Recent Developments in Statistical Methods for Stepped Wedge Cluster Randomized Trials

Fan Li
Assistant Professor
Department of Biostatistics
Center for Methods in Implementation and Prevention Science
Yale School of Public Health

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS
UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO

Colorado Clinical and Translational Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

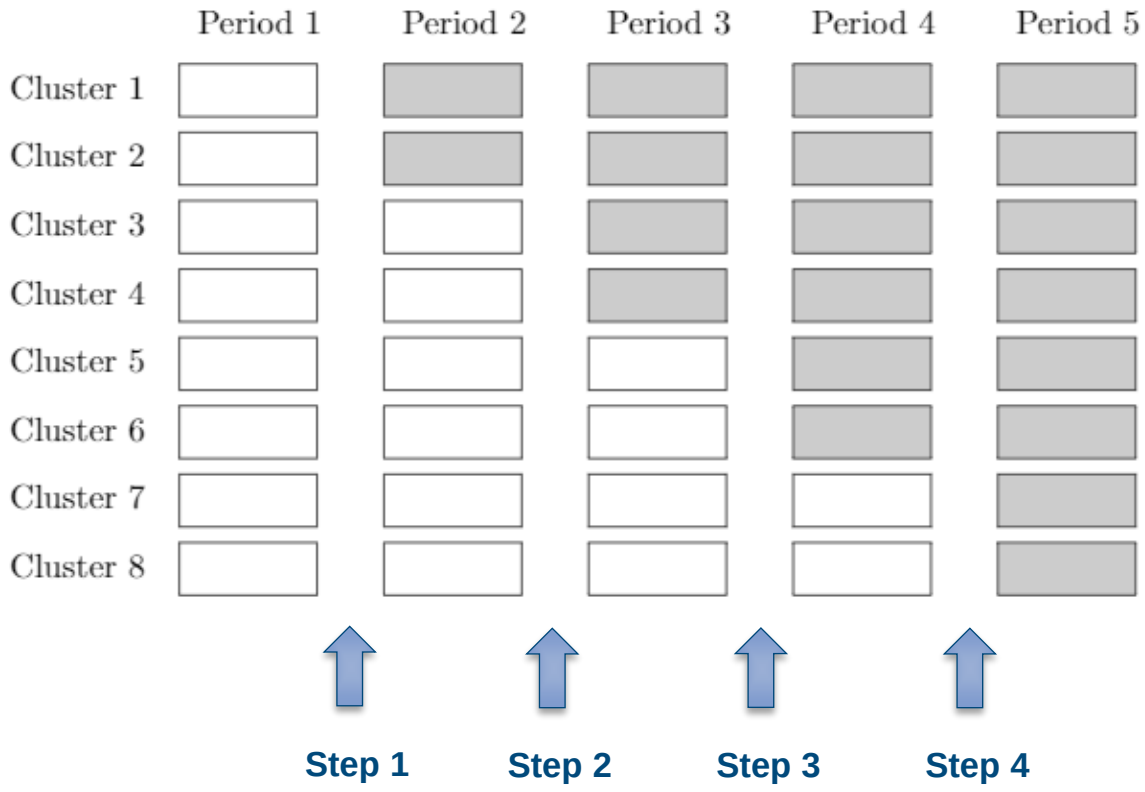
I. Introduction and Background

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Introduction

- Cluster randomized trials (CRTs) allocate clusters (hospitals, nursing homes, clinics etc.) of individuals to intervention groups
  - Minimize contamination
  - Administrative convenience
  - Usually in parallel design
- Stepped wedge (SW) design rolls out intervention in a staggered fashion
  - Cluster is the unit of randomization
  - Cluster randomized to each **step** or **wave**
  - Outcome measurements taken in each **period**



When is SW-CRT a good study design choice?

Four broad justification of using SW-CRT design *(Hemming and Taljaard, 2020 IJE)*

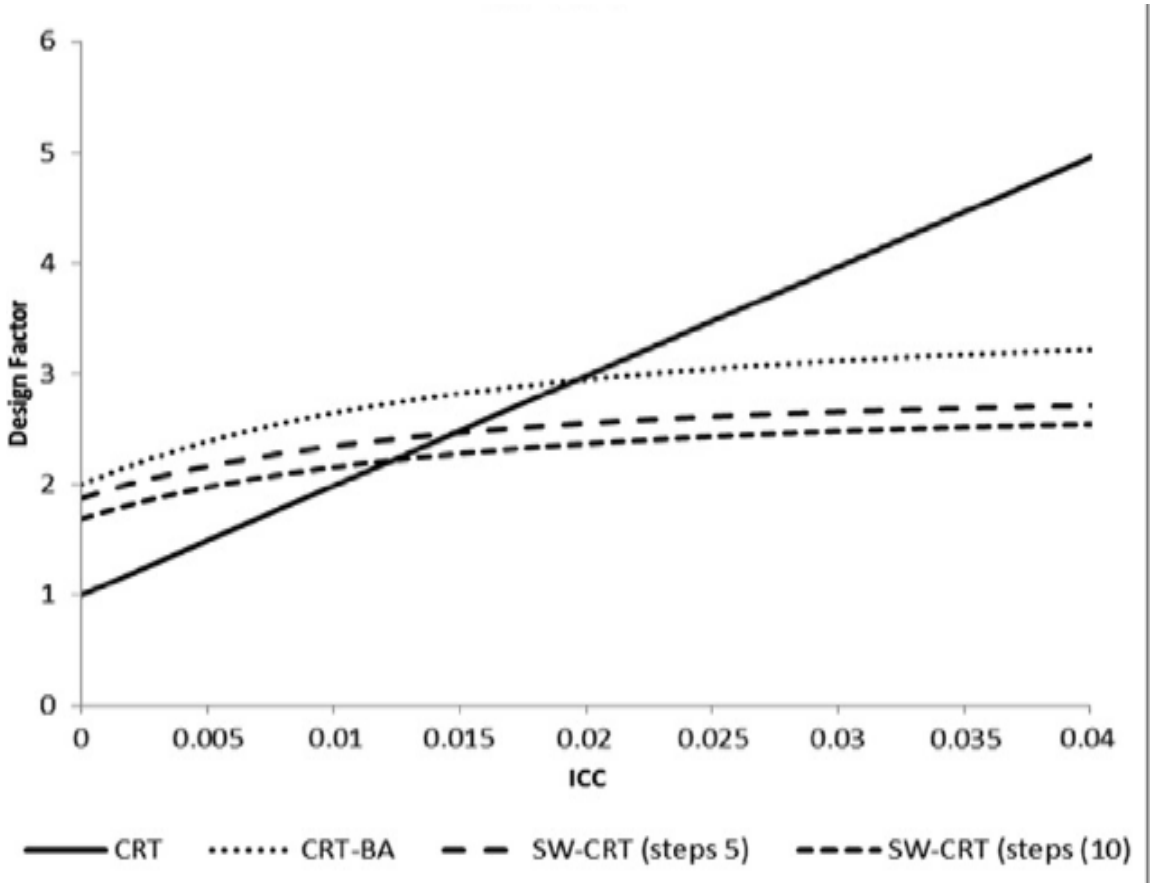
Can facilitate cluster recruitment when intervention **perceived to be effective** with minimum harm

Logistically feasible design by staggering the roll-out

Provides a means to conduct a randomized evaluation with **full roll-out**

Within-cluster before-after comparisons **can increase statistical power**

e.g., comparison of design effect *(Hemming and Taljaard, JCE 2016)*



COPRH Con
Colorado Pragmatic Research in Health Conference

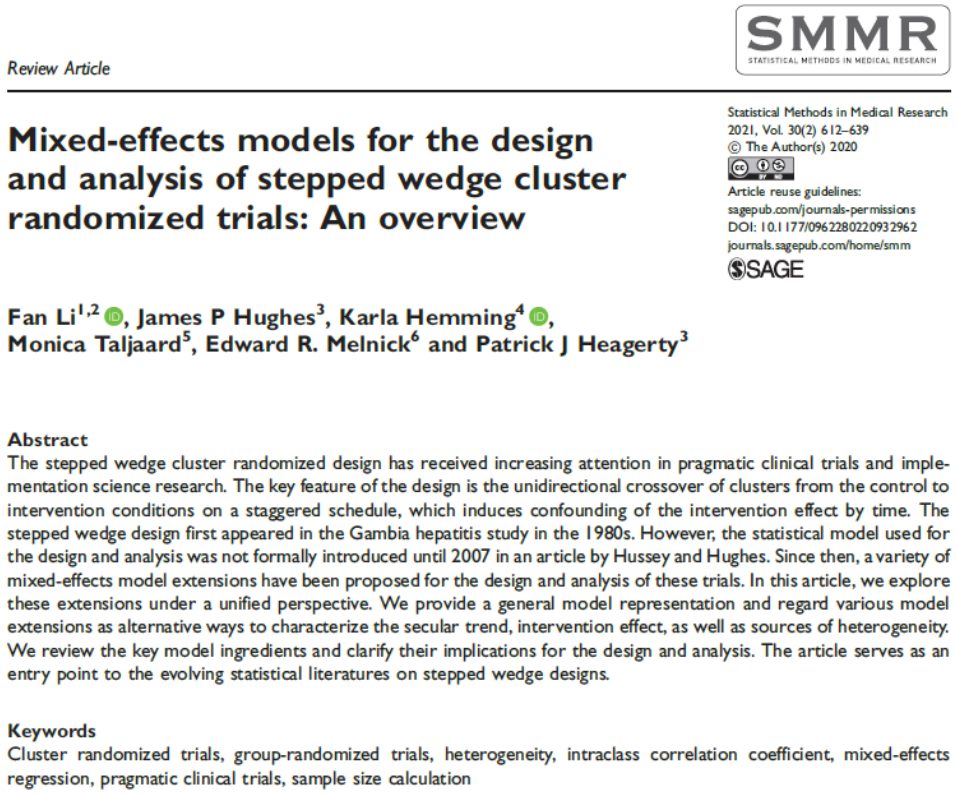
ACCORDS

CCTSI

Caveats for SW-CRTs

- But SW-CRTs can also be prone to risks of biases
  - Identification and recruitment biases
    - common to all CRTs with post-randomization recruitment
  - Complex and heterogeneous **secular trend** even in the absence of intervention
  - Risks associated with extremely small number of clusters *(Taljaard et al., 2016 Clinical Trials)*
    - Caution against 6 clusters or fewer
  - Other implementation challenges (longer duration, retaining participants etc.)
- Decision to adopt an SW-CRT deserves a comprehensive evaluation by weighing potential benefits against risks in each specific trial context

