RECENT ADVANCES IN ADDRESSING DESIGN AND ANALYSIS CHALLENGES OF CLUSTER RANDOMIZED TRIALS

Ed Mascha, Cleveland Clinic (chair/moderator/organizer/speaker)
Hrishikesh Chakraborty, Duke University
Emine Bayman, University of Iowa
Fan Li, Yale University

46th ANNUAL MEETING

May 18-21, 2025

VANCOUVER
CANADA

SCT
SOCIETY FOR CLINICAL TRIALS

RECENT ADVANCES IN ADDRESSING DESIGN AND ANALYSIS CHALLENGES OF CLUSTER RANDOMIZED TRIALS

- **Ed Mascha:** "Salient Design and Sample Size Features of Cluster-Randomized Crossover (CRXO) Trials"
- **Hrishikesh Chakraborty**: "Power Comparison for Different Hierarchical Composite Endpoints in Cluster Randomized Trials"
- **Emine Bayman**: "Interim Reassessment of ICC in a Cluster-Randomized Trial"
- **Fan Li :** "Model-robust standardization in cluster-randomized trials"

Format: 15 min talk + 5 min discussion

Salient Design and Sample Size Features of Cluster-Randomized Crossover (CRXO) Trials

Edward J Mascha, PhD
Head, Section of Biostatistics
Department of Quantitative Sciences
Cleveland Clinic

46th ANNUAL MEETING

May 18-21, 2025

VANCOUVER
CANADA

SCT
SOCIETY FOR CLINICAL TRIALS

Disclosures

- No Disclosures to declare

Today

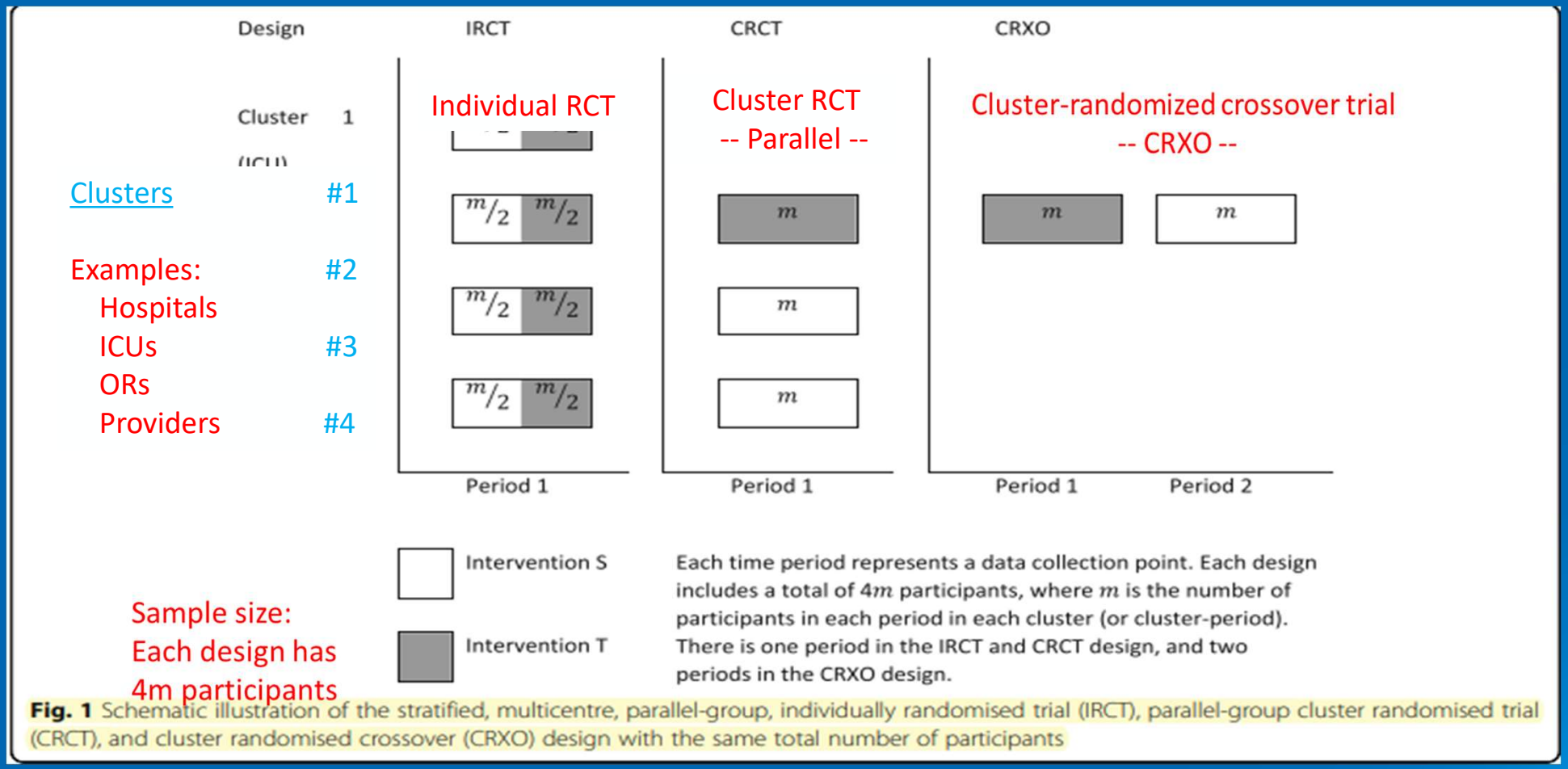
- Cluster randomized trial designs-- overview
- Cluster-randomized multiple crossover design -- CRXO
- Randomization, Analysis, Sample size calculations
- Examples
- Conclusions

Cluster Randomized Trials -- overview

SCT | 46TH ANNUAL MEETING

Hemming (2020) International Journal of Epidemiology, 2020, 979–995
doi: 10.1093/ije/dyz237

Cluster Randomized Trials -- CRXO



Sample size:
Each design has
4m participants

Fig. 1 Schematic illustration of the stratified, multicentre, parallel-group, individually randomised trial (IRCT), parallel-group cluster randomised trial (CRCT), and cluster randomised crossover (CRXO) design with the same total number of participants

Cluster-randomized crossover (CRXO) design

Combines features of:

1) **Parallel cluster randomized trial**

-- *clusters* are randomized to an *arm* (not patients as in IRCT)
-- treat cluster as random effect; adjust for ICC in sample size calcs

2) **Individual patient crossover** trial, and **stepped wedge**

-- gain power with multiple measurements (on same or new units)
-- randomize *clusters to sequence* (test tx-by-sequence interaction)
-- adjust for period effect (test tx-by-period interaction… differential carryover)
-- treat *period within cluster* (or cluster x period interaction) as random effect

CRXO Design: when appropriate?

When intervention *can* be *turned on and off*: Not an irreversible systemic change
-- if intervention *cannot* easily be turned off, stepped wedge may be good option

When patient outcome can be measured *before* start of *next* period (short-term outcomes)

Common CRXO designs:

1) Cross-sectional design [new units each period]
2) Cohort design [repeated measurements on same units across periods]
3) Combination

CRXO Design: How to randomize clusters?

Do not simply randomize clusters in first period, then alternate interventions each period.
Instead, ***randomize each cluster to a sequence*****.**
Optimal scheme: e.g., every 4 periods, randomize each cluster to 1 of these patterns:

ABBA
BAAB
AABB
BBAA

CRXO Design: How to randomize?

Do not simply randomize clusters in first period, then 'alternate' interventions each period.

Use optimal scheme: e.g., every 4 periods, randomize each cluster to 1 of these sequences:

ABBA
BAAB
AABB
BBAA

This design is:

1) **uniform within sequence**: each treatment appears same number of times within each sequence
2) **uniform within period**: each treatment appears same number of times within each period
3) **strongly balanced with respect to first-order carryover effects** because each treatment precedes every other treatment, *including itself*, the same number of times.

References: Grantham et al (2019). How many times should a cluster randomized crossover trial cross over? *Stat Med* 38 (25).

Reed (2012). Four Period Crossover Designs. *Journal of Modern Applied Statistical Methods* 11(1).

CRXO Design: How to randomize?

Do not simply randomize clusters in first period, then 'alternate' interventions each period.

Use optimal scheme: e.g., every 4 periods, randomize each cluster to 1 of these sequences:

ABBA
BAAB
AABB
BBAA
ABAB
BABA

This design is:

1) **uniform within sequence**: each treatment appears same number of times within each sequence
2) **uniform within period**: each treatment appears same number of times within each period
3) **strongly balanced with respect to first-order carryover effects** because each treatment precedes every other treatment, *including itself*, the same number of times.

References: Grantham et al (2019). How many times should a cluster randomized crossover trial cross over? *Stat Med* 38 (25).

Reed (2012). Four Period Crossover Designs. *Journal of Modern Applied Statistical Methods* 11(1).

Analysis methods for CRXO design

Typically, a generalized linear mixed effects model.

Fixed effects: treatment, period, sequence.

Random effects: cluster, period (cluster), patient [e.g, if cohort design]

Sample Size Calcs for CRXO

A CRXO Sample Size Calculation Depends on two intracluster correlations:

within-cluster within-period correlation: WPC (or *ρ)* and
within-cluster between-period correlation: BPC (or *η*).

Sample Size Calcs for CRXO

A CRXO Sample Size Calculation Depends on two intracluster correlations:

within-cluster within-period correlation: WPC (or *ρ*) and
within-cluster between-period correlation: BPC (or *η*).

The ratio of BPC/WPC, called the **Cluster-Autocorrelation Coefficient (CAC), also has a strong effect on the sample size, as we'll see.**

Note: Today we'll consider the case where the BPC is a constant across periods.
Especially in studies with many crossovers, that may not be true –
i.e., the BPC, and thus the CAC, may decrease as distance between periods increases.

Hemming et al (IJE 2020)

Sample Size Calcs for CRXO

$\sigma_C^2 =$ variance between time periods (within *clusters*)
$\sigma_T^2 =$ variance between subjects (within *time* period and *cluster*)
$\sigma_e^2 =$ variance between clusters

- Then define WPC and BPC in terms of variance components:

$$\boldsymbol{WPC} = \frac{\sigma_C^2+\sigma_T^2}{\sigma_C^2+\sigma_e^2+\sigma_T^2} \qquad \boldsymbol{BPC} = \frac{\sigma_C^2}{\sigma_C^2+\sigma_e^2+\sigma_T^2}$$

Martin (Trials, 2016)

$$\boldsymbol{WCC} = BPC + \frac{1}{n_{tp}}(WPC - BPC)$$

where n_{tp} is number of time periods in study.

Sample size calculations – binary endpoint

Apply appropriate design effect to the usual sample size for a chi-square test.
We know that 1 + (*m*-1) *ρ* is design effect for parallel CRT, where m is cluster size and *ρ* = ICC

For CRXO, redefine *m*, *ρ* = WPC; add *η* = BPC.

$$N_{CRXO} = 2(z_{\alpha/2} + z_{1-\beta})^2 \frac{p_A(1-p_A) + p_B(1-p_B)}{(p_A - p_B)^2}(1 + (m-1)\rho - m\eta), \qquad (1)$$

where $z_{\alpha/2}$ and z_{β} are the standard normal values for upper tail probabilities of $\alpha/2$ and β;
α is significance level; $1 - \beta$ is power to detect $(p_A - p_B)$; p_A, p_B proportions with outcome in each arm;

m is the number of participants per cluster-period
ρ is the within-cluster within-period correlation (WPC)
η is the within-cluster between-period correlation (BPC)

Total N = s * t * k * m
[s= # arms, t= #periods, k= #clusters, m= cluster size/period]

Sample size calculations – binary endpoint

Apply an appropriate design effect to the usual sample size for a chi-square test.

