b)$ and $c_{ij} \sim \mathcal{N}(0, \sigma^2_c)$, then $\rho_x = \sigma^2_b / (\sigma^2_b + \sigma^2_c)$.

[2]Tong G, Esserman D, Li F. Accounting for unequal cluster sizes in designing cluster randomized trials to detect treatment effect heterogeneity. *Stat Med*. 2022 Apr 15;41(8):1376-96.

Covariate ICC

- Empirical evidence of substantial variation in distribution of potential effect modifiers across clusters even in multi-center studies
- As an example, $\rho_x \approx 0.08$ for age and $\rho_x \approx 0.22$ for racial group in a completed multi-center trial
- Concept of covariate ICC dates back to 1997[3]
- Generally unrealistic to assume $\rho_x = 0$ as in individually randomized trials

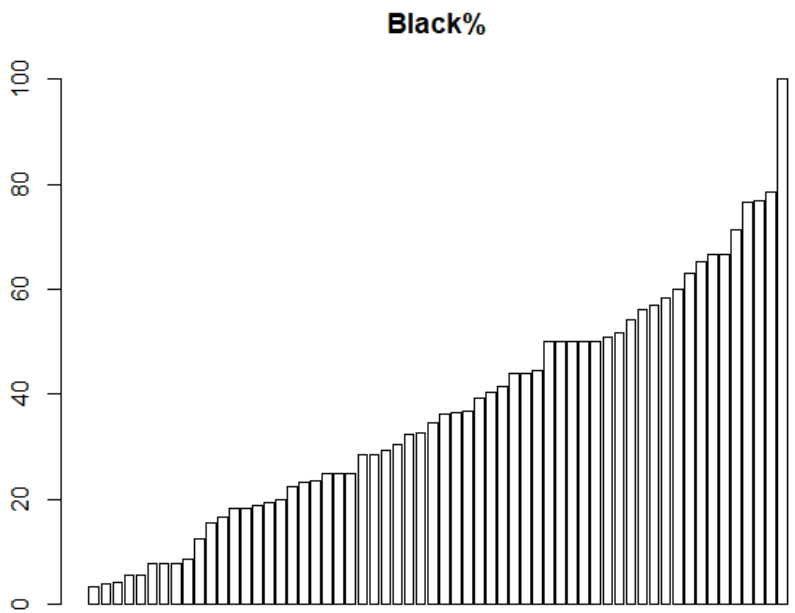


Figure: Variation of % black in the HF-ACTION multi-center trial with 82 sites

[3]Raudenbush SW (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychol. Methods.*

What is the variance for $\hat{\beta}_4$?

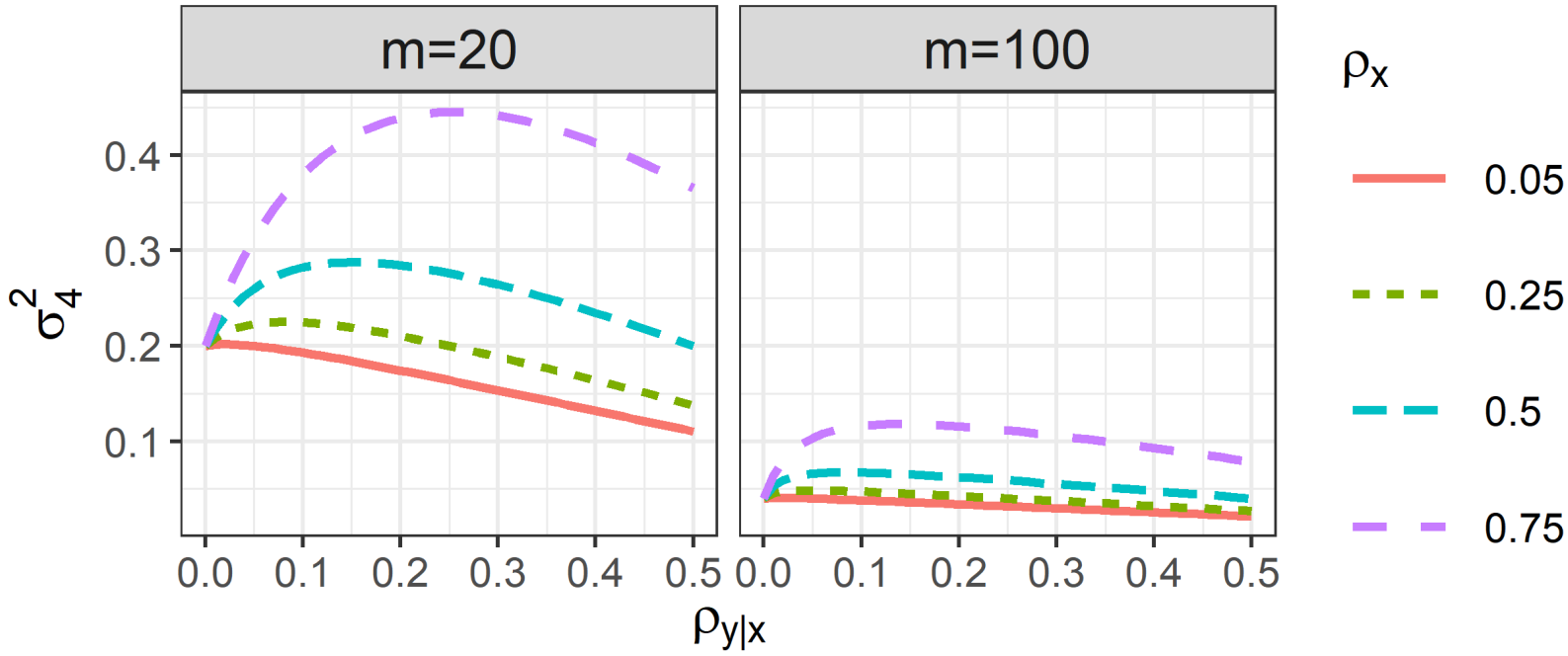
- For design purposes, we derive expression of the HTE estimator, under the linear mixed ANCOVA model[4]

$$var(\hat{\beta}_4) = \frac{4\sigma^2_{y|x}}{nm\sigma^2_x} \times \underbrace{\frac{(1-\rho_{y|x})\{1+(m-1)\rho_{y|x}\}}{1+(m-2)\rho_{y|x}-(m-1)\rho_x\rho_{y|x}}}_{\text{DE}(m)}$$

- **Interpretation**: variance of HTE estimator in individually randomized trial $\times$ design effect, DE(m)
  - DE(m) depends on both outcome-ICC and covariate-ICC
  - larger variance of X_{ij} and smaller covariate-ICC lead to smaller variance (larger power)

[4]Yang S, Li F, Starks MA, Hernandez AF, Mentz RJ, Choudhury KR (2020). Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. *Statistics in Medicine*. 39(28), 4218-4237

Variance as a function of outcome ICC



- Variance can be quadratic in $\rho_{y|x}$, stationary point obtained at

$$\tilde{\rho}_{y|x} = \frac{\sqrt{(1-\rho_x)\{1+(m-1)\rho_x\}}-1}{(1-\rho_x)(m-1)-1} \in [0,1)$$

- As $\rho_x \to 0$ or $m \uparrow$, $\tilde{\rho}_{y|x} \to 0$
- **A Message**: holding other parameters constant, larger $\rho_{y|x}$ may even lead to larger power for testing HTE

Design effect

- The usual design effect in CRTs for studying average treatment effect is unbounded and increases indefinitely with larger m
- $\mathrm{DE}(\infty) = (1 - \rho_{y|x})/(1 - \rho_x)$ is a finite constant
  - depending on the relative magnitude of the two ICCs, the limit of the design effect may be either $\geq$ or $\leq$ than 1
  - the limit of the design effect decreases as $\rho_{y|x} \uparrow$ and $\rho_x \downarrow$
- If $\rho_x = \rho_{y|x}$, there is no effect due to residual clustering in studying HTE, because $\mathrm{DE}(m) = 1$ for any m
- **A message**: CRTs tend to have larger total sample sizes than individually randomized trials, but may also have an increased chance to detect HTE with adequate power
  - the formula provides a tool to formally assess this

Cluster-level effect modifier

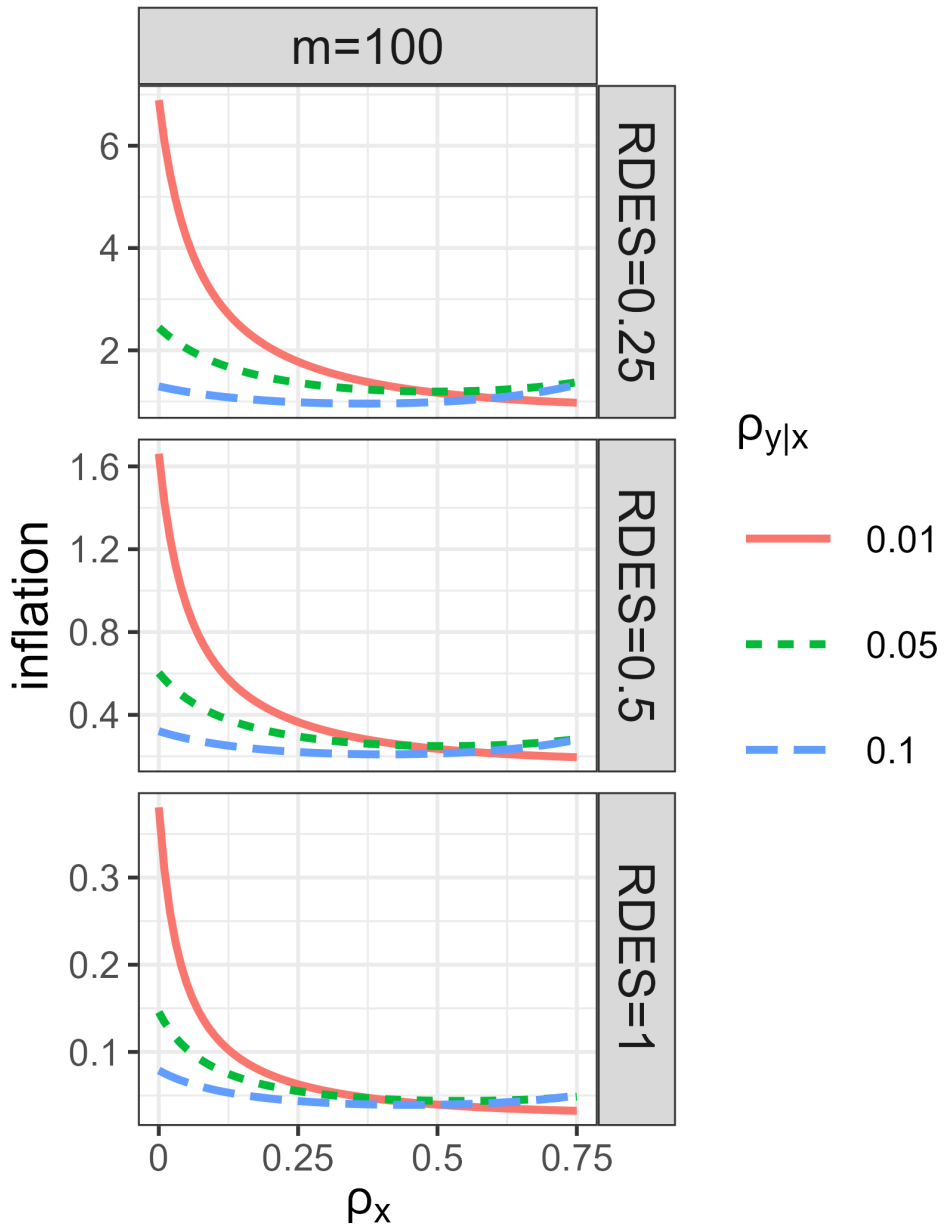
- What if we wish to study effect modification by geographical location or cluster characteristics?
- This is obtained as a special case with $\rho_x = 1$
- Variance of the HTE estimator

$$var(\hat{\beta}_4) = \frac{4\sigma^2_{y|x}}{nm\sigma^2_x} \times \underbrace{\{1 + (m-1)\rho_{y|x}\}}_{\text{DE}(m)}$$

