Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

On Small-Sample Inference in Multiphase Stepped Wedge Cluster Randomized Trial (MSW-CRT) With Binary Outcomes

Jing Peng (Jing.Peng@osumc.edu)
The Ohio State University Wexner Medical Center
Abigail Shoben
The Ohio State University
Pengyue Zhang
Indiana University
Philip M. Westgate
University of Kentucky
Soledad Fernandez
The Ohio State University Wexner Medical Center

Research Article

Keywords: multi-phase stepped wedge design, small sample size, variance correction, add-on effect

Posted Date: October 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-993153/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

On small-sample inference in multiphase stepped wedge cluster randomized trial (MSW-CRT) with binary outcomes

Jing Peng [2] *, Abigail Shoben[1], Pengyue Zhang[3], Philip M. Westgate[4] and Soledad Fernandez[2]

1. Department of Biostatistics, College of Public Health, Ohio State University

2. Department of Biomedical Informatics, Ohio State University

3. Department of Biostatistics & Health Data Sciences, Indiana University

4. Department of Biostatistics, College of Public Health, University of Kentucky

*: to whom the correspondence should be addressed (Jing.Peng@osumc.edu).

Abstract

Background

The stepped wedge cluster randomized trial (SW-CRT) design is now preferred for many health-related trials because of its flexibility on resource allocation and clinical ethics concerns. However, as a necessary extension of studying multiple interventions, multiphase stepped wedge designs (MSW-CRT) have not been studied adequately. Since estimated intervention effect from Generalized estimating equations (GEE) has a population-average interpretation, valid inference methods for binary outcomes based on GEE are preferred by public health policy makers.

Methods

We form hypothesis testing of add-on effect of a second treatment based on GEE analysis in an MSW-CRT design with limited number of clusters. Four variance-correction estimators are used to adjust the bias of the sandwich estimator. Simulation studies have been used to compare the statistical power and type I error rate of these methods under different correlation matrices.

Results

We demonstrate that an average estimator with $t(I-3)$ can stably maintain type I error close to the nominal level with limited sample sizes in our settings. We show that power of testing the add-on effect depends on the baseline event rate, effect sizes of two interventions and the number of clusters. Moreover, by changing the design with including more sequences, power benefit can be achieved.

Conclusions

For designing the MSW-CRT, we suggest using more sequences and checking event rate after initiating the first intervention via interim analysis. When the number of clusters is not very large in MSW-CRTs, inference can be conduct using GEE analysis with an average estimator with $t(I-3)$ sampling distribution.

Keywords: multi-phase stepped wedge design, small sample size, variance correction, add-on effect.

Background

As firstly applied in the hepatitis B (HBV) study (Hall et al., 1987), stepped wedge cluster randomized trials (SW-CRTs) have gained much popularity based on both intervention initiation resources and clinical ethics perspectives. Since SW-CRTs allow to roll out the intervention to two or more communities in stages, they are especially useful under labor and financial constraints (Cook and Campbell, 1979). Additionally, the participating communities in SW-CRTs keep receiving the intervention once they cross over from the placebo group to the intervention group. Eventually, all the participants will receive the intervention. Therefore, the SW-CRT is more clinically ethical than the parallel design or the crossover design when the intervention is believed to be beneficial (Brown and Lilford, 2006; Hemming et al., 2015a). The traditional SW-CRT design is the cluster randomized trial (CRT) that measures participant-level outcomes (observation level) within units such hospital, schools, etc. (cluster level). In the conventional SW-CRT design, the clusters are randomly assigned to initiate the intervention at pre-determined time points (Figure 1A). Moreover, clusters can be grouped in sequences for randomization and transition from the placebo to the intervention. Comparison between sequences can help ameliorate the impact of the contamination between clusters (Hemming et al., 2015b).

A

0	A	A	A
0	0	A	A
0	0	0	A

B

0	A	A	A	A+B	A+B	A+B
0	0	A	A	A	A+B	A+B
0	0	0	A	A	A	A+B

Figure 1A. Traditional stepped wedge cluster randomized trial with one intervention, 3 sequences (rows) and 4 periods (columns) are used. Multiple clusters can be grouped in each sequence, **B.** The multiphase stepped wedge cluster randomized trial (MSW-CRT) used in the simulations of this study. Treatment B will be implemented only after all the clusters in the sequences receive treatment A.

Traditional SW-CRTs utilize a single-phase design by considering only one intervention during the study. However, different interventions may be combined to treat participants (Parker et al., 2016; Pears et al., 2016; Salmon et al., 2011), so the MSW-CRT design becomes a necessary extension to study effects from multiple interventions. Lyons et al. (2017) then introduced the MSW-CRT. They pointed out that this design can help reduce the study time and cost, increase engagement among participants and improve efficiency. Moreover, it allows assessment of possible interaction effects between multiple treatments and requires fewer number of clusters compared with running two separate trials. Figure 1B shows a simplest case of MSW-CRTs—treatment B will be sequentially implemented to each sequence after treatment A. This design therefore will allow for estimation of the add-on effect of the second treatment (treatment B). The inference of continuous outcomes in MSW-CRTs is completed by our group and under review, in which the closed-form sample size and power formulas are derived based on the linear mixed model. The properties between design parameters and statistical power are shown in that study, which provide general guidance for the MSW-CRT study design. Even though we have investigated the continuous outcomes, the inference method for binary outcomes in MSW-CRTs is still less well developed. Binary outcomes, however, are very common in health-related trials, and different from the continuous ones on modeling and analysis.

Since the responses observed from the same cluster are usually correlated, two types of statistical models are commonly used for analysis: conditional and marginal models. Though they are distinctive on other aspects like model assumptions, one important difference is the interpretation of parameters (Preisser et al., 2003). For SW-CRTs, while the treatment effects based on the conditional model are interpreted as cluster-specific changes, the treatment effects estimated from the marginal model are interpreted as the population-average changes. Therefore, generalized linear mixed models are often used for inferences on the conditional effects (Hussey and Hughes, 2007; Hemming et al., 2015; Hooper et al., 2016; Woertman et al., 2013). However, in health population sciences research, the population level interpretation from marginal models may be preferred for making health policies (Li et al., 2018). To evaluate the statistical power of testing the treatment effects in SW-CRT designs, a marginal model is commonly assumed with a constant exchangeable intraclass correlation (ICC) and a logit link for binary responses (Crespi

et al., 2009). GEE is then used to fit this model and make inferences (Li et al., 2018; Li, 2020; Scott et al., 2017; Ford and Westgate, 2020).