We know that $1 + (m\text{-}1)\rho$ is design effect for parallel CRT, where *m* is cluster size and ρ = ICC

For CRXO, redefine *m*, *ρ = WPC;* *add* **η = BPC.**

$$N_{CRXO} = 2(z_{\alpha/2} + z_{1-\beta})^2 \frac{p_A(1-p_A) + p_B(1-p_B)}{(p_A - p_B)^2}(1 + (m-1)\rho - m\eta), \qquad (1)$$

where $z_{\alpha/2}$ and z_{β} are the standard normal values for upper tail probabilities of $\alpha/2$ and β;
α is significance level; $1 - \beta$ is power to detect $(p_A - p_B)$; p_A, p_B proportions with outcome in each arm;

m is the number of participants per cluster-period

ρ is the within-cluster within-period correlation (WPC)

η is the within-cluster between-period correlation (BPC)

NOTE: if BPC (η) = 0, design effect = $1 + (m-1)$WPC, **same as CRT**

if BPC = WPC, design effect = $1 -$ WPC, **so N is ≤ IRTC**

Ex 1: ORACLE CRXO – air filter trial

The ORACLE Cluster-Randomized Multiple Crossover Trial (Cleveland Clinic)

- **Intervention**: Operating Room Supplemental Air Filtration and Sterilization versus sham filter (off)
  Outcome: Binary composite of serious **surgical site infections**, complications, death within 30 days post-op

Clusters: 61 Operating rooms
Time periods: 36 1-month periods

Randomization scheme: randomize ORs to 4-month sequences: AABB, BBAA, ABBA, or BAAB

Assumptions: WPC = .10 BPC = 0.08
P0 = .065, P1 = .0585 (RR = 0.90)

Maximum Required Sample size: 109,487

Internal pilot study: re-assess "nuisance parameters" of control group incidence, BPC, WPC

Interim analyses: 4 planned using group sequential design [gamma -4 efficacy, -2 futility]
-- trial crossed futility boundary at 3rd interim

As BPC -> WPC, N decreases *below* that of IRCT (row 2)
As BPC -> ZERO, N increases to that of CRT (row 7)

P0=.065 p1=.0585, 36 periods, m=50 per cluster-period

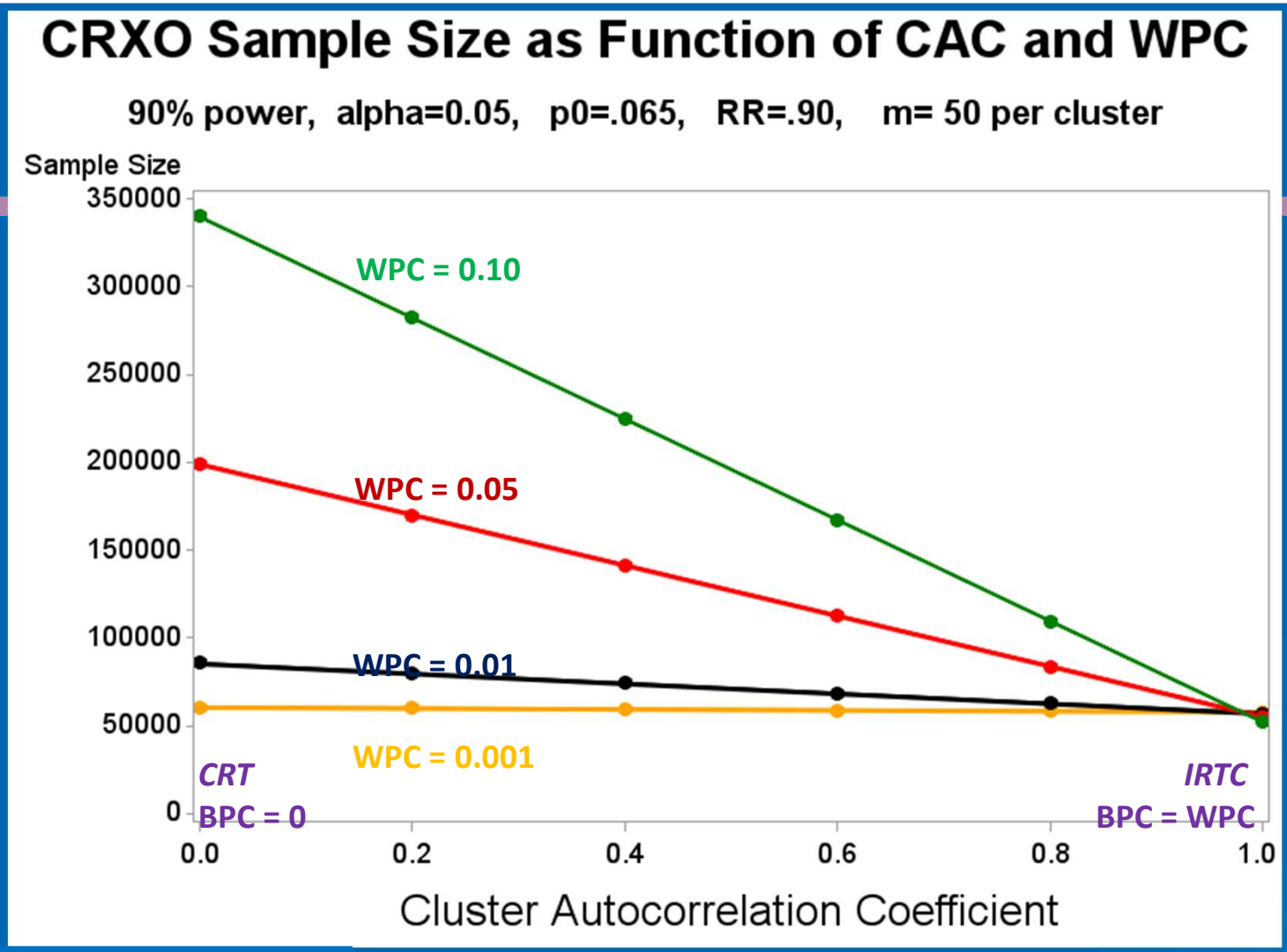
Row	WPC	BPC	CAC= BPC / WPC	N_IRCT	N_CRT	N_CRXO	N_clusters
1	.0	.00	.	57,625	57,626	57,625	33
2	.10	.10	1.0	57,625	339,988	51,862	29
3	**.10**	**.08**	**.80**			**109,487**	**61**
4	.10	.06	.60			167,111	93
5	.10	.04	.40			224,735	125
6	.10	.02	.20			282,360	157
7	.10	.00	.00	57,625	339,988	339,984	189

CRXO SAMPLE SIZE AS FUNCTION OF CAC =BPC/WPC

CAC = Cluster Autocorrelation Cofficient = BPC / WPC

Has huge effect on sample size, independent of actual WPC

CRXO Sample Size as Function of CAC and WPC

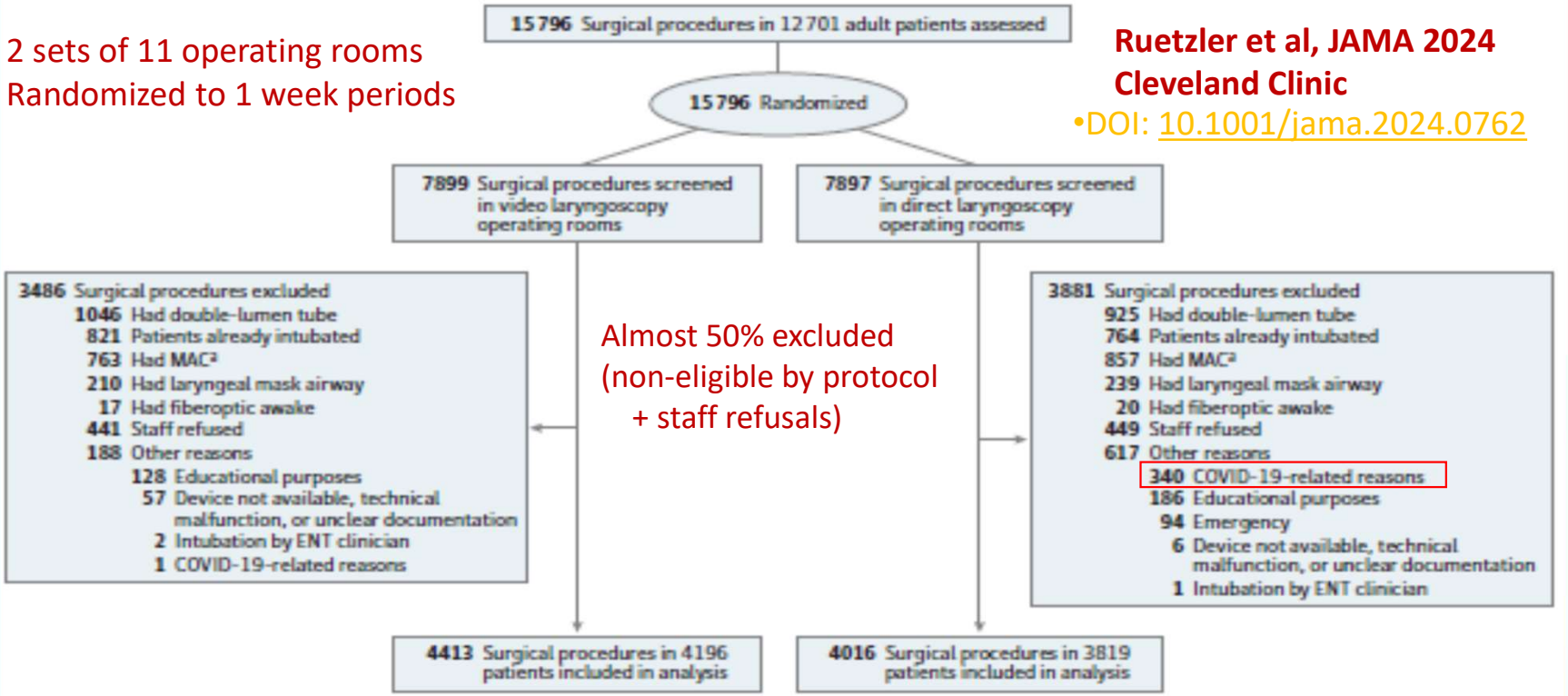


Ex 2. Video vs Direct Laryngoscopy for Endotracheal Intubation in Operating Room: A Cluster Randomized Multiple Crossover Clinical Trial

2 sets of 11 operating rooms
Randomized to 1 week periods

Ruetzler et al, JAMA 2024
Cleveland Clinic
•DOI: 10.1001/jama.2024.0762

Figure. Study Flow Diagram



ENT indicates ears, nose, and throat; MAC, monitored anesthesia care.

