HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1450–1458. 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]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina 27710, U.S.A.

[2]Duke Global Health Institute, Durham, North Carolina 27708, U.S.A.

[3]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

Summary.

In stepped wedge cluster randomized trials, intact clusters of individuals switch from control to intervention from a randomly assigned period onwards. Such trials are becoming increasingly popular in health services research. When a closed cohort is recruited from each cluster for longitudinal follow-up, proper sample size calculation should account for three distinct types of intraclass correlations: the within-period, the inter-period, and the within-individual correlations. Setting the latter two correlation parameters to be equal accommodates cross-sectional designs. We propose sample size procedures for continuous and binary responses within the framework of generalized estimating equations that employ a block exchangeable within-cluster correlation structure defined from the distinct correlation types. For continuous responses, we show that the intraclass correlations affect power only through two eigenvalues of the correlation matrix. We demonstrate that analytical power agrees well with simulated power for as few as eight clusters, when data are analyzed using bias-corrected estimating equations for the correlation parameters concurrently with a bias-corrected sandwich variance estimator.

Keywords

Finite sample correction; Generalized estimating equations (GEE); Group randomized trials; Matrix-adjusted estimating equations (MAEE); Power; Sandwich estimator

1. Introduction

Cluster randomized trials (CRTs) are designed to evaluate the effect of an intervention administered at the cluster level. Common reasons for conducting such trials include minimizing treatment contamination between individuals in the same cluster, facilitating administrative convenience, and avoiding ethical issues (Murray, 1998). In the traditional

* frank.li@duke.edu.

7.Supplementary Materials

Web Appendices referenced in Section 2–4, along with R code implementing the estimating equation methods, are available with this article at the *Biometrics* website on Wiley Online Library.

two-arm parallel design, half of the clusters are assigned to each arm, and the intervention is implemented concurrently in the treated clusters. Although frequently used in practice, a parallel design may not always be logistically feasible with limited resources (Turner et al., 2017). The stepped wedge design combats this resource limitation by switching clusters to intervention in a staggered fashion. In a stepped wedge design, each cluster starts from the control condition and crosses over to receive intervention from a randomly assigned period onwards (Hussey and Hughes, 2007). Individual responses within each cluster will be assessed during each period based on a cross-sectional sample or a closed cohort, until all clusters are exposed to the intervention. The stepped wedge design is sometimes considered more ethically acceptable when the intervention is believed to be superior than the standard care, and it has been increasingly used in health care research to evaluate the effect from changes in the way that health services are delivered or in the training that health care professionals received.

A distinguishing feature of a CRT is that responses within the same cluster are more similar than those from different clusters, because each cluster is usually not formed at random but rather through some natural connections among its members. The intraclass correlation coefficient provides a quantitative assessment of this within-cluster similarity, and the statistical implications on the sample size due to the intraclass correlation have been well studied, particularly in parallel designs (Murray, 1998). When a stepped wedge CRT involves cross-sectional measurements on different sets of individuals, two types of intraclass correlations have been recognized, the within-period and inter-period correlations (Martin et al., 2016). When a closed cohort is recruited from each cluster for longitudinal follow-up, an additional within-individual correlation should be considered in the design and analysis, since repeated measurements are taken for the same individuals (Hughes et al., 2015). There are two commonly used statistical models which account for these different types of correlations, conditional and marginal models. Although each modeling approach has its advantages, an important distinction between them is the difference in interpretation of the regression parameters (Preisser et al., 2003). In a stepped wedge trial, the treatment effect from a marginal model describes how the average response changes across the subsets of population defined by the treated and control cluster-periods. By contrast, the treatment effect from a conditional model is interpreted as the average change in responses from control to intervention conditional on the unobserved random effects; in other words, this interpretation applies to a conceptual population of cluster-periods possessing the same values of some latent variables. Correspondingly, the design and analysis of stepped wedge CRTs have mostly been based on (generalized) linear mixed models; see, for instance, Hussey and Hughes (2007), Woertman et al. (2013); Hemming et al. (2015); Hooper et al. (2016). However, since stepped wedge CRTs are often used in health care research to inform policy decisions, marginal models carry a straightforward population-averaged interpretation and may be preferred. Accordingly, this article proposes methods for designing stepped wedge CRTs for analysis with marginal models, with an emphasis on cohort studies.