Objectives and Goals



Review Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview

Statistical Methods in Medical Research
2021, Vol. 30(2) 612–639

DOI: 10.1177/0962280220932962
journals.sagepub.com/home/smm
SAGE

Fan Li[1,2], James P Hughes[3], Karla Hemming[4], Monica Taljaard[5], Edward R. Melnick[6] and Patrick J Heagerty[3]

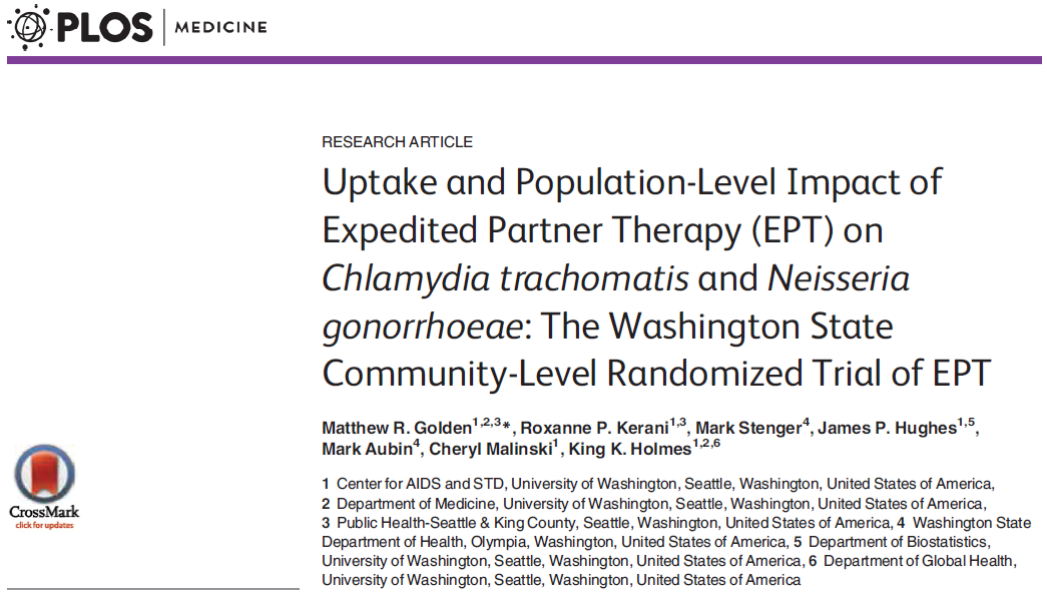
Abstract
The stepped wedge cluster randomized design has received increasing attention in pragmatic clinical trials and implementation science research. The key feature of the design is the unidirectional crossover of clusters from the control to intervention conditions on a staggered schedule, which induces confounding of the intervention effect by time. The stepped wedge design first appeared in the Gambia hepatitis study in the 1980s. However, the statistical model used for the design and analysis was not formally introduced until 2007 in an article by Hussey and Hughes. Since then, a variety of mixed-effects model extensions have been proposed for the design and analysis of these trials. In this article, we explore these extensions under a unified perspective. We provide a general model representation and regard various model extensions as alternative ways to characterize the secular trend, intervention effect, as well as sources of heterogeneity. We review the key model ingredients and clarify their implications for the design and analysis. The article serves as an entry point to the evolving statistical literatures on stepped wedge designs.

Keywords
Cluster randomized trials, group-randomized trials, heterogeneity, intraclass correlation coefficient, mixed-effects regression, pragmatic clinical trials, sample size calculation

- If proceed with the SW-CRT design:
  - What are variants of SW-CRTs?
  - What are available statistical methods and tools to assist in the **design** and **analysis**?
    - Method of analysis *(Li et al., 2020 SMMR)*
    - Sample size determination
    - Two inter-connected aspects
  - What are recommended practices?
    - CONSORT extension to SW-CRTs
  - What are the remaining issues?

Main types of SW-CRT

- Repeated cross-sectional design
  - enrolls new participants from each cluster during each period



PLOS | MEDICINE

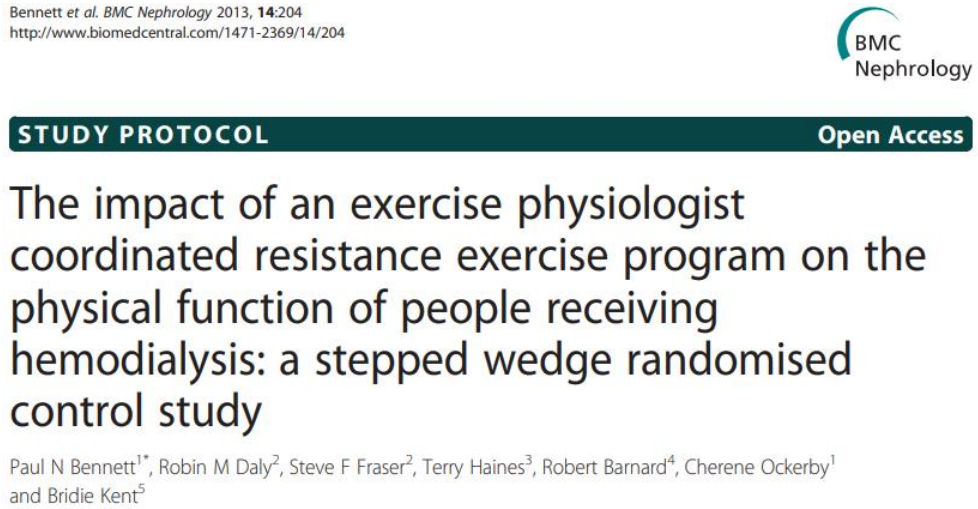
RESEARCH ARTICLE

Uptake and Population-Level Impact of Expedited Partner Therapy (EPT) on *Chlamydia trachomatis* and *Neisseria gonorrhoeae*: The Washington State Community-Level Randomized Trial of EPT

Matthew R. Golden[1,2,3]*, Roxanne P. Kerani[1,3], Mark Stenger[4], James P. Hughes[1,5], Mark Aubin[4], Cheryl Malinski[1], King K. Holmes[1,2,6]

1 Center for AIDS and STD, University of Washington, Seattle, Washington, United States of America, 2 Department of Medicine, University of Washington, Seattle, Washington, United States of America, 3 Public Health-Seattle & King County, Seattle, Washington, United States of America, 4 Washington State Department of Health, Olympia, Washington, United States of America, 5 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America, 6 Department of Global Health, University of Washington, Seattle, Washington, United States of America

- Closed-cohort design
  - identifies a cohort at the beginning of the study and schedules repeated follow-up outcome assessments for the same cohort



Bennett *et al. BMC Nephrology* 2013, **14**:204
http://www.biomedcentral.com/1471-2369/14/204