- DE(m) is precisely the classic design effect
- Not surprising because $W_i X_i$ is a cluster-level covariate (within-cluster contrasts no longer contribute to β_4)
- Variance can be used to develop sample size formula
  - Extensive computer simulations done to validate (simple) formulas

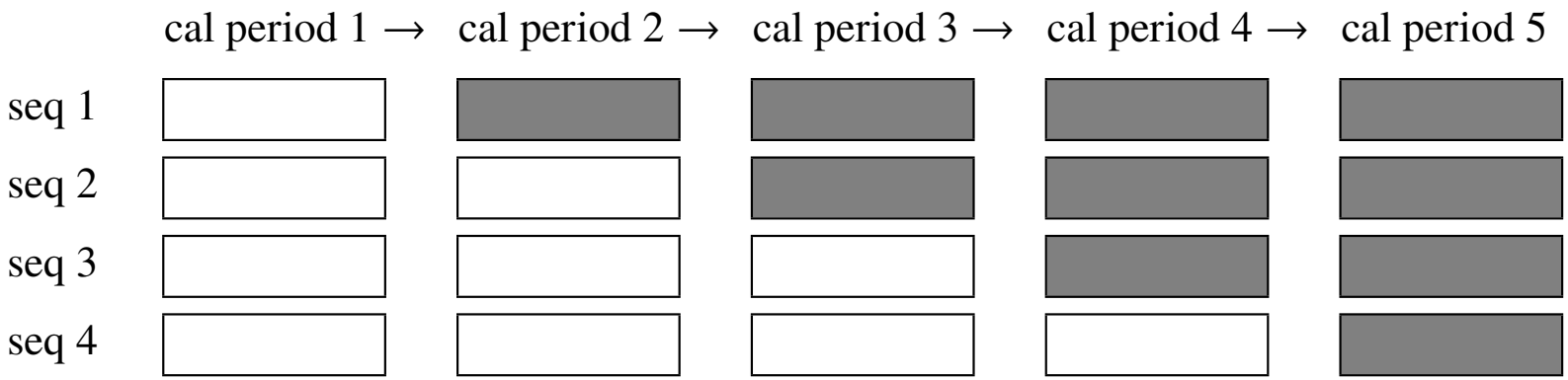
How much more do we need?

- Compare ratio of sample size required for testing HTE versus that for testing an overall effect
- ratio of detectable effect size (RDES)
- Toy example: set variance of covariate and outcome to be 1
  - when the outcome ICC is minimal (close to zero), the inflation factor is larger
  - when the outcome ICC increases, the inflation factor becomes much more "reasonable"
  - "*in CRTs, we are compensating clustering with a larger sample size anyways*"



Stepped wedge cluster randomized trials

- As an example of longitudinal CRT
- Stepped wedge cluster randomized trials (SW-CRTs) sequentially transition experiment units (clusters) from control to the intervention conditions



  - each sequence (seq) can include multiple clusters
  - calendar period (cal period) often defined evenly
- SW-CRTs are increasingly popular in pragmatic trials
  - ensure full roll-out of an intervention during the study period
  - feasible to allocate finite resources at multiple calendar periods

Data structure with effect modifier

(a) *Cross-sectional stepped wedge design*

	period $j=1$	*period* $j=2$	*period* $j=3$	*period* $j=4$
cluster $i=1$	$\{\boldsymbol{X}_{11k}, Y_{11k}; k \in \mathbb{S}_{11}\}$	$\{\boldsymbol{X}_{12k}, Y_{12k}; k \in \mathbb{S}_{12}\}$	$\{\boldsymbol{X}_{13k}, Y_{13k}; k \in \mathbb{S}_{13}\}$	$\{\boldsymbol{X}_{14k}, Y_{14k}; k \in \mathbb{S}_{14}\}$
cluster $i=2$	$\{\boldsymbol{X}_{21k}, Y_{21k}; k \in \mathbb{S}_{21}\}$	$\{\boldsymbol{X}_{22k}, Y_{22k}; k \in \mathbb{S}_{22}\}$	$\{\boldsymbol{X}_{23k}, Y_{23k}; k \in \mathbb{S}_{23}\}$	$\{\boldsymbol{X}_{24k}, Y_{24k}; k \in \mathbb{S}_{24}\}$
cluster $i=3$	$\{\boldsymbol{X}_{31k}, Y_{31k}; k \in \mathbb{S}_{31}\}$	$\{\boldsymbol{X}_{32k}, Y_{32k}; k \in \mathbb{S}_{32}\}$	$\{\boldsymbol{X}_{33k}, Y_{33k}; k \in \mathbb{S}_{33}\}$	$\{\boldsymbol{X}_{34k}, Y_{34k}; k \in \mathbb{S}_{34}\}$
cluster $i=4$	$\{\boldsymbol{X}_{41k}, Y_{41k}; k \in \mathbb{S}_{41}\}$	$\{\boldsymbol{X}_{42k}, Y_{42k}; k \in \mathbb{S}_{42}\}$	$\{\boldsymbol{X}_{43k}, Y_{43k}; k \in \mathbb{S}_{43}\}$	$\{\boldsymbol{X}_{44k}, Y_{44k}; k \in \mathbb{S}_{44}\}$
cluster $i=5$	$\{\boldsymbol{X}_{51k}, Y_{51k}; k \in \mathbb{S}_{51}\}$	$\{\boldsymbol{X}_{52k}, Y_{52k}; k \in \mathbb{S}_{52}\}$	$\{\boldsymbol{X}_{53k}, Y_{53k}; k \in \mathbb{S}_{53}\}$	$\{\boldsymbol{X}_{54k}, Y_{54k}; k \in \mathbb{S}_{54}\}$
cluster $i=6$	$\{\boldsymbol{X}_{61k}, Y_{61k}; k \in \mathbb{S}_{61}\}$	$\{\boldsymbol{X}_{62k}, Y_{62k}; k \in \mathbb{S}_{62}\}$	$\{\boldsymbol{X}_{63k}, Y_{63k}; k \in \mathbb{S}_{63}\}$	$\{\boldsymbol{X}_{64k}, Y_{64k}; k \in \mathbb{S}_{64}\}$

(b) *Closed-cohort stepped wedge design*

	period $j=1$	*period* $j=2$	*period* $j=3$	*period* $j=4$
cluster $i=1$	$\{\boldsymbol{X}_{11k}, Y_{11k}; k \in \mathbb{S}_{1}\}$	$\{Y_{12k}; k \in \mathbb{S}_{1}\}$	$\{Y_{13k}; k \in \mathbb{S}_{1}\}$	$\{Y_{14k}; k \in \mathbb{S}_{1}\}$
cluster $i=2$	$\{\boldsymbol{X}_{21k}, Y_{21k}; k \in \mathbb{S}_{2}\}$	$\{Y_{22k}; k \in \mathbb{S}_{2}\}$	$\{Y_{23k}; k \in \mathbb{S}_{2}\}$	$\{Y_{24k}; k \in \mathbb{S}_{2}\}$
cluster $i=3$	$\{\boldsymbol{X}_{31k}, Y_{31k}; k \in \mathbb{S}_{3}\}$	$\{Y_{32k}; k \in \mathbb{S}_{3}\}$	$\{Y_{33k}; k \in \mathbb{S}_{3}\}$	$\{Y_{34k}; k \in \mathbb{S}_{3}\}$
cluster $i=4$	$\{\boldsymbol{X}_{41k}, Y_{41k}; k \in \mathbb{S}_{4}\}$	$\{Y_{42k}; k \in \mathbb{S}_{4}\}$	$\{Y_{43k}; k \in \mathbb{S}_{4}\}$	$\{Y_{44k}; k \in \mathbb{S}_{4}\}$
cluster $i=5$	$\{\boldsymbol{X}_{51k}, Y_{51k}; k \in \mathbb{S}_{5}\}$	$\{Y_{52k}; k \in \mathbb{S}_{5}\}$	$\{Y_{53k}; k \in \mathbb{S}_{5}\}$	$\{Y_{54k}; k \in \mathbb{S}_{5}\}$
cluster $i=6$	$\{\boldsymbol{X}_{61k}, Y_{61k}; k \in \mathbb{S}_{6}\}$	$\{Y_{62k}; k \in \mathbb{S}_{6}\}$	$\{Y_{63k}; k \in \mathbb{S}_{6}\}$	$\{Y_{64k}; k \in \mathbb{S}_{6}\}$

LM-ANCOVA model

- Extending the LM-ANCOVA model to formulate the HTE test in cross-section designs:

$$Y_{ijk} = \beta_{1j} + \beta_2 W_{ij} + \beta_{3j} X_{ijk} + \beta_4 W_{ij} X_{ijk} + \gamma_i + u_{ij} + \epsilon_{ijk},$$

  - β_{1j} is the secular trend when the effect modifier is $X_{ijk} = 0$
  - β_{3j} is the differential secular trend modified by covariates
  - β_2 is the main effect, $\boldsymbol{\beta_4}$ is the focus on the interaction test
  - $\gamma_i \sim \mathcal{N}(0, \sigma^2_\gamma)$, $u_{ij} \sim \mathcal{N}(0, \sigma^2_u)$, and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_\epsilon)$
  - nested exchangeable correlation structure differentiating within-period and between-period outcome ICCs

- The classic constant treatment effect model, generalized to a constant effect modification model

- Simple treatment effect structure as a practical starting point

  - focus on covariate-based heterogeneity, but not exposure-time heterogeneity

Variance expression

- Under a cross-sectional design, the variance

$$\text{var}^{CS}(\widehat{\beta}_4) = \frac{\sigma^2_{y|x}/\sigma^2_x}{I\text{tr}(\mathbf{\Omega}_W)} \times \frac{J^2}{(J-1)(1-\tau_W)(\zeta_3-\zeta_2)\left(\lambda_2^{-1}-\lambda_3^{-1}\right)+J\theta^{CS}(J,N)},$$

  - λ_2, λ_3, leading eigenvalues of Y-correlation matrix
  - ζ_2, ζ_3, leading eigenvalues of X-correlation matrix
  - $\theta^{CS}(J,N) = J(N-1)\lambda_1^{-1}\zeta_1 + (J-1)\lambda_2^{-1}\zeta_2 + \lambda_3^{-1}\zeta_3$
  - $\mathbf{\Omega}_W$: covariance matrix of the intervention vector under a specific design
  - τ_W: is the generalized ICC of the intervention sequence

- Covariate and outcome ICCs (within-period and between-period ICCs) affect variance through eigenvalue expressions in a nonlinear fashion

- Can extend to closed-cohort designs[5]

[5] Li F, Chen X, Tian Z, Wang R, Heagerty PJ (2024). Planning stepped wedge cluster randomized trials to detect treatment effect heterogeneity. *Statistics in Medicine*. 43(5):890-911.

Other cluster randomized designs and issues?