2. GEE Analyses of Stepped Wedge Designs

We consider a cohort stepped wedge design with I clusters and T periods, where a closed cohort of individuals from each cluster are identified at the start of the trial and followed up

for repeated measurements. Let y_{ijk} be the response of individual $k = 1, \ldots, N_i$ from cluster $i = 1, \ldots, I$ during period $j = 1, \ldots, T$. A complete design is assumed such that measurements are taken for all individuals during each period (Hemming et al., 2015). A step is defined as the pre-planned time point when at least one cluster crosses over from control to intervention. We denote the total number of steps by S ($2 \leq S < T$), and the number of clusters that cross over at each step by m_s such that $\sum_{s=1}^{S} m_s = I$. We assume there are $b \geq 1$ baseline measurements taken for each individual under the control condition, and $c_s \geq 1$ follow-up measurements taken for each individual after step s but prior to step $s + 1$ (or end of study). Therefore, we associate each measurement time point with a distinct period and define the total number of periods $T = b + \sum_{s=1}^{S} c_s$. A standard stepped wedge design is given by $b = c_s = 1$ for all s, and $T = S + 1$ ($T \geq 3$). A schematic illustration of a standard design is given in Figure 1.

Denote μ_{ijk} as the marginal mean response, which is related to the intervention and period effects via the following generalized linear model

$$g(\mu_{ijk}) = \beta_j + X_{ij}\delta, \quad (1)$$

where g is a link function, β_j is the jth period effect, X_{ij} is the treatment indicator of cluster i in period j ($X_{ij} = 1$ if cluster i receives intervention in period j and 0 otherwise), and δ is the marginal intervention effect on the link function scale. We further let $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_T, \delta)'$ be the vector of parameters in mean model (1), $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT})'$ be the treatment sequence associated with cluster i (i.e., a sequence of zeros followed by a sequence of ones), and $v(\mu_{ijk})$ be the variance function. To characterize the degree of similarity between individual responses taken from each cohort, we employ the correlation structure with the following specification suggested by Preisser et al. (2003): (i) the within-period correlation, α_0, that measures the similarity between responses from different individuals within the same cluster during the same period ($\text{corr}(y_{ijk}, y_{ijk'}) = \alpha_0$ for $k \neq k'$); (ii) the inter-period correlation, α_1, that measures the similarity between responses from different individuals within the same cluster but across periods ($\text{corr}(y_{ijk}, y_{ij'k'}) = \alpha_1$ for $j \neq j'$ and $k \neq k'$); (iii) the within-individual correlation, α_2, that measures the similarity between responses from the same individual across periods ($\text{corr}(y_{ijk}, y_{ij'k}) = \alpha_2$ for $j \neq j'$). Although no additional covariates are included in model (1), such an extension is straightforward.

Let $\boldsymbol{y}_i = (y_{i11}, y_{i12} \ldots, y_{iTN_i})'$ and $\boldsymbol{\mu}_i = (\mu_{i11}, \mu_{i12} \ldots, \mu_{iTN_i})'$ be the $TN_i \times 1$ response vector and marginal mean vector of cluster i, respectively, where TN_i is the total number of observations in cluster i. We use generalized estimating equations (GEE; Liang and Zeger, 1986) to estimate the intervention effect in equation (1). We define $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\theta}'$, and let $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i \boldsymbol{A}_i^{1/2}$ be a working covariance matrix for $\boldsymbol{y}_i$, where $\boldsymbol{A}_i$ is the TN_i-dimensional diagonal matrix with elements $\phi v(\mu_{ijk})$ with ϕ representing the dispersion parameter; $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is a working correlation matrix that may vary across clusters but is specified by the common parameter $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$. We can succinctly write the working correlation for cluster i as

$$\boldsymbol{R}_i = (1-\alpha_0+\alpha_1-\alpha_2)\boldsymbol{I}_{TN_i} + (\alpha_2-\alpha_1)\boldsymbol{J}_T \otimes \boldsymbol{I}_{N_i} + (\alpha_0-\alpha_1)\boldsymbol{I}_T \otimes \boldsymbol{J}_{N_i} + \alpha_1 \boldsymbol{J}_{TN_i}, \quad (2)$$