BMC Nephrology

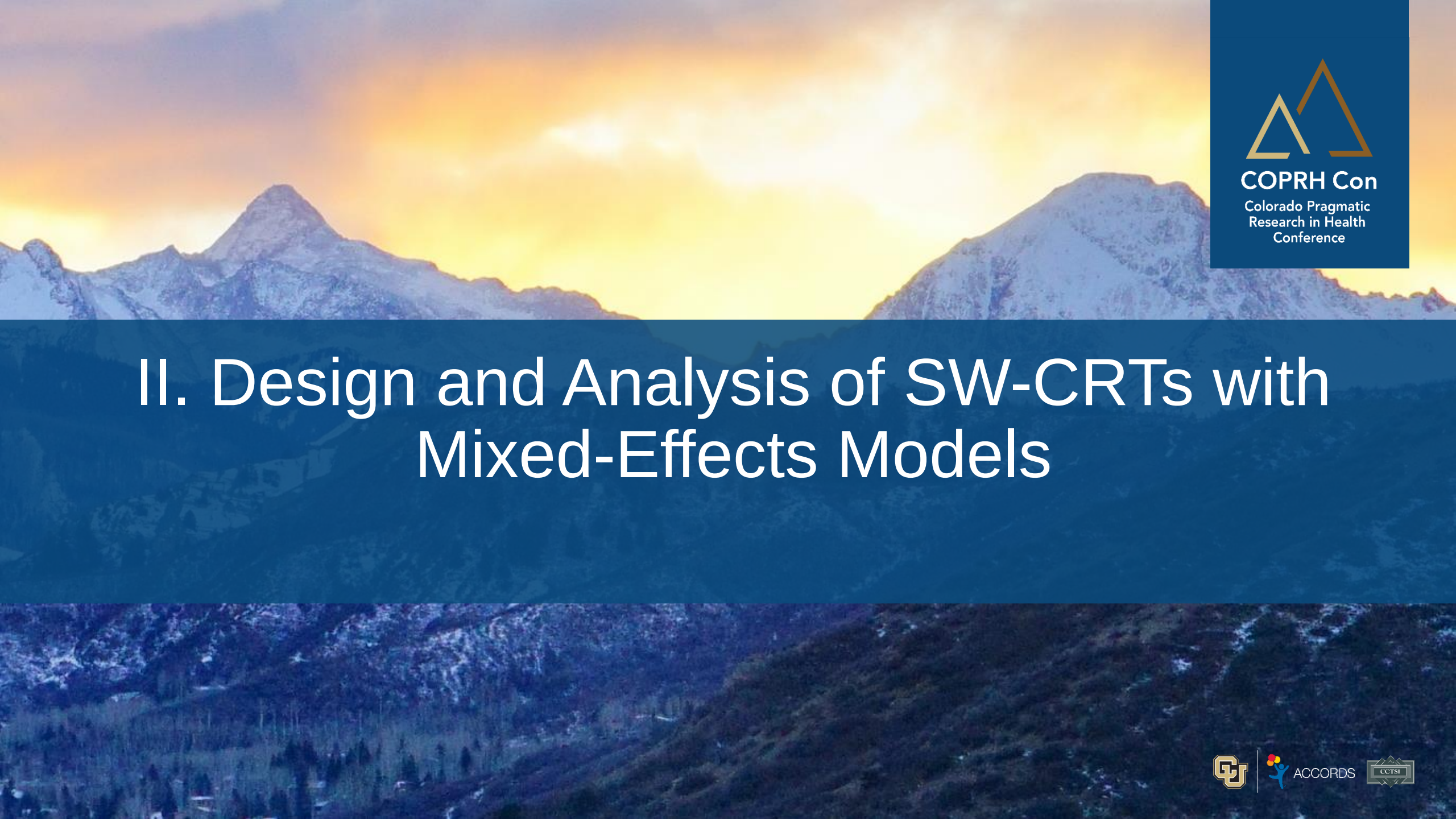
STUDY PROTOCOL Open Access

The impact of an exercise physiologist coordinated resistance exercise program on the physical function of people receiving hemodialysis: a stepped wedge randomised control study

Paul N Bennett[1*], Robin M Daly[2], Steve F Fraser[2], Terry Haines[3], Robert Barnard[4], Cherene Ockerby[1] and Bridie Kent[5]

Main types of SW-CRT – Cont'd

- Repeated cross-sectional design implicitly assumes observed population is **representative** of the target study population
  - Violation to which could result in selection/recruitment bias
  - Often used when randomizing facilities in health care systems
- Closed-cohort design can require a strong effort in **retaining participants**
  - Informative drop-out, or outcomes truncation by "death" leads to selection bias
  - Can require a smaller total sample size compared to repeated cross-sectional designs due to correlations between repeated outcome measures
- Open-cohort design is a third option *(Copas et al., 2015 Trials)*



II. Design and Analysis of SW-CRTs with Mixed-Effects Models

ACCORDS

Analytical models – mixed-effects models

- Unique features of SW-CRTs requires more complex considerations on analytical models than those in a parallel CRT
- Mixed-effects models
  - Fixed-effects to control for discrete-time secular trend
  - Intervention effect
  - Random-effects to account for clustering
- Key ingredient of a mixed-effects model *(Li et al., 2020 SMMR)*

$$g[\mu_{ijk}(s)] = \underbrace{\boldsymbol{F}^0(j)'\boldsymbol{\beta}}_{\text{secular trend}} + \underbrace{F_i^1(j,s)\Delta(j,s)}_{\text{intervention effect}} + \underbrace{\boldsymbol{R}_{ik}(j,s)'\boldsymbol{\alpha}_i}_{\text{heterogeneity}}.$$

- Widely accessible from standard software; most used in SW-CRTs *(Barker et al. 2016, BMC Med. Res. Methodol.)*

The simplest “discrete time” linear mixed model

- **Model 1**: Hussey and Hughes (2007) developed the **random-intercept model** for cross-sectional SW-CRT designs:

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \epsilon_{ijk}$$
$$\alpha_i \sim N(0, \tau_\alpha^2), \qquad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$$

  - μ is the overall mean
  - β_j is fixed categorical secular trend (time effect)
  - δ is the intervention effect
  - α_i is the random cluster effect
  - ϵ_{ijk} is the independent error

- Between-cluster heterogeneity is induced by a single α_i term, capturing the cluster-specific departure from the average is assumed to be **homogeneous** across time periods and intervention sequences

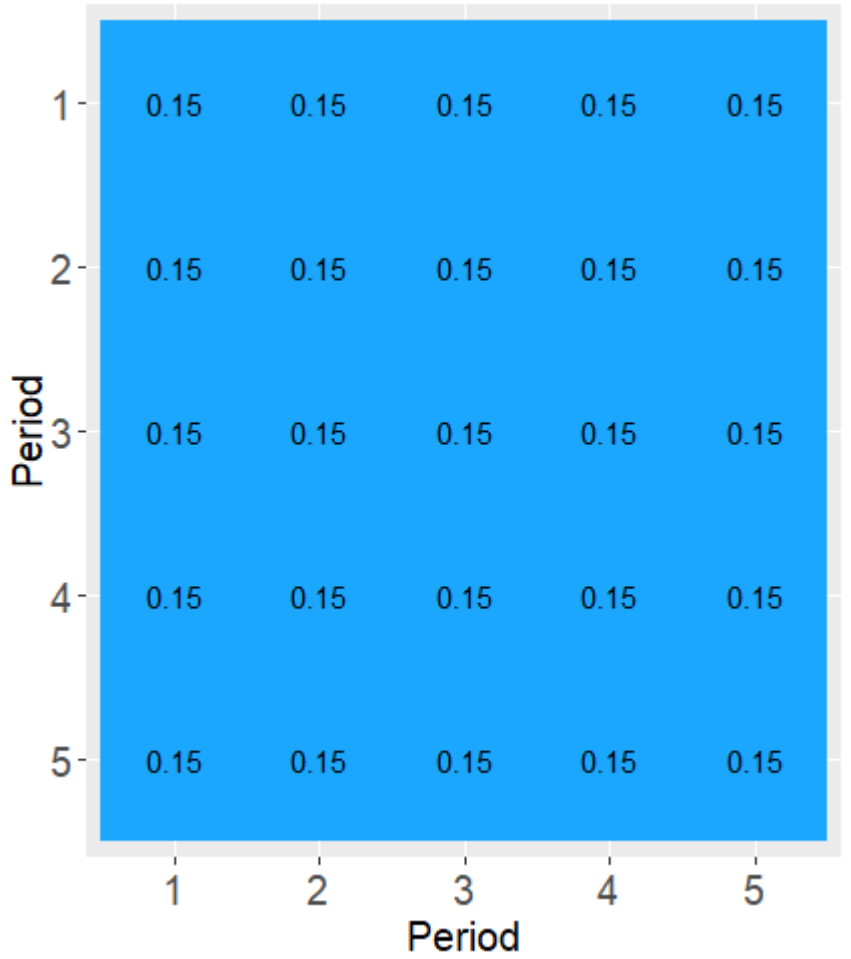
Simple exchangeable ICC structure

- Intraclass correlation coefficient (ICC) defined as proportion of outcome total variance explained by between-cluster heterogeneity

$$\alpha_0 = \frac{\tau_\alpha^2}{\tau_\alpha^2 + \sigma_\epsilon^2}$$

- In model 1, a common α_0 is implied for both observations **within** the same time period and **between** different periods
- Simple sample size formula and design effect are available *(Hussey and Hughes, 2007; Woertman et al.; 2013)*

For example, when $\alpha_0 = 0.15$



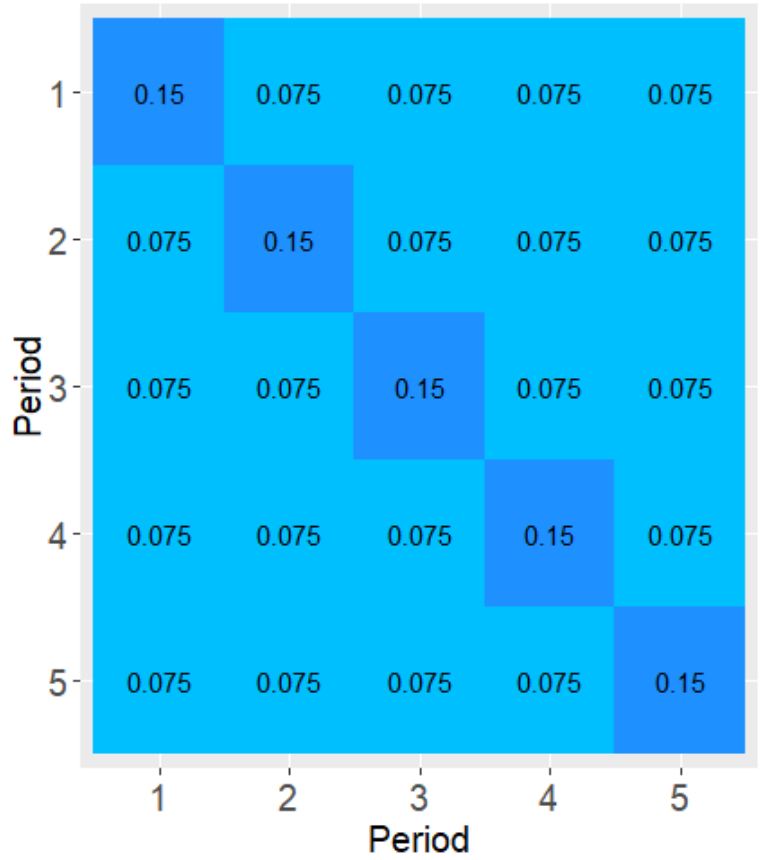
COPRH Con
Colorado Pragmatic Research in Health Conference