[a] MAC is a type of anesthesia service in which an anesthesia clinician continually monitors and supports the patient's vital functions; diagnoses and treats clinical problems that occur; administers sedative, anxiolytic, or analgesic medications if needed; and converts to general anesthesia if required.

SCT

Adjust for confounding – as warranted

Table 1. Patient and Procedure Characteristics (N = 8429)

Characteristic	Unadjusted			Adjusted[a]		
	Video (n = 4413)	Direct (n = 4016)	ASD[a]	Video (n = 4419)	Direct (n = 4005)	ASD[a]
Age, mean (SD), y	63 (13)	63 (14)	0.017	63 (13)	63 (14)	0.000
Height, mean (SD), cm	173 (10)	173 (10)	0.013	173 (10)	173 (10)	0.003
Weight, mean (SD), kg	86 (21) [4412]	84 (20) [4015]	0.124	85 (21)	85 (20)	0.016
BMI, mean (SD)	29 (6.5) [4412]	28 (6.0) [4015]	0.130	28 (6.3)	28 (6.1)	0.019
Sex, No. (%)						
Male	2880 (65)	2599 (65)	0.011	2870 (35)	2615 (65)	0.007
Female	1533 (35)	1417 (35)		1549 (35)	1390 (35)	
Admission category, No. (%)						
Elective	3711 (84)	3424 (85)	0.032	3744 (85)	3392 (85)	0.001
Emergency	702 (16)	592 (15)		675 (15)	613 (15)	
Surgical procedure type, No. (%)						
Cardiac	3317 (75)	2934 (73)	0.051	3279 (74)	2977 (74)	0.008
Vascular	559 (13)	560 (14)		581 (13)	530 (13)	
Other	474 (11)	453 (11)		489 (11)	436 (11)	
Thoracic	63 (1.4)	69 (1.7)		70 (1.6)	62 (1.5)	
Level of first intubation clinician, No. (%)[c]						
CRNA	1672 (38)	1611 (40)	0.070	1724 (40)	1558 (39)	0.008
Resident	1357 (31)	1157 (29)		1314 (30)	1198 (30)	
Fellow	603 (14)	582 (15)		622 (14)	557 (14)	
SRNA	639 (15)	536 (13)		616 (14)	563 (14)	
Attending anesthesiologist	113 (2.6)	112 (2.8)		119 (2.7)	106 (2.6)	
Medical student	29 (0.66)	18 (0.45)		24 (0.55)	22 (0.56)	
ASA physical status, No. (%)[c]						
1 (Healthy)	6 (0.14)	14 (0.35)	0.049	9 (0.20)	10 (0.24)	0.002
2 (Mild systemic illness)	74 (1.7)	94 (2.3)		89 (2.0)	79 (2.0)	
3 (Severe systemic illness)	694 (16)	668 (17)		715 (16)	644 (16)	
4 (Life-threatening systemic illness)	3605 (82)	3211 (80)		3574 (81)	3244 (81)	
5 (Not expected to survive without the operation)	34 (0.77)	29 (0.72)		33 (0.75)	29 (0.73)	
Modified Mallampati score, No./total No. (%)[a]						
I (Soft palate, uvula, pillars visible)	1078/4385 (25)	1100/3992 (28)	0.122	1150 (26)	1039 (26)	0.000
II (Soft palate, major part of uvula visible)	2583/4385 (59)	2406/3992 (60)		2610 (60)	2374 (60)	
III (Soft palate, base of uvula visible)	660/4385 (15)	453/3992 (11)		584 (13)	528 (13)	
IV (Only hard palate visible)	64/4385 (1.5)	24/3992 (0.60)		47 (1.1)	39 (0.99)	
History of difficult intubation within past 5 y, No. (%)	138 (3.1)	112 (2.8)	0.020	131 (3.0)	119 (3.0)	0.001

"direct" laryngoscopy providers sometimes insisted on crossing over to "video" group

-- during COVID the Video intervention thought safer

IPTW (stabilized) to attain balance.

Adjust for confounding – as warranted

Table 1. Patient and Procedure Characteristics (N = 8429)

Characteristic	Unadjusted			Adjusted[a]		
	Video (n = 4413)	Direct (n = 4016)	ASD[b]	Video (n = 4419)	Direct (n = 4005)	ASD[b]
Age, mean (SD), y	63 (13)	63 (14)	0.017	63 (13)	63 (14)	0.000
Height, mean (SD), cm	173 (10)	173 (10)	0.013	173 (10)	173 (10)	0.003
Weight, mean (SD), kg	86 (21) [4412]	84 (20) [4015]	0.124	85 (21)	85 (20)	0.016
BMI, mean (SD)	29 (6.5) [4412]	28 (6.0) [4015]	0.130	28 (6.3)	28 (6.1)	0.019
Sex, No. (%)						
Male	2880 (65)	2599 (65)	0.011	2870 (35)	2615 (65)	0.007
Female	1533 (35)	1417 (35)		1549 (35)	1390 (35)	
Admission category, No. (%)						
Elective	3711 (84)	3424 (85)	0.032	3744 (85)	3392 (85)	0.001
Emergency	702 (16)	592 (15)		675 (15)	613 (15)	
Surgical procedure type, No. (%)						
Cardiac	3317 (75)	2934 (73)	0.051	3279 (74)	2977 (74)	0.008
Vascular	559 (13)	560 (14)		581 (13)	530 (13)	
Other	474 (11)	453 (11)		489 (11)	436 (11)	
Thoracic	63 (1.4)	69 (1.7)		70 (1.6)	62 (1.5)	
Level of first intubation clinician, No. (%)[c]						
CRNA	1672 (38)	1611 (40)	0.070	1724 (40)	1558 (39)	0.008
Resident	1357 (31)	1157 (29)		1314 (30)	1198 (30)	
Fellow	603 (14)	582 (15)		622 (14)	557 (14)	
SRNA	639 (15)	536 (13)		616 (14)	563 (14)	
Attending anesthesiologist	113 (2.6)	112 (2.8)		119 (2.7)	106 (2.6)	
Medical student	29 (0.66)	18 (0.45)		24 (0.55)	22 (0.56)	
ASA physical status, No. (%)[c]						
1 (Healthy)	6 (0.14)	14 (0.35)	0.049	9 (0.20)	10 (0.24)	0.002
2 (Mild systemic illness)	74 (1.7)	94 (2.3)		89 (2.0)	79 (2.0)	
3 (Severe systemic illness)	694 (16)	668 (17)		715 (16)	644 (16)	
4 (Life-threatening systemic illness)	3605 (82)	3211 (80)		3574 (81)	3244 (81)	
5 (Not expected to survive without the operation)	34 (0.77)	29 (0.72)		33 (0.75)	29 (0.73)	
Modified Mallampati score, No./total No. (%)[d]						
I (Soft palate, uvula, pillars visible)	1078/4385 (25)	1100/3992 (28)	0.122	1150 (26)	1039 (26)	0.000
II (Soft palate, major part of uvula visible)	2583/4385 (59)	2406/3992 (60)		2610 (60)	2374 (60)	
III (Soft palate, base of uvula visible)	660/4385 (15)	453/3992 (11)		584 (13)	528 (13)	
IV (Only hard palate visible)	64/4385 (1.5)	24/3992 (0.60)		47 (1.1)	39 (0.99)	
History of difficult intubation within past 5 y, No. (%)	138 (3.1)	112 (2.8)	0.020	131 (3.0)	119 (3.0)	0.001

"direct" laryngoscopy providers sometimes insisted on crossing over to "video" group

-- during COVID the Video intervention thought safer

IPTW (stabilized) to attain balance.

Conclusions for CRXO Designs

CRXO especially useful when:

1) Intervention can be turned on and off, and fairly easily
2) Outcomes measurable short-term (within the period)
3) Low potential for recruitment bias – post-randomization

Efficient – Power Increases with:

1) Decreasing WPC (within-cluster within-period correlation), and
2) Increasing Cluster Autocorrelation (BPC / WPC)
3) More powerful than parallel CRT as long as BPC > 0

Special Features

* Reduced logistical complexity over IRCTs (waive consent if ethically appropriate)
* Finding similar clusters or matching clusters not necessary – making within-cluster comparisons
* Reduced cost vs IRCT
* Provider compliance can be an issue

BPC = within-cluster between-period correlation

References

Arnup SJ, McKenzie JE, Hemming K, Pilcher D, Forbes AB. Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial. Trials. 2017. doi: 10.1186/s13063-017-2113-2.

Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McKenzie JE. The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality. Trials 2016;17:575. doi:10.1186/s13063-016-1685-6.

Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. Int J Epidemiol. 2020. doi: 10.1093/ije/dyz237.

Hemming K, Taljaard M, Weijer C, Forbes A B. Use of multiple period, cluster randomised, crossover trial designs for comparative effectiveness research *BMJ* 2020; 371 :m3800 doi:10.1136/bmj.m3800

Li F, Forbes AB, Turner EL, Preisser JS. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. *Statistics in Medicine*. 2019; 38: 636–649. **https://doi.org/10.1002/sim.7995**

Martin, J., Girling, A., Nirantharakumar, K. *et al.* Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. *Trials* 17, 402 (2016).

Parienti JJ, Kuss O. Cluster-crossover design: a method for limiting clusters level effect in community-intervention studies. Contemp Clin Trials. 2007;28(3):316–23.

References (2)

Byrne JE, Rodriguez-Patarroyo FA, Mascha EJ, Han Y, Bravo M, Bloomfield MR, Rao SM, Sessler DI. Cooling vest improves surgeons' thermal comfort without affecting cognitive performance: a randomised cross-over trial. Occup Environ Med. 2023 Jun;80(6):339-345. doi: 10.1136/oemed-2022-108457. Epub 2023 May 4. PMID: 37142418; PMCID: PMC10313985.

Reed (2012). “Higher-order crossover designs are useful because they allow a treatment effect to be estimated even in the presence of a carry-over effect, they provide estimates of intra-subject variability and they draw inference on the carryover effect (Chow & Lu, 1992)”

Discussion

CRXO : compared to other designs

Characteristic	Individual RCT (IRCT)	CRT	CRXO	Stepped Wedge (SW-CRT)
Balance at baseline?	**Expected** with randomization	**Needs many clusters,** matched/similar pairs	Crossover **avoids** issue - w/in cluster comps - no need to match	**Similar to CRXO**, but interventions **ratio not constant** across periods
Sample size	CRXO larger if BPC/WPC < 1	CRXO smaller if BPC > 0 .	IRCT < CRXO < CRT	Only 1 crossover/cluster so **CRXO has more info** in same time frame
Generalizability	If multicenter, very strong	Assumes tx, control clusters similar	Strong if heterogeneity across clusters	Strong if heterogeneity across clusters
Enrollment time	Slow since no crossover	Slow since no crossover	Fast due to crossover	Faster due to crossover. **Slower than CRXO** if long subject observation time
Feasibility/ logistics	Need to consent each patient	Challenging – clusters ramp up same time	Make sure measuring provider compliance since many switches	Easier: One-by-one ramp up for clusters

Power Comparison for Different Hierarchical Composite Endpoint in Cluster Randomized Trial

Hrishikesh Chakraborty, DrPH

Department of Biostatistics and Bioinformatics
Duke Clinical research Institute (DCRI)
Duke University

Acknowledgments: Ying Sun and Dr. Roland Matsouaka

Overview

1. Cluster Randomized Trial and ICC
2. Hierarchical Composite Outcome – Rank-based methods
3. Example: HiLo Trial
4. Rank-based Methods : Cluster Trial
5. Simulation and Results

Cluster Randomized Trial and ICC

Cluster Randomized Trial

Cluster Randomized Trial (CRT) randomize groups such as clinicians, families, medical practices, schools and communities rather than individuals.