Though some SW-CRTs (Li et al. 2018; Ford and Westgate, 2020; Li, 2020; Thompson et al. 2020) have been conducted using marginal models with binary outcomes, binary responses from MSW-CRTs have not been studied. Therefore, the focus of this paper is to demonstrate appropriate inference methods for the analysis of binary subject-level outcomes arising from MSW-CRTs. In the following three sections, we will introduce the GEE analysis for MSW-CRT designs with binary outcomes and conduct simulations to assess the performance of inference methods using Mancl and DeRouen (2001) (MD estimator), Kauermann and Carroll (2001) (KC estimator), average of MD and KC estimator (Ford and Westgate, 2017) and Fay and Graubard (2001) (FG estimator). The Simulation study section will describe our simulation settings, hypothesis, and planned analysis.

Methods

GEE analyses of MSW-CRT

In this study, we consider a cross sectional stepped wedge design with a placebo phase and two consecutive intervention phases, A and B. For example, Figure 1B shows an MSW-CRT with 7 periods and 3 sequences, where each cross section defines one period. Each sequence consists of group of clusters receiving the same treatment in every period. In this example, we have two interventions: A and B, while 0 refers to placebo or control. In this example design intervention B is initiated only after all the clusters receive intervention A. We denote the binary outcome of interest to be Y_{ijk}, representing the measurement for the kth individual in the ith cluster at the jth period. The marginal means $\pi_{ij} = E(Y_{ijk})$ can be modeled as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + A_{ij}\theta_1 + B_{ij}\theta_2 + t_j, \;\; i = 1, 2, \dots, I; j = 1, 2, \dots, T; k = 1, 2, \dots, N_i, \qquad (1)$$

where a_{ij} and b_{ij} are the treatment indicators of receiving the intervention A or B (a_{ij}= 1 if the cluster i receives treatment A at period j, and a_{ij}= 0 otherwise), respectively; and μ is the log odds of the baseline event rate. Therefore, θ_1 is the treatment effect of A; θ_2 is the add-on effect of treatment B; and t_j is the period effect (t_1 = 0 for identifiability). Considering the multi-phase design may usually happen when the first intervention is shown to be beneficial, our model is designed to answer the specific question: is it more beneficial by adding the second intervention B than using the first intervention A alone? Therefore, we did not include the interaction term between two interventions in our model. We denote the parameter vector as $\boldsymbol{\beta} = (\mu, t_2, \dots, t_T, \theta_1, \theta_2)'$ in the mean model (1). For cluster i, the vector of outcomes is $\boldsymbol{Y_i} = \left(Y_{i11}, Y_{i12}, \dots, Y_{iTN_i}\right)'$, and thus the marginal mean $TN_i \times 1$ vector is $\boldsymbol{\pi_i} = (\pi_{i1}, \pi_{i1}, \dots, \pi_{iT})'$. The $TN_i \times (T+2)$ design matrix $\boldsymbol{X_i}$ can be written based on model (1) and the pre-defined design. We can then write model (1) as:

$$logit(\boldsymbol{\pi_i}) = \boldsymbol{X_i\beta}, \quad i = 1, 2, \dots, I.$$

The marginal variance in our case is $v\left(\pi_{ij}\right) = \pi_{ij}\left(1 - \pi_{ij}\right)$, and we can obtain $\widehat{\boldsymbol{\beta}}$ by solving:

$$\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{Y_i} - \boldsymbol{\pi_i}) = 0. \qquad (2)$$

where $\boldsymbol{D_i} = \frac{\partial \boldsymbol{\pi_i}}{\partial \boldsymbol{\beta^T}}$, and $\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i \boldsymbol{A}_i^{\frac{1}{2}}$ is the working covariance matrix for $\boldsymbol{Y_i}$, with $\boldsymbol{A}$ as the $TN_i \times TN_i$ diagonal matrix with elements $v\left(\pi_{ij}\right)$ (setting the dispersion parameter to be 1) and $\boldsymbol{R}_i$ as the working correlation matrix. Similar to the Hussey and Hughes model (Hussey and Hughes, 2007) for continuous outcomes, we assume common exchangeable structure with a uniform correlation parameter α. We just consider cross sectional sampling here, for making the constant correlation assumption reasonable (Li et al., 2021). We note that the constant correlation is restrictive, and it is suggested using multiple correlation parameters for longitudinal CRT designs (Li, 2020; Hemming et al., 2020). For example, different within-period correlation α_0 and intra-period correlation α_1 can be included for sample size calculation. Specifically, α_0 is used for quantifying the similarity between individuals in the same cluster at the same period, while α_1 is used to represent the similarity between individuals in the same cluster across

different periods. Therefore, we studied the impact of different correlation structures in the following sections.

Since the variance function is related to the mean for the binary responses, the analytical solution of equation (2) cannot be obtained, and iterative reweighted least squares (IRLS) is used to obtain a numerical solution. When the number of clusters is large enough, the estimator $\widehat{\boldsymbol{\beta}}$ can be assumed to be multivariate normally distributed with mean $\boldsymbol{\beta}$. The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be estimated using the model-based variance:

$$\widehat{\boldsymbol{\Sigma}_{\mathbf{1}}} = \left\{\sum_{i=1}^{I} \boldsymbol{D}_i'(\widehat{\boldsymbol{\beta}})\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\alpha}})\boldsymbol{D}_i(\widehat{\boldsymbol{\beta}})\right\}^{-1} \tag{3}$$

or the empirical sandwich estimator:

$$\widehat{\boldsymbol{\Sigma}_{\mathbf{2}}} = \widehat{\boldsymbol{\Sigma}_{\mathbf{1}}} \left\{\sum_{i=1}^{I} \boldsymbol{D}_i'(\widehat{\boldsymbol{\beta}})\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\alpha}})\boldsymbol{r}_i\boldsymbol{r}_i'\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\alpha}})\boldsymbol{D}_i(\widehat{\boldsymbol{\beta}})\right\}^{-1} \widehat{\boldsymbol{\Sigma}_{\mathbf{1}}}\ , \tag{4}$$