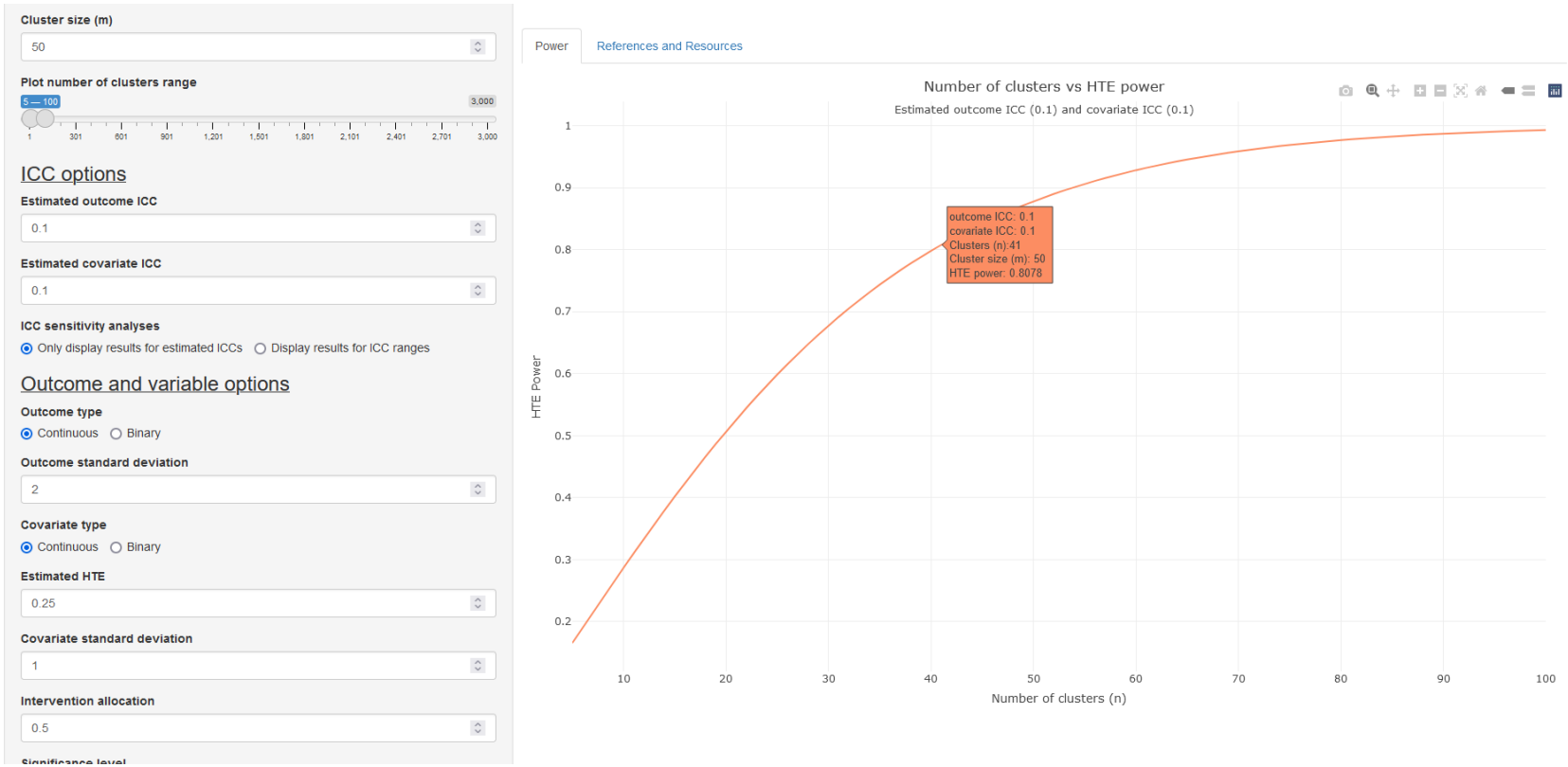
Topic	Note	Reference
Two-level CRT		
Unequal cluster size	*quantify efficiency loss	Tong et al. (2022) *Stat Med*
Binary outcome	*variance depends on mean *ratio effect measures	Maleyeff et al. (2023) *Stat Med*
Optimal maximin design	*unknown ICC	Ryan et al. (2023) *Stat Med*
Missing outcome	*sample size adjusting for missingness	Tong et al. (2023) *BMC Med Res Method*
Subgroup-specific effect	*power for subgroup effect	Wang et al. (2024) *Prevention Science*
Three-level CRT	*within-/between-subcluster ICC *level of randomization	Li et al. (2022) *Biostatistics*
Individually randomized group treatment trial	*between-arm heterogeneity *no covariate ICC	Tong et al. (2023) *Stat Med*
Cluster randomized crossover trial	*improving efficiency over 2-level CRTs *unequal cluster sizes	Wang et al. (2024) *SMMR*

- Reference can be found at our PCORI project webpage
  `https://www.pcori.org/research-results/2021/developing-new-methods-assessing-heterogeneity-treatment-effect-cluster-randomized-trials`

Preliminary version of the software

- To assist in the implementation of these methods in practice, our team (led by Mary Ryan, PhD) is currently developing a free R shiny app that implements study design calculation for **confirmatory HTE** analysis with a **single** effect modifier
  - **Output 1**: Cluster size versus power
  - **Output 2**: Number of clusters versus power
  - **Output 3**: Cluster size versus number of clusters
- Easy to use interface, and URL at `https://cluster-hte.shinyapps.io/shinyapp/`
- In the process of developing a software tutorial to explain some of the common considerations for using the tool

The CRT HTE Calculator[6]



[5]URL: `https://cluster-hte.shinyapps.io/shinyapp/`

Summary - Why heterogeneity?

- Pragmatic trials likely recruit from the "usual" primary care clinics where the study results will be applied and include typical patients seeking health care
  - *The flexible inclusion of a range of clusters and patients to mimic real-world practice necessarily induces more heterogeneity, an aspect that should be reflected at the design stage and which invites studying associated variation in treatment effects*
- The availability of analytical expressions for HTE estimator clarifies key aspects (insights) of data generating process (ρ_x and $\rho_{y|x}$) that drive the study power
  - a simulation-based procedure, however, requires assumptions on non-essential parameters (e.g. main effects parameters)
  - computational concerns
- Can also be used as a tool to provide a context to interpret secondary findings

Summary - Design parameters

- Accurate knowledge of outcome ICC is a common challenge in designing CRTs
  - an increasing number of publications reporting ICCs from existing databases
- Requiring an additional covariate ICC (ρ_x)
  - covariates are available (perhaps more available) in existing data
  - sensitivity analysis on range of ICCs
  - Maximin designs—optimal design that protect from efficiency loss in the worse case scenario[7]
  - URL: `https://mary-ryan.shinyapps.io/HTE-MMD-app/`
- Design & Statistics Core + Technical Data Core (IMPACT Collaboratory) reporting empirical estimates (Ouyang et al. 2024+)

[7] Ryan M, Esserman DA, Li F (2023). Maximin optimal cluster randomized designs to detect treatment effect heterogeneity. *Statistics in Medicine*.

Thank You!

Rui Wang, PhD

HARVARD MEDICAL SCHOOL
Associate Professor of Population Medicine

HARVARD PILGRIM HEALTH CARE INSTITUTE
Director, Division of Biostatistics

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH
Associate Professor in the Department of Biostatistics

- Design, monitoring, and analysis of randomized trials, with a recent emphasis on cluster randomized trials

Detecting Treatment Effect Heterogeneity in Cluster Randomized Trials

Rui Wang

Department of Population Medicine
Harvard Pilgrim Health Care Institute and Harvard Medical School

May 20, 2024
SCT Annual Meeting

Acknowledgements

- Lara Maleyeff , Chia-Rui Chang, Sebastien Haneuse (Harvard)
- Fan Li (Yale)
- Wenbin Lu (NC State)
- This work is in part supported by
  - A Patient-Centered Outcomes Research Institute Award ME-2020C3-21072
  - The National Institute of Allergy and Infectious Diseases T32 AI007358 and R01 AI136947

The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the views of the NIH, PCORI® or its Board of Governors or Methodology Committee.

Treatment effect heterogeneity in CRTs

- Treatment effect heterogeneity: nonrandom, explainable treatment effects that vary across individual or cluster subpopulations
- Can occur both at the cluster and individual level
  - Cluster level: hospitals may vary in terms of their quality of care and physician experiences, leading to differential responses to treatment
  - Individual level: a surgical intervention may lead to greater benefit in younger participants
- Understanding treatment effect heterogeneity is important to guide the implementation of effective interventions and the development of future intervention strategies
- Starks et al. (2019) found that only 4 out of 64 CRTs published between 2010 and 2016 reported subgroup analysis based on demographic information

Test for treatment-by-covariate interactions

- A common practice for exploring treatment effect heterogeneity is to test for treatment-by-covariate interactions
- In CRTs, this can be done by:
  - fitting a linear mixed effects model or a marginal model fitted by generalized estimating equations
  - conducting a statistical test for one or more effect modifiers
- Some limitations have been noted:
  - Rely on parametric assumptions
  - Testing for multiple effect modifiers: inflated type I error due to multiplicity, may have low power
  - Testing for a large number of effect modifiers: model may not be identifiable
- Alternative strategies?

Identify subgroups with enhanced treatment effects

Notation and setting:

- Two-arm parallel CRT of M clusters with continuous outcome Y_{ij}, $i = 1, \ldots, M$, $j = 1, \ldots, n_i$.
- $A_i \in \{0, 1\}$ be the treatment indicator for cluster i; the treatment assignment probability is known: $P(A_i = 1) = p_A$.
- Cluster-level baseline covariates Z_i
- Individual-level baseline covariates X_{ij}

A semi-parametric change-plane model

$$\mathbb{E}[Y_{ij} \mid A_i, Z_i, X_{ij}] = \phi(Z_i, X_{ij}) + \tau A_i I\left(\{1, Z_i^{\mathrm{T}}, X_{ij}^{\mathrm{T}}\}\theta \geq 0\right), \qquad (1)$$

- Extension of Fan et al. (2017) to clustered data
- $\phi(Z_i, X_{ij})$ is an unknown baseline mean function for patients under the control condition ($A_i = 0$)
- Places no assumptions on the baseline mean function $\phi(Z_i, X_{ij})$
- The change-plane $I(\{1, Z_i^T, X_{ij}^T\}\theta \geq 0)$ defines the existence of a patient subgroup with an enhanced treatment effect, τ
- Our goal: Testing if there exists a subgroup with an enhanced treatment effect, $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$.

A doubly robust test

- When θ is known, a doubly robust estimating equation for τ is given by (Tsiatis 2006, Robins and Rotnizky 2001):

$$\sum_{i=1}^{M} \mathbf{I}\left(Z_i, X_i; \theta\right) V_i^{-1} \left\{A_i - \pi(Z_i)\right\} \left\{Y_i - \phi_i(Z_i, X_i) - \tau A_i \mathbf{I}\left(Z_i, X_i; \theta\right)\right\} = 0, \tag{2}$$

- Under H_0: $\tau = 0$, we consider the score test statistic

$$\sum_{i=1}^{M} \psi_1 \left(Y_i, A_i, Z_i, X_i, \hat{\eta}, \hat{\beta}; \theta\right) \tag{3}$$

$$= \sum_{i=1}^{M} \mathbf{I}_i \left(Z_i, X_i; \theta\right) V_i^{-1} \left\{A_i - \pi(Z_i, \hat{\eta})\right\} \left\{Y_i - \phi_i(Z_i, X_i; \hat{\beta})\right\}$$

- The proposed test is valid if either the propensity score model π (always correct in randomized trials) or the mean function ϕ is correctly specified

Unknown θ

- θ is only identifiable under the alternative hypothesis when $\tau \neq 0$
- The standard asymptotic testing framework is not directly applicable (Davies 1987)
- Consider a supremum of squared score test statistics (Fan et al., 2017):

$$T_M = \sup_{\theta \in \Theta} \frac{\left\{ \sum_{i=1}^{M} \psi_1 \left(Y_i, A_i, Z_i, X_i, \hat{\eta}, \hat{\beta}; \theta \right) \right\}^2}{M \hat{V}_S(\theta)}, \tag{4}$$

  where $\hat{V}_S(\theta)$ is a consistent estimator for the asymptotic variance of $M^{-1/2} \sum_{i=1}^{M} \psi_1 \left(Y_i, A_i, Z_i, X_i, \hat{\eta}, \hat{\beta}; \theta \right)$ under the null hypothesis
- Use a numerical method to find the maximum over a unit ball in $\mathbb{R}^{p+1}$