where $\boldsymbol{J}_u = \mathbf{1}_u \mathbf{1}_u'$ is a $u \times u$ matrix of ones, $\boldsymbol{I}_u$ is the $u \times u$ identity matrix. Both the diagonal and off-diagonal $N_i \times N_i$ blocks of $\boldsymbol{R}_i$ are of the exchangeable form given by $(1-\alpha_0)\boldsymbol{I}_{N_i} + \alpha_0 \boldsymbol{J}_{N_i}$ and $(\alpha_2-\alpha_1)\boldsymbol{I}_{N_i} + \alpha_1 \boldsymbol{J}_{N_i}$, respectively. If the off-diagonal blocks are viewed as scalar elements, $\boldsymbol{R}_i$ assumes an exchangeable form; therefore $\boldsymbol{R}_i$ is termed block exchangeable. In Web Appendix A, we show that $\boldsymbol{R}_i$ has four distinct eigenvalues,

$$\begin{aligned}
\lambda_1 &= 1-\alpha_0+\alpha_1-\alpha_2, \\
\lambda_2 &= 1-\alpha_0-(T-1)(\alpha_1-\alpha_2), \\
\lambda_{i3} &= 1+(N_i-1)(\alpha_0-\alpha_1)-\alpha_2, \\
\lambda_{i4} &= 1+(N_i-1)\alpha_0+(T-1)(N_i-1)\alpha_1+(T-1)\alpha_2.
\end{aligned}$$

The combinations of $(\alpha_0, \alpha_1, \alpha_2)$ for which $\boldsymbol{R}_i$ is positive definite can be efficiently determined from the set of linear constraints, $\min\{\lambda_1, \lambda_2, \lambda_{i3}, \lambda_{i4}\} > 0$.

The GEE estimator $\hat{\boldsymbol{\theta}}$ is obtained by solving $\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$. In practice, any consistent estimator of $\boldsymbol{\alpha}$ may be used without affecting the consistency of $\hat{\boldsymbol{\theta}}$ (Liang and Zeger, 1986). When the dimension of the correlation matrix is large ($TN_i \geq 2000$), fast computation is achieved by directly calculating the following expression and hence avoiding numeric matrix inversion (see Web Appendix A for a derivation)

$$\boldsymbol{R}_i^{-1} = \frac{1}{\lambda_1}\boldsymbol{I}_{TN_i} - \frac{\alpha_2-\alpha_1}{\lambda_1\lambda_2}\boldsymbol{J}_T \otimes \boldsymbol{I}_{N_i} - \frac{\alpha_0-\alpha_1}{\lambda_1\lambda_{i3}}\boldsymbol{I}_T \otimes \boldsymbol{J}_{N_i} + \left\{\frac{(\alpha_2-\alpha_1)(\alpha_0-\alpha_1)}{\lambda_1\lambda_2\lambda_{i3}} + \frac{\alpha_2\alpha_0-\alpha_1}{\lambda_2\lambda_{i3}\lambda_{i4}}\right\}\boldsymbol{J}_{TN_i}.$$

To reduce the finite sample bias of the correlation parameter estimates, we use an additional set of matrix-adjusted estimating equations (MAEE) of Preisser et al. (2008) to estimate $\boldsymbol{\alpha}$. For continuous responses, we further propose to use the bias-corrected moment-based estimator for $\boldsymbol{\phi}$. For binary responses, $\boldsymbol{\phi}$ is usually set to be 1. Here, we focus on the marginal mean model parameters, and defer the related technical details of MAEE to Web Appendix B. When the number of clusters $\boldsymbol{I}$ is sufficiently large (≥ 40), $\hat{\boldsymbol{\theta}}$ is approximately multivariate normal with mean $\boldsymbol{\theta}$ and covariance estimated by the model-based estimator $\widehat{\boldsymbol{\Sigma}}_1^{-1} = \{\sum_{i=1}^{I} \boldsymbol{D}_i'(\hat{\boldsymbol{\theta}})\boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\alpha}})\boldsymbol{D}_i(\hat{\boldsymbol{\theta}})\}^{-1}$, or by the sandwich estimator $\widehat{\boldsymbol{\Sigma}}_1^{-1}\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_1^{-1}$ where

$$\widehat{\Sigma}_0 = \sum_{i=1}^{I} C_i D_i'(\hat{\theta}) V_i^{-1}(\hat{\alpha}) B_i r_i r_i' B_i' V_i^{-1}(\hat{\alpha}) D_i(\hat{\theta}) C_i, \quad (3)$$