where $\boldsymbol{r}_i = \boldsymbol{Y_i} - \widehat{\boldsymbol{\pi}}_i$ is the residual vector of the ith cluster. The statistical consistency of the model-based variance estimator relies on the correct specification of the working correlation structure, so the robust sandwich variance estimator is preferred. However, several studies have shown that when the number of clusters is limited, the fitted mean $\widehat{\boldsymbol{\pi}}_i$ tends to be close to $\boldsymbol{Y_i}$, leading to negatively biased sandwich estimators due to the small residuals (Mancl and DeRouen, 2001; Lu et al., 2007; Murray et al., 2004). To correct this bias, several correction methods have been proposed for valid inference. In this paper, we considered three commonly used bias correction estimators: Mancl and DeRouen (2001) (MD estimator), Kauermann and Carroll (2001) (KC estimator), and Fay and Graubard (2001) (FG estimator). We also include another new method by averaging the MD and KC estimators (Average estimator), which has been shown to achieve competitive performance than the above three estimators (Ford and Westgate, 2017). We choose standard normal for a Wald test or a t test using t distribution with degrees of freedom $I - 3$, which is the number of clusters minus the number of cluster-level regression parameters. Tests with degrees of freedom $I - 3$ have shown good performance in parallel CRT settings (Ford and Westgate, 2017; Li and Redden, 2015), while it leads to

conservative tests using $I - p$ as the degrees of freedom, with p as the total number of regression parameters (Ford and Westgate, 2020).

Simulation study

In this section, we describe the simulation studies under the two correlation structures. In section 1.1, datasets are simulated by specifying correlation structure for the outcomes as common exchangeable. In this section, we also introduce the studying questions, simulation settings (section 1.1.1), data generating process (section 1.1.2) and the analysis method (section 1.1.3). In section 1.2, we simulate data under the common block exchangeable correlation structure with two correlation parameters. Simulations are designed by including the two correlation parameters. We describe the setting difference in this section, while studying questions, data generation and analysis methods are the same with those in section 1.1.

1.1 Inference with a constant correlation parameter

In this section, the correlation structure is set to be exchangeable with a uniform parameter. We define the effect size as $1 - e^{\theta_i}$, and it can be interpreted as the extent to which the intervention decreases the odds of events. For instance, $e^{\theta_1} = 0.75$ means that treatment A decreases the odds of death by 25%, relative to the control group. We propose simulations to answer the following three questions:

1. For the intervention with null effect size ($e^{\theta_i} = 1$), is the type I error at its nominal level 0.05?
2. For the intervention with non-zero effect size ($e^{\theta_i} < 1$), how does the statistical power change with different simulation parameter values? Which tests can achieve the predictive power under different settings? Is there a big power loss by using conservative tests?
3. What is the bias of the different variance estimators (compared to the sample variance obtained from the Monte Carlo replicates) under different settings?

1.1.1 Simulation settings

To study the type I error, the parameters in six different scenarios are set as shown in Table 1.

Table 1: Six simulation scenarios for studying type I error of testing treatment effect of intervention B using a constant correlation parameter. In the 6 scenarios, we vary number of clusters in each sequence (I), different coefficients of variation for cluster sizes (cv), different ICCs (α), different baseline event rates (baseline), different designs in Figure 2 (Design), and varying effect sizes (1-effect size), respectively.

Scenario	I	cv	α	baseline	design (Figure 2)	1- effect size
1	12,18,36,60	0.4	0.2	0.2	1B	$e^{\theta_1} = 0.75, e^{\theta_2} = 1$
2	18	(0,1,0.1)	0.2	0.2	1B	
3	18	0.4	(0.01,0.1,0.02)	0.2	1B	
4	18	0.4	0.2	(0.1,0.5,0.1)	1B	
5	24	0.4	0.2	0.2	2A-2F	
6	18	0.4	0.2	0.2	1B	$e^{\theta_1} = (0.4,0.8,0.05)$, $e^{\theta_2} = 1$

For simplicity, we set the number of clusters to be the same in each sequence, and all the clusters in the same sequence receive the same intervention(s) in every period. In each scenario, one parameter is varied to study the impact on empirical test sizes. For example, in scenario 2, (0,1,0.1) indicates that the coefficient of variation of the cluster size (cv) is changed from 0 to 1 by 0.1 increments. Other variables of interest are the number of clusters (I), ICC (α), baseline event rate, designs with different number of sequences (see Figure 2), and effect sizes. We run a Monte Carlo simulation 5,000 times for each cv, and then compare the empirical type I error rate for testing the null add-on treatment effect. In this way, we can get an idea that how the type I error could change with parameters. Similarly, six scenarios are shown in Table S1 (see the Appendix) to study the empirical power of testing the add-on effect. All the parameter settings in Table S1 are the same with Table 1 except the effect size for intervention B. We use null effect size for intervention B in scenarios in Table 1 but nonzero effect sizes in Table S1.

A

Sequence	Periods							
	1	2	3	4	5	6	7	8
1	0	A	A	A	A+B	A+B	A+B	A+B
2	0	0	A	A	A	A+B	A+B	A+B

B

Sequence	Periods							
	1	2	3	4	5	6	7	8
1	0	A	A	A	A+B	A+B	A+B	A+B
2	0	0	A	A	A	A+B	A+B	A+B
3	0	0	0	A	A	A	A+B	A+B

C

Sequence	Periods							
	1	2	3	4	5	6	7	8
1	0	A	A	A	A+B	A+B	A+B	A+B
2	0	0	A	A	A	A+B	A+B	A+B
3	0	0	0	A	A	A	A+B	A+B
4	0	0	0	0	A	A	A	A+B

D

Sequence	Periods					
	1	2	3	4	5	6
1	0	A	A	A+B	A+B	A+B
2	0	0	A	A	A+B	A+B

E

Sequence	Periods					
	1	2	3	4	5	6
1	0	A	A	A+B	A+B	A+B
2	0	0	A	A	A+B	A+B
3	0	0	0	A	A	A+B

F

Sequence	Periods					
	1	2	3	4	5	6
1	0	A	A+B	A+B	A+B	A+B
2	0	0	A	A+B	A+B	A+B
3	0	0	0	A	A+B	A+B
4	0	0	0	0	A	A+B

Figure 2A-C. Designs used in the simulations. Designs with two or three sequences allow for more periods after all clusters receive the interventions. Designs with four sequences (**C**) will initiate intervention B before all clusters receive treatment A. **D-F** are the designs used in the simulation when the total study period is limited.