Testing and subgroup detection

- We use a resampling method to compute the critical values of the limiting null distribution
- When the null hypothesis is rejected, the change-plane parameter θ can be estimated by

$$\hat{\theta} = \arg\sup_{\theta \in \Theta} \frac{\left\{ \sum_{i=1}^{M} \psi_1 \left(Y_i, A_i, Z_i, X_i, \hat{\eta}, \hat{\beta}; \theta \right) \right\}^2}{M \hat{V}_S(\theta)} \tag{5}$$

- The identified subgroup with an enhanced treatment effect is $\left\{ i, j : \left\{ 1, Z_i^{\mathrm{T}}, X_{ij}^{\mathrm{T}} \right\} \hat{\theta} \geq 0 \right\}$

Simulation studies

Evaluate type I error and power for a range of settings:

- The baseline mean model is linear/non-linear
- Varying the enhanced treatment effect parameter τ
- Varying the magnitude of intraclass correlation and correlation structure in the data generation process (exchangeable vs. non-exchangeable)
- Choice of the working correlation matrix (independence vs. exchangeable)
- Impact of variable cluster sizes
- In comparison to a marginal mean model with linear treatment x covariate interaction terms

Doubly-robust property

Type I error is controlled even when the baseline mean model ϕ is misspecified

Table. Empirical type I error based on 3,000 simulated data

		B-model I (linear)				B-model II (non-linear)			
		size 0.05		size 0.1		size 0.05		size 0.1	
M/n	Corr.	LM	GAM	LM	GAM	LM	GAM	LM	GAM
100/10	Ind.	0.041	0.039	0.090	0.091	0.042	0.043	0.094	0.086
	Exc.	0.045	0.038	0.095	0.083	0.049	0.043	0.103	0.102
200/10	Ind.	0.052	0.044	0.103	0.096	0.043	0.044	0.089	0.097
	Exc.	0.044	0.043	0.096	0.090	0.047	0.044	0.096	0.095
500/10	Ind.	0.049	0.056	0.104	0.106	0.047	0.042	0.092	0.099
	Exc.	0.044	0.043	0.096	0.090	0.047	0.044	0.096	0.095

Power is affected by the mean model specification

Table. Empirical Power

			B-model I				B-model II			
			size 0.05		size 0.1		size 0.05		size 0.1	
τ	M/n	Corr.	LM	GAM	LM	GAM	LM	GAM	LM	GAM
0.5	100/10	Ind.	32.9	34.6	47.9	47.3	27.2	35.5	39.5	49.8
		Exc.	36.1	33.0	48.1	48.6	25.3	36.7	37.7	51.3
	200/10	Ind.	69.1	66.3	79.3	75.1	53.8	68.7	66.3	78.5
		Exc.	71.1	69.2	81.7	79.8	51.6	68.7	64.0	80.2
	500/10	Ind.	98.7	99.1	99.5	99.7	93.1	97.7	96.5	98.8
		Exc.	99.2	98.6	99.4	99.4	93.4	98.6	96.8	99.4

Choice of working correlation matrix

In the presence of variable cluster sizes, using an exchangeable (compared to independence) working correlation matrix leads to a higher power

Table. Empirical Power

			B-model I				B-model II			
			size 0.05		size 0.1		size 0.05		size 0.1	
τ	M	Corr.	LM	GAM	LM	GAM	LM	GAM	LM	GAM
0.5	100	Ind.	17.5	19.9	29.6	30.9	13.6	18.0	25.8	29.5
		Exc.	21.9	27.3	35.2	40.5	16.0	25.2	25.0	38.1
	200	Ind.	37.3	39.1	49.9	51.8	30.3	39.2	41.9	53.7
		Exc.	56.6	54.6	68.5	67.8	35.4	54.2	48.7	67.1
	500	Ind.	80.8	83.6	89.1	89.7	74.4	83.0	84.0	90.2
		Exc.	94.6	93.4	97.0	96.7	83.5	96.0	90.1	97.5

Exploratory HTE analysis

- In the change-plane analysis, the variables included in the change-plane are pre-specified
- Including more variables in the change-plan specification can be computationally challenging due to the need to search the superemum of squared score-type statistics over the possible space
- Next: a permutation test for detecting individual-level treatment effect heterogeneity

Individual-level treatment effect heterogeneity in CRTs

- Let $Y_{ij}(a)$ be the potential outcome for individual j in cluster i had cluster i been assigned to treatment $a \in \{0, 1\}$
- With the Stable Unit Treatment Value Assumption, the observed outcome

$$Y_{ij} = A_i Y_{ij}(1) + (1 - A_i) Y_{ij}(0),$$

- Let $\Delta_{ij} = Y_{ij}(1) - Y_{ij}(0)$ be the individual-specific treatment effect
- Consider the null hypothesis: $H_0 : Y_{ij}(1) = Y_{ij}(0) + \Delta_1^*$ for all i, j, where Δ_1^* is a constant

Permutation test

- Permutation tests generally rely on the exchangeability of observations under a specific null hypothesis
- Permutation tests for testing the null hypothesis of no treatment effect in CRTs have been well-studies (Braun and Feng 2001, Wang and De Gruttola 2017, Li et al., 2016, Rabideau and Wang 2021)
- Testing the null hypothesis of no treatment effect heterogeneity is more challenging due to the potential presence of the main effect
- Permutation methods for individually-randomized trials have been proposed (Wang et al., 2015, Foster et al, 2016, Ding et al., 2016)

Permutation test: Sharp null

For the idealized scenario when $\Delta_1^* = \Delta$ is known:

1. Calculate the test statistic for the observed data $t = t(\mathbf{A}, \mathbf{Y})$
2. For $b = 1, \ldots, B$, permute treatment indicators on the cluster level in accordance with the randomization scheme
3. Let $\mathbf{A}^b$ be a possible treatment assignment. Then compute:
   1. The transformed outcomes $\widetilde{\mathbf{Y}}^b = \mathbf{Y} - \Delta\mathbf{A} + \Delta\mathbf{A}^b$, and
   2. the test statistic $t^b = t(\mathbf{A}^b, \widetilde{\mathbf{Y}}^b)$.
4. Compare the test statistic t with its null distribution

$$p(\Delta) = p(t^b \geq t). \tag{6}$$

$\mathbf{1}_n$ a vector of n 1s, $\mathbf{A} = (A_1\mathbf{1}_{n_1}^\top, \ldots, A_I\mathbf{1}_{n_I}^\top)^\top$,$\mathbf{Y} = (Y_{11}, \ldots, Y_{In_I})^\top$

Permutation test: Unknown Δ

- In practice, however, Δ is usually unknown and needs to be estimated
- Natural approach: use any consistent estimator of Δ and simply **plug in (PI)** this value (Wang et al., 2015, Ding et al., 2016)
- **Idea**: Maximize over a $(1-\gamma)$-level confidence interval (CI) for Δ (Berger et al., 1994; Ding et al., 2016):

$$p_{sup} = \sup_{\Delta' \in CI_\gamma} p(\Delta') + \gamma.$$

  - Guarantees test validity as long as CI is valid; could be conservative

Choice of test statistic

We consider two test statistics:

- **Shifted Komologorov-Smirnov (K-S)**: compares marginal CDFs, shifted by Δ
- **Generalized additive mixed model (GAMM)-adjusted K-S**: comparing marginal CDFs while adjusting for baseline covariates
  - Fit a GAMM relating outcome to baseline covariates in the control group only; use the residuals as input in the K-S statistic
  - If covariates are predictive of outcome, expect to improve power

Simulation studies

1. Do these tests maintain the correct size under a variety of settings?
2. How powerful are the proposed permutation tests compared with a standard model-based LR test? How does the choice of test statistic (adjusted vs. not) impact power?
3. We consider:
   - **permutation tests** with:
     - the true Δ (RT-T),
     - the plug-in (RT-PI) estimator of Δ,
     - marginalized over a 99.999% CI for Δ (RT-CI) using the shifted K-S test statistic
   - the **model-based** LR test of interaction term(s) using LMM (LMM-LRT) with interaction between and main effects of A_i and $\boldsymbol{X}_{ij}$ and a cluster-level intercept

Performance of testing procedures: Type I error

Do the tests maintain the correct size in finite samples?

Main eff. of X_{ki}	Dist. of errors	I	Empirical type I error RT-T	RT-PI	RT-CI	LMM-LRT
Linear	Normal	20	0.046	0.048	0.029	0.047
Linear	Normal	100	0.061	0.061	0.042	0.060
Linear	Log normal	20	0.053	0.052	0.029	0.063
Linear	Log normal	100	0.051	0.048	0.044	0.056
Nonlinear	Normal	20	0.048	0.054	0.022	0.093
Nonlinear	Normal	100	0.049	0.049	0.036	0.084
Nonlinear	Log normal	20	0.051	0.053	0.028	0.080
Nonlinear	Log normal	100	0.045	0.049	0.035	0.100

- Plug-in performs similarly to true Δ, both maintain correct size
- CI can be conservative (less so as I increases)
- LMM-LRT is robust to distributional misspecification, not robust to mean model misspecification

[1]

[1] RT-T: randomization test w/ true Δ (unavailable in practice); PI: plug-in; CI: confidence interval marginalization; LMM-LRT: likelihood ratio test from a linear mixed model w/ interaction term

Power to detect heterogeneity

Dim. of X_{ij}	Form of Interaction	Empirical Power: Shifted K-S RT-PI	Shifted K-S RT-CI	GAMM-adjusted K-S RT-PI	GAMM-adjusted K-S RT-CI	LMM-LRT
1	Linear	0.663	0.615	0.889	0.858	1.000
1	Oscillating	0.982	0.975	0.996	0.995	0.117
1	Parabolic	0.947	0.939	0.996	0.992	0.234
2	Linear	0.411	0.383	0.890	0.862	0.698
2	Oscillating	0.193	0.174	0.973	0.971	0.033
2	Parabolic	0.352	0.329	0.902	0.886	0.041

- LMM-LRT has high power when assumptions are met and dimension is 1
- LMM-LRT has low power when mean model assumptions are not met and when dimension > 1
- Using GAMM-adjusted K-S yields higher power than using Shifted K-S (as expected) for both PI and CI

[2] K-S: Kolomogorov-Smirnov test statistic (compares marginal CDFs), RT-PI:

Pain program for active coping and training (PPACT) study

- Longitudinal CRT to evaluate cognitive behavioral therapy (CBT) vs. usual care for treating chronic pain (DeBar et al., 2022)
- Primary outcome was self-reported pain impact on PEGS scale (pain intensity and interference with enjoyment of life, general activity, and sleep)
- 816 patients in 106 clusters (median cluster size: 9, range: [3, 13]) at 12 month visit
- Reduced PEGS score in CBT arm (5.52 vs. 6.15), indicating a modest effect
- Consider 25 potential effect modifiers:
  - Age, sex, disability benefits, smoking status, BMI, alcohol misuse, drug misuse, four chronic co-morbid conditions, twelve nonmalignant chronic pain types, morphine dose per day, and benzodiazepine use

Detecting HTE in PPACT

Test	Specified effect modifiers?	Controls for covariates?	p-value
RT-PI (Shifted K-S)	No	No	0.01
RT-CI (Shifted K-S)	No	No	0.07
RT-PI (GAMM-Adj. K-S)	No	Yes	0.001
RT-CI (GAMM-Adj. K-S)	No	Yes	0.003
LMM-LRT	Yes	Yes	0.55

- PI yields smaller p-values than CI
- Adjusted yields smaller p-values than unadjusted
- Randomization test is able to detect heterogeneity where LMM-LRT does not

