

Methods: Mind the Gap
Webinar Series

Design and Analysis Considerations for Assessing Heterogeneity of Treatment Effects in Cluster-Randomized Trials

Presented by:

Noorie Hyun, Ph.D.

Brian Williamson, Ph.D.

Kaiser Permanente Washington Health



National Institutes of Health
Office of Disease Prevention

Design and analysis considerations for assessing heterogeneity of treatment effects in cluster-randomized trials

Brian D. Williamson, PhD

Noorie Hyun, PhD

NIH ODP Methods: Mind the Gap Webinar Series, 5 September 2025

Part 1: Design considerations for subgroup analyses in cluster-randomized trials

Acknowledgments

This work was done in collaboration with:



Yates Coley



Clarissa Hsu



Courtney McCracken



Andrea Cook



and was funded by the Patient-Centered Outcomes Research Institute (COVID-2021C2-1368)

Heterogeneity of treatment effects: what and why

Heterogeneity of treatment effects (HTE):

- differential effect of exposure/intervention across subgroups
- does not necessarily imply a causal or biological interpretation

If HTE is present:

- resources can be prioritized to maximize benefit and reduce risks [Knol & VanderWeele (2012)]
- can determine whether intervention **remediates** or **exacerbates** health disparities [Petkovic et al. (2020)]

Case study: ENSPIRE trial

ENSPIRE: Engaging Staff to Improve COVID-19 Vaccination Response at Long-term Care Facilities (NCT05449418)

Background:

- Rates of COVID-19 disease and death higher in long-term care than general population [Chidambaram (2022)]
- Lag in COVID-19 vaccinations among long-term care workers
- Long-term care (LTC) workers often low-wage, represent diverse social, cultural, racial, and ethnic backgrounds

Design:

- Enrolled 40 LTC centers, encompassing approximately 4000 staff members
- Recruited from 2 US states (Georgia and Washington)
- Recruited from urban (Atlanta or Seattle) and suburban/rural locations

Case study: ENSPIRE trial

Primary outcomes:

1. LTC center COVID-19 booster rate (**cluster-level**)
 - chosen due to concerns about data capture
2. LTC center staff net promoter score (**individual-level**)

Intervention: booster promotion materials co-designed with LTC staff

"Usual care"
(Standard materials)



VS.

"Co-design"
(Materials designed by teams of staff)



 THE ENSPIRE STUDY

Case study: ENSPIRE trial

Co-design: tailoring messages to specific audiences

- In ENSPIRE, focused on [language, cultural, and ethnic affinity](#) groups
- Measured race and ethnicity allowing 57 total options

HTE analysis natural for interventions, particularly co-designed:

- Planned analyses based on state, urban vs suburban/rural, baseline booster uptake
- Self-reported race and ethnicity both interesting and important, however:
 - neither sites nor other entity had comprehensive and accurate race and ethnicity data for staff
 - staff survey would only capture a portion of staff members
 - also, didn't know what this would look like with a cluster-level outcome!

Individual-level data structure

Data on n_c staff members in each cluster $c = 1, \dots, K$:

- A_{ci} : indicator of intervention assignment (cluster-level)
- $Z_{ci} \in \mathbb{R}^q$: cluster-level subgroup variables
- $W_{ci} \in \mathbb{R}^p$: baseline covariates
- Y_{ci} : binary outcome

HTE with individual-level outcomes

Suppose we are interested in HTE by a single binary variable W_{ci} .

Target estimand: the difference in the risk difference based on levels of W_{ci} .

Posit linear regression model:

$$E(Y_{ci} \mid A_{ci} = a, Z_{ci} = z, W_{ci} = w) = \beta_0 + \beta_1 a + \beta_2 z + \beta_3 w + \beta_4 aw$$

Under standard assumptions, β_4 quantifies HTE.

Cluster-level data structures

In many settings, we observe cluster-level data:

- $\bar{Y}_c := n_c^{-1} \sum_{i=1}^{n_c} Y_{ci}$: cluster-level outcome proportion
- A_c : indicator of intervention assignment
- $Z_c \in \mathbb{R}^q$: cluster-level subgroup variables
- $X_c \in \mathbb{R}^p$: (possibly aggregated) baseline covariates

Possible options for aggregating an individual-level covariate W_{cj} :

- $X_{cj} = V_{cj} - \frac{1}{K} \sum_{c=1}^K V_{cj}$, where $V_{cj} = \frac{1}{n_c} \sum_{i=1}^{n_c} W_{cji}$
- $X_{cj} = I(V_{cj} > t)$ for threshold t

Aggregated data: complications

Individual-level effects **may not be identifiable** based on aggregate data.

- Often referred to as **ecological bias** [Wakefield (2008)]
- Exposure-outcome association can be biased: [Greenland & Morgenstern (1989)]
 - by aggregated confounders
 - by aggregated HTE variables

Compositional data can cause **challenges in interpretation**:

- Arise when aggregating mutually-exclusive binary variables
- Changing the value of one proportion necessarily changes at least one other

HTE with cluster-level outcomes

Now, outcome $\bar{Y}_c$ is a proportion; interest in HTE by single variable X_c .