1.1.2 Data generation

For data simulation, the cluster mean at each period π_{ij} is specified by model 1 with $t_j = 0$ because we do not focus on exploring the impact of the period effect. We simulate correlated individual-level binary outcomes from each cluster using the EP method (Emrich and Piedmonte, 1991), which can be easily implemented in the R package "mvtBinaryEP" (By and Qaqish, 2011). The algorithm first simulates a multivariate normal variable and then dichotomizes the coordinates to get the binary outcomes. We note here that since the correlation matrix for simulating multivariate normal data is obtained by solving equations with π_{ij} and α, the original R program may fail if the resulting correlation matrix is not positive semidefinite. In that case, we replace it by the nearest positive definite matrix using the algorithm proposed by Higham (2002).

1.1.3 Analysis

Considering MSW-CRTs are beneficial as add-on trials, we focus on testing the add-on effect of the second treatment B in this study. Therefore, we evaluate test sizes of the following hypothesis testing strategy:

$$H_0: \theta_2 = 0 \; v.s. \; H_A: \theta_2 \neq 0.$$

For each dataset simulated under the 12 different settings in Tables 1 and S1, model 1 is used for GEE analysis. We compare Wald tests with the four variance correction estimators: Mancl and DeRouen (2001) estimator (MD_w), Kauermann and Carroll (2001) estimator (KC_w), Fay and Graubard (2001) estimator (FG_w), and average of MD and KC estimator (Average_w) (Ford and Westgate, 2017). Additionally, we choose $t(I-3)$ as another sampling distribution for further coverage adjustment (MD_t and FG_t), since these two estimators are observed to overestimate variances in our simulations. Here we have three cluster-level parameters in model 1 (μ, θ_1, θ_2) with $t_j = 0$ for all j, so we use $I-3$ as the degrees of freedom. Power calculation can be done by first computing the model-based variance using parameter values in Table S1:

$$\boldsymbol{\Sigma_1} = \{\sum_{i=1}^{I} \boldsymbol{D}_i'(\boldsymbol{\beta}) \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{D}_i(\boldsymbol{\beta})\}^{-1},$$

and then:

$$\mathbf{1}-\boldsymbol{\beta}=\boldsymbol{\Phi}\left(\boldsymbol{Z}_{\frac{\boldsymbol{\alpha}}{\mathbf{2}}}+\frac{|\boldsymbol{\theta}_{\mathbf{2}}|}{\sqrt{\boldsymbol{var}\left(\widehat{\boldsymbol{\theta}_{\mathbf{2}}}\right)}}\right)=\boldsymbol{\Phi}\left(\boldsymbol{Z}_{\frac{\boldsymbol{\alpha}}{\mathbf{2}}}+\frac{|\boldsymbol{\theta}_{\mathbf{2}}|}{\sqrt{\boldsymbol{\Sigma}_{\mathbf{1}(\boldsymbol{T}+\mathbf{2},\boldsymbol{T}+\mathbf{2})}}}\right) \quad (\mathbf{5}),$$

where $\boldsymbol{\Sigma}_{\mathbf{1}(\boldsymbol{T}+\mathbf{2},\boldsymbol{T}+\mathbf{2})}$ is the $(T+2, T+2)$ element of $\boldsymbol{\Sigma}_{\mathbf{1}}$.

1.2 Inference with two correlation parameters

Even though constant ICC can be used under restrictive assumptions, it has been argued that the within-period and intra-period correlations should be treated as two types of intraclass correlations and common block exchangeable correlation structure is preferred in GEE analysis (Martin et al., 2016; Li et al., 2018; Ford and Westgate, 2020). Thus, to study the impact on design and analysis of assuming a different correlation structure, we simulated data under common block exchangeable correlation structure. That means $\boldsymbol{R}_i$ is characterized by two parameters in this case: α_0 − within-period correlation quantifying the similarity between responses from different individuals in the same cluster in the same period, and α_1 − intra-period correlation quantifying the similarity between responses from different individuals in the same cluster but in different periods. In other words, $corr\left(Y_{ijk}, Y_{ijk'}\right)=\alpha_0$ and $corr\left(Y_{ijk}, Y_{ij'k'}\right)=\alpha_1$. Specifically, previous parallel CRTs reported small values of α_0 (Murray et al., 1998) and the intra-period correlation α_1 was usually smaller than α_0 (Martin et al., 2016).

Twelve simulation settings are proposed to study type I error and empirical power when the working correlation is incorrectly specified as common exchangeable with a uniform parameter. Tables 2 and S2 (see the Appendix) show the simulation settings, in which we change the correlation parameter to be both α_0 and α_1, but keep the other configurations the same as in Tables 1 and S1, respectively. Specifically, we use (α_0, α_1) = (0.03, 0.015) for scenarios 1, 2, 4, 5, and 6. Also, we study a range of different correlation values with α_0 = 0.03, 0.06 or 0.1 and α_1= (0.01, α_0, 0.01). For example, when α_0 = 0.03, then three correlation schemes are considered: (0.03, 0.01), (0.03, 0.02), (0.03, 0.03). Similarly, we can get other 16 specifications of the correlation values. The details on the 19 combinations of the two correlation parameters can be found in Table S3 in the Appendix. For computing the theoretical power under these

settings, we still use equation (5), but the correlation matrix for computing $\mathbf{\Sigma_1}$ should be common block exchangeable.

Table 2: Six simulation scenarios for studying type I error of testing treatment effect of intervention B using two correlation parameters in the common block exchangeable correlation structure (α_0, α_1).