- In a simple randomized trial, an individual is the unit of randomization, but in a CRT, a group is the unit of randomization

Advantages:

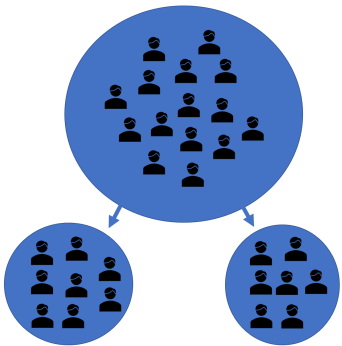
- Clustering is often used for practical reasons
- Used to avoid contamination
- Less costly

Disadvantages:

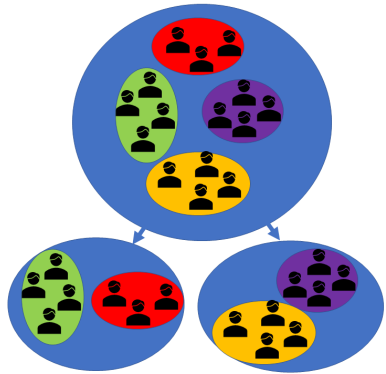
- Need to recruit more study participants
- Sample size calculation and analyses are more complicated

Cluster-Randomized Trial

Simple Randomized Trial



Cluster-Randomized Trial



Difficulties of Sample Size Calculation for CRT

Two major difficulties arise in a sample size calculation for clustered studies.

- The number of units in each cluster, called cluster size, tends to vary with a certain distribution
- The observations within each cluster are correlated

It is well known that the application of standard sample size approaches to cluster randomization designs may lead to

- Seriously under-powered studies
- Spurious statistical significance

Intra-cluster Correlation Coefficient (ICC)

The ICC, $\rho(0 \leq \rho \leq 1)$, is a measure of the relatedness of clustered data. It accounts for the relatedness of clustered data by comparing the variance within clusters with the variance between clusters.

- $\rho = 1$, all responses within a cluster are identical
- $\rho = 0$, no correlation of responses within a cluster

The intra-cluster correlation coefficient (ICC) ρ can be interpreted as the proportion of total variance in the data that is due to clusters.

Hierarchical Composite Outcome – Rank-based methods

Hierarchical Composite Outcome (HCO)

Hierarchical Composite Outcomes

Hierarchical composite outcome (HCO) is a combination of multiple endpoints as a single measure. This is done by pre-eliciting a hierarchy of outcomes based on their clinical importance to account for patients' disease experience.

- At least two outcome components measured from the same patients
- An ordering (usually decreasing order of clinical significance) of the possible outcome combinations
- Advantages: Improves study power by increasing expected number of events
- Challenges: Hard to interpret – the treatment's actual benefit may not be evident among individual components

Generalized pairwise comparison (Rank-based methods)

- Finkelstein-Schoenfeld
- Win Ratio (Unmatched/Matched)
- Net benefit (Proportion in Favor of Treatment)
- Worst Rank Score

Example: HiLo Trial

Hilo Trial

HiLo: A Pragmatic, Randomized Trial of Phosphate Management for Patients on Maintenance Hemodialysis

Setting & Participants

HiLo

Pragmatic, cluster-randomized trial

4,400 patients receiving thrice-weekly hemodialysis in 80-120 dialysis facilities

Intervention

'Hi' phosphate target (≥6.5 mg/dl)

vs

'Lo' phosphate target (<5.5 mg/dl)

Follow-up: 27-45 months

Interventions to reach phosphate targets at the discretion of the dietitians & providers

Novel Design Features

- Extensive stakeholder engagement with patients, dietitians, nephrologists
- Hierarchical composite outcome of all-cause mortality & hospitalizations
- Pragmatic trial with liberal eligibility criteria
- Electronic informed consent (eConsent)
- Real-world data collection from EHR

CONCLUSION: HiLo will address the question of what serum phosphate target to use in hemodialysis while advancing methods for pragmatic clinical trials in nephrology.

Daniel L. Edmonston, Tamara Isakova, Laura M. Dember, et al (2020)

@AJKDonline | DOI: 10.1053/j.ajkd.2020.10.008

AJKD AMERICAN JOURNAL OF KIDNEY DISEASES

Example: HiLo Trial

Kidney failure patients receive dialysis to filter their blood. High phosphate levels (hyperphosphatemia) lead to poor clinical outcomes, so we'd like to evaluate the effectiveness of phosphate restriction and medication in maintaining patient health.

Study Overview

- Cohort: 4,400 patients undergoing hemodialysis from 80-120 dialysis facilities
- **Intervention**: prescriptions and dietary recommendations to achieve either
  - "Hi" phosphate target ($\geq$ 6.5 mg/dL)
  - "Lo" phosphate target ($\leq$ 5.5 mg/dL)
- **Outcomes**:
  1. time to all-cause mortality
  2. total hospitalization events divided by total follow-up time

Edmonston DL, Isakova T, Dember LM, Brunelli S, Young A, Brosch R, Beddhu S, **Chakraborty H**, Wolf M (2021). Design and Rationale of HiLo: A Pragmatic, Randomized Trial of Phosphate Management for Patients Receiving Maintenance Hemodialysis. American Journal of Kidney Diseases, 77(6): 920-930.e1

Potential Problems in HiLo Trial

How do we evaluate both hospitalization rate and time-to-death?

- Hospitalization alone with death as a censoring event ignores that death is a competing risk (some patients die without prior hospitalizations)
- Evaluating a time-to-event hospitalization-death composite would weight even minor hospitalizations equal to death
- Evaluating multiple outcomes components separately would mean adjusting our Type-I error rate (e.g. with Bonferonni correction) and therefore needing to increase the overall sample size or reduce power

Composite Outcomes

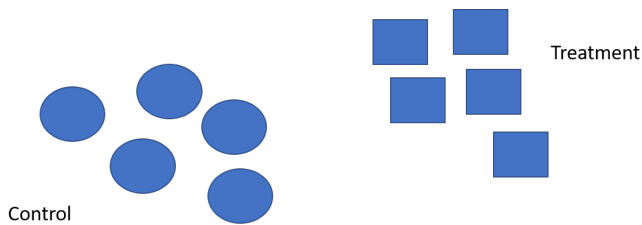
Taking into account both events jointly, assigning them ordinal importance (death more clinically significant than hospitalization) and condensing into a single aggregate measure prevents the above problems.

Rank-based Methods : Cluster Trial

Finkelstein-Schoenfeld Method

One of the earliest composite outcome methods (1999).

We compare all possible pairs of patients regardless of treatment/control status.



In the above graph, there are $\binom{10}{2} = 45$ pairs of patients.

- We collect K outcome components for each patient and choose a ranking for the K components (from most to least important).

Finkelstein-Schoenfeld Method: Notation

- There are $N = n + m$ patients total (n treatment and m control patients)
- X_i and Y_j $(i, j = \{1, \ldots, N\}, i \neq j)$ denote any two patients in the study group
- Suppose: Patient #1: $X_i = (X_{i1}, \ldots, X_{iK})$ and Patient #2: $Y_j = (Y_{j1}, \ldots, Y_{jK})$ where, $k \in \{1, \ldots K\}$ indexes the ranked outcome components from most important to least important

Favorable Operator

$X_i \succ Y_j$ if the following two conditions hold:

- $\exists l \in \{1, \ldots K\}$ s.t. X_{il} is more favorable than Y_{jl}
- $\forall h < l$ the comparison between X_{ih} and Y_{jh} is not informative

$X_i \succ Y_j \Rightarrow X_i$ is the "winner"
$X_i \prec Y_j \Rightarrow X_i$ is the "loser"
$X_i \bowtie Y_j \Rightarrow X_i$ and Y_j are "tied"

Finkelstein-Schoenfeld Method: Pair-wise Scores

We denote the winner quantitatively by assigning the pair of individuals (X_i, Y_j) a score. If we're interested in a score indicating whether X_i wins:

$$\phi(X_i, Y_j) = \begin{cases} +1 \text{ if } X_i \text{ wins } (X_i \succ Y_j) \\ -1 \text{ if } X_i \text{ loses } (X_i \prec Y_j) \\ 0 \text{ if } X_i \text{ and } Y_j \text{ tie } (X_i \bowtie Y_j) \end{cases}$$

For a tie to occur, either X_i or Y_j is censored for all outcome components or all components are equally favorable.? The summary score for a observation X_i given other observations Y_j for $i = \{1, \ldots N\}$ is:

$$U_i = \sum_{i \neq j}^{N} \phi(X_i, Y_j) \qquad \text{where, } E(U_i) \text{ is the true score for the } i\text{th patient}$$

Larger positive score means the patient X_i wins more, and larger negative score means X_i loses more.

Finkelstein-Schoenfeld Method: Cluster Trial

Wilcoxon Rank Sum test for cluster data (Rosner et al. 2003) to test U_i and incorporate clustering effect. Let $U_{i,j,m}$ be the score for the j^th individual in the i^th clusters of size m, where $m = 2, \ldots, m_{max}$, $i = 1, \ldots, k_m, j = 1, \ldots, m$. Let k_m be the total number of clusters of size m and k be the total number of all clusters.

Moreover, we can use a Mixed Model and include the effect of clustering. The form of the model would be:

$$U_{ij} = \beta_0 + \beta_1 X_{ij} + b_i + e_{ij}$$

where $U_{i,j,m}$ is the score for the j^{th} patient in the i^{th} clusters, and X_{ij} is the treatment received by the same patient. b_i is the random effect of cluster i. β_0 is the fixed intercept, β_1 is the fixed effect of the treatment, and e_{ij} is the error term.

Win Ratio (Unmatched)

For each pair of patients, the results can be classified into five categories:

- (a) The patient in the treatment group died first
- (b) The patient in the control group died first
- (c) The patient in the treatment group had more hospitalizations
- (d) The patient in the control group had more hospitalizations
- (e) None of the above (tie)

Categories (a) (b) take priority over (c) (d).

Let N_a, N_b, N_c, N_d, N_e denotes the numbers of matched pairs in categories (a) (b) (c) (d) and (e), respectively.

$N_b + N_d = N_w$ is the number of "wins" for the new treatment.

$N_a + N_c = N_L$ is the number of "losses" in the new treatment.