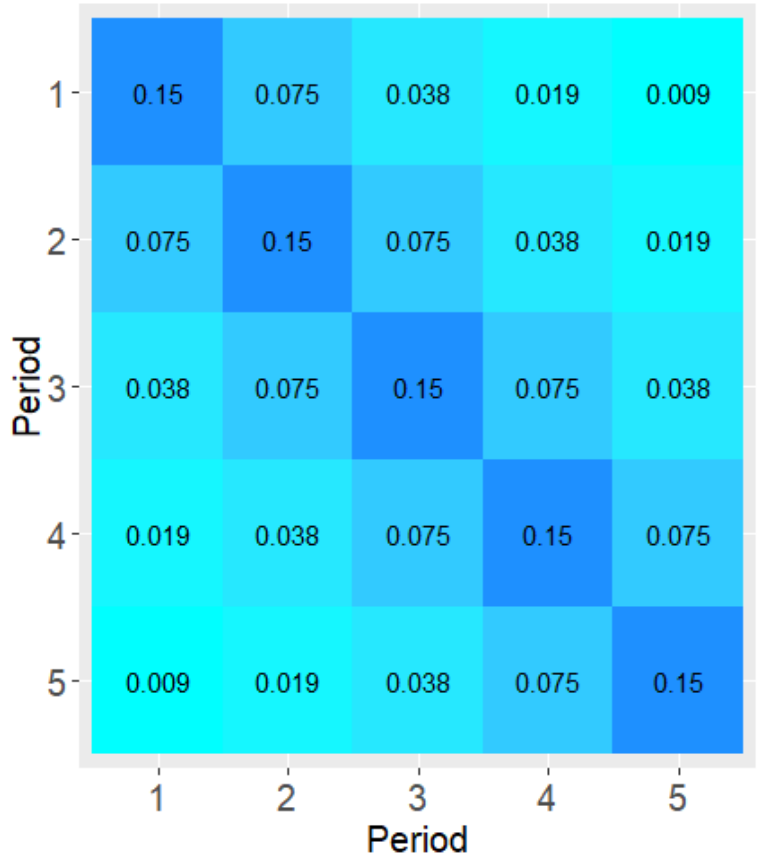
ACCORDS

CCTSI

Modifications to accommodate “decay”

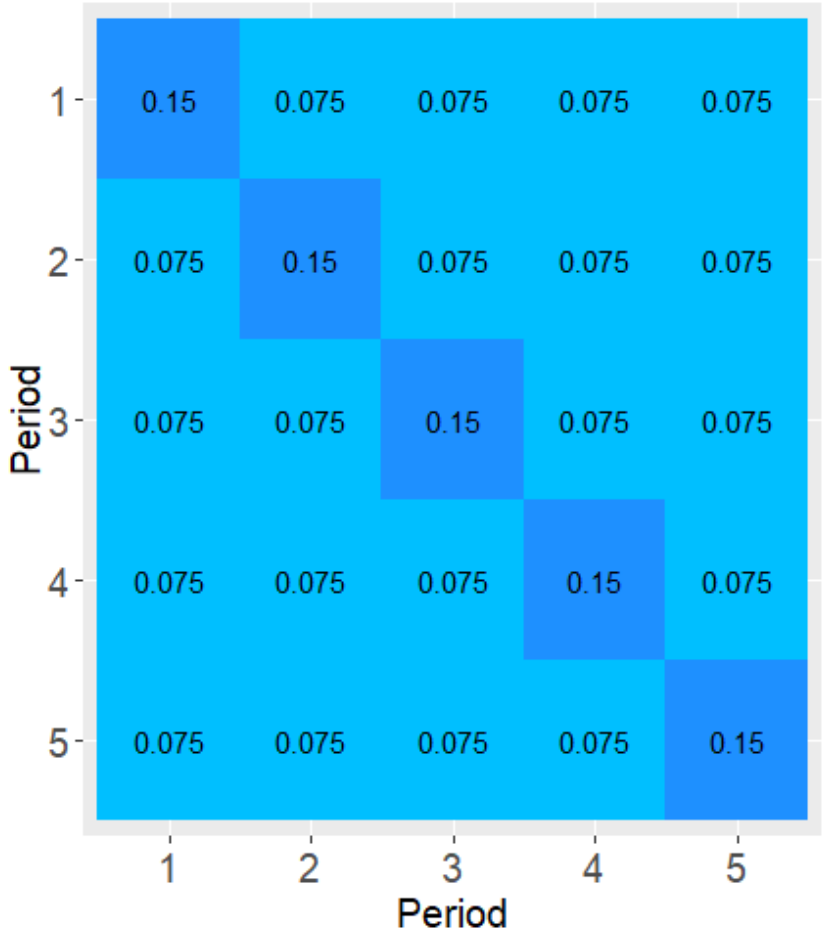
- Within-period ICC can be stronger than between-period ICC, and more flexible models should address the possibility that the strength of correlations may decay over time
- Two examples, when the within-period ICC remains $\alpha_0 = 0.15$





COPRH Con
Colorado Pragmatic Research in Health Conference

Nested exchangeable ICC structure



- **Model 2**: includes an additional random cluster-by-time interaction *(Hooper et al. 2016 Stat Med)*

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + c_{ij} + \epsilon_{ijk}$$

$$\alpha_i \sim N(0, \tau_\alpha^2), \qquad c_{ij} \sim N(0, \tau_c^2), \qquad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$$

- Between-period ICC (α_1) differs from the within-period ICC (α_0) but stays constant over time
- Quantify the between-period ICC decay by **cluster autocorrelation coefficient (CAC)**

$$\text{CAC} = \alpha_1/\alpha_0 = 0.5$$

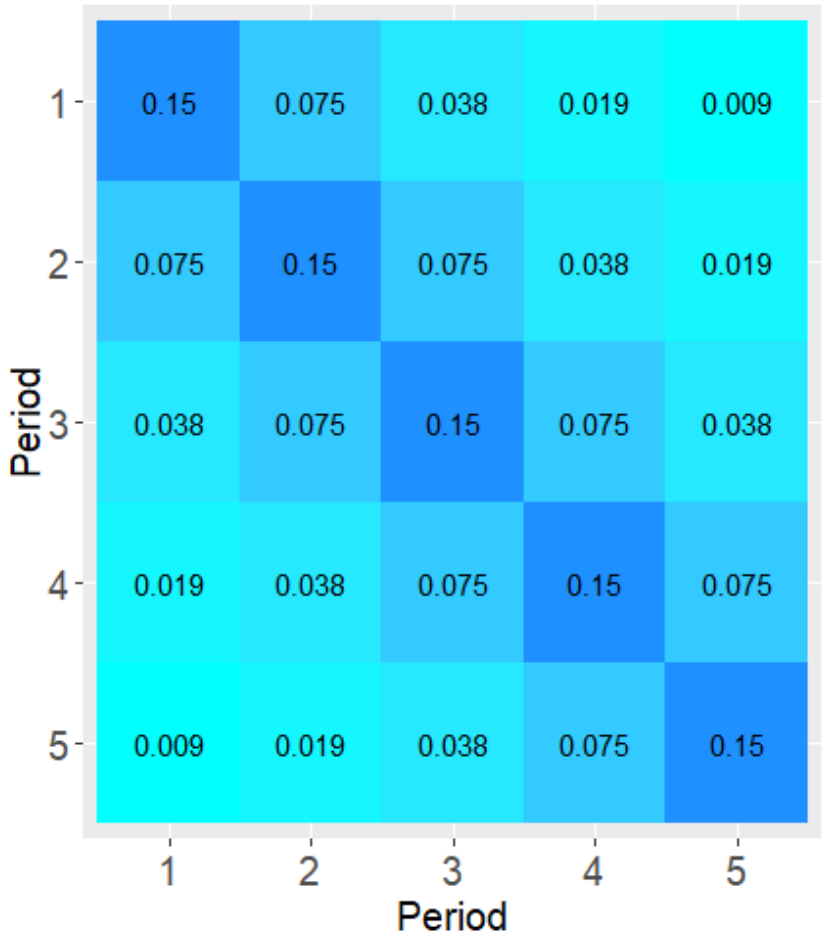
- Analytical sample size formula available, **requires CAC**

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

CCTSI

Exponential decay ICC structure



- **Model 3**: includes an additional random cluster-by-time interaction *(Kasza et al. 2019 SMMR)*

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + CP_{ij} + \epsilon_{ijk}$$
$$\left(CP_{i1}, \ldots, CP_{iJ}\right)' \sim N(0, \tau^2_{CP}\boldsymbol{R}), \qquad \epsilon_{ijk} \sim N(0, \sigma^2_\epsilon)$$

- $\boldsymbol{R}$ is the first-order auto-regressive (AR1) matrix
- Between-period ICC ($\alpha_1^{(j,k)}$) decays at an exponential rate over time
- CAC measures the rate of decay per period
- Unlike the previous models which are easy to fit in SAS and R, the exponential decay model is more difficult to fit (currently only possible in SAS)

Implications for correlation mis-specification

- No consensus on how to choose a best fitting model yet
- Under-specification (omitting a necessary decay parameter) **results in bias** of the (model-based) **variance** of the treatment effect estimator *(Kasza and Forbes, 2019 SMMR)*
    - Assuming model 1 (CAC =1) when model 3 holds will underestimate variance (p-value too small, CI too narrow)
    - Assuming model 2 when model 3 holds will usually underestimate variance (but can go both ways)
    - Impact depends on strength of correlation decay, within-period ICC and cluster period sizes
- Over-specification (including a decay unnecessarily) **does not lead to bias**

Practical considerations in applying mixed-effects models

- Other model extensions include **random intervention model** and **random time coefficient model**
- Can be preferable due to the ability to flexibly specify random heterogeneity structure
- Available in standard software
  - *PROC MIXED, PROC GLIMMIX (SAS)*
  - *nlme, lme4 (R)*
- Model-based variance (standard software output) can be **biased** if random-effects structure mis-specified
- **Interpretation** of treatment effect also depends on the random-effects structure, especially for logistic mixed models with a binary outcome