Scenario	I	cv	(α_0, α_1)	baseline	design (Figure 2)	1- effect size
1	12,18,36,60	0.4	(0.03,0.015)	0.2	1B	$e^{\theta_1} = 0.75, e^{\theta_2} = 1$
2	18	(0,1,0.1)	(0.03,0.015)	0.2	1B	
3	18	0.4	α_0=0.03,0.06,0.1 α_1=(0.01, α_0,0.01)	0.2	1B	
4	18	0.4	(0.03,0.015)	(0.1,0.5,0.1)	1B	
5	24	0.4	(0.03,0.015)	0.2	2A-2F	
6	18	0.4	(0.03,0.015)	0.2	1B	$e^{\theta_1} = (0.4,0.8,0.05)$, $e^{\theta_2} = 1$

Results

Inference with a constant correlation parameter

Figure 3A-F shows the empirical type I error rate for testing null treatment effect of B in the 6 scenarios shown in Table 1. We find that Average_t controls the type I error rate best overall, while the Wald tests MD_w, FG_w, KC_w and Average_w lead to inflated type I error. KC_t produces empirical type I error close to nominal level when the coefficient of variation of cohort size is small (cv < 0.3). The other two t tests FG_t and MD_t produce empirical type I error lower than expected, leading to conservative inference. As the number of clusters increases, all the correction methods control type I error rate better. When the total number of clusters is 60, all the correction methods have very similar empirical type I error rate. Similar results can be seen by increasing the baseline event rate. Other factors (ICC, designs, effect size) are much less influential on type I error.

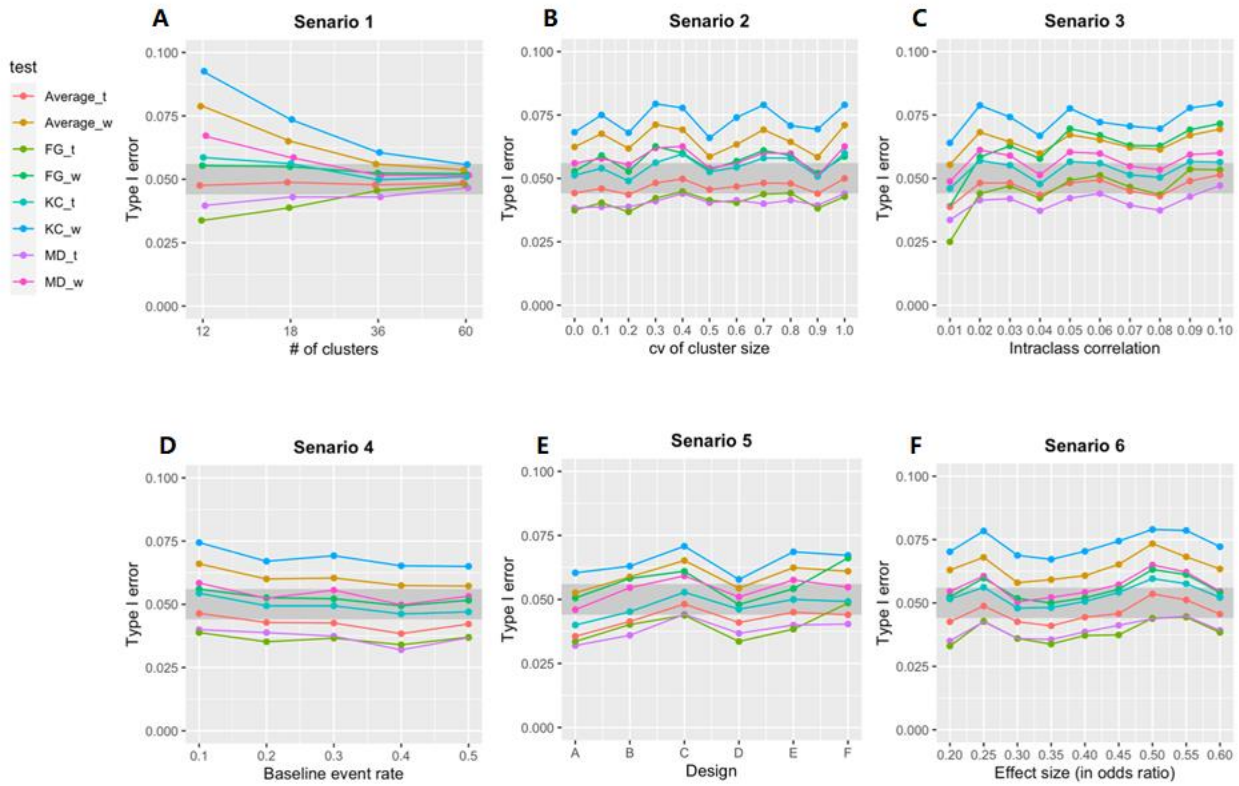

Figure 3A-F. The empirical type I error rate for testing the treatment effect of B in six different scenarios in Table 1 using the true working correlation structure. The gray band in each plot represents 95% confidence interval of the nominal type I error rate 0.05.

Figure 4A-F shows the empirical power with the theoretical power (from equation 5) for testing non-zero treatment effect of B using 5,000 simulated datasets. The Wald tests are not included for power comparison due to inflated type I error rates (Figure 3). We observed that when the working correlation is correctly specified, KC_t is closest to the theoretical value. Other tests (MD_t, FG_t, Average_t) are just slightly less powerful, indicating no significant power loss when we use conservative tests under these settings.