The win ratio R_w is defined as

$$R_w = \frac{N_w}{N_L}$$

Win Ratio (Unmatched)

In this research, we used the R package BuyseTest (Ozenne 2019) to conduct the test for the unmatched win ratio. The variance of R_w (denoted by V) was estimated by the asymptotic U-statistic theory (Buyse 2010; Péron *et al.* 2018). V was inflated by the design effect to incorporate the effect of clustering:

$$V(R_w)' = V(R_w) \times [1 + \ (\overline{m} - 1) \times \rho \ ]$$

where $\overline{m}$ is the average cluster size (consider cluster variation). The factor $[1 + (\overline{m} \ - 1)\, \rho]$ is the design effect to inflate variance for clustering. ρ is the intracluster correlation coefficient (ICC).

Win Ratio (Matched)

Let's denote the proportion as:

$$p_w = \frac{N_w}{N_w + N_L}$$

According to the two-sample t test of proportion, for clustered data, we have

$$V(p_w) = \frac{p_w(1-p_w)}{N_w + N_L}[1 + (\overline{m} - 1)\rho]$$

where $\overline{m}$ is the average cluster size (consider cluster variation). The factor $[1 + (\overline{m} - 1)\rho]$ is the design effect to inflate variance for clustering. ρ is the intracluster correlation coefficient (ICC).

Where the standard version:

$$z = \frac{p_w - 0.5}{\sqrt{V(p_w)}}$$

yielding the required p-value in an approximate normal distribution.

Net benefit (Proportion in Favor of Treatment)

Using the same notation, with $N_b + N_d = N_w$ the number of "wins" for the new treatment and $N_a + N_c = N_L$ the number of "losses" in the new treatment, the net treatment benefit is equal to:

$$D_w = \frac{N_w - N_L}{N_w + N_L}$$

Similarly to the unmatched win ratio, we used R package BuyseTest to conduct the test for the win difference.

Worst Rank Score

In the setting, we divide patients into 5 categories or buckets.

1. The patients who experience death.
2. The patients who do not experience death, and have more than 2 hospitalizations
3. The patients who do not experience death, and have less or equal to 2 hospitalizations
4. The patients without death and hospitalization, and loss to follow-up during study period (censor)
5. The patients without death and hospitalization, and followed until study ends (administrative censor)

Within each bucket, we can rank the patients based on their survival time, or time to the first hospitalization, or time to loss to follow-up. Let D_i denotes the survival time of patient i, H_{gi} denotes the time to the first hospitalization of patient i

Worst Rank Score

Let D_i denotes the survival time of patient i, H_{gi} denotes the time to the first hospitalization of patient i in bucket 2, H_{li} denotes the time to the first hospitalization of patients i in bucket 3, C_{ci} denotes the maximum follow-up time of patient i in bucket 4, and C_{ai} denotes the maximum follow-up time of patient i in bucket 5. Let $\delta_{1i} = 1,\ \delta_{2i} = 1,\ \delta_{3i} = 1,\ \delta_{4i} = 1,\ \delta_{5i} = 1$ denote patient i falls in bucket 1, 2, 3, 4, 5, respectively. Otherwise $\delta_{bi} = 0$. To define the worst-rank endpoint, we define 8 constants:

$$\xi_1 = \min(H_{gi}) - 1, \quad \eta_1 = \xi_1 - max(D_i)$$

$$\xi_2 = \min(H_{li}) - 1, \quad \eta_2 = \xi_2 - \max(H_{gi})$$

$$\xi_3 = \min(C_{ci}) - 1, \quad \eta_3 = \xi_3 - max(H_{li})$$

$$\xi_4 = \min(C_{ai}) - 1, \quad \eta_4 = \xi_4 - max(C_{ci})$$

Worst Rank Score

Such that
$D_i + \eta_1 + \eta_2 + \eta_3 + \eta_4 < \xi_1 + \eta_2 + \eta_3 + \eta_4 < H_{gi} + \eta_2 + \eta_3 + \eta_4 < \xi_2 + \eta_3 + \eta_4 <$
$H_{li} + \eta_3 + \eta_4 < \xi_3 + \eta_4 < C_{ci} + \eta_4 < \xi_4 < C_{ai}$

The worst-rank endpoint is defined as:

$$\widetilde{X}_i = \delta_{1i}\left(\eta_1 + \eta_2 + \eta_3 + \eta_4 + D_i\right) + \delta_{2i}\left(\eta_2 + \eta_3 + \eta_4 + H_{gi}\right) + \delta_{3i}\left(\eta_3 + \eta_4 + H_{li}\right) + \delta_{4i}\left(\eta_4 + C_{ci}\right) +$$

Patients who died receive the lowest rank compared to the survivors whereas those who had a shorter time to first hospitalization perform worse. The worst-rank score can be evaluated by the Wilcoxon Rank Sum test.

In this study, we used clustered Wilcoxon Rank Sum to adjust for clustering effect.

Simulation and Results

HiLo Power Simulations

4400 patients to reject the null hypothesis of no difference between treatment arms for the hierarchical composite outcome.

Power for the Components of the hierarchical Composite Outcome:

- All-cause mortality: The sample of 4400 will provide $\geq 80\%$ power to detect a HR of 0.85 on all-cause mortality
  - ICC=0.001

- All-cause hospitalization rate: The sample of 4400 will provide $> 80\%$ power to detect at least a 7% difference in hospitalization rates (Total number of hospitalizations divided by total length of follow-up)
  - ICC=0.003

HiLo Power Simulations set-up: Cluster Randomized Trial

The HiLo simulation set-up for power calculation.

1. Generate Monte Carlo correlated time to all-cause mortality and time to all-cause hospitalization simulated data using Moran's algorithm for 4400 adults using each unique combination of number of clusters, cluster sizes, and cluster size variations
2. 4400 adults with end-stage renal disease from 120-150 facilities with a facility (cluster) size of 20 and 60 with number of clusters of 220 and 74 and cluster size variations of 10%, 25%, and 50%
   - 35% of patients have no hospitalizations
3. ICC for all-cause mortality [0.001 and 0.01] and time to all-cause hospitalization [0.001 and 0.01]
4. Create all 5 hierarchical composite scores and test the hypothesis
5. Repeat the above process 5,000 times for each combination
6. To estimate power, divide the number of iterations that had a significant result ($P < 0.05$) by the total of 5000

Simulation results: Power for varying cluster sizes, number of clusters, and ICC values

Simulation Results: Power

ICC Values (all-cause mortality & hospitalization)	# of clusters with average cluster size	Cluster Size Variation	Finkelstein-Schoenfeld method	Unmatched Win Ratio	Unmatched Net Treatment Benefit	Matched Win Ratio	Worst Rank Score
0.001 and 0.003	220 clusters (cluster size 20)	10%	0.866	0.996	0.996	0.216	0.758
		25%	0.866	0.990	0.990	0.226	0.740
		50%	0.850	1.000	1.000	0.224	0.724
	74 clusters (cluster size 60)	10%	0.832	0.996	0.996	0.234	0.678
		25%	0.778	0.992	0.994	0.206	0.644
		50%	0.692	0.994	0.994	0.220	0.578
0.01 and 0.01	220 clusters (cluster size 20)	10%	0.860	0.992	0.992	0.246	0.720
		25%	0.866	0.998	0.998	0.244	0.744
		50%	0.866	1.000	1.000	0.244	0.748
	74 clusters (cluster size 60)	10%	0.796	0.986	0.986	0.218	0.654
		25%	0.754	0.994	0.994	0.196	0.612
		50%	0.666	0.998	0.998	0.196	0.526

Simulation results

- The Unmatched Win Ratio and Unmatched Net Treatment Benefit has the most power in all scenarios
- Finkelstein-Schoenfeld method works better for large number of clusters with smaller cluster sizes
- The power of the Worst Rank Score is less, especially when cluster variation becomes larger
- We do not have enough power for the Matched Win Ratio

Thank You

Interim Reassessment of ICC in a Cluster-Randomized Trial

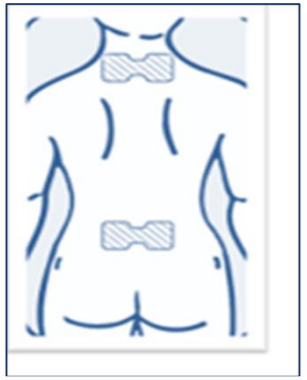
Emine Özgür Bayman, PhD
Associate Professor of Biostatistics and Anesthesia
University of Iowa
Deputy Director of the CTSDMC

CT DMC
Clinical Trials Statistical and Data Management Center

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

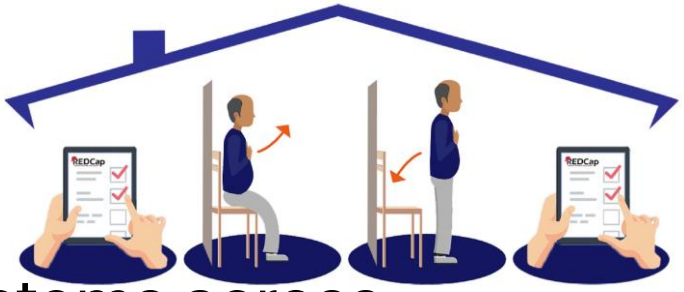
FM-TIPS: Fibromyalgia TENS in Physical Therapy Study

- Transcutaneous electrical nerve stimulation (TENS) is a nonpharmacological intervention that provides an electrical current through the skin to produce analgesia.
- FM-TIPS is a cluster randomized pragmatic clinical trial where PT clinics were randomized to TENS or No TENS, stratified by healthcare system and clinic size (large vs small).



FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Methods



- Enroll patients with fibromyalgia from 6 healthcare systems across seven states in 28 physical therapy clinics.
- We **hypothesized** that adding TENS, compared to no-TENS, to standard physical therapy care for individuals with FM would reduce movement-evoked pain.
- **Primary Outcome:** Change in Movement-Evoked Pain (Baseline to 60 days). Pain was measured during the sit and stand test at home using 0-10 Numeric Rating Scale, NRS.

Sample Size

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

- For this trial comparing TENS+PT vs. PT-alone, sample size was originally determined such that a statistical test at the 0.05 significance level would be able to detect a difference of at least 1.0 in mean change in pain (in numerical rating scale of 0 to 10), with an assumed standard deviation (SD) of 2, with 80% power.
- PT clinics to be randomized to TENS vs No TENS. Thus, an estimate of the intra-cluster correlation (ICC) was needed for the sample size calculation.
- There was no estimate of the ICC value available at the time of study design.

Intraclass Correlation Coefficient (ICC)

- In a pilot study examining a non-pharmacological intervention for pain by our group, we have shown an ICC of 0.01.
- DeBar and colleagues used an ICC of 0.002 for sample size calculation for a chronic pain population.
- Adams et al. showed widely varying ICCs between data sets, but the majority of patient-reported outcomes were below 0.095.

Intraclass Correlation Coefficient (ICC)

Complete 6-month data from 456 participants in 9 to 12 clinics per arm would allow for a conservative ICC estimate of 0.12 to 0.14, reaching 80% power for the primary analyses.