[3] K-S: Kolomogorov-Smirnov test statistic (compares marginal CDFs), RT-PI: randomization test w/ plug-in estimate of Δ; CI: confidence interval marginalization; LMM-LRT: likelihood ratio test from a linear mixed model w/ interaction terms

An illustration of subgroup detection

- Separate linear mixed models to examine the interaction effect of gender and pain counts

$$Y_{ij} = \beta_0 + \beta_A A_i + \beta_{AX} A_i X_{ij} + \gamma_i + \varepsilon_{ij}, \tag{7}$$

  p-value = 0.09 (Gender) and 0.37 (Pain counts)

- The change-plane analysis:

$$\mathbb{E}[Y_{ij} \mid A_i, Z_i, X_{ij}] = \phi(Z_i, X_{ij}) + \tau A_i I\left(\{1, Z_i^{\mathrm{T}}, X_{ij}^{\mathrm{T}}\}\theta \geq 0\right), \tag{8}$$

  Testing $H_0 : \tau = 0$, p-value = 0.005

- $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_{\text{gender}}, \hat{\theta}_{\text{pain counts}}) = (-0.61, -0.75, 0.27)$.
- The identified subgroup (36.7% of participants) included male patients with paint counts greater than 2 and female patients with pain counts greater than 5
- Marginal treatment effect: -0.64; estimated treatment effect in the enhanced treatment effect subgroup: -1.03

Summary

- Exploratory HTE analysis: a permutation test for detecting treatment effect heterogeneity in CRTs
  - Approaches for dealing with unknown Δ (main treatment effect) and baseline mean adjustment increases power
  - Can be adapted to multi-arm and stepped wedge CRTs
  - Rejecting the null hypothesis provides evidence about treatment effect heterogeneity, but it does not identify effect modifiers
- A test for semi-parametric subgroup identification
  - Doubly-robust: its validity requires the correct specification of either the propensity score model or the baseline mean model
  - Targets a change-plane alternative
  - Can be computationally intensive as the number of candidate effect modifiers increases
  - Would be useful to build in a variable selection procedure

References

- Berger, R. L., Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. Journal of the American Statistical Association, 89(427), 1012-1016.
- Braun, T. M., Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. Journal of the American Statistical Association, 96(456), 1424-1432.
- Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 74(1), 33-43.
- DeBar, L., Mayhew, M., Benes, L., Bonifay, A., Deyo, R. A., Elder, C. R., ... Vollmer, W. M. (2022). A primary careâbased cognitive behavioral therapy intervention for long-term opioid users with chronic pain: a randomized pragmatic trial. Annals of Internal Medicine, 175(1), 46-55.
- Ding, P., Feller, A., Miratrix, L. (2016). Randomization inference for treatment effect variation. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(3), 655-671.
- Fan, A., Song, R., & Lu, W. (2017). Change-plane analysis for subgroup detection and sample size calculation. Journal of the American Statistical Association, 112(518), 769-778.
- Foster, J. C., Nan, B., Shen, L., Kaciroti, N., Taylor, J. M. (2016). Permutation testing for treatmentâcovariate interactions and subgroup identification. Statistics in biosciences, 8, 77-98.

References

- Li, F., Lokhnygina, Y., Murray, D. M., Heagerty, P. J., DeLong, E. R. (2016). An evaluation of constrained randomization for the design and analysis of groupârandomized trials. Statistics in medicine, 35(10), 1565-1579.
- Rabideau, D. J., Wang, R. (2021). Randomization-based confidence intervals for cluster randomized trials. Biostatistics, 22(4), 913-927.
- Robins, J. M., Rotnitzky, A. (2001). Comment on the Bickel and Kwon article,âInference for semiparametric models: Some questions and an answerâ. Statistica Sinica, 11(4), 920-936.
- Starks, M. A., Sanders, G. D., Coeytaux, R. R., Riley, I. L., Jackson, L. R., Brooks, A. M., ... Hernandez, A. F. (2019). Assessing heterogeneity of treatment effect analyses in health-related cluster randomized trials: a systematic review. PLoS One, 14(8), e0219894.
- Tsiatis, A. A. (2006). Semiparametric theory and missing data (Vol. 4). New York: Springer.
- Wang, R., De Gruttola, V. (2017). The use of permutation tests for the analysis of parallel and steppedâwedge clusterârandomized trials. Statistics in medicine, 36(18), 2831-2843.

Avi Kenny, PhD

DUKE UNIVERSITY

Assistant Professor of Biostatistics & Bioinformatics

DUKE GLOBAL HEALTH INSTITUTE

Assistant Research Professor of Global Health

- Stepped wedge cluster randomized trials, survival analysis using machine learning tools, vaccine clinical trials, and global health program evaluation

Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect

Society for Clinical Trials Conference
45th Annual Meeting, 2024

Avi Kenny

May 20th, 2024

Agenda

1. Background on stepped wedge design and analysis
2. The problem of a time-varying treatment effect
3. Analysis of stepped wedge data when the treatment effect varies over time
4. Data analysis examples

(No relevant disclosures)

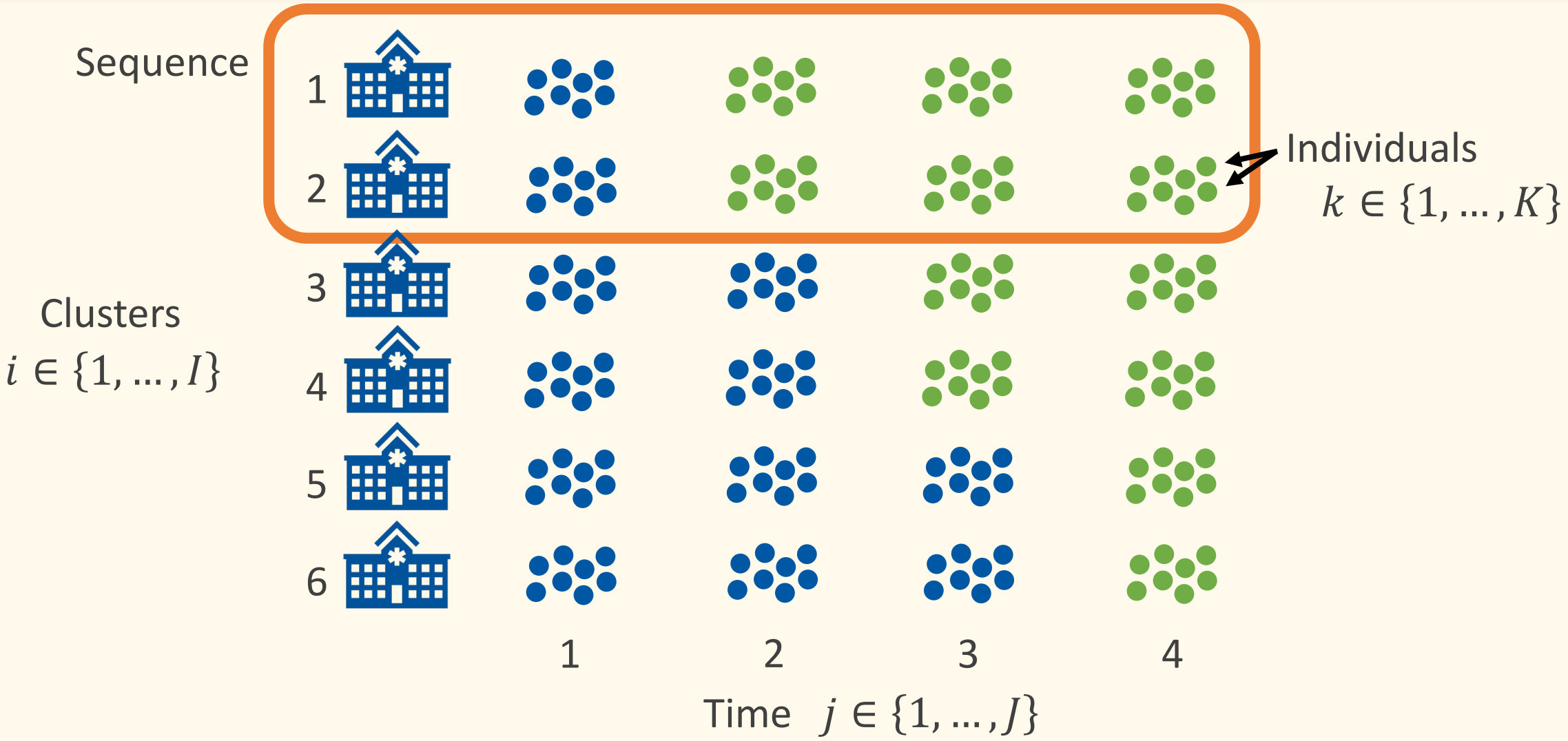
Background on stepped wedge design and analysis

What is a stepped wedge trial?

Hypothetical research question: what is the effect of implementing a pre-surgical checklist on surgery outcomes (e.g., infection)?

Month: 0

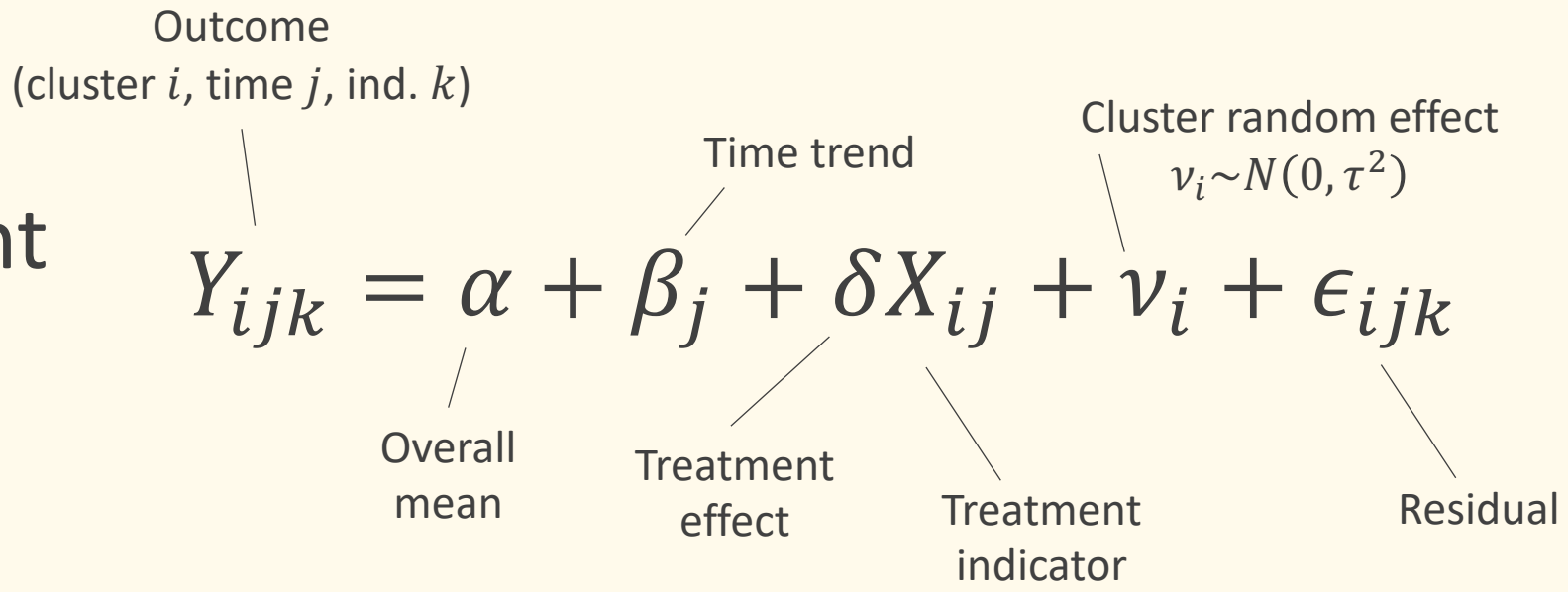
Background on stepped wedge design and analysis