Sample size for SW-CRTs with mixed-effects models

Sample size algorithms more complex than parallel CRTs

- more design parameters (# of clusters, cluster-period sizes, and ICCs)
- If not model 1, need **CAC (decay)** in addition to the within-period ICC

A select summary of available tools for computing power based on (linear) mixed models

- Linear mixed model approximation for binary outcomes may be inaccurate *(Zhou et al., Biostatistics, 2020)*

Software	Outcome	Feature
steppedwedge (Stata) Hemming and Girling (2014)	Continuous	Model 1
	Binary	Linear mixed model approximation
SWSamp (R) Baio et al. (2015, Trials)	Continuous	Model 1
	Binary	Linear mixed model approximation
swCRTdesign (R) Vodal et al. (2020)	Continuous	Model 1-2 and others
	Binary	Linear mixed model approximation
swdpwr (R) **%swdpwr (SAS Macro)** Chen et al. (2021+)	Continuous	Model 1-2 (allow cohort designs)
	Binary	Linear mixed probability model (**with the correct binomial variance**)
Shiny CRT Calculator (Hemming et al., IJE 2020)	Continuous	Model 1-3 (allow cohort designs)
	Binary	Linear mixed model approximation

Sample size calculation requires ICC and CAC

- Recommended to calculate using routinely collected data
- Published trials reporting ICCs and CAC (similar population and outcome)
- Databases or publications that report lists of ICCs
  - Clinical outcomes: ICCs typically ≤ 0.05
  - Process measures: ICC typically larger, up to 0.15
  - CAC from 0.6 to 0.8 considered reasonable
- Sensitivity analysis across a range of plausible values

Martin *et al. Trials* (2016) 17:402
DOI 10.1186/s13063-016-1532-9

Trials

RESEARCH Open Access

Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care

James Martin*, Alan Girling, Krishnarajah Nirantharakumar, Ronan Ryan, Tom Marshall and Karla Hemming

Abstract

Background: Clustered randomised controlled trials (CRCTs) are increasingly common in primary care. Outcomes within the same cluster tend to be correlated with one another. In sample size calculations, estimates of the intra-cluster correlation coefficient (ICC) are needed to allow for this nonindependence. In studies with observations over more than one time period, estimates of the inter-period correlation (IPC) and the within-period correlation (WPC) are also needed.

Methods: This is a retrospective cross-sectional study of all patients aged 18 or over with a diagnosis of type-2 diabetes, from The Health Improvement Network (THIN) database, between 1 October 2007 and 31 March 2010. We report estimates of the ICC, IPC, and WPC for typical outcomes using unadjusted and adjusted generalised linear mixed models with cluster and cluster by period random effects. For binary outcomes we report on the proportions scale, which is the appropriate scale for trial design. Estimated ICCs were compared to those reported from a systematic search of CRCTs undertaken in primary care in the UK in type-2 diabetes.

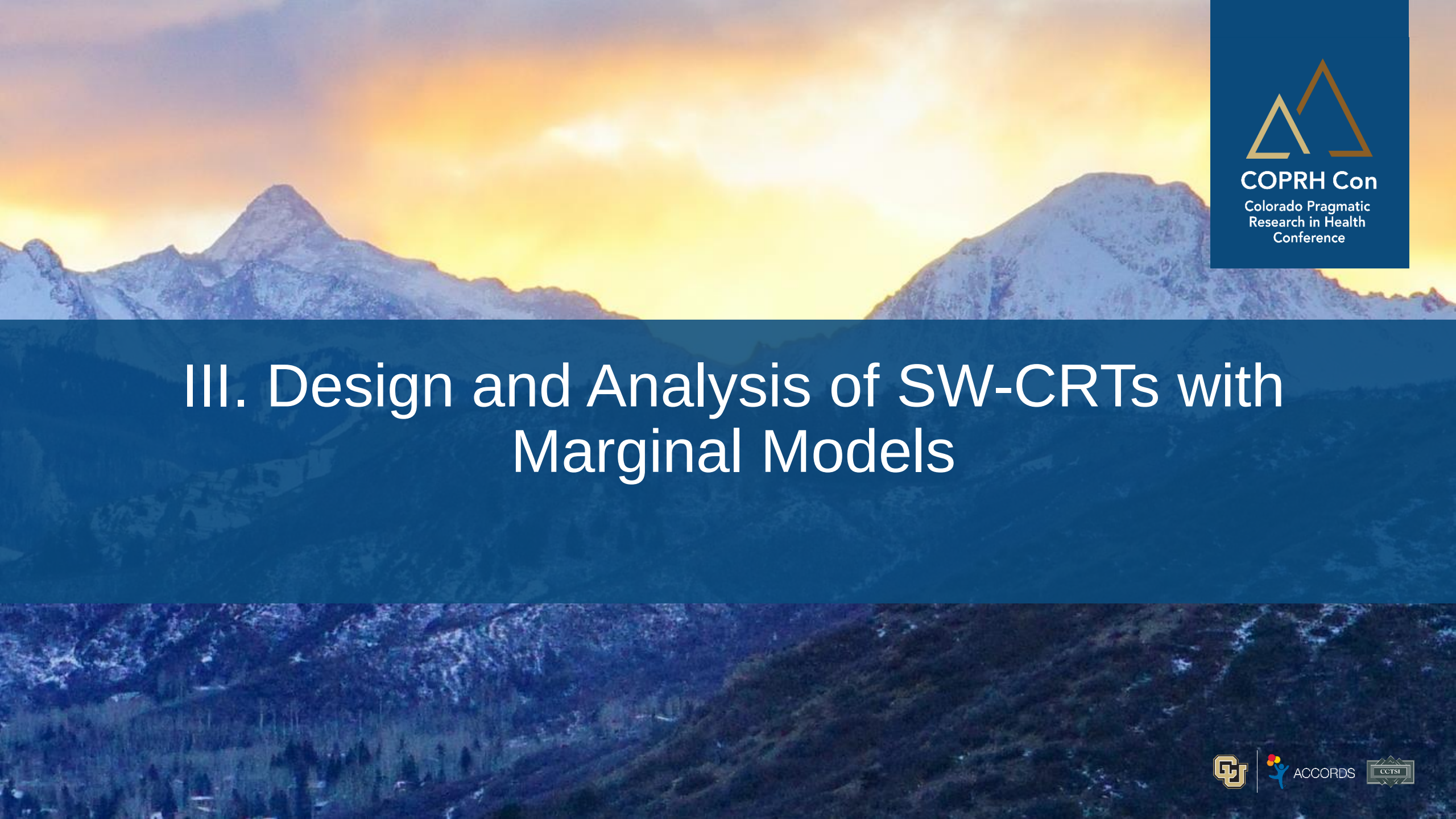
Results: Data from 430 general practices, with a median [IQR] number of diabetics per practice of 241 [150–351], were analysed. The ICC for HbA1c was 0.032 (95 % CI 0.026–0.038). For a two-period (each of 12 months) design, the WPC for HbA1c was 0.035 (95 % CI 0.030–0.040) and the IPC was 0.019 (95 % CI 0.014–0.026). The difference between the WPC and the IPC indicates a decay of correlation over time. Following dichotomisation at 7.5 %, the ICC for HbA1c was 0.026 (95 % CI 0.022–0.030). ICCs for other clinical measurements and clinical outcomes are presented. A systematic search of ICCs used in the design of CRCTs involving type-2 diabetes with HbA1c (undichotomised) as the outcome found that published trials tended to use more conservative ICC values (median 0.047, IQR 0.047–0.050) than those reported here.

COPRH Con Colorado Pragmatic Research in Health Conference

ACCORDS CCTSI

CONSORT extension to SW-CRTs

- CONSORT item 17a: Outcomes and estimation *(Hemming et al., 2018 BMJ)*
  - *CONSORT cluster extension* – Results at the individual or cluster level as applicable and a coefficient of **intracluster correlation** for each primary outcome
  - *Extension for SW-CRTs* – For each primary and secondary outcome, results for each treatment condition, and the estimated effect size and its precision; **any correlations (or covariances)** and time effects estimated in the analysis.
- Reporting any estimated ICCs (and their uncertainty) can be informative for the planning of future trials (CONSORT SW-CRT extension item 7)
- Relatively few studies recognize CAC, and few empirical estimates are currently available (e.g., CLOUD databank study from the Monash group; in press at *Clinical Trials*)



III. Design and Analysis of SW-CRTs with Marginal Models

Alternative to mixed-effects models – marginal models

- Relatively rich literature on SW-CRTs based linear mixed models with a continuous outcome *(Li et al. 2020 SMMR)*

- Generalized linear mixed models with non-identity link?
  - Requires “more effort” for fitting complex random-effects models (e.g., exponential decay)
  - ICC may be defined as a complex function of variance components
  - Interpretation of variance components depends on the choice of link function, and hard to standardize from a reporting perspective

- Marginal models can be attractive because *(Preisser et al., 2003 Stat Med)*
  - **Population-averaged interpretation** of regression parameters – policy implications
  - Separately specify mean model and ICC model – **ICC defined on the natural scale of outcome; easier to standardize reporting**
  - Robust sandwich variance (*empirical option in SAS GLIMIIX*) accounts for ICC model misspecification

Essential ingredients of marginal models for SW-CRTs

- Generalized linear mean model

$$g\{E(Y_{ijk})\} = \mu + \beta_j + \delta X_{ij}$$

  - g is link function
  - μ is the overall mean
  - β_j is fixed categorical secular trend (time effect)
  - δ is the intervention effect