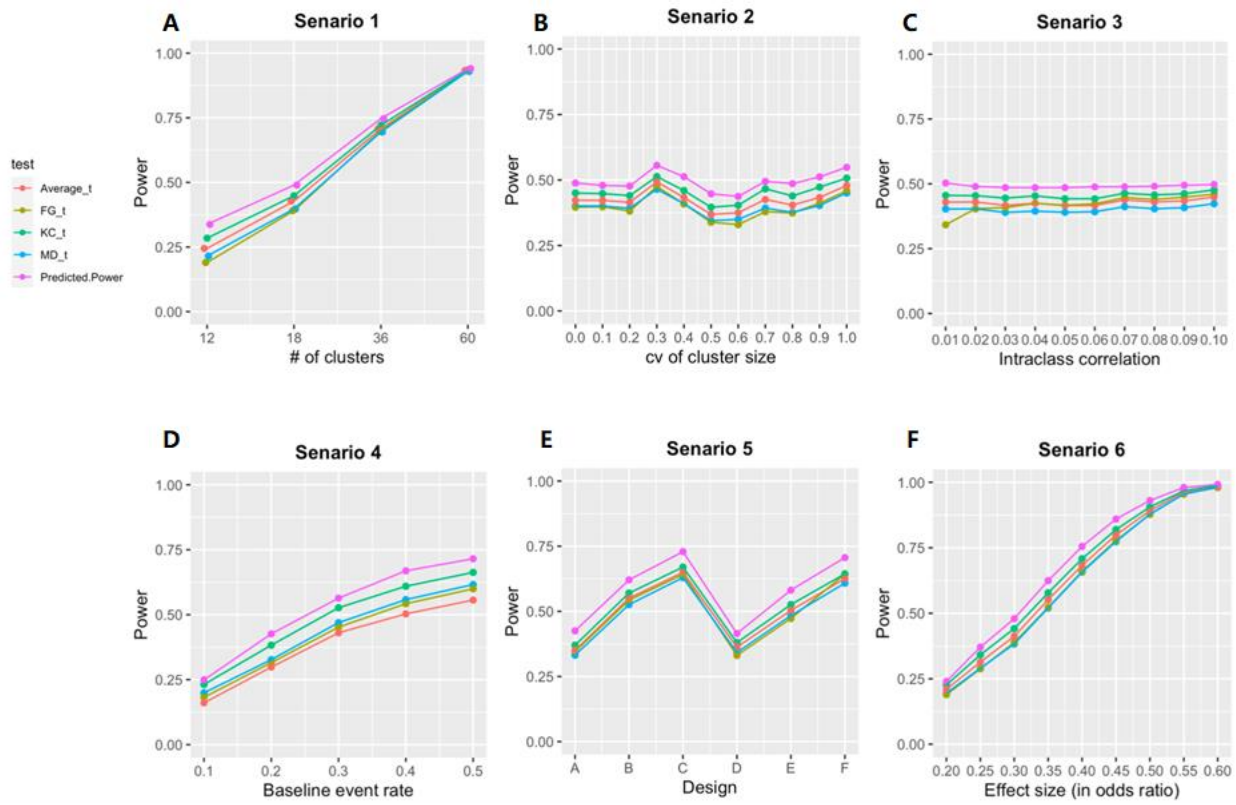

Figure 4A-F. The statistical power for testing non-zero treatment effect of B in six different scenarios in Table S1 using the true working correlation structure. The empirical power of the eight tests is shown with the predicted power calculated from equation (5).

Overall, we observe that greater power can be achieved when the number of clusters (Figure 4A), baseline event rate (Figure 4D) or effect size (Figure 4F) increases. Specifically, Figure 4E shows the power comparison when we use the different designs shown in Figure 2. Adding more sequences results in greater power, which can be seen by observing that power increases from design A to B to C (and likewise from D to E to F). On the other hand, extra measurements with more periods at the end of the study do not show much power benefit – there is little change in power comparing designs A and D, B and E, or C and F. In contrast, the variability of cluster size (cv) does not have an obvious impact on power or type I error rate (Figure 4B), and variation of the intraclass correlation does not change power significantly (Figure 4C).

Inference with two correlation parameters

When we misspecify the working correlation structure in analysis, we observed generally similar patterns (Figure 5B) to Figure 5A with the correct working correlation structure. The sandwich estimator (SW) is robust to a misspecified working correlation structure but still biased because of the small sample size. In general, the other three bias-correction variance estimators (FG, MD, KC) are less biased. Distributions of the FG and the MD estimators are quite similar and slightly overcorrect, while the KC and the Average estimators do not correct enough bias of the SW estimator. We only show data for the distribution of the variance estimates in scenario 1 with 18 clusters, but we note that results for other scenarios and parameter values are similar (data not shown).

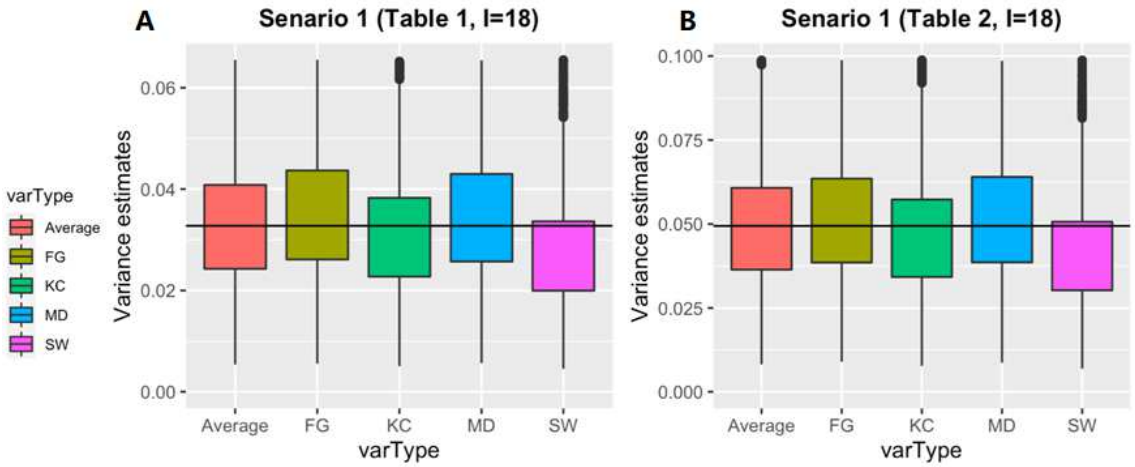

Figure 5A. Boxplots of five types of estimates in scenario 1 with 18 clusters using the true working correlation (with a uniform correlation parameter). The horizontal solid line represents the value of the sample variance of 5,000 estimates. **B.** Boxplots of five types of estimates in scenario 1 with 18 clusters using a misspecified working correlation (common block exchangeable with two parameters). The horizontal solid line represents the value of the sample variance of 5,000 estimates.

Results for the empirical type I error rate and power using the eight different tests are generally consistent with those from the constant ICC model (see Appendix Figures S1 and S2). As an example, we show the type I error and empirical power results for scenario 1 in Figure 6A and 6B. The change pattern in each figure from different sample sizes is the same with Figure 3A or Figure 4A. In addition, power is slightly reduced with this misspecification when comparing results shown in Figure 6B and Figure 4A. Also, we observed a smaller difference in power

among the 6 bias-correction tests in Figure 6B than Figure 4A. Additionally, the magnitude of correlation parameters shows great impact on power. Increasing within-period correlation reduces power, while increasing inter-period correlation gains more power (see the Appendix Figure S2).