Target estimand: difference in difference of proportions based on values of X_c .

Posit linear regression model:

$$E(\bar{Y}_c \mid A_c = a, Z_c = z, X_c = x) = \alpha_0 + \alpha_1 a + \alpha_2 z + \alpha_3 x + \alpha_4 ax.$$

Under standard assumptions, α_4 quantifies HTE.

If W represents a single binary variable, then aggregation determines interpretation:

- Using mean-centered proportion: difference in difference **for 1-unit increase from the cluster-level mean proportion**
- Using threshold: difference in difference **comparing “high” to “low”**

If W is compositional, this is further complicated:

- how are the categories transformed to enable model fitting?
- what does a one-unit change imply?

Simulation setup

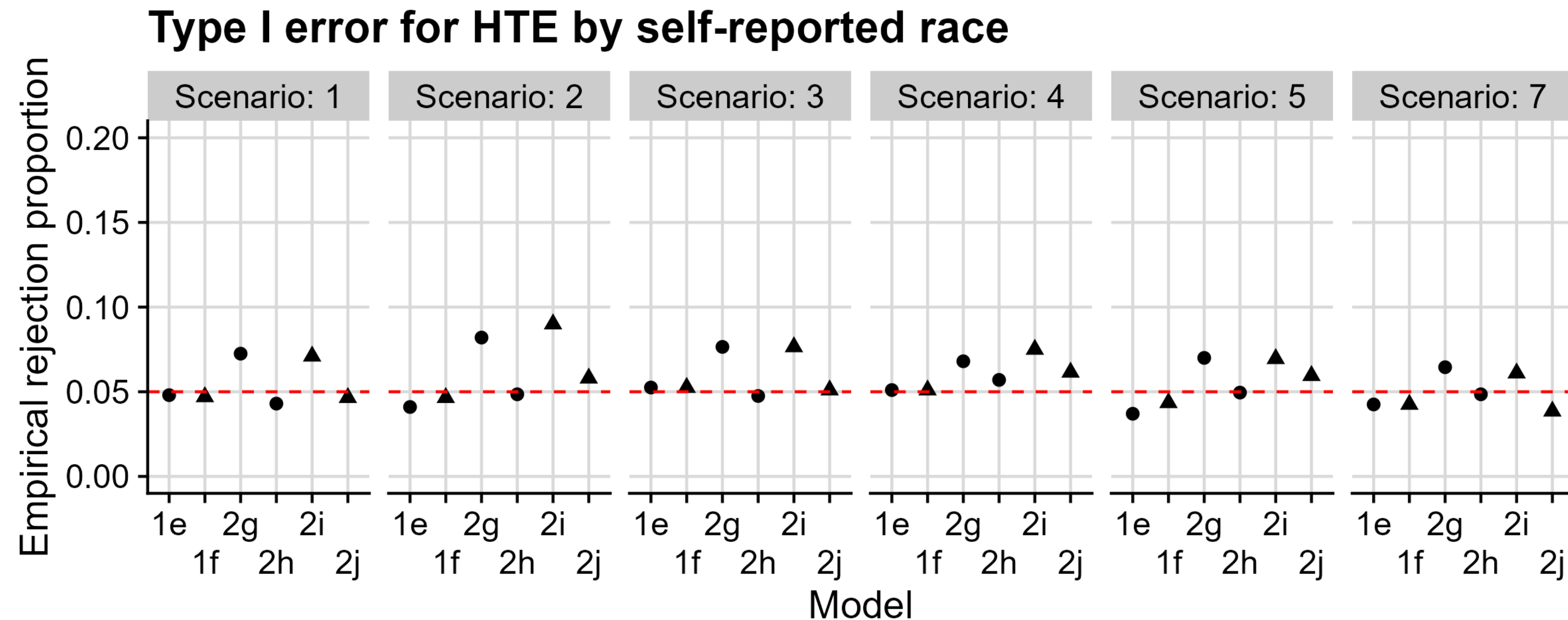
Data mimic ENSPIRE trial:

- Sample $K \in \{30,40,80\}$ clusters within two regions
- Clusters are either urban or suburban/rural
- Randomization constrained to balance on region, urban vs suburban/rural
- Each center has 100 staff members, who each:
 - have record of receiving COVID-19 booster vaccine
 - have self-reported race (recorded as a single binary variable)
- Define cluster-level booster rate $\bar{Y}_c$
- Define two aggregations of self-reported race:
 - center-level proportion
 - indicator of whether proportion > 0.5

