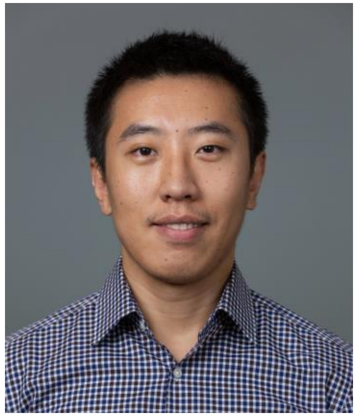
NIA IMPACT COLLABORATORY

TRANSFORMING DEMENTIA CARE

Methods for Designing Cluster Randomized Trials to Detect Treatment Effect Heterogeneity



Fan Li, PhD

Assistant Professor of Biostatistics

Yale School of Public Health

Housekeeping

- All participants will be muted
- Enter **all questions** in the Zoom **Q&A/chat box** and send to Everyone
- Moderator will review questions from chat box and ask them at the end
- Want to continue the discussion? Associated podcast released about 2 weeks after Grand Rounds
- Visit impactcollaboratory.org
- Follow us on Twitter & LinkedIN:

@IMPACTcollab1 https://www.linkedin.com/company/65346172

Methods for Designing Cluster Randomized Trials to Detect Treatment Effect Heterogeneity

Fan Li

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

🌍 https://lifan90.com/

NIA IMPACT Collaboratory, Grand Rounds
Feb 16, 2023

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.
- 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

Learning objective

- Understand the sample size requirements for testing treatment effect heterogeneity in cluster randomized trials
- Be aware of tools for designing cluster randomized trials
- A call for involving statisticians at the outset to design cluster randomized trials
  - stayed tuned for the IMPACT Design & Statistics Core Health Equity Best Practices Training Module

Outline

- 1. Introduction
- 2. Planning cluster randomized trials for assessing treatment heterogeneity
  - 2.1 Demystifying a sample size formula
  - 2.2 Software tool and an example
- 3. Additional considerations
- 4. Discussion

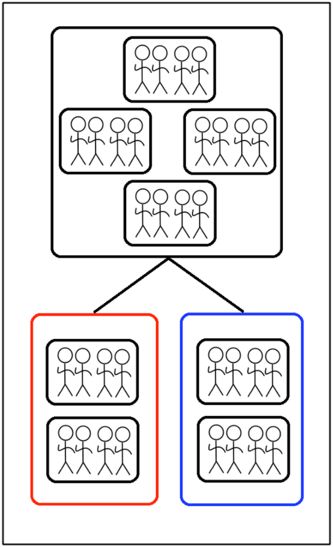
1. Introduction

Cluster randomized trials

- Cluster randomized trials (CRTs) randomize entire clusters/groups of individuals to treatment conditions
    - avoid contamination
    - administrative and logistical considerations
- Increasingly seen in pragmatic trials for AD/ADRD population
- Essential task in planning studies is to ensure adequate power for detecting a clinically meaningful 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

- 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

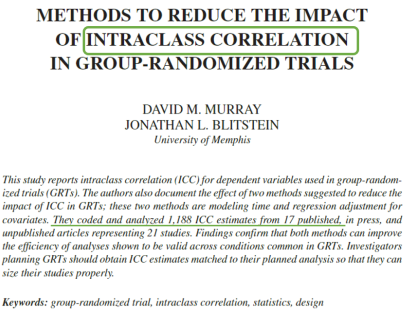
$$\rho_y = \frac{\text{between-cluster variance}}{\text{total variance}}$$

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

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

- Often available from published literature, existing database, or pilot data

Published ICC estimates

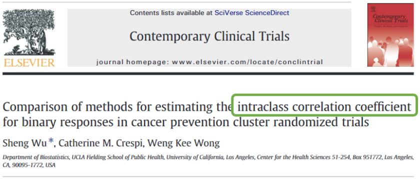


METHODS TO REDUCE THE IMPACT OF INTRACLASS CORRELATION IN GROUP-RANDOMIZED TRIALS

DAVID M. MURRAY
JONATHAN L. BLITSTEIN
University of Memphis

This study reports intraclass correlation (ICC) for dependent variables used in group-randomized trials (GRTs). The authors also document the effect of two methods suggested to reduce the impact of ICC in GRTs; these two methods are modeling time and regression adjustment for covariates. They coded and analyzed 1,188 ICC estimates from 17 published, in press, and unpublished articles representing 21 studies. Findings confirm that both methods can improve the efficiency of analyses shown to be valid across conditions common in GRTs. Investigators planning GRTs should obtain ICC estimates matched to their planned analysis so that they can size their studies properly.