ICC	ICC values by clinics (C), per treatment arm; subjects per clinic (total subjects)			
	C=9	C=10	C=11	C=12
0.050	11 (99)	9 (90)	8 (88)	7 (84)
0.100	23 (207)	16 (160)	13 (143)	11 (132)
0.110	30 (270)	20 (200)	15 (165)	12 (144)
0.120	--	25 (250)	17 (187)	14 (168)
0.125	--	28 (280)	19 (209)	14 (168)
0.130	--	--	21 (231)	16 (192)
0.135	--	--	24 (264)	17 (204)
0.140	--	--	27 (297)	19 (228)
0.145	--	--	--	21 (252)
0.150	--	--	--	23 (276)

Intraclass Correlation Coefficient (ICC)

- A conservative estimate of ICC value of 0.14 was assumed with an equal number of participants per clinic.
- To account for up to a 24% drop-out rate, we aimed to enroll a total of 600 participants (300 per arm).

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Interim Assessment of ICC

- We planned to conduct an interim assessment of the ICC value after enrolling the first 200 participants.
- An interim assessment after half of the patients report their day 60 outcome would be ideal, but not feasible to allow enough time for necessary approvals from the study's sponsor for a potential sample size reduction.
- The interim assessment would allow us to account for the study design (such as # of clinics, # of patients per clinic), but would not assess the effect of treatment on the primary outcome measure.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Interim Assessment of ICC

- An interim reassessment of the mean and SD of the primary outcome could have been considered.
- We did not take this approach because the minimal clinically important difference in the outcome was available in the existing literature and from previous studies.
- Since there were limited preliminary data available at the time of the interim reassessment, we relied on the original estimate of the mean and SD and focused on re-estimating the ICC to support a reassessment of the sample size.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Observed Numbers

- The number of participants per clinic was assumed to be the same in the original sample size calculation.
- Due to variable enrollment per clinic, at the time of the interim assessment, we decided to account for variation by including the coefficient of variation (CV = SD/mean) of subjects per clinic in the sample size calculation.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Enrollment Projections

At the time of interim assessment, there were 28 (14/grp) active clinics with 26 clinics (13/ grp) having ≥1 participant completing day 60.

We calculated enrollment projections under different scenarios.

1. Enroll with the same trajectory, no cap on each clinic's enrollment number.
2. Enroll with the same trajectory, cap the enrollment number at 20 participants per clinic.
3. Enroll with the same trajectory, cap the enrollment number at 25 participants per clinic.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Enrollment Projections, By the End of Dec 2024

	No Cap	Cap 20	Cap 25
Mean ± SD	22 ± 13	16 ± 5.5	18 ± 7.4
CV	0.60	0.35	0.41
Projected Enrollment Total	571	414	467
Total (N/grp) Enrollment After Drop-Out	448 (224)	324 (162)	366 (183)

Based on 13 clinics per group.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Interim ICC Calculation

- We re-estimated the ICC using a modeling approach aimed at evaluating the relative contributions of different sites to the variance, in comparison with other study characteristics.
- Used generalized linear mixed model with type I sums of squares to estimate ICC.
- To maintain blinding to treatment effect in the interim reassessment, we did not include the main effect of the treatment.
- We included the interaction between clinic size and treatment in the model.

Interim ICC Calculation

- Any sample size reassessment should account for uncertainty in the new ICC estimate.
- The jackknife method allows for the estimation of the SE of the ICC without making parametric assumptions about the data.
- Using the jackknife, we perform calculations based on a leave-one-out resampling of the data wherein resampling of clusters, rather than individuals, is used to account for cluster randomization.
- In FM-TIPS, this corresponded to calculating the ICC while leaving one clinic out at a time and assessing the influence of each clinic.
- These ICC estimates were then used to calculate the jackknife-based SE.

Enrollment Projections

- When the interim ICC recalculation was conducted to determine the current ICC value based on the adjusted model, consistent with the SAP (228 enrolled subjects, 183 mITT, 144 with day 60), the ICC value was 0.05, with the jackknife estimate of the SE of ICC being 0.07.
- We recalculated the sample size, assuming a conservative estimate of the CV of 0.60.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Sample Size Recalculation

Power	N Per Grp (Total)	N of Clinic / Grp	N Per Clinic	ICC
0.92	169 (338)	13	13	0.05
0.94	182 (364)	14	13	0.05
0.81	169 (338)	13	13	0.10
0.84	182 (364)	14	13	0.10
0.78	169 (338)	13	13	0.12
0.81	182 (364)	14	13	0.12

Assuming CV = 0.6, a difference of 1.0 in mean change in pain, SD = 2, and α = 0.05, using a t-test with the degrees of freedom based on the number of clusters.

Sample Size Recalculation

- Accounting for the variable sample size per clinic (CV = 0.6) and using 13 clinics/group, the required sample size was estimated at 450 enrolled subjects to obtain a final number of 342 participants completing day 60 based on a 24% drop-out rate.
- This sample size provides
  - >90% power for ICC = 0.05,
  - >80% power for ICC = 0.10 and
  - 78% power for ICC of 0.12 (ICC + 1SE = 0.05 + 0.07).

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Sample Size Recalculation

- Based on this interim ICC reassessment, the study sample was reduced from 600 to 450, assuming an ICC value of 0.10.
- The sample size recalculation was presented to and approved by the NIAMS and the DSMB.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Conclusions

- When a reliable estimate of the ICC is unavailable during the planning of a cluster randomized trial, calculating the ICC from interim trial data is a viable option.
- In FM-TIPS, conducting an interim reassessment allowed us to estimate a new ICC based on actual trial data.
- This approach not only provided a point estimate of the ICC but also yielded the CV of enrollment per clinic for use in the sample size recalculation.

Conclusions

- The jackknife method is a very simple way to estimate the standard error (SE) of the ICC value to assess variability.
- It estimates the SE without making any parametric assumptions about the data and performs calculations based on a leave-one-out resample of the data.

Conclusions

- We not only had the point estimate (±jackknife SE) of the ICC, but also the CV of enrollment per site, which were used in the sample size recalculation during the interim assessment.
- The choice of timing for the internal pilot is important as it could be too early (25%), at the midpoint (50%), or at a late point (75%).
- Van Schie and Moerbeek recommend calculating the ICC after enrolling 50% of the participants.
- We proposed the sample size re-estimation at the time of study design.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Conclusions

- In addition to the timing of the interim assessment, it is also essential to preplan the interim reassessment and incorporate it into the SAP.
- Interim reassessment of sample size may pose logistical challenges. The interim ICC estimate in FM-TIPS was smaller than the initial estimate, which enabled us to reduce the sample size. It is possible that a trial's sample size may need to be increased if the interim ICC estimate is larger than expected.
- In such situations, operational adaptations such as securing additional funding to enroll more patients may be necessary and should be considered in advance.

REFERENCES

- Post AA, Dailey DL, Bayman EO, et al. FM-TIPS Protocol: A Multisite Embedded Pragm Trial. PhysTher, 2022
- Van Schie S, Moerbeek M. Re-estimating sample size in cluster RTs with active recruitment within clusters. Stat Med. 2014;33(19):3253-68.
- Dailey DL, Vance CGT, Rakel BA, et al. Transcutaneous Electrical Nerve Stimulation Reduces Movement-Evoked Pain and Fatigue: A Randomized, Controlled Trial. Arthritis & rheumatology (Hoboken, NJ) 2020;72:824-36.
- Lafontant D-E, McCabe ME, Heagerty PJ, Turner EL, Zimmerman MB, Bayman EO. Interim Reassessment of Sample Size in Cluster Randomized Trials, NIH Collaboratory, 2024.

- FM-TIPS is supported by the NIH through the NIH HEAL Initiative (UG3AR076387) from NIAMS.

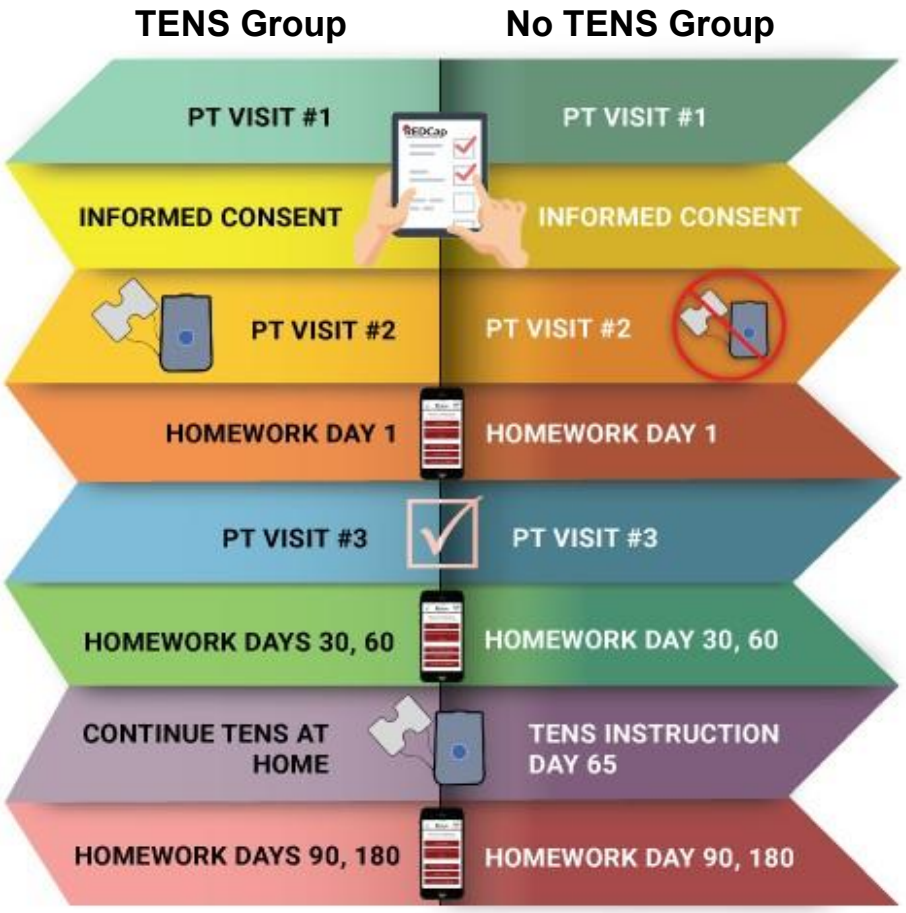
FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

HEAL
NIH · Helping to End Addiction Long-term

Observed Final ICC in FM-TIPS

- The final observed ICC: 0.01
- The final coefficient of variation: 0.49
- The final Jackknife estimate of the SE of ICC: 0.02
- Number of clinics: 14 & 15 with mITT, 13 + 13 with any day 60 data
- N per clinic: 13.5

FM-TIPS Timeline of Events



Conclusions

- Larger clinics are not necessarily enrolling faster than smaller ones.
- Conducting an internal pilot study for the ICC calculation, rather than determining variability at the design stage of the study, enabled us to assess ICC based on the actual data.

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Intraclass Correlation Coefficient (ICC)

- The dropout rate in the previous randomized clinical trial from our group, in the FAST study, was 12% at 30-day follow-up.
- Follow-up is 60 days in the proposed study, a higher drop-out rate.
- The drop-out rate includes those patients who complete the informed consent form but do not complete the baseline assessment as a part of day 1 home activities, as well as those patients who start the study and drop out prior to day 60.
- To account for the up to 24% drop-out rate, the original sample size to enroll was inflated to a total of 600 participants (300 per arm).