Fit both individual-level and cluster-level regression models

Tested for HTE by self-reported race, investigated type I error and power

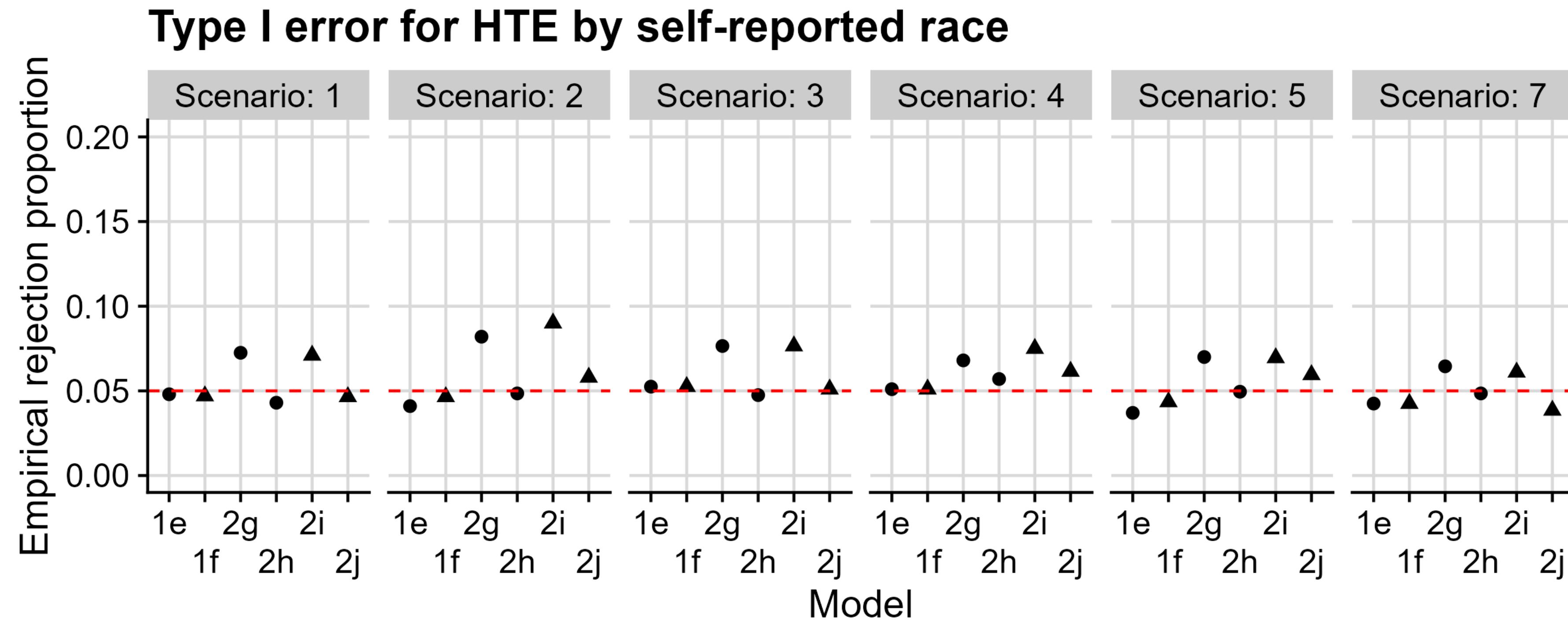
Results: type I error



Scenarios encode presence of true differences in booster rate or HTE:

- 1,2,5: no difference in booster rate by self-reported race, no HTE
- 3,4,7: difference in booster rate by self-reported race, no HTE