Keywords: *group-randomized trial, intraclass correlation, statistics, design*



Contents lists available at SciVerse ScienceDirect

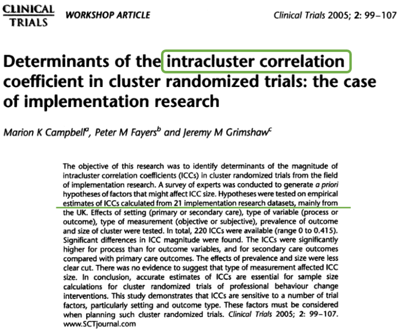
Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials

Sheng Wu *, Catherine M. Crespi, Weng Kee Wong

Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, Center for the Health Sciences 51-254, Box 951772, Los Angeles, CA, 90095-1772, USA

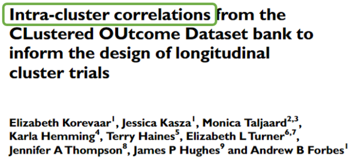


CLINICAL TRIALS WORKSHOP ARTICLE *Clinical Trials* 2005; **2:** 99–107

Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research

Marion K Campbell[a], Peter M Fayers[b] and Jeremy M Grimshaw[c]

The objective of this research was to identify determinants of the magnitude of intracluster correlation coefficients (ICCs) in cluster randomized trials from the field of implementation research. A survey of experts was conducted to generate *a priori* hypotheses of factors that might affect ICC size. Hypotheses were tested on empirical estimates of ICCs calculated from 21 implementation research datasets, mainly from the UK. Effects of setting (primary or secondary care), type of variable (process or outcome), type of measurement (objective or subjective), prevalence of outcome and size of cluster were tested. In total, 220 ICCs were available (range 0 to 0.415). Significant differences in ICC magnitude were found. The ICCs were significantly higher for process than for outcome variables, and for secondary care outcomes compared with primary care outcomes. The effects of prevalence and size were less clear cut. There was no evidence to suggest that type of measurement affected ICC size. In conclusion, accurate estimates of ICCs are essential for sample size calculations for cluster randomized trials of professional behaviour change interventions. This study demonstrates that ICCs are sensitive to a number of trial factors, particularly setting and outcome type. These factors must be considered when planning such cluster randomized trials. *Clinical Trials* 2005; 2: 99–107. www.SCTjournal.com

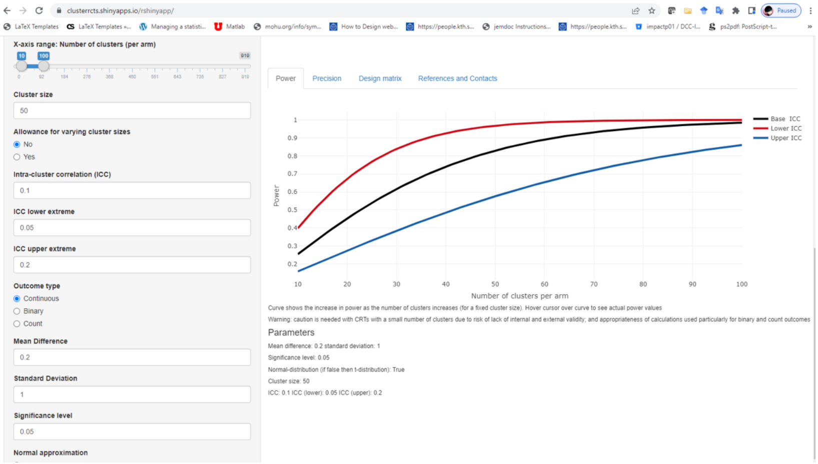


Intra-cluster correlations from the CLustered OUtcome Dataset bank to inform the design of longitudinal cluster trials

Elizabeth Korevaar[1], Jessica Kasza[1], Monica Taljaard[2,3], Karla Hemming[4], Terry Haines[5], Elizabeth L Turner[6,7], Jennifer A Thompson[8], James P Hughes[9] and Andrew B Forbes[1]