FM-TIPS
Fibromyalgia TENS In Physical Therapy Study

Model-robust standardization in cluster-randomized trials

Fan Li

Department of Biostatistics
Yale School of Public Health

https://lifan90.com/

The Society for Clinical Trials (SCT) 46th Annual Meeting
May, 2025

Acknowledgement

- Research in this presentation was supported by a Patient-Centered Outcomes Research Institute Award® (PCORI® Award ME-2022C2-27676) and the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R00AI173395.

A forgotten message in Section 1.5 of Donner and Klar

- Cluster randomized trials (CRTs) are increasingly popular
- While we spent considerable efforts in getting the "right" model and right **unit of analysis**, we have rarely reflected upon the **unit of inference** (Hemming and Taljaard, 2022, IJE)
  - unit of analysis—level of data observation contributing to data analysis
  - unit of inference—level of the statistical hypothesis
  - **they do not have to agree!**
- The level of statistical hypothesis (unit of inference) is written as early as in 2000, on page 13 of Donner and Klar[1]

> *"The target of inference in such studies (CRTs) could be at either the individual level or community level . . . These examples show the importance of investigators explicitly formulating and stating the hypothesis under test in cluster randomization trials. . . "*

[1] Donner A, Klar N, Klar NS. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000 Sep.

Estimands as a solution

- Increasing interest in defining **model-free** estimands at the outset
- Substantial progress has been made in individually randomized trials[2]

> **Estimand**: *A precise description of the treatment effect investigators aim to estimate from a study. An estimand comprises five aspects: (i) population; (ii) treatment conditions; (iii) endpoint; (iv) summary measure (e.g. difference in means, risk ratio, etc.); and (v) how intercurrent events are to be handled.*
> ***(The Estimands Framework)***

- Need to interpret (and refine) this framework in CRTs
- We will primarily focus on the "summary measure" concept in defining & estimating average treatment effect (ATE) estimands for CRTs

[2] Kahan BC, Cro S, Li F, Harhay MO. Eliminating ambiguous treatment effects using estimands. *American Journal of Epidemiology*. 2023 Jun;192(6):987-94.

Averaging over what population?

Consider a hypothetical cluster randomized trial with

Cluster	Cluster size N_i	True (constant) causal effect in cluster i
1	10	5
2	10	5
3	10	5
4	100	1
5	100	1
6	100	1

- treatment effect as an average across all individuals in the trial

$$\frac{(10)(5) + (10)(5) + (10)(5) + (100)(1) + (100)(1) + (100)(1)}{10 + 10 + 10 + 100 + 100 + 100} = 1.4$$

- treatment effect as an average across clusters in the trial

$$\frac{5 + 5 + 5 + 1 + 1 + 1}{6} = 3$$

- difference can be substantial depending on how treatment effect varies according to cluster size variation[3]

[3] Kahan BC, Li F, Copas AJ, Harhay MO. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology*. 2023 Feb 1;52(1):107-18.

Formalizing the ATE estimands

- Pursue the potential (counterfactual) outcomes framework
- Individual j in cluster i has two potential outcomes $\{Y_{ij}(1), Y_{ij}(0)\}$, and N_i the number of individuals per cluster
  - individual-average treatment effect (individual-ATE)

$$\text{i-ATE} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{N_i} \{Y_{ij}(1) - Y_{ij}(0)\}}{\sum_{i=1}^{m} N_i}$$

  - cluster-average treatment effect (cluster-ATE)

$$\text{c-ATE} = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{\sum_{j=1}^{N_i} \{Y_{ij}(1) - Y_{ij}(0)\}}{N_i} \right]$$

- Literature: e.g. Imai et al. (2009, Stat Sci); Su and Ding (2021, JRSSB); Kahan et al. (2022, IJE); Wang et al. (2022, Cont Clin Trials); Balzer et al. (2023, Biostatistics); Wang et al. (2024, JASA)
  - marginal estimands that are free of models
  - we focus on difference scale (absolute scale) for simplicity, but can easily extend to ratio scale

Mathematical difference

- **Mathematical** difference under informative cluster size (ICS)
  - ICS—N_i affects outcome or marginally correlated with the treatment effect in each cluster
  - cluster size is an effect modifier (**example**: odds ratio is 0.75 and 0.5 in small and large clusters)
  - can depend on scale of measure
- ICS can occur when factors that differ between small and large clusters also affect the potential outcome or interact with the treatment group
  - **example**: differences in staff experience or levels of care between larger and smaller hospitals
  - **example**: differences in socioeconomic status between larger urban schools compared with smaller rural schools
- An inherit feature of estimands in CRTs that cannot be "adjusted away"[4]

[4] Kahan BC, Li F, Blette B, Jairath V, Copas A, Harhay M. Informative cluster size in cluster-randomised trials: A case study from the TRIGGER trial. *Clinical Trials*. 2023 Dec;20(6):661-9.

Conceptual difference

- **Conceptual** difference due to unit of inference (not unit of analysis)
  - individual-ATE $\Rightarrow$ the same estimand you would have estimated had this been an individually-randomized trial
  - cluster-ATE $\Rightarrow$ a unique by-product of cluster randomization
- Some reflections:
  - **Rule of thumb 1**: if this is a cluster-level intervention, then cluster-ATE may be of interest
  - **Rule of thumb 2**: an individual-level intervention but cluster randomization is used, then individual-ATE is likely of interest
  - **Rule of thumb 3**: more generally, if the intent is to mimic an individually-randomized trial, then individual-ATE is likely of interest
  - **Rule of thumb 4**: they both can be of interest, when estimated to be quite different, prompts reflection on informative cluster size

Estimand-aligned analysis

- Ensuring the right unit of inference, regardless of the unit of analysis
- Give equal weight to each individual or each cluster?

Estimand	Example recipe (no covariates)
Individual-average treatment effect	**Cluster-level analysis** $\Rightarrow$ analyze cluster-level summaries using a weighted regression model (weights equal to N_i); **Individual-level analysis** $\Rightarrow$ GEE with a working independence correlation structure;
Cluster-average treatment effect	**Cluster-level analysis** $\Rightarrow$ Analyze cluster-level summaries using regression model (unweighted); **Individual-level analysis** $\Rightarrow$ GEE with a working independence correlation structure, but with inverse cluster-size weights equal to $1/N_i$;

- Caveat: need to work out the probability limit for each possible choice model ☹
- Caveat: coefficient from mixed models target neither cluster- or individual-ATE under ICS ☹ [5]

[5] Kahan BC, Blette B, Harhay M, Halpern S, Jairath V, Copas A, Li F (2024). Demystifying estimands in cluster randomised trials. *Statistical Methods in Medical Research*. 33(7), 1211-1232.

A path forward—standardization

- Standardization is a classic idea in epidemiological research,[6] and often allows us to obtain the risk estimates (after confounding adjustment) among a pre-specified target population
  - also known as a g-compuatation formula in observational research
- The key idea is to come up with an outcome model $Y_{ij} \sim A + \boldsymbol{X}_{ij}$, and use this model to predict $Y_{ij}(1)$ and $Y_{ij}(0)$
  - once we know $Y_{ij}(1)$, $Y_{ij}(0)$, we can plug into the estimands formula
- If the outcome model is correct, the inference is estimand-aligned
- If the outcome model is incorrect, there is **NO** guarantee that the final estimator will be consistent to c-ATE or i-ATE, despite its good intention

[6] Zou GY. Assessment of risks by predicting counterfactuals. *Statistics in Medicine*. 2009 Dec 30;28(30):3761-81.

Model-robust standardization

- **Proposal**: inherit the idea of standardization, but leverage cluster randomization to build in robustness[7]
- **Step 1**: predict each cluster's mean potential outcome under $a = 0, 1$

$$\overline{Y}_i^*(a) = \underbrace{\widehat{\mathbb{E}}(\overline{Y}_i | A_i = a, \boldsymbol{X}_i)}_{\text{regression prediction}} + \frac{I(A_i = a)}{(1/2)} \underbrace{\left(\overline{Y}_i - \widehat{\mathbb{E}}(\overline{Y}_i | A_i = a, \boldsymbol{X}_i)\right)}_{\text{mean residual}}$$

$\overline{Y}_i$ is the mean outcome, $\boldsymbol{X}_i$ the design matrix for cluster i

- cluster i is randomized to treatment ($A_i = 1$)

$$\begin{cases} \overline{Y}_i^*(1) = \widehat{\mathbb{E}}(\overline{Y}_i | A_i = 1, \boldsymbol{X}_i) + 2 \times \left(\overline{Y}_i - \widehat{\mathbb{E}}(\overline{Y}_i | A_i = 1, \boldsymbol{X}_i)\right) \\ \overline{Y}_i^*(0) = \widehat{\mathbb{E}}(\overline{Y}_i | A_i = 0, \boldsymbol{X}_i) \end{cases}$$

- cluster i is randomized to control ($A_i = 0$)

$$\begin{cases} \overline{Y}_i^*(1) = \widehat{\mathbb{E}}(\overline{Y}_i | A_i = 1, \boldsymbol{X}_i) \\ \overline{Y}_i^*(0) = \widehat{\mathbb{E}}(\overline{Y}_i | A_i = 0, \boldsymbol{X}_i) + 2 \times \left(\overline{Y}_i - \widehat{\mathbb{E}}(\overline{Y}_i | A_i = 0, \boldsymbol{X}_i)\right) \end{cases}$$

[10] Li F, Tong J, Fang X, Cheng C, Kahan BC, Wang B (2025+). Model-robust standardization in cluster-randomized trials. *Working paper*.

Model-robust standardization - cont'd

- **Step 2**: standardize to cluster-average treatment effect

$$\text{c-ATE} = \frac{1}{m}\sum_{i=1}^{m}\left\{\overline{Y}_i^*(1) - \overline{Y}_i^*(0)\right\},$$

standardize to individual-average treatment effect

$$\text{i-ATE} = \frac{\sum_{i=1}^{m} N_i\left\{\overline{Y}_i^*(1) - \overline{Y}_i^*(0)\right\}}{\sum_{i=1}^{m} N_i}$$

- Comparing to model-based standardization:
  - model-based standardization uses

$$\begin{cases}\overline{Y}_i^*(1) = \widehat{\mathbb{E}}(\overline{Y}_i|A_i = 1, \boldsymbol{X}_i)\\ \overline{Y}_i^*(0) = \widehat{\mathbb{E}}(\overline{Y}_i|A_i = 0, \boldsymbol{X}_i)\end{cases}$$

  - only one minor change—adding 2 $\times$ cluster-level residual

Worked example 1: Unit of analysis = cluster

- Requires collapsing the individual-level observations as cluster-specific means

$$\overline{Y}_i = \beta_0 + \beta_1 A_i + \boldsymbol{\beta}_2^T \overline{\boldsymbol{X}}_i + \epsilon_i \tag{1}$$

- **Step 1**: predict each cluster's mean potential outcome

$$\begin{aligned}
\overline{Y}_i^*(a) &= \widehat{\mathbb{E}}(\overline{Y}_i | A_i = a, \boldsymbol{X}_i) + \frac{I(A_i = a)}{(1/2)} \left(\overline{Y}_i - \widehat{\mathbb{E}}(\overline{Y}_i | A_i = a, \boldsymbol{X}_i) \right) \\
&= \left(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\boldsymbol{\beta}}_2^T \overline{\boldsymbol{X}}_i \right) + \frac{I(A_i = a)}{(1/2)} \left(\overline{Y}_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\boldsymbol{\beta}}_2^T \overline{\boldsymbol{X}}_i \right) \right)
\end{aligned}$$