Results: type I error

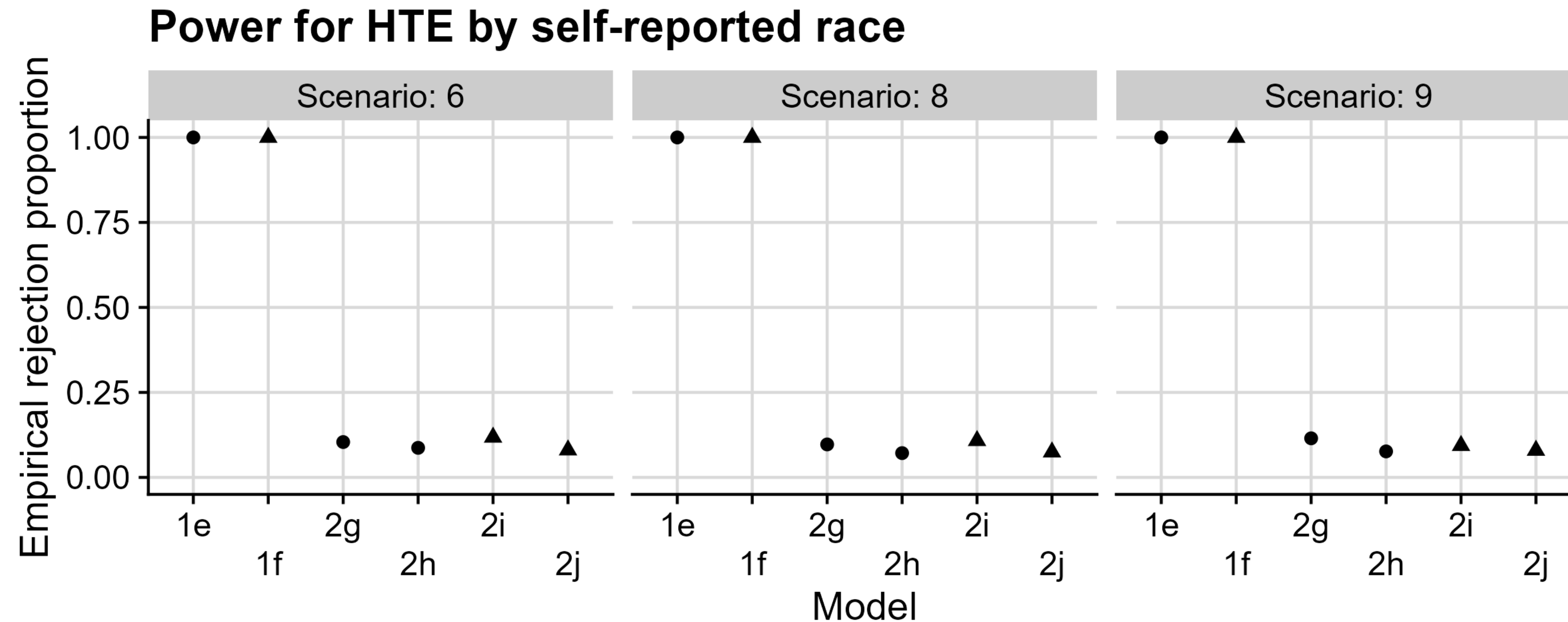


- Interaction variable(s)
- Self-reported race
 - ▲ Self-reported race and urban status

Models encode individual- (1) vs cluster-level (2), adjustment and interaction variables:

- 1e,1f: adjust for self-reported race, allow HTE by self-reported race (1f also by urban)
- 2g,2i: adjust for center-level race proportion, allow HTE by self-reported race (2i also by urban)
- 2h,2j: adjust for proportion > 0.5 , allow HTE by self-reported race (2j also by urban)

Results: power

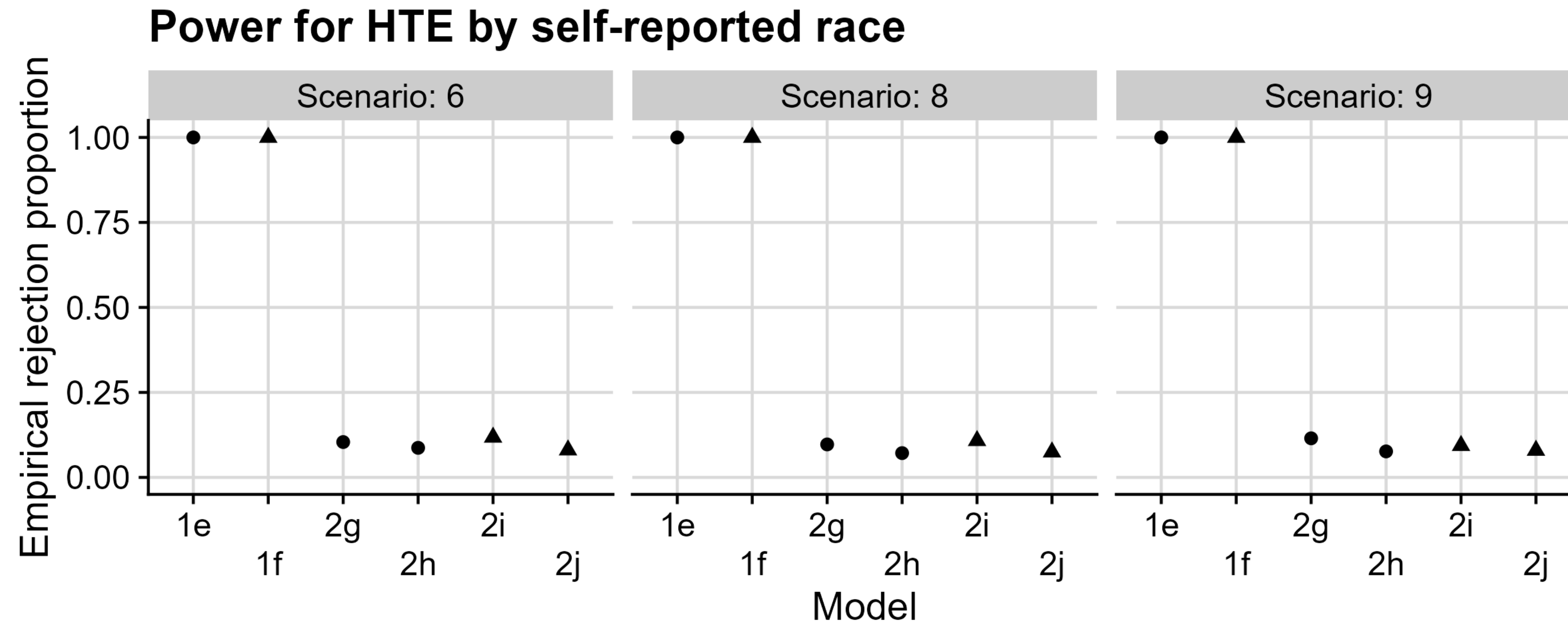


- Interaction variable(s)
- Self-reported race
 - ▲ Self-reported race and urban status