Background on stepped wedge design and analysis

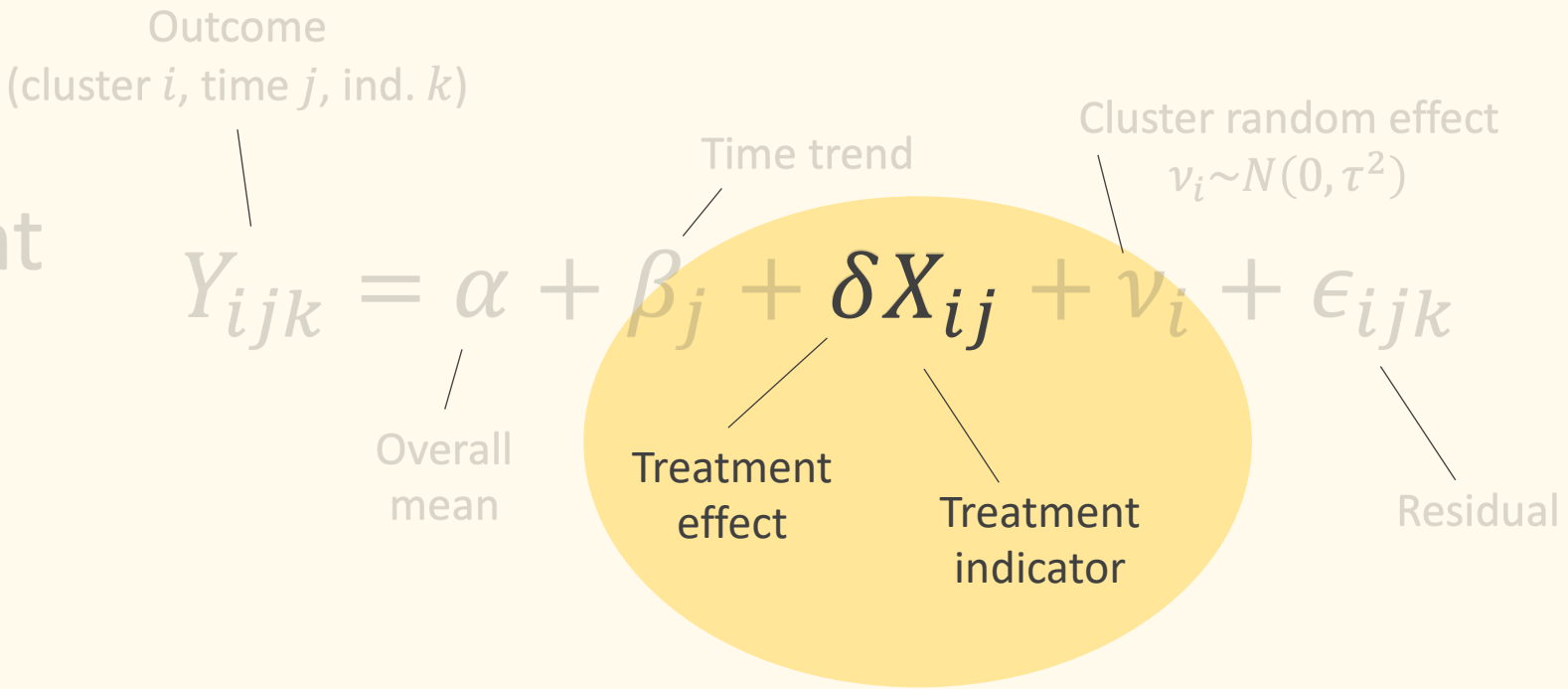
Immediate treatment (IT) model:

(Hussey & Hughes 2007)



Background on stepped wedge design and analysis

Immediate treatment (IT) model:
(Hussey & Hughes 2007)



The problem of a time-varying treatment effect

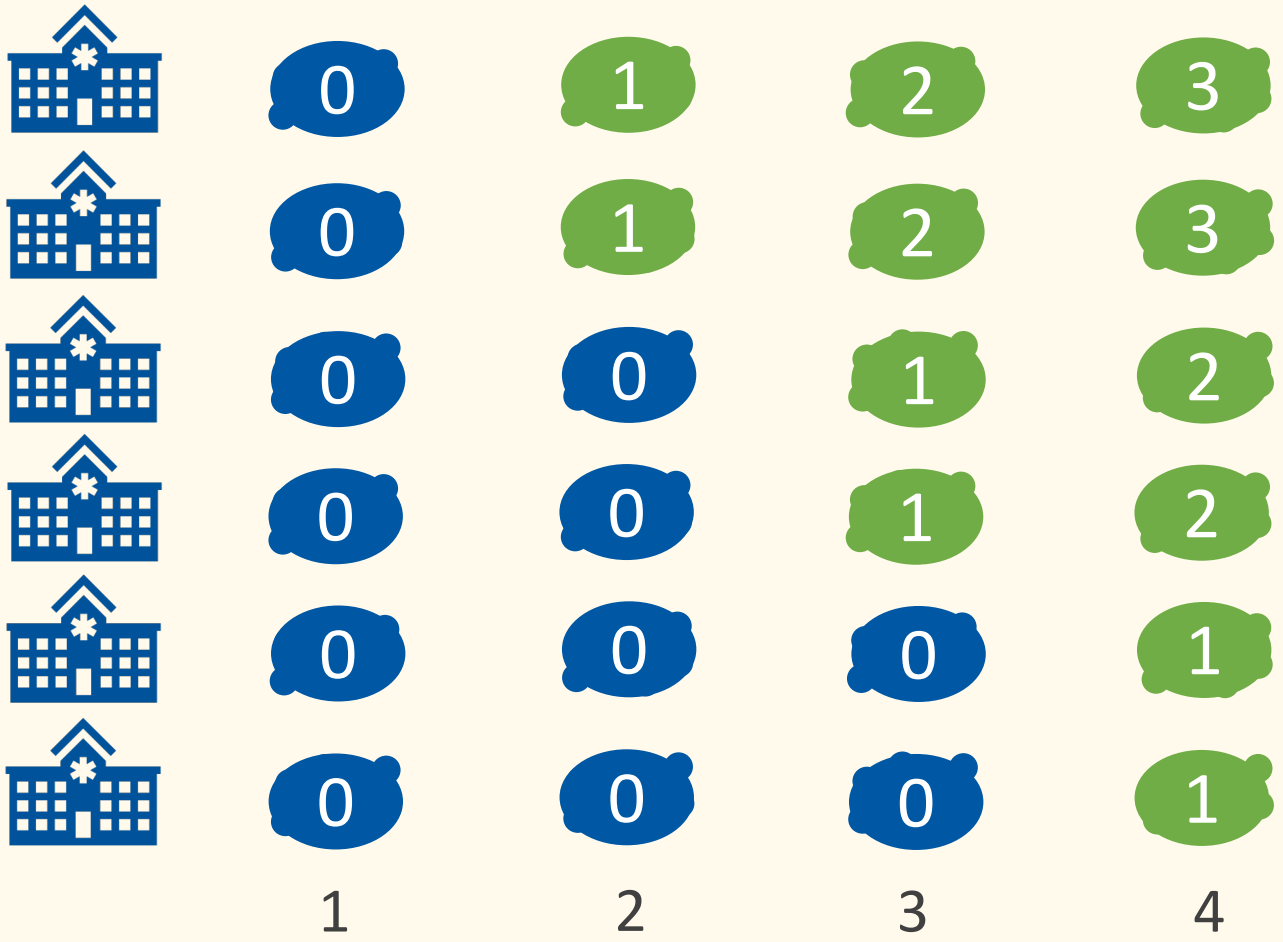
"Exposure time" = Time since the start of the intervention

"Time-varying treatment effect models"

Duke UNIVERSITY
nt of Biostatistics & Bioinformatics
Duke Global Health Institute

The problem of a time-varying treatment effect

It is critical to distinguish between exposure time versus calendar time.



The problem of a time-varying treatment effect

Immediate treatment model:

$$Y_{ij} = \alpha + \beta_j + \delta X_{ij} + \nu_i \qquad (1)$$

δ is a *number*

Time-varying treatment effect model:

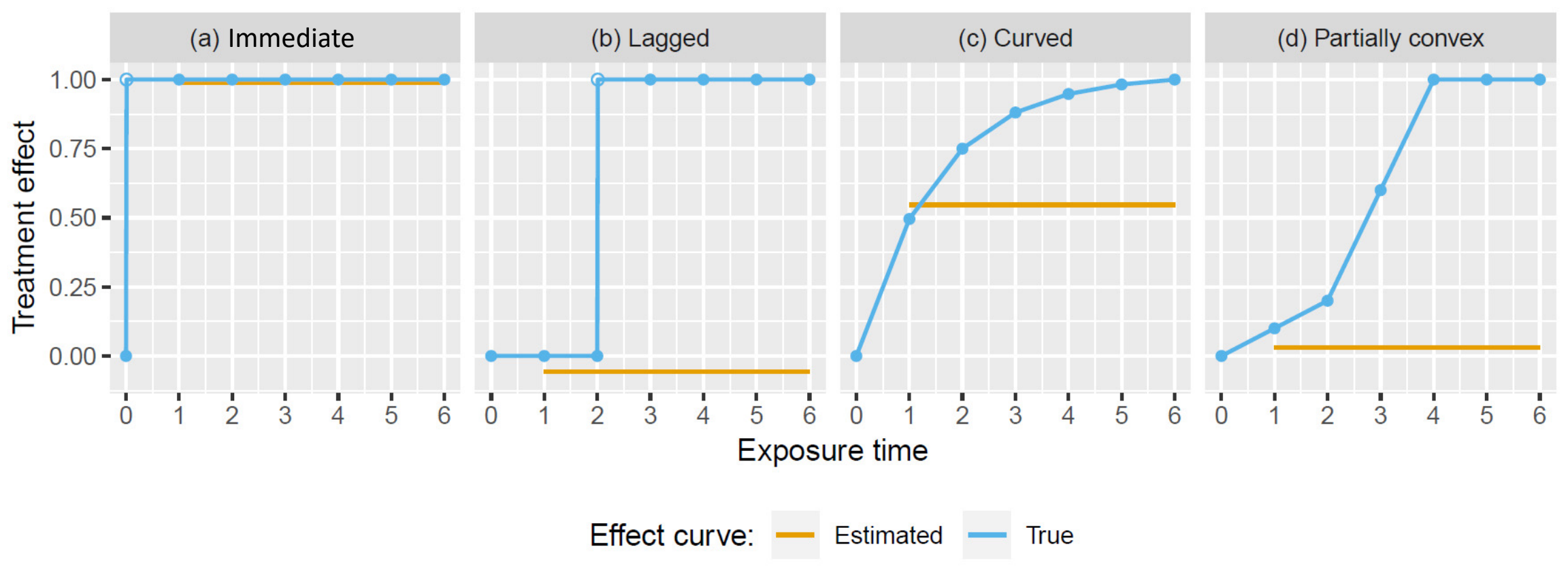
$$Y_{ij} = \alpha + \beta_j + \delta(s_{ij}) X_{ij} + \nu_i \qquad (2)$$

δ is a *function*

Exposure time

Our key question: What happens if data are generated from (2) but analyzed with to (1)?