The Shiny CRT Calculator[1]

(Hemming et al. 2018 IJE)



X-axis range: Number of clusters (per arm): 10 – 100

Cluster size: 50

Allowance for varying cluster sizes: ● No ○ Yes

Intra-cluster correlation (ICC): 0.1

ICC lower extreme: 0.05

ICC upper extreme: 0.2

Outcome type: ● Continuous ○ Binary ○ Count

Mean Difference: 0.2

Standard Deviation: 1

Significance level: 0.05

Normal approximation

Curve shows the increase in power as the number of clusters increases (for a fixed cluster size). Hover cursor over curve to see actual power values

Warning: caution is needed with CRTs with a small number of clusters due to risk of lack of internal and external validity; and appropriateness of calculations used particularly for binary and count outcomes

Parameters

Mean difference: 0.2 standard deviation: 1

Significance level: 0.05

Normal-distribution (if false then t-distribution): True

Cluster size: 50

ICC: 0.1 ICC (lower): 0.05 ICC (upper): 0.2

[0] URL: `https://clusterrcts.shinyapps.io/rshinyapp/`

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 and other health-equity variables
  - clinical characteristics: baseline value of outcomes
- How to plan such a CRT?
  - address the question of how different the treatment works in different subpopulations?
- What are methods or simple tools like the Shiny CRT that enables convenient sample size & power calculation for heterogeneity of treatment effect (HTE) analysis in a CRT?

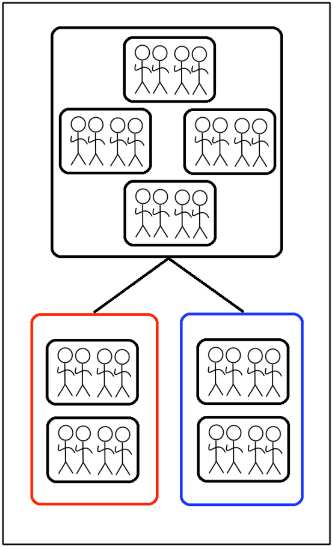
Scope

- We focus on explained treatment effect heterogeneity with measured baseline cluster-level or individual-level covariates
    - in contrast to unexplained treatment effect heterogeneity, such as those modeled by a random treatment effect by cluster
- We focus on confirmatory heterogeneity of treatment effect (HTE) anlayses that are hypothesis-driven with pre-specified effect modifiers
    - sets us apart from exploratory HTE analysis that is mostly data-driven and without pre-specification
- An existing systematic review reported that 16 out of 64 CRTs examined HTE among demographic patient subgroups, but noticed a lack of guidance on HTE for CRTs[2]

[2] Starks MA et al. (2019). Assessing heterogeneity of treatment effect analyses in health-related cluster randomized trials: a systematic review. *PloS one*.

A hypothetical example - cont'd

- 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?



2.1 Demystifying a sample size formula

Testing an overall effect

- 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 $\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2)$ and $\xi_{ij} \sim \mathcal{N}(0, \sigma_\xi^2)$

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

Testing treatment effect difference

- Baseline covariates are collected in CRTs, some of which are effect modifiers 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 $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$
- For binary X_{ij} (race), β_4 encodes difference in treatment effect among white and non-white patients – HTE parameter ($\mathcal{H}_0 : \beta_4 = 0$) – interaction test
- Essentially a linear mixed analysis of covariance (ANCOVA) model

Central question

- **Central question**: Are we able to design CRTs to sufficiently power the interaction test on HTE based on the linear mixed ANCOVA model?
  - what are key design parameters that drive the statistical power for testing $\mathcal{H}_0 : \beta_4 = 0$?
  - interaction test is known to be under-powered in individually randomized trials, but it remains unknown whether those earlier lessons learned can be directly applied to CRTs
  - is there a simple design effect to help us evaluate the power of interaction test in CRTs?

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
- 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
    - 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 $\rho_x = \sigma^2_b / (\sigma^2_b + \sigma^2_c)$.