Scenarios encode presence of true differences in booster rate or HTE:

- 6: difference in booster rate by self-reported race, HTE by self-reported race
- 8: difference in booster rate by urban and self-reported race, HTE by self-reported race
- 9: difference in booster rate by urban and self-reported race, HTE by urban and self-reported race

Results: power



- Interaction variable(s)
- Self-reported race
 - ▲ Self-reported race and urban status

Models encode individual- (1) vs cluster-level (2), adjustment and interaction variables:

- 1e,1f: adjust for self-reported race, allow HTE by self-reported race (1f also by urban)
- 2g,2i: adjust for center-level race proportion, allow HTE by self-reported race (2i also by urban)
- 2h,2j: adjust for proportion > 0.5 , allow HTE by self-reported race (2j also by urban)

Closing thoughts

High power, controlled type I error to detect HTE in individual-level analysis

- where possible, this approach is key for identifying and remedying disparities

Low power to detect HTE by an aggregated individual-level predictor in cluster-level analysis

- even in simple setting with only two possible categories!
- likely exacerbated in more realistic settings
- elevated type I error when using proportion in some settings

However, meaningful estimated differences across groups can still be acted upon.

Power calculations should aggregate simulated individual-level data, rather than simulating cluster-level data directly.

For more details, including more realistic settings:



Part 2: Increased risk of type I errors for detecting HTE in cluster-randomized trials using mixed effects models

Acknowledgement



Jennifer Bobb



Abisola Idu



Andrea Cook

Funded by the National Institute of the NIH On Drug Abuse (Award Number UG1DA040314)

Motivation Study

Primary Care Opioid Use Disorder treatment (PROUD) trial

- Pragmatic Two-Arm (1:1) Cluster Randomized Trial
- Clinic-level Intervention: implementation of the Massachusetts Model of nurse care management for opioid use disorder (OUD)
- Aim: to increase OUD treatment in diverse primary care settings in the United States
- 12 clinics (clinic: cluster unit)
- Median 122 patients per clinic (IQR: 97-230) at baseline

Generalized linear mixed-effects model (GLMM)

For subject j in cluster i ,

Y_{ij} : number of days of utilization of acute care services (count data)

To estimate heterogeneity of treatment effect (HTE) by sex, we fit the model:

$$\log\{E(Y_{ij} \mid U_i)\} = \beta_0 + \beta_1 * Y_{0ij} + \beta_{trt} * trt_i + \beta_{sex} * sex_{ij} + \beta_{mod}(trt_i * sex_{ij}) + U_i,$$

for $U_i \sim N(0, \sigma^2)$

- $trt_i = 1$ if PROUD; 0 if usual care.
- $sex_{ij} = 1$ if male; 0 if female.

Result from the fitted GLMM

Fixed effects	Beta	Standard Error	P-value*
Intercept	-5.38	0.26	<0.001
Intervention PROUD	0.21	0.37	0.58
Male	0.015	0.03	0.60
Baseline Value	0.27	0.003	<0.001
Effect modification (Intervention × Male)	-0.14	0.04	0.01
Intervention effect by subgroup	Effect (95% Confidence interval)		
Within men	0.072 (-0.75, 0.89)		
Within women	0.21 (-0.61, 1.03)		

* Small size correction: subtracting the number of cluster-level parameters for fixed effects from the number of clusters

Research gaps

- **Barr et al. (2013)** suggested that, for testing interactions in linear mixed-effects models, the models should have by-cluster random slopes for any interactions, where **all factors comprising the interactions are within-cluster**.
- **Research gaps:** (1) these prior papers were targeted to psychometrics field rather than CRTs; (2) Non-continuous outcomes and small numbers of clusters have not been considered.

GEE vs. GLMM vs. GLMM2

- **Outcome types:** continuous, count and binary
- Generalized Estimating Equation (**GEE**) with independent working correlation,
- Generalized Linear Mixed-Effect Model (**GLMM**) with **within-clinic correlation**
- GLMM with **clinic-by-subgroup correlation (GLMM2)**

For subject j in cluster i,

$$\textbf{GEE: } g\{E(Y_{ij})\} = \beta_0 + trt_i * \beta_{trt} + sex_{ij} * \beta_{sex} + (trt_i * sex_{ij}) * \beta_{mod}$$

$$\textbf{GLMM: } g\{E(Y_{ij} | U_i)\} = \beta_0 + trt_i * \beta_{trt} + sex_{ij} * \beta_{sex} + (trt_i * sex_{ij}) * \beta_{mod} + \textcolor{blue}{U_i},$$
$$\textcolor{blue}{U_i} \sim N(\mathbf{0}, \sigma^2)$$

$$\textbf{GLMM2: } g\{E(Y_{ij} | U_{i,s_{ij}})\} = \beta_0 + trt_i * \beta_{trt} + sex_{ij} * \beta_{sex} + (trt_i * sex_{ij}) * \beta_{mod} + \textcolor{blue}{U_{0i}} + \textcolor{blue}{U_{1i}} * \textcolor{blue}{sex_{ij}},$$
$$(\textcolor{blue}{U_{0i}}, \textcolor{blue}{U_{1i}}) \sim MVN(\mathbf{0}, \Sigma)$$

where g()=link function; trt= 1 if PROUD, 0 otherwise; sex_{ij} =1 if male, 0 otherwise.