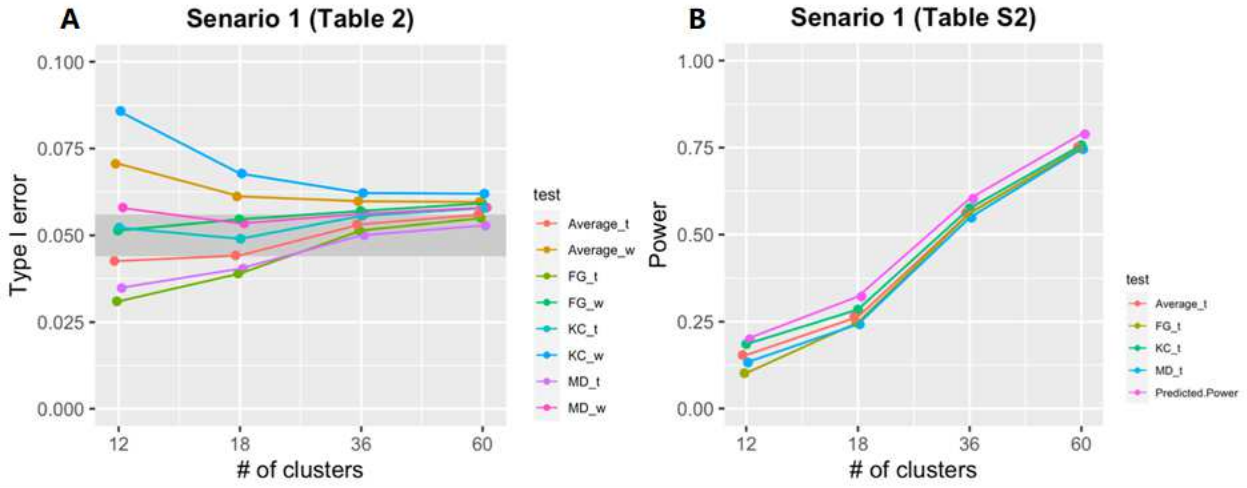

Figure 6A. The empirical type I error rate in scenario 1 using misspecified working correlation (common block exchangeable with two parameters) in Table 2. **B.** The empirical power in scenario 1 using misspecified working correlation (common block exchangeable with two parameters) in Table S2.

Discussion

In this study, we build the marginal model for binary outcomes in MSW-CRT designs. We compare the performance of inference methods based on GEE and 5 variance estimators with t or normal distributions, which can be used as guidance for implementation of MSW-CRT designs. We argue that this is especially helpful when investigators are interested in testing the add-on effect of a second intervention after the first beneficial treatment. From our simulations, when the inter-period and within-period correlations differ, GEE analysis with incorrectly specified correlation structure leads to incorrect inference using the Wald test with model-based variance or sandwich estimator. Therefore, we suggest the use of other bias correction inference methods for inference when the number of clusters is not enough (>60 in our settings). We found that average of MD and KC estimator (Ford and Westgate, 2017) with sampling distribution

$t(I-3)$ can adjust the bias and maintain type I error rate close to the nominal level. Kauermann and Carroll (2001) estimator with $t(I-3)$ can also perform well on controlling type I error rate when the cohort sizes for each cluster are with less variation (coefficient of variation <0.3). This conclusion is also consistent with results in Ford and Westgate (2017). Also, we find the sandwich estimator is usually smaller than the MD estimator, which is close to the FG estimator but larger than the KC estimator. Moreover, the type I error rates produced by Average_t test are stable and robust to different parameter values and working correlation structure misspecification. For testing the second intervention effect, increasing the number of clusters, baseline event rate and effect size can increase the empirical power, and thus reduce the required sample size. Since the effect size of the first intervention determines the event rate before the initiation of the second intervention, power and required sample size for the second intervention depend on the effect size of the first treatment.

Another finding of this study is that when the working correlation structure is misspecified, we loss power and more data are needed to test the effectiveness of the intervention. Even though we do not observe a major impact of changing the correlation magnitude when the inter-period and within-period correlations are equal, power will decrease as the difference $\alpha_0 - \alpha_1$ becomes large. Our simulations show that the four tests (MD_t, FG_t, KC_t and Average_t) perform quite differently on power when the (α_0, α_1) combination changes. We found that when the difference $\alpha_0 - \alpha_1$ is largest, i.e., we are furthest away from the constant correlation assumption, these four tests are much less conservative and closer to the theoretical power values. Conversely, they lose more power as the difference between the two correlation parameters decreases.

In our study, we use 6 designs shown in Figure 2 to investigate the impact of using more sequences when the total number of clusters is fixed. Since the within-period comparison may help gain statistical power, more sequences may increase power. From our simulations, we conclude that when implementation resources are sufficient, we can gain power by using more sequences, even though we must initiate the second intervention before all the clusters receive the first intervention. Another concern in this design is that the cost of more sequences in a limited resources community-based study is high. For example, considering the designs D and F in Figure 2, the design shown in F will simultaneously implement two different interventions in different sequences (periods 3, 4 and 5), while in D, the interventions are initiated in different

periods for each sequence. In practice, this usually means higher travel costs and more implementation resources on the field, which could be unfeasible and delay the trial progress.

Conclusion

Based on the above discussion, when the number of clusters is not very large in MSW-CRTs, inference can be conduct using GEE analysis with average of MD and KC estimator (Ford and Westgate, 2017) with $t(I-3)$. When the cohort size does not change a lot across periods, Kauermann and Carroll (2001) estimator with $t(I-3)$ can also be used for inference. For designing the MSW-CRT, we suggest using more sequences when implementation resources are sufficient, which means we may initiate the second intervention before all participants receive the first intervention. Moreover, interim analysis is also recommended to check event rate after initiating the first intervention, because the add-on effect of the second intervention will be hard to detect if the event rate is too low. Future work on MSW-CRT includes exploring the combination effect of different parameters and design aspects on power, and the effect of incorporating decaying correlations. Even though the simulation studies are limited compared with scenarios in practice, this work provides a basic guidance for MSW-CRT designs, which can save recruiting cost and time than implementing separate SW-CRT designs.