- Working correlation (ICC) model:

$$\boldsymbol{R}_i = corr(\boldsymbol{Y}_i)$$

- Where $\boldsymbol{Y}_i = (Y_{i11}, Y_{i12}, \ldots, Y_{iTN})'$ is the collection of all outcomes in a cluster over all periods

- Estimation and inference of treatment effect and ICCs via the method of **Generalized Estimating Equations (GEE)**

Assume $J = 3$ periods and $(n_{i1}, n_{i2}, n_{i3}) = (2, 2, 3)$ observations for cluster i. Define $\boldsymbol{Y}_i = (Y_{i11}, Y_{i12} | Y_{i21}, Y_{i22} | Y_{i31}, Y_{i32}, Y_{i33})^T$. WP-ICC: α_0; BP-ICC: α_1; decay parameter: ρ

Nested exchangeable (NEX)

$$\begin{pmatrix} 1 & \alpha_0 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_0 & 1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & 1 & \alpha_0 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_0 & 1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & 1 & \alpha_0 & \alpha_0 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_0 & 1 & \alpha_0 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 & \alpha_0 & \alpha_0 & 1 \end{pmatrix}$$

Exponential decay (ED)

$$\begin{pmatrix} 1 & \alpha_0 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho^2 \\ \alpha_0 & 1 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho^2 \\ \alpha_0\rho & \alpha_0\rho & 1 & \alpha_0 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho \\ \alpha_0\rho & \alpha_0\rho & \alpha_0 & 1 & \alpha_0\rho & \alpha_0\rho & \alpha_0\rho \\ \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho & \alpha_0\rho & 1 & \alpha_0 & \alpha_0 \\ \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho & \alpha_0\rho & \alpha_0 & 1 & \alpha_0 \\ \alpha_0\rho^2 & \alpha_0\rho^2 & \alpha_0\rho & \alpha_0\rho & \alpha_0 & \alpha_0 & 1 \end{pmatrix}$$

Sample size for marginal models

- Unlike other contexts, ICC parameters are of interest in SW-CRTs for many reasons
  - Whatever the link function for the mean model, ICCs in marginal models are on the natural scale of outcome
  - Easier to standardize reporting and plug in for sample size calculation

- Sample size formulas and algorithms available for **continuous** and **binary** outcomes, for several correlation models (nested exchangeable and decay) and even for cohort designs

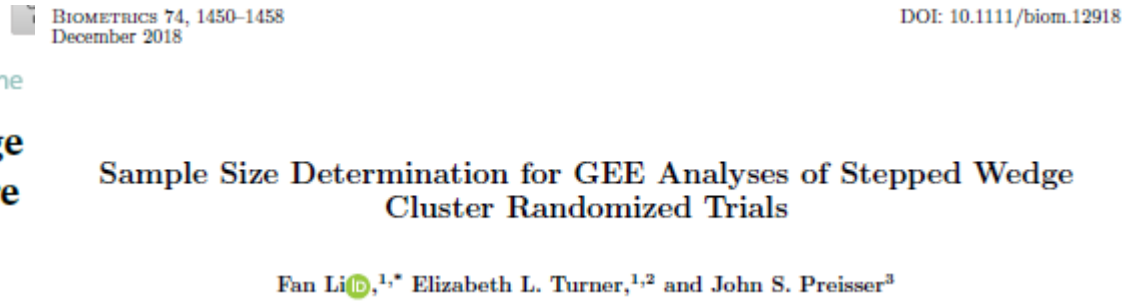

DOI: 10.1002/sim.8415

RESEARCH ARTICLE

WILEY Statistics in Medicine

Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure

Fan Li[1,2]


Biometrics 74, 1450–1458
December 2018

DOI: 10.1111/biom.12918

Sample Size Determination for GEE Analyses of Stepped Wedge Cluster Randomized Trials

Fan Li,[1,*] Elizabeth L. Turner,[1,2] and John S. Preisser[3]

Sample size tools for marginal models

- With continuous outcomes, sample size and power calculation based on GEE are no different from linear mixed models

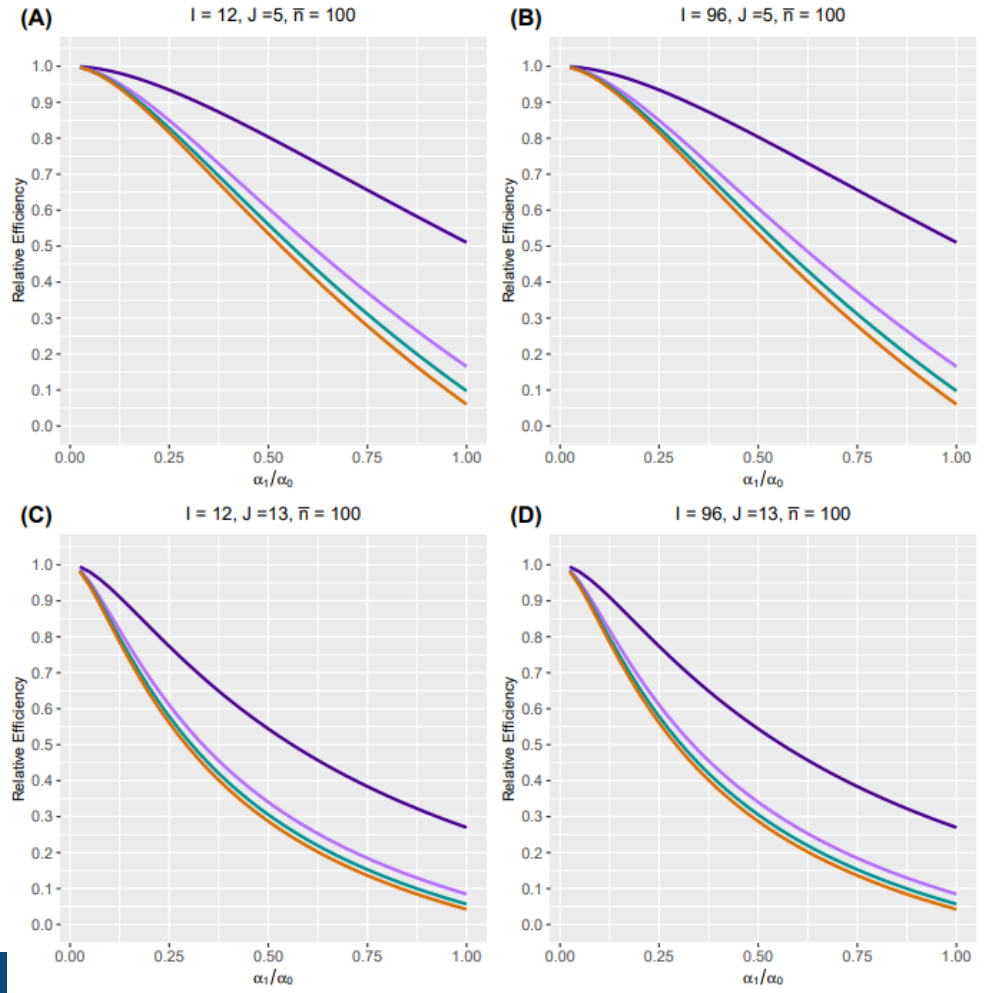
- With binary outcomes, sample size and power calculation based on GEE are currently only implemented in *swdpwr* R package and *%swdpwr* SAS macro *(Chen et al., 2021+)*
  - Currently assume the nested and block exchangeable ICC models *(Li et al., 2018 Biometrics)*

- More software tools are **under development**, with a focus on **binary/count** outcomes and **decaying** correlation models
  - Forthcoming Stata package integrated in the *power* command: *power swgee (Gallis et al., 2021+)*
  - Forthcoming SAS macro and R packages addressing incomplete stepped wedge designs, and cluster size variations

Potential issues with marginal models in SW-CRTs

- Design and analysis should be consistent – same analysis model for sample size and primary analysis
- Despite conceptual advantages, there can be operational challenges
  - **Choice** of working correlation model
  - **Computational scalability** with enormous cluster sizes in pragmatic trials
  - **Concerns** on small-sample validity of the robust sandwich variance
  - **Software** for simultaneously estimating treatment effects and ICCs with GEE

Choice of working correlation model



- **Independence** working correlation
  - Computationally convenient
  - Unbiased, and correlation fixed by sandwich variance
  - Many existing software
- **Nested exchangeable** or **exponential decay** working correlation
  - Requires more computational effort
  - Not as many software
- Independence working correlation leads to inefficient treatment effect estimator even with equal cluster sizes (Tian et al. 2021+; relative efficiency curve shown)

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Choice of working correlation model – implications on design

- Number of clusters required for **Washington State EPT study** under true (independence) ICC models (Tian et al, 2021+)
  - # of periods = 5
  - Mean cluster period sizes ≈ 300
  - Within-period ICC = 0.007; CAC = 0.5 (Nested exchangeable); CAC = 0.7 (exponential decay)
  - Coefficient of variation (CV) measuring between cluster variability in sizes