Covariate ICC

- Empirical evidence of substantial variation in distribution of potential effect modifiers across clusters
- 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

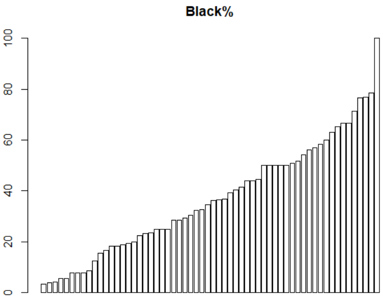


Figure: Variation of % black in the HF-ACTION multi-center trial with 82 sites

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

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

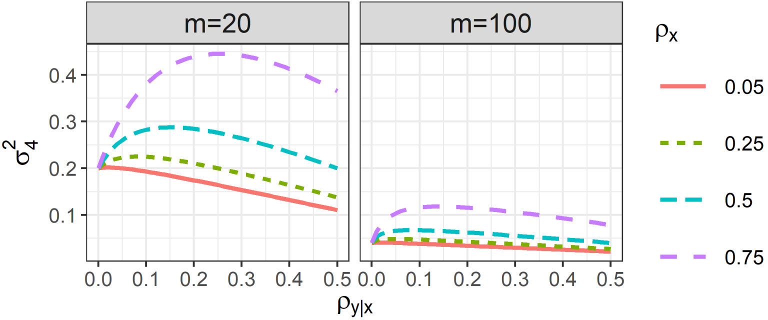
- For design purposes, we derive expression of the HTE estimator, under the linear mixed ANCOVA model[4]

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

- **Interpretation**: variance of HTE estimator in individually randomized trial × 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 studying HTE

Design effect

- The usual design effect in CRTs for studying average treatment effect is unbounded and increases indefinitely with larger m
- $\text{DE}(\infty) = (1 - \rho_{y|x})/(1 - \rho_x)$ 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 $\text{DE}(m) = 1$ 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

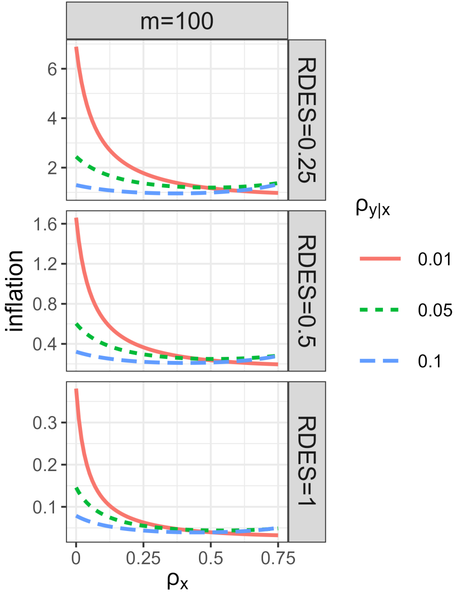
- What if we wish to study effect modification by geographical location or cluster characteristics?
- This is obtained as a special case with $\rho_x = 1$
- Variance of the HTE estimator

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

- DE(m) now looks like our 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*"

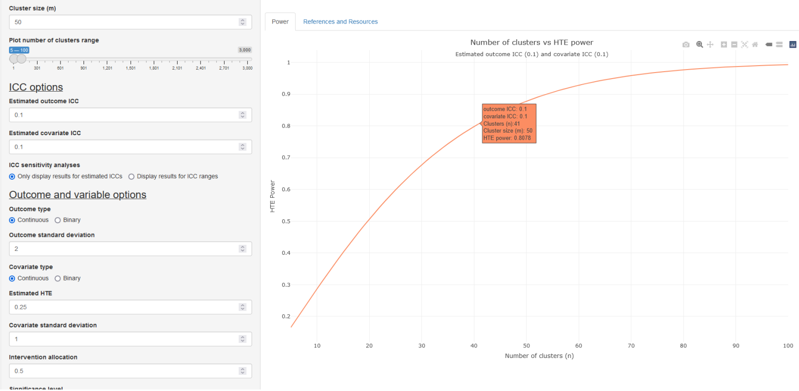


2.2 Software tool and an example

Any tools available?