List of abbreviations

MSW-CRT : multiphase stepped wedge cluster randomized trial

SW-CRT : stepped wedge cluster randomized trial

GEE : generalized estimating equations

CRT : cluster randomized trial

ICC : exchangeable intraclass correlation

MD : Mancl and DeRouen (2001)

KC : Kauermann and Carroll (2001)

FG : Fay and Graubard (2001)

IRLS : iterative reweighted least squares

MD_w : wald test with MD estimator

KC_w : wald test with KC estimator

average_w : wald test with average of MD and KC estimator

FG_w : wald test with FG estimator

MD_t : test using MD estimator and $t(I-3)$ distribution

KC_t : test using KC estimator and $t(I-3)$ distribution

average_t : test using the average estimator and $t(I-3)$ distribution

FG_t : test using FG estimator and $t(I-3)$ distribution

SW : the sandwich estimator

cv : variation of the cluster size

References

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. BMC Medical Research Methodology, 6(54).

By, K. and Qaqish, B. (2011). mvtBinaryEP: Generates Correlated Binary Data. R package version 1.0.1.

Cook, T. D. and Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Houghton Mifflin Company, MA.

Crespi, C. M., Wong, W. K., and Mishra, S. I. (2009). Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster-randomized trials. Statistics in medicine, 28(5), 814-827.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. The American Statistician, 45(4), 302–304.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. Biometrics, 57(4), 1198–1206.

Ford, W. P. and Westgate, P. M. (2017). Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. Biometrical Journal, 59(3), 478-495.

Ford, W. P. and 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.

Hall, A., Inskip, H., Loik, F., Day, N., O'Connor, G., Bosch, X., Muir, C., Parkin, M., Munoz, N., Tomatis, L., et al. (1987). The gambia hepatitis intervention study. Cancer Res, 47(21), 5782–5787.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., and Lilford, R. J. (2015a). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. Bmj, 350, h391.

Hemming, K., Lilford, R., and Girling, A. J. (2015b). Stepped-wedge cluster randomized controlled trials: a generic framework including parallel and multi-level designs. Statistics in Medicine, 34(2): 181-196.

Hemming, K., Taljaard, M., Weijer, C., & Forbes, A. B. (2020). Use of multiple period, cluster randomised, crossover trial designs for comparative effectiveness research. bmj, 371.

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. IMA journal of Numerical Analysis, 22(3), 329–343.

Hooper, R., Teerenstra, S., de Hoop, E., and Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in medicine, 35(26), 4718–4728.

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials, 28(2), 182–191.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association, 96(456), 1387–1396.

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., Hughes, J. P., Hemming, K., Taljaard, M., Melnick, E. R., & Heagerty, P. J. (2021). Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. Statistical Methods in Medical Research, 30(2), 612-639.

Li, F., Turner, E. L., and Preisser, J. S. (2018). Sample size determination for gee analyses of stepped wedge cluster randomized trials. Biometrics, 74(4), 1450–1458.

Li, P. and Redden D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. Statistics in Medicine, 34, 281-196.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13–22.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I., and Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimat- ing equations. Biometrics, 63(3), 935–941.

Lyons, V. H., Li, L., Hughes, J. P., and Rowhani-Rahbar, A. (2017). Proposed variations of the stepped-wedge design can be used to accommodate multiple interventions. Journal of clinical epidemiology, 86, 160–167.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. Biometrics, 57(1), 126–134.

Martin, J., Girling, A., Nirantharakumar, K., Ryan, R., Marshall, T., and 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), 402.

Murray, D. M. et al. (1998). Design and analysis of group-randomized trials, volume 29. Oxford University Press, USA.

Murray, D. M., Varnell, S. P., and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. American journal of public health, 94(3), 423–432.

Parker, A. G., Hetrick, S. E., Jorm, A. F., Mackinnon, A. J., McGorry, P. D., Yung, A. R., Scanlan, F., Stephens, J., Baird, S., Moller, B., et al. (2016). The effectiveness of sim- ple psychological and physical activity interventions for high prevalence mental health problems in young people: a factorial randomised controlled trial. Journal of affective disorders, 196, 200–209.

Pears, S., Bijker, M., Morton, K., Vasconcelos, J., Parker, R. A., Westgate, K., Brage, S., Wilson, E., Prevost, A. T., Kinmonth, A.-L., et al. (2016). A randomised controlled trial of three very brief interventions for physical activity in primary care. BMC Public Health, 16(1), 1033.

Preisser, J. S., Young, M. L., Zaccaro, D. J., and 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.

Preisser, J. S., Lu, B., and Qaqish, B. F. (2008). Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. Statistics in medicine, 27(27), 5764–5785.

Rochon, J. (1998). Application of gee procedures for sample size calculations in repeated measures experiments. Statistics in medicine, 17(14), 1643–1658.

Salmon, J., Arundell, L., Hume, C., Brown, H., Hesketh, K., Dunstan, D. W., Daly, R. M., Pearson, N., Cerin, E., Moodie, M., et al. (2011). A cluster-randomized controlled trial to reduce sedentary behavior and promote physical activity and health of 8-9 year olds: The transform-us! study. BMC Public Health, 11(1), 759.

Scott, J. M., deCamp, A., Juraska, M., Fay, M. P., and Gilbert, P. B. (2017). Finite- sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Statistical methods in medical research, 26(2), 583–597.

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, 0962280220958735.

Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L., and Teerenstra, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. Journal of clinical epidemiology, 66(7), 752–758.

Westgate, P. M. (2013). On small-sample inference in group randomized trials with binary outcomes and cluster-level covariates. Biometrical Journal 55, 789–806.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are generated from simulations. R codes are available from the corresponding author on request if necessary.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institute of Health (NIH) [1UM1-DA049417 and 3UL1-TR002733 to Jackson, Rebecca D. and Winhusen, Theresa M.].

Authors' contributions

All authors contributed to the research problem development. JP did all the analysis and drafted the manuscript. AS, SF provided clinical context and guidance. PZ and PW validated the results in the paper. All authors comment the manuscript and approved the final version.

Acknowledgments

We acknowledge The Ohio State University College of Arts and Sciences (http://go.osu.edu/unitycompute) for providing computational resources.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MSWCRTAppendix.docx