Simulation Study

Scenario description	True model	Covariance Σ for $(U_{i,0}, U_{i,1})$ or variance σ^2 for U_i	Number of clusters, N
Scenario 1: ideal settings for GLMM2 in terms of adequate random effect and number of clusters	GLMM2	$\begin{pmatrix} 0.2 & 0.13 \\ 0.13 & 0.1 \end{pmatrix}$, $\begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}$ and $\begin{pmatrix} 0.25 & 0.18 \\ 0.18 & 0.5 \end{pmatrix}$ for continuous, count and binary outcomes, respectively.	N=50 and 100 with fixed cluster size 100
Scenario 2: small random effect but adequate number of clusters	GLMM2	Covariances are downsized by 10% of the variances in Scenario 1 with the same correlation degrees. These covariances reflect PROUD data application.	N=50 and 100 with fixed cluster size 100
Scenario 3: adequate random effect but small number of clusters	GLMM2	The same covariances of Scenario 1.	N=12 with fixed cluster 100 and N=12 with varying cluster size of (25, 50, 100, 150, 300)
Scenario 4: to evaluate the robustness of GLMM2 when GLMM2 is over fitting.	GLMM	σ^2 is 0.2, 0.5 and 0.5 for continuous, count and binary outcomes, respectively. e.g. $\Sigma = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	N=50 with fixed cluster size 100 and N=12 with varying cluster size of (25, 50, 100, 150, 300)

Simulation Study (2)

Scenario description	True model	Covariance Σ for $(U_{i,0}, U_{i,1})$ or variance σ^2 for U_i	Number of clusters, N
Scenario 5: to evaluate the robustness of GLMM2 when GLMM2 is over fitted, and subgroup is three-level categorical.	GLMM	The same covariances of Scenario 1.	N=50 with fixed cluster size 100 and N=12 with varying cluster size of (25, 50, 100, 150, 300)
Scenario 6: to evaluate the robustness of GLMM2 when GLMM2 is misspecified.	GLMM2 with heterogeneous covariances between both arms for subgroup-specific random effects	$c_t \begin{pmatrix} 0.2 & 0.13 \\ 0.13 & 0.1 \end{pmatrix}, c_t \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}$ and $c_t \begin{pmatrix} 0.25 & 0.18 \\ 0.18 & 0.5 \end{pmatrix}$ for continuous, count and binary outcomes, respectively, where $c_t = 1$ for intervention arm and $c_t=1.4$ for control arm.	N=50 with fixed cluster size 100 and N=12 with varying cluster size of (25, 50, 100, 150, 300)
Scenario 7: to compare power of GLMM and GLMM2 for detecting HTE.	GLMM	The same covariances of Scenario 1.	N=50 with fixed cluster size 100 and N=12 with varying cluster size of (25, 50, 100, 150, 300)

Simulation Scenario 1: ideal setting for GLMM2 (GLMM2 is the true model.)

	N=50 with fixed cluster size 100	N=100 with fixed cluster size 100
Models	Type I error rate	Type I error rate
Continuous		
GLMM2	0.05	0.05
GLMM	0.22	0.21
GEE	0.06	0.05
Count		
GLMM2	0.07	0.06
GLMM	0.47	0.47
GEE	0.09	0.06
Binary		
GLMM2	0.06	0.06
GLMM	0.19	0.18
GEE	0.07	0.06

Simulation Scenario 2: small random effect (GLMM2 is the true model.)

	N=50 with fixed cluster size 100	N=100 with fixed cluster size 100
Models	Type I error rate	Type I error rate
Continuous		
GLMM2	0.037 (248)	0.036 (183)
Two-step	0.056	0.044
GLMM	0.092 (0)	0.085 (0)
GEE	0.053	0.043
Count		
GLMM2	0.055 (400)	0.039 (286)
Two-step	0.067	0.05
GLMM	0.085 (1)	0.077 (0)
GEE	0.06	0.04
Binary		
GLMM2	0.056 (278)	0.038 (218)
Two-step	0.053	0.05
GLMM	0.071 (75)	0.054 (28)
GEE	0.076	0.045

(Number of singular fits out of 1,000 replicates)

Simulation Scenario 3: small number of clusters (GLMM2 is the true model.)