- The variance expressions are relatively simple to work out the calculations in computer software
  - involve a biostatistician at the design stage
  - “design trumps analysis”
- Our team (led by Mary Ryan, PhD) is currently developing a free R shiny app that implements the above study design calculation
  - previous slides provide a guide to design parameters
  - **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/`
- Still being developed/refined (future software tutorial)

The CRT HTE Calculator[5]



Cluster size (m)

50

Plot number of clusters range

5 — 100 3,000

ICC options

Estimated outcome ICC

0.1

Estimated covariate ICC

0.1

ICC sensitivity analyses

● Only display results for estimated ICCs ○ Display results for ICC ranges

Outcome and variable options

Outcome type

● Continuous ○ Binary

Outcome standard deviation

2

Covariate type

● Continuous ○ Binary

Estimated HTE

0.25

Covariate standard deviation

1

Intervention allocation

0.5

Significance level

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

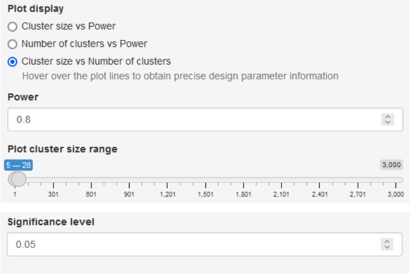
UMDEX

- **Objective**: Obtain the requires sample size for detecting HTE in the context of the design of the Umeå Dementia and Exercise (UMDEX) study[6]
- **Setting**: Two-arm CRT targeting individuals aged 65 or above with a dementia diagnosis, Mini-Mental State Examination (MMSE) score of 10 or greater, and dependence in Activities of Daily Living (ADLs), living in residential care facilities
  - 36 clusters were randomized (defined by the same wing, unit, or floor)
- **Intervention**: High-intensity functional exercise program versus seated control activity
- **Cluster Size**: The average cluster size $\overline{m} = 20$

[6] Toots A et al (2016). Effects of a high-intensity functional exercise program on dependence in activities of daily living and balance in older adults with dementia. *JAGS*

UMDEX

- **Variables**: As an example, focus on Functional Independence Measure (FIM) outcome, and two potential effect modifiers measured at the individual level, level of cognitive impairment (continuous) and dementia type (binary, Alzheimer's versus non-Alzheimer's dementia)
- Consider two-sided tests with nominal 5% type I error rate and 20% type II error rate (80% power)



Plot display

○ Cluster size vs Power

○ Number of clusters vs Power

● Cluster size vs Number of clusters

Hover over the plot lines to obtain precise design parameter information

Power

0.8

Plot cluster size range

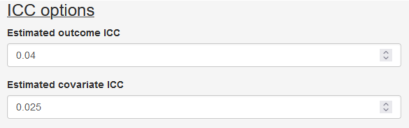
5 — 28

Significance level

0.05

UMDEX

- Effect modification with cognitive impairment level (MMSE)
  - covariate ICC $\rho_x = 0.025$, and the outcome ICC $\rho_{y|x} = 0.04$



ICC options

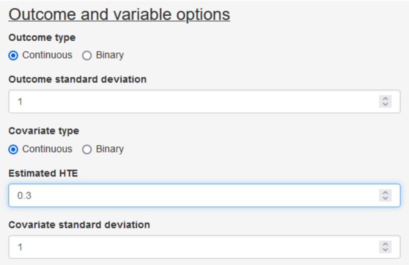
Estimated outcome ICC

0.04

Estimated covariate ICC

0.025

  - standardized HTE effect size, $\delta\sigma_x/\sigma_{y|x} = 0.3$, interpreted as the effect on standard deviation unit increase in covariate on standard deviation unit of the outcome



Outcome and variable options

Outcome type

● Continuous ○ Binary

Outcome standard deviation

1

Covariate type

● Continuous ○ Binary

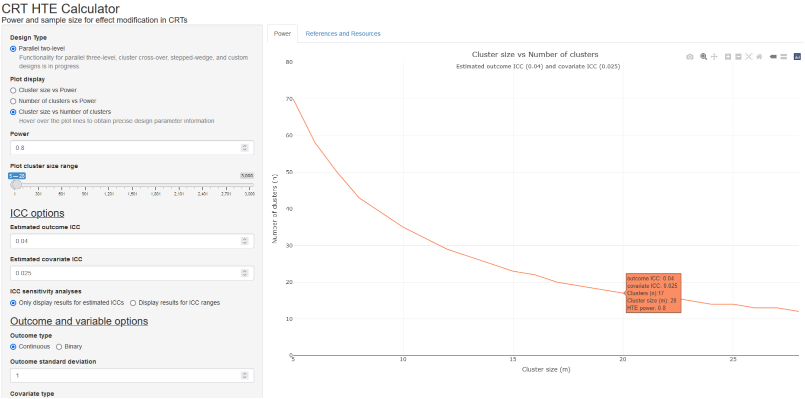
Estimated HTE

0.3

Covariate standard deviation

1

UMDEX



CRT HTE Calculator

Power and sample size for effect modification in CRTs

Design Type

● Parallel two-level

Functionality for parallel three-level, cluster cross-over, stepped-wedge, and custom designs is in progress.