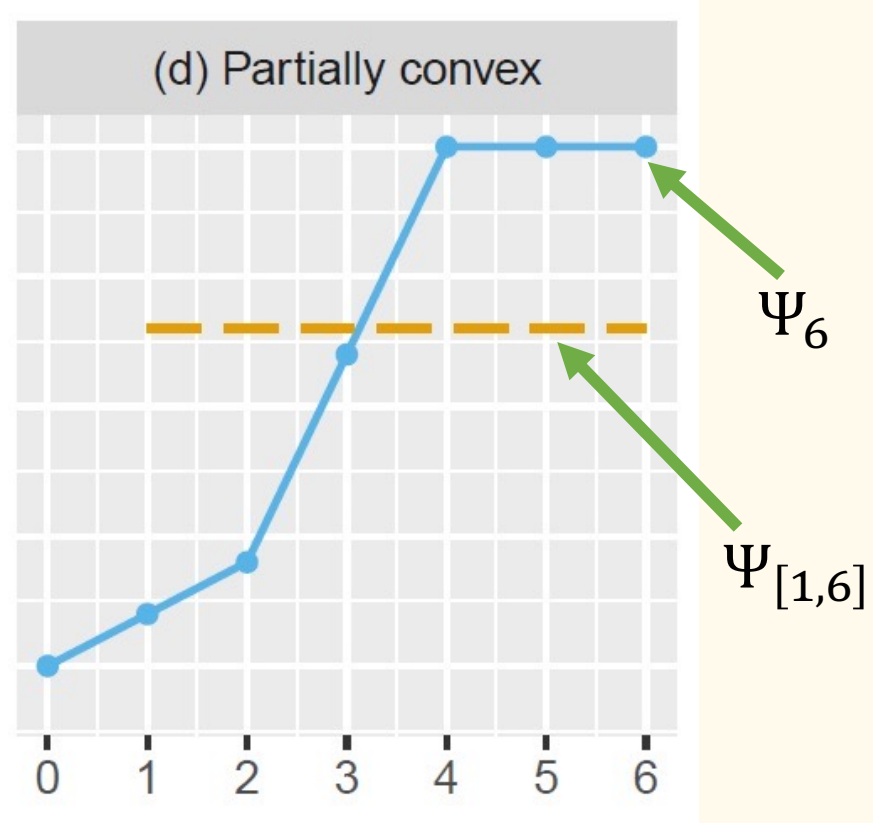
The problem of a time-varying treatment effect



So what do we do?

Analysis of stepped wedge data when the treatment effect varies over time

- When the treatment effect varies with exposure time, we must first define our estimands more precisely
- The *point treatment effect* (PTE) at exposure time s is the value $\Psi_s \equiv \delta(s)$
- The *time-averaged treatment effect* (TATE) between exposure times s and t is defined as $\Psi_{[s,t]} \equiv \frac{1}{t-s}\int_s^t \delta(r)dr$



Analysis of stepped wedge data when the treatment effect varies over time

Immediate treatment model:

$$Y_{ij} = \alpha + \beta_j + \delta X_{ij} + \nu_i \qquad (1)$$

Time-varying treatment effect model:

$$Y_{ij} = \alpha + \beta_j + \delta(s_{ij}) X_{ij} + \nu_i \qquad (2)$$

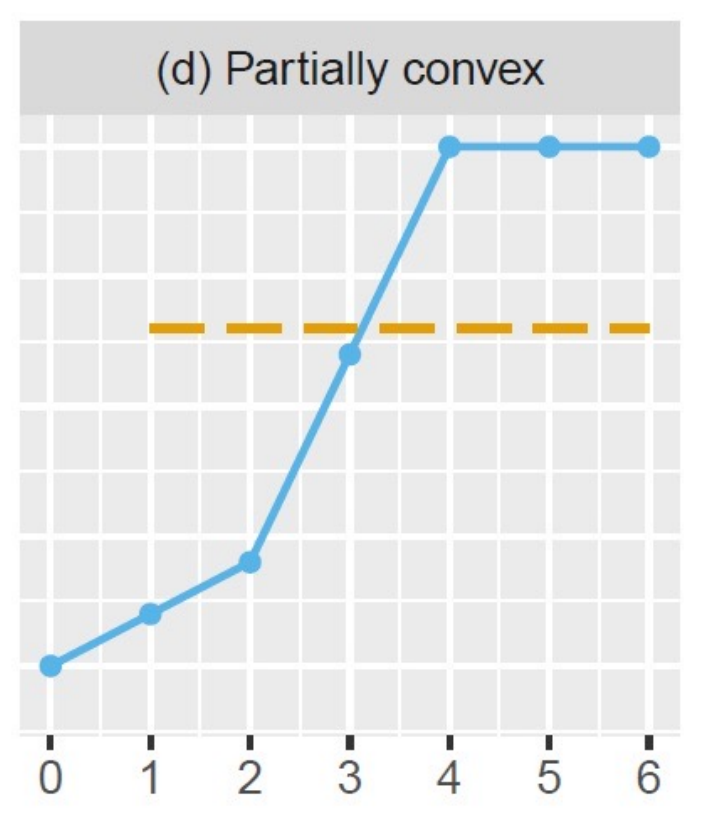
Our model must account for this term and involve a way to estimate the unknown effect curve $s \mapsto \delta(s)$

Exposure time indicator (ETI) model

$$Y_{ij} = \alpha + \beta_j + \sum_{s=1}^{J-1} \delta_s I(s_{ij} = s) + \nu_i$$

TATE estimator: $\hat{\Psi}_{[s,t]} \equiv \frac{\hat{\delta}_s + \hat{\delta}_{s+1} + \cdots + \hat{\delta}_t}{t-s}$

PTE estimator: $\hat{\Psi}_s \equiv \hat{\delta}_s$



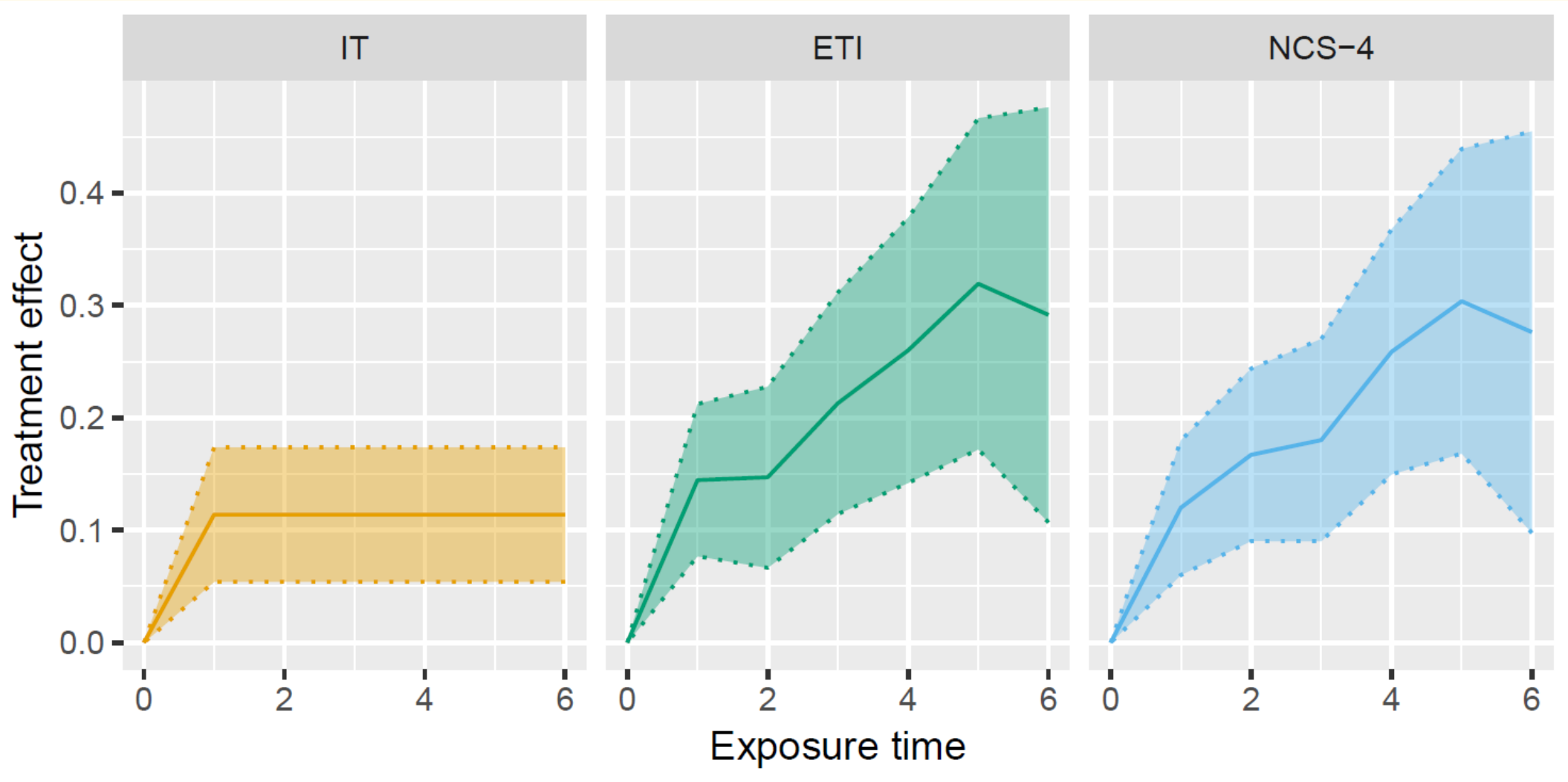
Other models to consider:

- Modeling the effect curve using a natural cubic spline
- Assuming that the full effect is reached by a certain number of time steps
- Assuming that the true effect curve is monotone
- Use a random effect term to account for time-varying treatment effects (Maleyeff et al. 2022)