	NC=12 with fixed cluster sizes 100	NC=12 with varying cluster sizes of (25, 50, 100, 150, 300)
Models	Type I error rate	Type I error rate
Continuous		
GLMM2 ^c	0.040 (159)	0.048 (181)
Two-step ^c	0.053	0.089
GLMM ^c	0.205	0.321
GEE ^d	0.067	0.072
Count		
GLMM2 ^e	0.041 (17)	0.034 (30)
Two-step ^{e,f}	0.045	0.044
GLMM ^f	0.356	0.449
GEE ^d	0.099	0.123
Binary		
GLMM2 ^f	0.060 (236)	0.068 (167)
Two-step ^f	0.09	0.144
GLMM ^f	0.138	0.21
GEE ^d	0.072	0.087

c = Satterthwaite; d = Fay and Graubard; e =between-within; f = Subtracting the number of cluster-level parameters for fixed effects from the number of clusters.

Simulation Scenario 4: GLMM is the true model, and the subgroup is binary.

	NC=50 with fixed cluster size 100	NC=12 with varying cluster sizes of (25, 50, 100, 150, 300)
Models	Type I error rate	Type I error rate
Continuous		
GLMM2	0.045 (402)	0.032 (574)
Two-step	0.046	0.038
GLMM	0.058	0.046
GEE	0.055	0.062
Count		
GLMM2	0.036 (273)	0.031 (387)
Two-step	0.04	0.042
GLMM	0.045	0.051
GEE	0.063	0.077
Binary		
GLMM2	0.039 (173)	0.041 (349)
Two-step	0.039	0.043
GLMM	0.046	0.041
GEE	0.069	0.072

Simulation Scenario 5: GLMM is the true model, and the sub-group level is 3.

Likelihood ratio test	NC=50 with fixed cluster size 100	NC=12 with varying cluster sizes of (25, 50, 100, 150, 300)
Models	Type I error rate	Type I error rate
Continuous		
GLMM2	0.041 (612)	0.037 (778)
Two-step	0.045	0.051
GLMM	0.052	0.055
Count		
GLMM2	0.044 (118)	0.033 (13)
Two-step	0.045	0.034
GLMM	0.050	0.048 (1 ^b)
Binary		
GLMM2	0.037 (35)	0.027 (116)
Two-step	0.037	0.032
GLMM	0.041	0.053

Simulation Scenario 7: GLMM2 with heterogeneous subgroup-specific covariance between arms.

	NC=50 with fixed cluster size 100	NC=12 with varying cluster sizes of (25, 50, 100, 150, 300)
Models	Type I error rate	Type I error rate
Continuous		
GLMM2	0.040 (1)	0.048 (129)
Two-step	0.040	0.080
GLMM	0.243	0.360
GEE	0.055	0.078
Count		
GLMM2	0.070 (0)	0.040 (26)
Two-step	0.070	0.043
GLMM	0.518	0.484
GEE	0.086	0.135
Binary		
GLMM2	0.074 (1)	0.063 (123)
Two-step	0.074	0.078
GLMM	0.219	0.210 (12b)
GEE	0.056	0.088