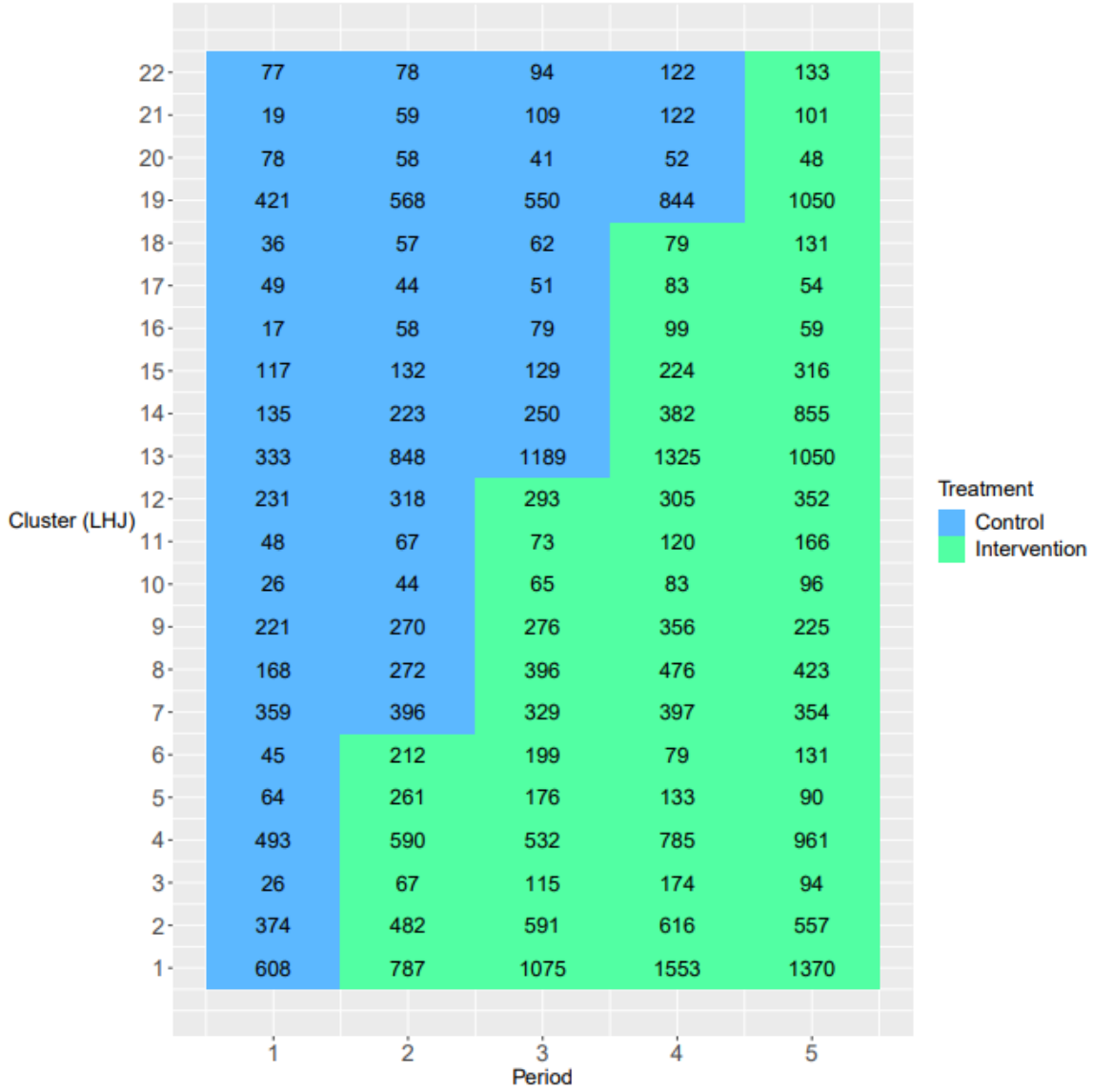
True correlation structure	CV	No within-cluster imbalance	Within-cluster imbalance pattern 2	Within-cluster imbalance pattern 4
Exchangeable	0	11 (31)	11 (32)	11 (33)
	0.25	11 (33)	12 (33)	12 (33)
	0.75	12 (43)	13 (38)	13 (38)
	1.25	13 (64)	17 (48)	17 (48)
Nested exchangeable	0	18 (25)	19 (26)	19 (27)
	0.25	18 (26)	19 (27)	19 (27)
	0.75	20 (34)	21 (32)	21 (32)
	1.25	24 (50)	26 (42)	26 (42)
Exponential decay	0	17 (27)	18 (28)	18 (29)
	0.25	18 (28)	18 (29)	18 (29)
	0.75	19 (37)	21 (34)	21 (34)
	1.25	22 (54)	26 (43)	26 (43)

Large cluster sizes in pragmatic SW-CRTs

- Estimating ICCs through GEE is a computationally challenging task with large cluster sizes

- In the Washington EPT study, the cluster sizes range from 277 to 5393; require enumeration of $\binom{5393}{2} \approx$ **14.5 million** residual cross-products terms in one cluster to form the GEE for ICC parameters

- Without individual-level covariates, we have developed a new GEE approach that takes only **cluster-period means** *(Li et al., 2021 Biostatistics)*
  - Simultaneously estimate treatment effects and (bias-corrected) ICCs along with their standard errors
  - Circumvent computational challenges as the new "cluster size" = number of periods
  - Implemented in a recent R package ***geeCRT***



COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS CCTSI

Small-sample correction to sandwich variance

- Intuition: the so-called sandwich variance (middle part) underestimates the true variance with limited number of clusters (<= 30)

$$\Omega_0 = \sum_{i=1}^{I} C_i D_i' V_i^{-1} B (y_i - \mu_i)(y_i - \mu_i)' B_i' V_i^{-1} D_i C_i,$$

  - **Review of SW-CRTs suggests median # of clusters is only 20.5** (Grayling et al., 2017 Trials)
- Active pursuit even for parallel CRTs *(Li and Redden, 2014 Stat Med; Ford and Westgate, 2017 Biom J)*

TABLE 3 Summary of bias-corrected sandwich variance estimators for $\hat{\theta}$

Label	Correction	C_i	B_i	References
BC0	none	I	I	Liang and Zeger[19]
BC1	less	I	$(I - H_i)^{-1/2}$	Kauermann and Carroll[28]
BC2	more	I	$(I - H_i)^{-1}$	Mancl and DeRouen[29]
BC3	less	$\text{diag}\{(1 - \min\{\zeta, [D_i' V_i^{-1} D_i \Omega_1^{-1}]_{jj}\})^{-1/2}\}$	I	Fay and Graubard[30]

- Investigations on bias-corrected sandwich variance for small SW-CRTs with converging recommendations *(Ford and Westgate, 2020 Stat Med; Thompson et al. 2020 SMMR)*

Small-sample correction to sandwich variance – Cont'd

Ref: *Ford and Westgate, 2020 Stat Med*

TABLE 1 Comparison of studies assessing type I error rates

	Scott et al[7]	Li et al[12]	Li[20]	This article
Motivating SW CRT Design	Cohort in Context of Vaccine Efficacy Trial	Cohort[a]	Cohort	Cross-sectional
Number of Time Periods	10	3-7	3-8	4, 8, 12
Outcome Type	Continuous	Continuous, Binary	Continuous	Binary
Outcome Level	Cluster-Period Means	Individual	Individual	Individual
True Corr Structure	AR-1	Block Exchangeable	Block AR-1[b]	Common Exchangeable, Hooper-Girling[c14,27,28]
Working Corr Structure	AR-1, Exchangeable	Block Exchangeable	Block AR-1	Common Exchangeable
N	10, 20, 50	8-25	9-24	3-24
m	NA (Equal)	4-25	5-24	Equal: 10-25, Unequal: min(*Negbin*(m_i, 0.5), 5)
m- Across Clusters	NA (Equal)	Equal	Equal	Both Equal/Unequal
m- Across Periods	NA (Equal)	Equal	Equal	Equal
BC SEs Compared	mFG/KC[d], MD, FG, MBN	KC, MD, FG	KC, MD, FG	KC, MD, FG, AVG
df Compared	∞ [e], d5 w/FG	∞, *N*-*p*	∞, *N*-*p*, *N*-2	*N*-2, *N*-*p*, PW
Recommended BC SE and df	FG with d5 df[f]	KC or FG, *N*-*p*	KC[g], *N*-2	AVG, *N*-2[h]

Abbreviations: AVG, Average of Mancl and DeRouen[18] and Kauermann and Carroll;[21] BC, bias-corrected; Corr, correlation; df, degrees of freedom; FG, Fay and Graubard;[19] *m*, Cluster Sizes; mFG, modified Fay and Graubard;[31] MBN, Morel et al;[32] MD, Mancl and DeRouen;[18] *N*, number of clusters; *p*, number of regression parameters; PW, Pan and Wall;[22] SE, standard error.

COPRH Con Colorado Pragmatic Research in Health Conference

ACCORDS CCTSI

Software tools for fitting marginal models in SW-CRTs

Several widely-used routines available in SAS and R

- **Does not always take two-level correlation models**
- Not always come with small-sample corrections

The cluster-period GEE is recently implemented in R (**geeCRT**), but currently only supports binary outcomes with logistic link function

- More components to be developed, allowing for other types of outcomes, and ICC models appropriate for SW-CRTs

Software	Feature	Comment
gee/geesmv/saws (R)	ICC model	Support EX but not NEX or ED
	Small-sample variance	MD, KC, FG, MBN and more
geeglm (R)	ICC model	EX and NEX, but not ED
	Small-sample variance	Not supported
GEECORR (SAS Macro)	ICC model	EX and NEX, can adapt for ED
	Small-sample variance	MD, KC
PROC GLIMMIX (SAS)	ICC model	Support EX but not NEX or ED
	Small-sample variance	MD, KC, FG, MBN
xtgeebcv (Stata) (Gallis et al., 2020)	ICC model	Support EX but not NEX or ED
	Small-sample variance	MD, KC, FG, MBN
geeCRT (R) (Yu et al., 2021)	ICC model	**Support EX, NEX and ED**
	Small-sample variance	MD, KC, FG (**for treatment effect and ICC**)

NEX: nested exchangeable / ED: exponential decay
MD: Mancl and DeRouen (2001, Biometrics)
KC: Kauermann and Carroll (2001, JASA)
FG: Fay and Graubard (2001, Biometrics)
MBN: Morel et al. (2003, Biom J)