Plot display

○ Cluster size vs Power

○ Number of clusters vs Power

● Cluster size vs Number of clusters

Hover over the plot lines to obtain precise design parameter information

Power

0.8

Plot cluster size range

5 — 28

ICC options

Estimated outcome ICC

0.04

Estimated covariate ICC

0.025

ICC sensitivity analyses

● Only display results for estimated ICCs ○ Display results for ICC ranges

Outcome and variable options

Outcome type

● Continuous ○ Binary

Outcome standard deviation

1

Covariate type

- Require $n = 17$ clusters

UMDEX

- Effect modification with dementia type (AD versus other)
  - marginal prevalence and the standard deviation of dementia type is 0.36 and 0.48
  - covariate ICC $\rho_x = 0.05$, and the outcome ICC $\rho_{y|x} = 0.04$
  - standardized HTE effect size, $\delta/\sigma_{y|x} = 0.5$, interpreted as the effect from change in dementia type on the standard deviation unit of the outcome
- Require $n = 27$ clusters

Sensitivity Analysis

		HTE (MMSE)		HTE (Dementia type)	
		cluster size		cluster size	
$\rho_{y\|x}$	ρ_x	10	20	10	20
0.01	0.01	35	17	55	27
	0.025	35	17	55	27
	0.05	35	18	55	27
	0.1	35	18	55	28
	0.2	35	18	55	28
0.04	0.01	35	17	54	27
	0.025	35	$\boxed{\mathbf{17}}$	54	27
	0.05	35	18	55	$\boxed{\mathbf{27}}$
	0.1	35	18	55	28
	0.2	36	19	57	29
0.1	0.01	33	16	52	26
	0.025	34	17	52	26
	0.05	34	17	53	26
	0.1	35	17	55	27
	0.2	37	19	58	29

- Varying key design parameters

3. Additional considerations

Unequal cluster sizes

- Equal cluster sizes m can be a strong assumption
- The impact of unequal cluster sizes on power has been studied for testing the average treatment effect in parallel CRTs
- Rule of thumb:
  - "*loss of efficiency due to variation of cluster sizes rarely exceeds 10 per cent and can be compensated by sampling 11 per cent more clusters*"[7]
- An explicit **correction factor** has been derived to quantify the variance inflation (depends on mean and coefficient of variation of cluster sizes, $\overline{m}$ and CV)

[7] van Breukelen GJ, Candel MJ, Berger MP (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*

Impact of cluster size variability

We are able to characterize a suitable correction factor for testing HTE due to unequal cluster sizes[8]

$$\underbrace{\left[1 - \mathrm{CV}^2 \frac{\overline{m}\rho_{y|x}(1-\rho_{y|x})(\rho_x - \rho_{y|x})}{\{1 + (\overline{m}-2)\rho_{y|x} - (\overline{m}-1)\rho_x\rho_{y|x}\}\{1 + (\overline{m}-1)\rho_{y|x}\}^2}\right]^{-1}}_{\text{Correction Factor } \theta_1(\mathrm{CV})}$$

- $\lim_{\overline{m}\to\infty} \theta_1(\mathrm{CV}) = 1$
- Given the CV rarely exceed one, when the average cluster size is not too small (e.g., < 20), unequal cluster sizes should have close to no impact on power for the HTE test with an individual-level effect modifier → smaller impact than studying ATE

[8] Tong G, Esserman DA, Li F (2022). Accounting for unequal cluster sizes in designing cluster randomized trials to detect treatment effect heterogeneity. *Statistics in Medicine*