- Standardization with **Step 2**
- Simple linear regression (`lm`) but without assuming (1) as the true data generating model

Worked example 2: Unit of analysis = individual

- A (almost most) popular approach is linear mixed model
- Individual-level data analysis accounting for intracluster correlation

$$Y_{ij} = \gamma_0 + \gamma_1 A_i + \boldsymbol{\gamma}_2^T \boldsymbol{X}_{ij} + b_i + \epsilon_i, \tag{2}$$

- **Step 1**: predict each cluster's mean potential outcome

$$\begin{aligned}\overline{Y}_i^*(a) &= \widehat{\mathbb{E}}(\overline{Y}_i | A_i = a, \boldsymbol{X}_i) + \frac{I(A_i = a)}{(1/2)} \left(\overline{Y}_i - \widehat{\mathbb{E}}(\overline{Y}_i | A_i = a, \boldsymbol{X}_i)\right) \\ &= \left(\widehat{\gamma}_0 + \widehat{\gamma}_1 a + \widehat{\boldsymbol{\gamma}}_2^T \overline{\boldsymbol{X}}_i\right) + \frac{I(A_i = a)}{(1/2)} \left(\overline{Y}_i - \left(\widehat{\gamma}_0 + \widehat{\gamma}_1 a + \widehat{\boldsymbol{\gamma}}_2^T \overline{\boldsymbol{X}}_i\right)\right)\end{aligned}$$

- Standardization with **Step 2**
- Multilevel regression (`nlme`, `lme4`, `SAS GLIMMIX`) but without assuming (2) as the true data generating model

Worked example 3: Unit of analysis = individual

- For binary outcomes, often use generalized linear mixed models
- Logistic linear mixed model

$$\text{logit}\{P(Y_{ij} = 1|A_i, \boldsymbol{X}_i, c_i)\} = \alpha_0 + \alpha_1 A_i + \boldsymbol{\alpha}_2^T \boldsymbol{X}_{ij} + c_i, \qquad (3)$$

- **Step 1**: prediction of $\widehat{\mathbb{E}}(\overline{Y}_i|A_i = a, \boldsymbol{X}_i)$ using

$$\frac{1}{N_i}\sum_{i=1}^{N_i}\int_0^{\infty} \text{expit}\left\{\widehat{\alpha}_0 + \widehat{\alpha}_1 A_i + \widehat{\boldsymbol{\alpha}}_2^T \boldsymbol{X}_{ij} + c\right\} f(c|0, \widehat{\sigma}_c^2)dc,$$

$$\approx \frac{1}{N_i}\sum_{i=1}^{N_i} \text{expit}\left\{\left(\widehat{\alpha}_0 + \widehat{\alpha}_1 a + \widehat{\boldsymbol{\alpha}}_2^T \boldsymbol{X}_{ij}\right) \Big/ \sqrt{\frac{\widehat{\sigma}_c^2 + \Pi^2/3}{\Pi^2/3}}\right\}$$

- Standardization with **Step 2**
- Multilevel regression (`lme4`, `SAS GLIMMIX`, `SAS NLMIXED`) but without assuming (3) as the true data generating model

Why model-robust?

- **Model robustness**: Under 1:1 randomization, each estimator converges to their respective target estimand (c-ATE or i-ATE) even when the model used for regression prediction is arbitrarily misspecified
  - emphasize unit of inference, regardless of the decision about unit of analysis nor even decisions about analysis model
  - these two decisions **DO NOT** introduce any ambiguity on the intended estimand and therefore does not affect bias
- This is in contrast to **model-based analysis**:
  - decision about unit of analysis nor analysis model will implicitly affect the intended estimand and hence bias
- Theoretical under-pinning:
  - Semiparametric efficiency theory and efficient influence function[8]
  - Counterparts have been studied in individually randomized trials & observational studies—AIPW estimator

[8] Wang, B., Park, C., Small, D.S. and Li, F., 2024. Model-robust and efficient covariate adjustment for cluster-randomized experiments. *Journal of the American Statistical Association*, pp.1-13.

What are possible, practical strategies for inference?

- Unfortunately, usual software does not spit out variance for the standardization estimator
- Cumbersome and impractical to derive an M-estimator variance for each (semi)-parametric working model
- Resampling?

	Bootstrap	Permutation	Jackknifing
Computation demand	heavy	heavy	manageable
Interval estimation	easy	hard	easy
Finite-sample performance	unstable	exact	stable

  - consider cluster-jackknife for inference as a general recipe
  - jackknifing connected to bias-corrected sandwich variance in GEE & LM

Simulation results under non-ICS (continuous)

Table: Simulation results with **non-informative** cluster size across 1000 iterations. LM: linear regression with cluster means; LMM: linear mixed model; EX: exchangeable working correlation; IND: independent working correlation. All methods used linear covariate adjustment.

Estimand	c-ATE = −3			i-ATE = −3		
	BIAS	MCSD	COV	BIAS	MCSD	COV
			model-based coefficient			
LM	0.1	0.17	94.5%	0.1	0.17	94.5%
LMM	0.2	0.17	94.4%	0.2	0.17	94.4%
GEE_{EX}	0.2	0.17	94.4%	0.2	0.17	94.4%
GEE_{IND}	0.1	0.17	93.2%	0.1	0.17	93.2%
			model-robust standardization			
LM	0.1	0.17	96.1%	0.1	0.17	96.4%
LMM	0.1	0.17	95.7%	0.1	0.17	96.2%
GEE_{EX}	0.1	0.17	95.6%	0.1	0.17	96.2%
GEE_{IND}	0.1	0.17	95.7%	0.1	0.17	96.5%

Simulation results under ICS (continuous)

Table: Simulation results with **informative** cluster size across 1000 iterations. LM: linear regression with cluster means; LMM: linear mixed model; EX: exchangeable working correlation; IND: independent working correlation. All methods used linear covariate adjustment.

Estimand	c-ATE = 4.48			i-ATE = 6.25		
	BIAS	MCSD	COV	BIAS	MCSD	COV
			model-based coefficient			
LM	0.0%	0.47	92.9%	-28.3%	0.47	3.9%
LMM	0.7%	0.47	93.1%	-27.8%	0.47	3.7%
GEE_{EX}	0.3%	0.47	92.7%	-28.0%	0.47	3.7%
GEE_{IND}	37.1%	0.51	5.3%	-1.6%	0.51	89.9%
			model-robust standardization			
LM	0.0%	0.47	95.4%	-2.8%	0.52	93.8%
LMM	0.2%	0.47	94.9%	-2.1%	0.52	93.8%
GEE_{EX}	0.2%	0.47	94.9%	-2.1%	0.52	93.8%
GEE_{IND}	0.6%	0.47	94.9%	-1.6%	0.51	93.6%

A new R package

`MRStdCRT`: execute model-robust standardization with cluster-jackknife variance estimates (and a test for ICS).

Declaration:

- Available at `https://github.com/deckardt98/Model-robust-standardization-CRT`

```
MRStdCRT_fit <- function(formula, data, clus_id, trt, prob, method,
                         family, corstr, scale, jack, alpha)
```

Example

```
> library(MRStdCRT)
> data(ppact)

> example <- MRStdCRT_fit(formula = PEGS ~ AGE + FEMALE + comorbid + Dep_OR_Anx + pain_count +
                          PEGS_bl + BL_benzo_flag + BL_avg_daily + satisfied_primary + cluster(n),
                          data = ppact,
                          clus_id = "CLUST",
                          trt = "INTERVENTION",
                          prob = prob,
                          method = "GEE",
                          corstr = "independence",
                          scale = "RD")
>
> example$estimate
     Estimate Std, Error CI lower CI upper ICS_tstat p value
cATE    0.907      0.028    0.852    0.962  -1.719   0.089
iATE    0.926      0.024    0.879    0.973
```

Discussion

- Additional scenarios explored in simulations, same findings (binary outcome, smaller # of clusters, constrained randomization)
- Assessing $\mathcal{H}_0$: c-ATE = i-ATE motivates a natural model-robust standardization ICS test (including in our software)
- Efficiency considerations
  - choice of outcome models CAN affect efficiency and power
  - model-robust standardization provides a natural vehicle for future simulations that compare different models in a "cleaner" way
    - GLMM or GEE?
    - cluster-level or individual-level analysis?
- efficient influence function permits the inclusion of machine learning methods for outcome regression and achieves the theoretical variance lower bound (yet to be included in package)[9]

[9] Wang, B., Park, C., Small, D.S. and Li, F., 2024. Model-robust and efficient covariate adjustment for cluster-randomized experiments. *Journal of the American Statistical Association*, pp.1-13.

Extra: Simulation results under non-ICS (binary)

- Target estimand: marginal log OR

Table: Simulation results with **non-informative** cluster size across 1000 iterations. GLM: logistic regression with cluster means; GLMM: logistic linear mixed model; EX: exchangeable working correlation; IND: independent working correlation. All methods used linear covariate adjustment.

Estimand	c-ATE = 0.65			i-ATE = 0.65		
	BIAS	MCSD	COV	BIAS	MCSD	COV
			model-based coefficient			
GLM	11.0	0.13	91.8%	11.0	0.13	91.8%
GLMM	15.6	0.13	86.7%	15.6	0.13	86.7%
GEE_{EX}	12.1	0.12	90.4%	12.1	0.12	90.4%
GEE_{IND}	13.1	0.13	90.0%	13.1	0.13	90.0%
			model-robust standardization			
GLM	−3.0	0.11	96.7%	−0.9	0.11	96.5%
GLMM	−2.8	0.11	96.2%	−0.8	0.11	95.9%
GEE_{EX}	−2.9	0.11	96.3%	−0.8	0.11	96.0%
GEE_{IND}	−2.9	0.11	96.3%	−0.8	0.11	95.7%

Extra: Simulation results under ICS (binary)

- Target estimand: marginal log OR

Table: Simulation results with **informative** cluster size across 1000 iterations. GLM: logistic regression with cluster means; GLMM: logistic linear mixed model; EX: exchangeable working correlation; IND: independent working correlation. All methods used linear covariate adjustment.

Estimand	c-ATE = 0.71			i-ATE = 0.97		
	BIAS	MCSD	COV	BIAS	MCSD	COV
			model-based coefficient			
GLM	−0.8	0.19	96.4%	−27.4	0.19	72.8%
GLMM	28.3	0.23	85.5%	−6.1	0.23	94.5%
GEE_{EX}	5.1	0.19	95.7%	−23.1	0.19	79.6%
GEE_{IND}	35.1	0.21	77.7%	−1.1	0.21	96.0%
			model-robust standardization			
GLM	−0.8	0.19	96.0%	−0.9	0.22	95.9%
GLMM	−0.8	0.19	96.0%	−0.9	0.22	95.8%
GEE_{EX}	−0.9	0.19	96.0%	−0.9	0.22	96.0%
GEE_{IND}	−1.1	0.19	96.3%	−1.1	0.21	96.0%