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

IV. Concluding Remarks

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS

Choice of analytical methods for SW-CRTs

- There is no consensus on “best” models for design and analysis
- Mixed-effects models and marginal models each have their pros and limitations; choice can depend on the desired interpretation and research question
  - Methods themselves are not competing, they offer complementary approaches to address the same problem (estimate treatment effect and report desired correlation parameters)
- Need to consider small sample corrections to maintain valid inference
- CONSORT extension to SW-CRTs recommended clear descriptions of assumptions and model specification; **and to report ICCs** (need more empirical estimates)
- Recommend to work with a statistician starting from the design stage

Common mis-conceptions

- **Mis-conception 1**: By choosing a stepped wedge design, I can avoid logistical challenges
    - While stepped wedge designs have advantages, they can be challenging to implement because
        - One needs to ensure all sites adhere to implementation schedule
        - Can increase the total duration of the study
        - Vulnerable to external interferences
        - Increase the data collection burden
- **Mis-conception 2:** I can run my cluster randomized trials with only 4 clusters as long as my sample size formula shows I have 80% power
    - While there are small-sample corrections that improve inference in the small sample setting
        - Relatively fewer simulations looked at <= 6 clusters for valid sample size and power performance
        - 4 clusters may not have the desired level of power if we start to bring in the concept of CAC (underlying CAC = 1 assumption is not always plausible, but is likely what is needed to observe the above results)
        - Limit generalizability of trial

Slide credited to Dr. Monica Taljaard

Entry point to understanding the evolving literature

International Journal of Epidemiology, 2020, 1043–1052
doi: 10.1093/ije/dyaa077
Advance Access Publication Date: 9 May 2020
Education Corner

Education Corner

Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice?

Karla Hemming [1,*] and Monica Taljaard[2,3]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, I
Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada,
Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

*Corresponding author: Institute of Applied Health Research, Public Health Building, University
Birmingham, B15 2TT, UK. E-mail: k.hemming@bham.ac.uk

Editorial decision 1 April 2020; Accepted 8 April 2020

Abstract

The stepped-wedge cluster randomized trial (SW-CRT) involves the sequentia
of clusters (such as hospitals, public health units or communities) from cont
vention conditions in a randomized order. The use of the SW-CRT is growing
the SW-CRT is at greater risks of bias compared with the conventional para
randomized trial (parallel-CRT). For this reason, the CONSORT extension fo

Review Article

Statistical Methods in Medical Research
2021, Vol. 30(2) 612–639

Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview

Fan Li[1,2], James P Hughes[3], Karla Hemming[4],
Monica Taljaard[5], Edward R. Melnick[6] and Patrick J Heagerty[3]

Abstract

The stepped wedge cluster randomized design has received increasing attention in pragm
mentation science research. The key feature of the design is the unidirectional crossover of
intervention conditions on a staggered schedule, which induces confounding of the inter
stepped wedge design first appeared in the Gambia hepatitis study in the 1980s. However,
the design and analysis was not formally introduced until 2007 in an article by Hussey and H
mixed-effects model extensions have been proposed for the design and analysis of these tri
these extensions under a unified perspective. We provide a general model representatio
extensions as alternative ways to characterize the secular trend, intervention effect, as well
We review the key model ingredients and clarify their implications for the design and anal
entry point to the evolving statistical literatures on stepped wedge designs.

Keywords

Cluster randomized trials, group-randomized trials, heterogeneity, intraclass correlatio
regression, pragmatic clinical trials, sample size calculation

International Journal of Epidemiology, 2020, 979–995
doi: 10.1093/ije/dyz237
Advance Access Publication Date: 22 February 2020
Original article

Methods

A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator

Karla Hemming [1]*, Jessica Kasza [2], Richard Hooper,[3]
Andrew Forbes[2] and Monica Taljaard[4,5]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK, [2]Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, VIC, Australia, [3]Pragmatic Clinical Trials Unit, Centre for Primary Care and Public Health, Queen Mary University of London, London, UK, [4]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada and [5]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

*Corresponding author. Institute of Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK. E-mail: k.hemming@bham.ac.uk

Editorial decision 10 October 2019; Accepted 11 November 2019

Abstract

It has long been recognized that sample size calculations for cluster randomized trials require consideration of the correlation between multiple observations within the same cluster. When measurements are taken at anything other than a single point in time, these correlations depend not only on the cluster but also on the time separation between measurements and additionally, on whether different participants (cross-sectional designs) or the same participants (cohort designs) are repeatedly measured. This is particularly relevant in trials with multiple periods of measurement, such as the cluster

References

- Hemming, K., & Taljaard, M. (2020). Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice?. International journal of epidemiology, 49(3), 1043-1052.
- Taljaard, M., Teerenstra, S., Ivers, N. M., & Fergusson, D. A. (2016). Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. Clinical Trials, 13(4), 459-463.
- Li, F., Hughes, J. P., Hemming, K., Taljaard, M., Melnick, E. R., & Heagerty, P. J. (2020). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. Statistical Methods in Medical Research, 096228022093296 2.
- Golden, M. R., Kerani, R. P., Stenger, M., Hughes, J. P., Aubin, M., Malinski, C., & Holmes, K. K. (2015). Uptake and population-level impact of expedited partner therapy (EPT) on Chlamydia trachomatis and Neisseria gonorrhoeae: the Washington State community-level randomized trial of EPT. PLoS Med, 12(1), e1001777.
- Bennett, P. N., Daly, R. M., Fraser, S. F., Haines, T., Barnard, R., Ockerby, C., & Kent, B. (2013). The impact of an exercise physiologist coordinated resistance exercise program on the physical function of people receiving hemodialysis: a stepped wedge randomised control study. BMC nephrology, 14(1), 1-7.
- Copas, A. J., Lewis, J. J., Thompson, J. A., Davey, C., Baio, G., & Hargreaves, J. R. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials, 16(1), 1-12.

References – Cont'd

- Barker, D., McElduff, P., D'Este, C., & Campbell, M. J. (2016). Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC medical research methodology, 16(1), 1-19.
- Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials, 28(2), 182-191.
- Hooper, R., Teerenstra, S., de Hoop, E., & Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in medicine, 35(26), 4718-4728.
- Kasza, J., Hemming, K., Hooper, R., Matthews, J. N. S., & Forbes, A. B. (2019). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Statistical methods in medical research, 28(3), 703-716.
- Kasza, J., & Forbes, A. B. (2019). Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. Statistical methods in medical research, 28(10-11), 3112-3122.
- Hemming, K., & Girling, A. (2014). A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. The Stata Journal, 14(2), 363-380.
- Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E., & Omar, R. Z. (2015). Sample size calculation for a stepped wedge trial. Trials, 16(1), 1-15.

References – Cont'd

- Voldal, E. C., Hakhu, N. R., Xia, F., Heagerty, P. J., & Hughes, J. P. (2020). swCRTdesign: An R Package for Stepped Wedge Trial Design and Analysis. Computer methods and programs in biomedicine, 196, 105514.
- Chen, J., Zhou, X., Li, F., & Spiegelman, D. (2020). swdpwr: A SAS Macro and An R Package for Power Calculation in Stepped Wedge Cluster Randomized Trials. arXiv preprint arXiv:2011.06031.
- Hemming, K., Kasza, J., Hooper, R., Forbes, A., & Taljaard, M. (2020). A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. International journal of epidemiology, 49(3), 979-995.
- Zhou, X., Liao, X., Kunz, L. M., Normand, S. L. T., Wang, M., & Spiegelman, D. (2020). A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. Biostatistics, 21(1), 102-121.
- Martin, J., Girling, A., Nirantharakumar, K., Ryan, R., Marshall, T., & Hemming, K. (2016). Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials, 17(1), 1-12.
- Hemming, K., Taljaard, M., McKenzie, J. E., Hooper, R., Copas, A., Thompson, J. A., ... & Grimshaw, J. M. (2018). Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. bmj, 363.
- Preisser, J. S., Young, M. L., Zaccaro, D. J., & Wolfson, M. (2003). An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Statistics in medicine, 22(8), 1235-1254.

References – Cont'd

- Li, F., Turner, E. L., & Preisser, J. S. (2018). Sample size determination for GEE analyses of stepped wedge cluster randomized trials. Biometrics, 74(4), 1450-1458.
- Li, F. (2020). Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. Statistics in medicine, 39(4), 438-455.
- Li, F., Yu, H., Rathouz, P. J., Turner, E. L., & Preisser, J. S. (2021). Marginal modeling of cluster-period means and intraclass correlations in stepped wedge designs with binary outcomes. Biostatistics.
- Tian, Z., Preisser, J., Esserman, D., Turner, E., Rathouz, P., & Li, F. (2021). Impact of unequal cluster sizes for GEE analyses of stepped wedge cluster randomized trials with binary outcomes. medRxiv.
- Ford, W. P., & Westgate, P. M. (2020). Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. Statistics in Medicine, 39(21), 2779-2792.
- Thompson, J. A., Hemming, K., Forbes, A., Fielding, K., & Hayes, R. (2020). Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: A simulation study. Statistical Methods in Medical Research, 09622802209
- Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L., & Teerenstra, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. Journal of clinical epidemiology, 66(7), 752-758.

References – Cont'd

- Gallis, J. A., Li, F., & Turner, E. L. (2020). xtgeebcv: a command for bias-corrected sandwich variance estimation for Gee analyses of cluster randomized trials. The Stata Journal, 20(2), 363-381.
- Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. Biometrics, 57(1), 126-134.
- Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association, 96(456), 1387-1396.
- Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics, 57(4), 1198-1206.
- Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. Biometrical Journal: journal of mathematical methods in biosciences, 45(4), 395-409.
- Shing, T. L., Preisser, J. S., & Zink, R. C. (2020). GEECORR: A SAS macro for regression models of correlated binary responses and within-cluster correlation using generalized estimating equations. arXiv preprint arXiv:2011.11592.

COPRH Con
Colorado Pragmatic Research in Health Conference

ACCORDS