Impact of cluster size variability - cont'd

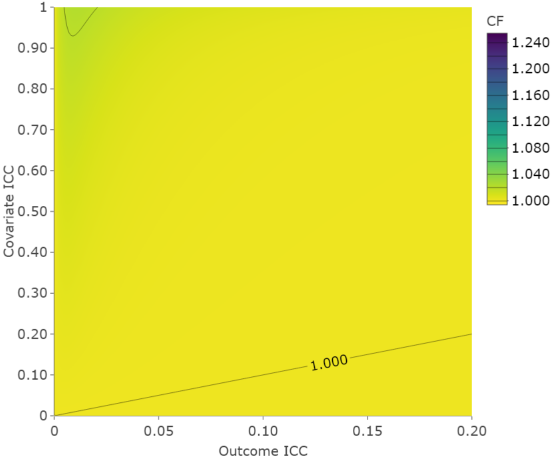
If we have a cluster-level effect modifier ($\rho_x = 1$), the correction factor becomes

$$\underbrace{\left[1 - \mathrm{CV}^2 \frac{\overline{m}\rho_{y|x}(1 - \rho_{y|x})}{\{1 + (\overline{m} - 1)\rho_{y|x}\}^2}\right]^{-1}}_{\text{Correction Factor } \theta_2(\mathrm{CV})}$$

- this is identical to the one derived in van Breukelen et al., (2007), except that we are using an adjusted outcome-ICC $\rho_{y|x}$
- power for studying cluster-level effect moderation more sensitive to cluster size variation

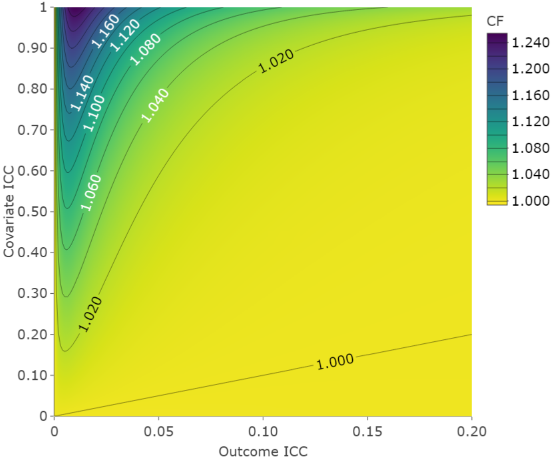
Visualizing correction factor

- Plotting Correction Factor (CF) with $\overline{m} = 100$
- Assuming a mild case with CV = 0.3
- CF is close to one
- Close to no impact of cluster size variation on power



Visualizing correction factor - cont'd

- Plotting Correction Factor (CF) with $\overline{m} = 100$
- Assuming an extreme case with CV = 0.9
- CF is close to one except when outcome ICC ($\rho_{y|x}$) is close to zero and covariate ICC (ρ_x) close to one
- Often adequate to assume equal cluster size



Extension to non-continuous outcomes

- Many CRTs assess binary (yes/no) outcomes
    - variance function of the outcome is an explicit function of the mean
- Effect measure of interest may be on the ratio scale (such as risk ratio or odds ratio)
- We have developed new methods for determining sample size and power for testing HTE in CRTs with non-continuous outcomes[9]

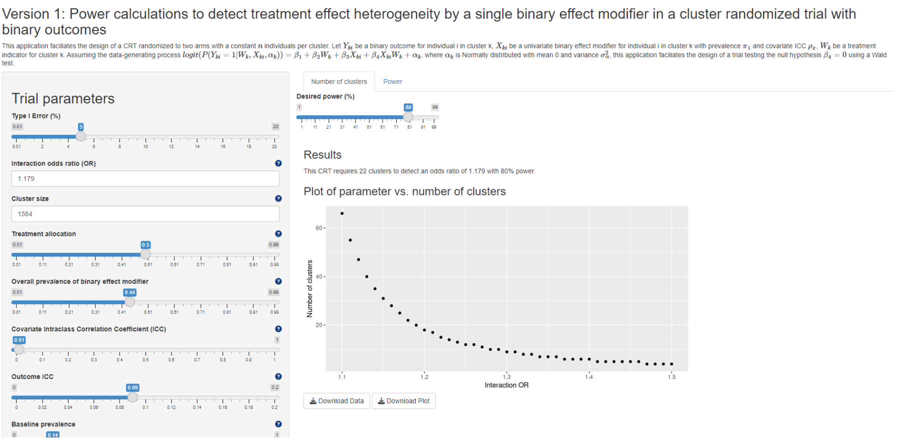
Outcome type	Effect measure	Dispersion	Variance	Link
continuous	mean difference	σ_ϵ^2	1	μ
binary	risk difference	1	$\mu(1-\mu)$	μ
binary	risk ratio	1	$\mu(1-\mu)$	$\log(\mu)$
binary	odds ratio	1	$\mu(1-\mu)$	$\log\left(\mu/\{1-\mu\}\right)$
count	rate difference	1	μ	μ
count	rate ratio	1	μ	$\log(\mu)$