Small Sample Size Correction Methods for GLMM and GLMM2

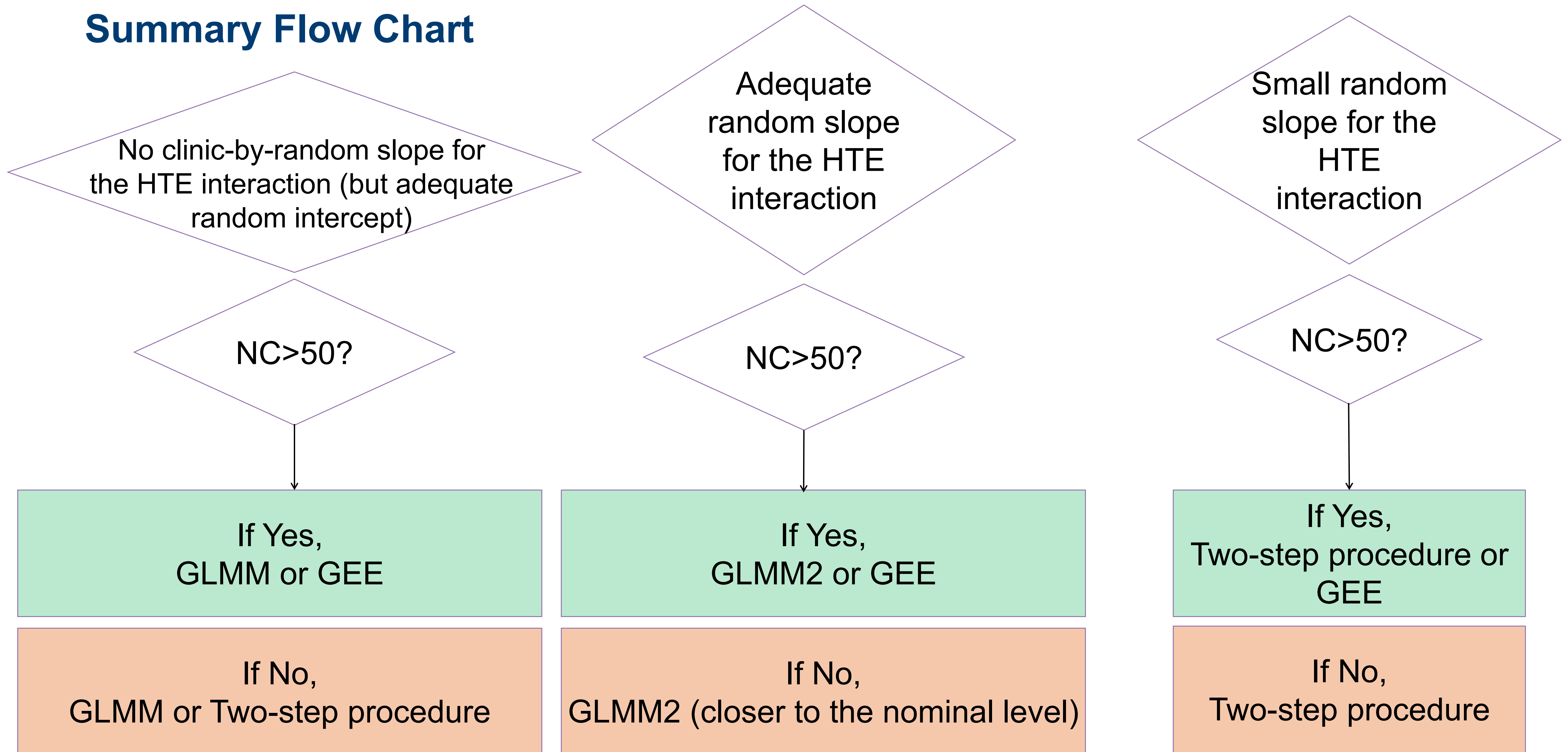
True model	Fitted model	Outcomes		
		Continuous	Count	Binary
GLMM2	GLMM2	Satterthwaite	Between-within denominator degrees of freedom (DDF) that is I-rank (X).	Subtracting the number of cluster-level parameters for fixed effects from the number of clusters
	GLMM		Subtracting the number of cluster-level parameters for fixed effects from the number of clusters	Subtracting the number of cluster-level parameters for fixed effects from the number of clusters
GLMM	GLMM2		Parametric bootstrap	Parametric bootstrap
	GLMM		Between-within DDF	Between-within DDF

The 'parameters' R-package was used; Fay-Graubard small sample size correction for GEE (saws R-package)

Comparison of models: application to the PROUD trial

	GLMM2			GLMM			GEE		
Fixed effects	Beta	Standard Error	P-value	Beta	Standard Error	P-value	Beta	Standard Error	P-value
Intercept	-5.34	0.3	<0.001	-5.38	0.26	<0.001	-4.94	0.09	<0.001
Intervention PROUD	-0.067	0.43	0.88	0.21	0.37	0.58	-0.2	0.25	0.19
Male	-0.09	0.32	0.79	0.015	0.03	0.60	-0.04	0.06	0.83
Baseline Outcome	0.28	0.003	<0.001	0.27	0.003	<0.001	0.28	0.02	<0.001
Effect modification (Inter × Male)	0.21	0.45	0.64	-0.14	0.04	0.01	0.02	0.14	0.93
Intervention effect by subgroup				Effect (95% Confidence interval)					
Within men	0.14 (-0.84, 1.13)			0.072 (-0.75, 0.89)			-0.18 (-0.60, 0.24)		
Within Women	-0.067 (-1.02, 0.89)			0.21 (-0.61, 1.03)			-0.20 (-0.51, 0.12)		

Summary Flow Chart



NC=Number of Clusters

References

Part 1:

- Knol MJ and VanderWeele TJ (2012). Recommendations for presenting analyses of effect modification and interaction. *International Journal of Epidemiology*.
- Petkovic et al. (2020). Reporting of health equity considerations in cluster and individually randomized trials. *Trials*.
- Chidambaram P (2022). Over 200,000 residents and staff in long-term care facilities have died from COVID-19. *Kaiser Family Foundation*.
- Wakefield J (2008). Ecologic studies revisited. *Annual Review of Public Health*.
- Greenland S & Morgenstern H (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology*.

Part 2:

- Nugent J and Kleinman K (2021), Type I error control for cluster randomized trials under varying small sample structures. *BMC Medical Research Methodology*.
- Barr D (2013). Random effects structure for testing interactions in linear mixed-effects models, *Frontiers in Psychology*.
- Fay M and Graubard BI (2001). Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. *Biometrics*.
- Qiu H, Cook AJ and Bobb JF (2024). Evaluating tests for cluster-randomized trials with few clusters under generalized linear mixed models with covariate adjustment: A simulation study, *Statistics in Medicine*.

Questions?

Brian Williamson:

 brian.d.williamson@kp.org

 <https://bdwilliamson.github.io/>

Noorie Hyun:

 noorie.hyun@kp.org

 <https://kpwashingtonresearch.org/our-research/our-scientists/hyun-noorie>