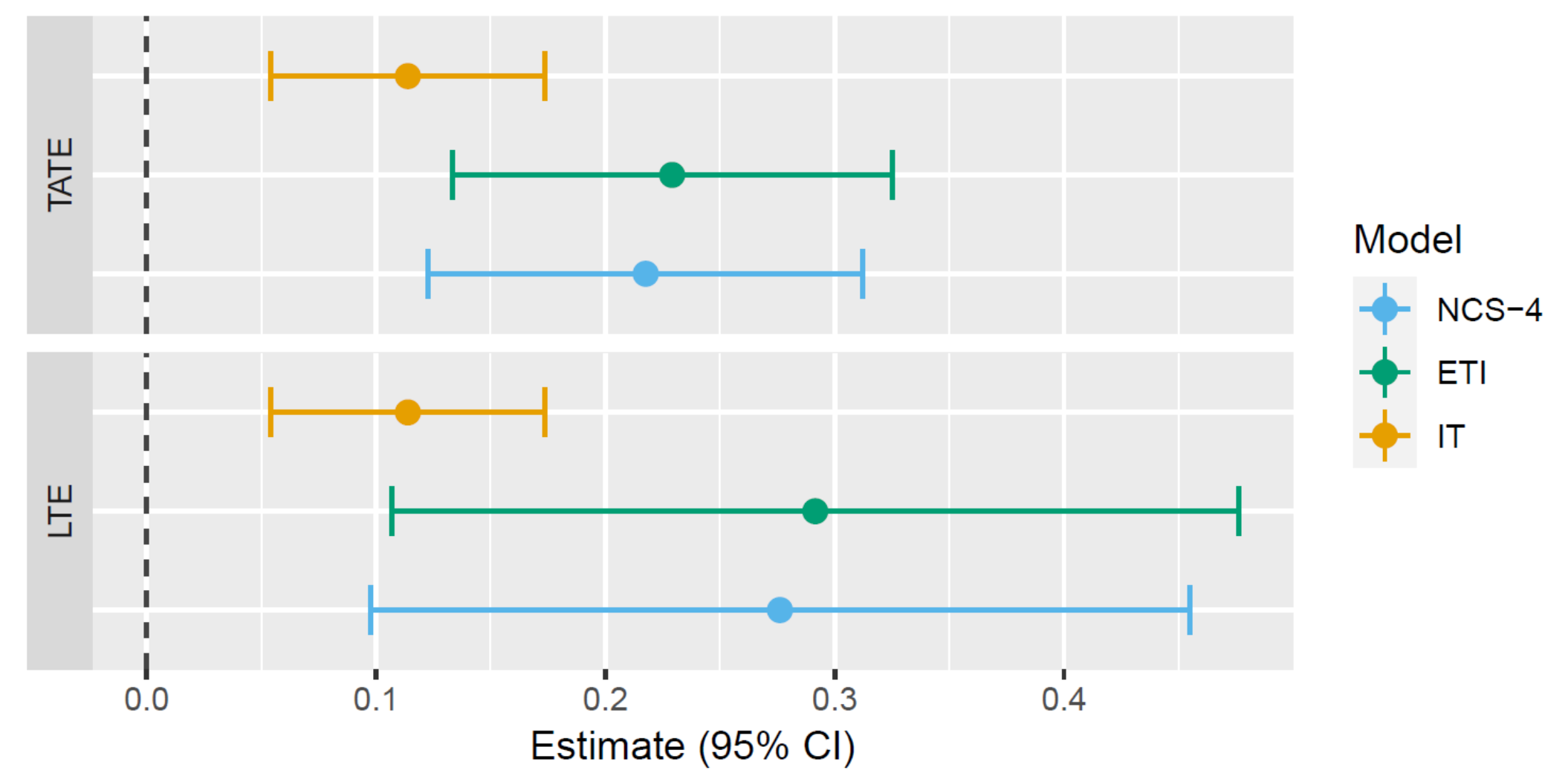
Data analysis examples

- The Australia Disinvestment Trial examined the impact of the removal of weekend health services from twelve hospital wards (Haines et al. 2017).
- The Washington State Community-Level Expedited Partner Treatment (EPT) Randomized Trial sought to test the effect of EPT, an intervention in which the sex partners of individuals with sexually transmitted infections are treated without medical evaluation, on rates of chlamydia and gonorrhea (Golden et al 2015).

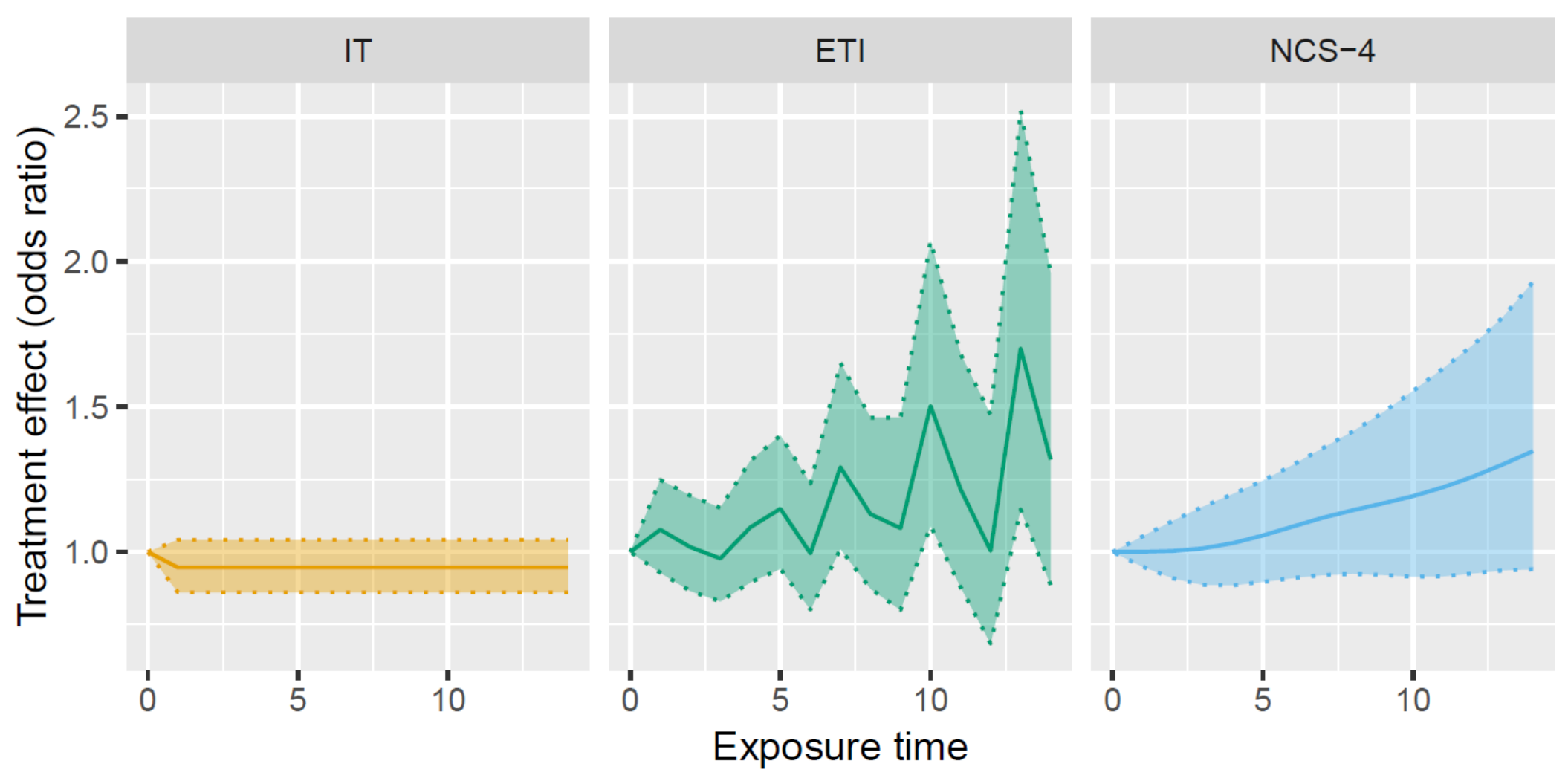
Data analysis example: Australia Disinvestment Trial



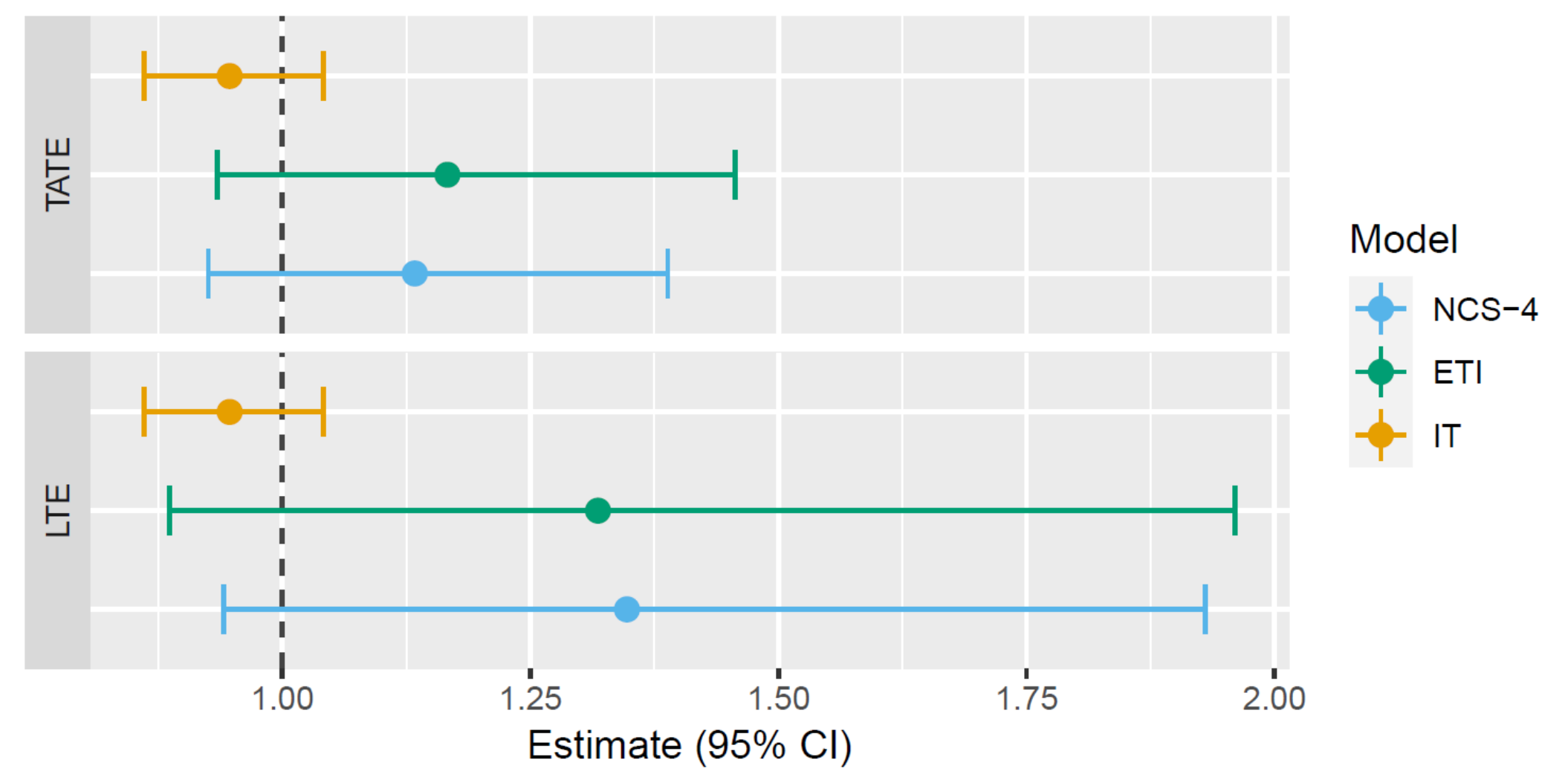
Data analysis example: Australia Disinvestment Trial



Data analysis example: WA EPT Trial



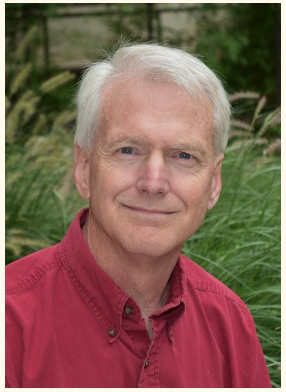
Data analysis example: WA EPT Trial



Conclusions

- If a stepped wedge trial is testing a treatment that varies as a function of exposure time, the use of a model that assumes an immediate treatment effect can lead to serious errors in both estimation and inference.
- We recommend that a model that accounts for time-varying treatment effects is ***always*** used in stepped wedge trials moving forward unless the researcher can defend the strong assumption that the treatment effect is immediate.
- Carefully consider whether the treatment effect is more likely to vary with exposure time or with calendar time.

Collaborators



James P. Hughes
Professor Emeritus, Biostatistics
University of Washington

Fan Xia
Assistant Professor,
Epidemiology & Biostatistics
University of California, San Francisco

Patrick J. Heagerty
Professor, Biostatistics
University of Washington

Emily C. Voldal
Staff Scientist
Fred Hutch Cancer Center

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.
- Hussey MA and Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 2007.
- Maleyeff L, Li F, Haneuse S, Wang R. Assessing exposure-time treatment effect heterogeneity in stepped-wedge cluster randomized trials. *Biometrics*, 2022.
- Haines TP et al. Impact of disinvestment from weekend allied health services across acute medical and surgical wards. *PLoS Medicine*, 2017.
- Golden MR et al. Uptake and population-level impact of expedited partner therapy (EPT) on chlamydia trachomatis and Neisseria gonorrhoeae: the Washington State community-level randomized trial of EPT. *PLoS Medicine*, 2015.

Questions?

Thank you for attending!

Invited Session 7
Novel Methods to Address Treatment Effect Heterogeneity in Cluster Randomized Trials

SCT | 45TH ANNUAL MEETING