[9] Maleyeff L, Wang R, Haneuse S, Li F (2023+). Sample size requirements for testing treatment effect heterogeneity in cluster randomized trials with binary outcomes. *Submitted*

Initial version of Shiny calculator (binary)[10]

(Maleyeff et al. 2023+)

.



Version 1: Power calculations to detect treatment effect heterogeneity by a single binary effect modifier in a cluster randomized trial with binary outcomes

This application faciliates the design of a CRT randomized to two arms with a constant n individuals per cluster. Let Y_{ki} be a binary outcome for individual i in cluster k, X_{ki} be a univariate binary effect modifier for individual i in cluster k with prevalence π_1 and covariate ICC ρ_x, W_k be a treatment indicator for cluster k. Assuming the data-generating process $logit(P(Y_{ki}=1|W_k, X_{ki}, \alpha_k)) = \beta_1 + \beta_2 W_k + \beta_3 X_{ki} + \beta_4 X_{ki} W_k + \alpha_k$, where α_k is Normally distributed with mean 0 and variance σ^2_α, this application faciliates the design of a trial testing the null hypothesis $\beta_4 = 0$ using a Wald test.

Trial parameters

Type I Error (%): 5

Interaction odds ratio (OR): 1.179

Cluster size: 1584

Treatment allocation: 0.5

Overall prevalence of binary effect modifier: 0.44

Covariate Intraclass Correlation Coefficient (ICC): 0.01

Outcome ICC: 0.09

Baseline prevalence: 0.14

Desired power (%): 80

Results

This CRT requires 22 clusters to detect an odds ratio of 1.179 with 80% power.

Plot of parameter vs. number of clusters

[10] URL: `https://laramaleyeff1.shinyapps.io/sample_size/`

Other cluster randomized designs?

Design	Additional questions to address
Individually randomized group treatment trials[11]	(1) arm-specific ICC (2) between-arm heterogeneity in variance (3) no covariate ICC
Multilevel cluster randomized trials[12]	(1) within- and between-subcluster ICC (outcome) (2) within- and between-subcluster ICC (covariate) (3) level of randomization
Multi-period (Stepped wedge) cluster randomized trials	(1) within- and between-period ICC (outcome) (2) within- and between-period ICC (covariate) (3) sampling design

- Ongoing efforts in developing these methods and final version of R shiny software will include all these designs

[11] Tong G, Taljaard M, Li F (2023+). Sample size considerations for assessing treatment effect heterogeneity in randomized trials with heterogeneous intracluster correlations and variances. *Submitted*.

[12] Li F, et al. (2022). Designing three-level cluster randomized trials to assess treatment effect heterogeneity. *Biostatistics*.

4. Discussion

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
- A tool to provide a context to interpret findings
  - the what-if question?

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[13]
  - URL: `https://mary-ryan.shinyapps.io/HTE-MMD-app/`
- Design & Statistics Core + Technical Data Core (IMPACT Collaboratory) reporting such estimates in ongoing work

[13] Ryan M, Esserman DA, Li F (2023+). Maximin optimal cluster randomized designs to detect treatment effect heterogeneity. *Submitted*.

Final consideration

- In many cases, a binary effect modifier is of interest
- We acknowledge our current focus on sample size requirements for testing the difference between subgroup average treatment effects, rather than those for testing the subgroup average treatment effects
  - question 1: does intervention work in a specific subpopulation
  - question 2: whether intervention works differently between subpopulations (the heterogeneity question)
- Addressing question 1 is an ongoing efforts
  - in principle requires a larger subgroup sample size
  - insight is, variance of subgroup average treatment effect estimator is a weighted combination of that of the overall effect estimator and that of the interaction effect estimator
  - weight depends on subgroup proportion

Thank You!