and $r_i = y_i - \hat{\mu}_i$ is the residual vector of cluster i. In large samples ($I \geq 40$), the sandwich estimator provides valid inference regardless of the correct specification of R_i, while the consistency of the model-based variance estimator is dictated by the correct specification of the correlation structure. In equation (3), setting $C_i = I_{T+1}$ and $B_i = I_{TN_i}$ gives the uncorrected sandwich estimator of Liang and Zeger (1986), which is referred to as BC0. When I is small, BC0 tends to be biased downwards and may inflate the type I error rate (Kauermann and Carroll, 2001). We define the cluster leverage by $H_i = D_i \Sigma_1^{-1} D_i' V_i^{-1}$; setting $C_i = I_{T+1}$ and $B_i = \left(I_{TN_i} - H_i\right)^{-1/2}$ gives the bias-corrected variance estimator of Kauermann and Carroll (2001), or BC1. Setting $C_i = I_{T+1}$ and $B_i = \left(I_{TN_i} - H_i\right)^{-1}$ gives the bias-corrected variance estimator of Mancl and DeRouen (2001), or BC2. Setting $C_i = \text{diag}\{(1 - \min\{r, [Q_i]_{jj}\})^{-1/2}\}$ and $B_i = I_{TN_i}$, where $Q_i = D_i' V_i^{-1} D_i \Sigma_1^{-1}$, gives the bias-corrected variance of Fay and Graubard (2001), or BC3. The bound parameter r usually takes the default value 0.75 to avoid over-correction. Since the elements of the cluster leverage matrix are between 0 and 1, we have BC0 < BC1 < BC2 (Preisser et al., 2008). Further, BC3 tends to be close to BC1 (Scott et al., 2017).

For a parallel CRT with binary responses, Lu et al. (2007) found that bias correction for α by MAEE slightly improved the coverage probability of the normality-based confidence interval of θ using the model-based variance, but such a correction had a negligible effect on procedures that used the sandwich estimators, that is, BC0, BC1, and BC2. On the other hand, MAEE could substantially reduce the bias of the correlation estimator $\hat{\alpha}$ (Preisser et al., 2008). For a three-level CRT with binary responses, Teerenstra et al. (2010) reported that the use of MAEE with a t-test for θ based on the model-based variance estimator or BC1 maintained the nominal test size and provided power levels close to analytical predictions. We investigate in Section 4 the performance of these tests for a cohort stepped wedge design with both continuous and binary responses.

Although our presentation focuses on a cohort design, an application to a cross-sectional design is straightforward. In a cross-sectional design, the correlation structure may depend only on α_0 and α_1 since usually different sets of individuals are assessed for each cluster at different periods, and α_2 is no longer required. Therefore, the block exchangeable correlation structure R_i reduces to the nested exchangeable structure introduced in Teerenstra et al. (2010), and the above GEE procedure can be adapted by setting $\alpha_2 = \alpha_1$.

3. Statistical Power and Sample Size

Suppose that we are interested in testing the null hypothesis of no intervention effect $H_0 : \delta = 0$ using a two-sided test. Based on mean model (1), the asymptotic distribution of $\sqrt{I}(\hat{\delta} - \delta)$ is normal with mean zero and variance determined by the $(T + 1, T + 1)$th element of $\text{cov}\{\sqrt{I}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\}$. A normality-based z-test statistic for H_0 uses $\hat{\delta}/\{\text{var}(\hat{\delta})\}^{1/2}$ and is compared to the standard normal distribution. Asymptotically, the power to detect an intervention effect of size $\delta \neq 0$ with a nominal type I error rate α is

$$\text{power} = \Phi\left(z_{\alpha/2} + |\delta|/\sqrt{\text{var}(\hat{\delta})}\right), \quad (4)$$

where Φ is the standard normal distribution function and $z_{\alpha/2}$ is the associated upper $\alpha/2$-th quantile. To account for the uncertainty in estimating the asymptotic variance of $\hat{\delta}$, we may alternatively use $\hat{\delta}/\{\text{var}(\hat{\delta})\}^{1/2}$ as a t-statistic, which is compared to the t-distribution with $I - (T + 1)$ degrees of freedom. Then the power to detect an intervention effect of size δ is modified to be

$$\text{power} = \Phi_{t, I-(T+1)}\left(t_{\alpha/2, I-(T+1)} + |\delta|/\sqrt{\text{var}(\hat{\delta})}\right), \quad (5)$$

where $\Phi_{t,n}$ is the cumulative t-distribution function with n degrees of freedom and $t_{\alpha/2,n}$ is the corresponding upper $\alpha/2$th quantile. Because critical values associated with the z-test are closer to zero, it is expected that the z-test is more likely to result in anti-conservative inference for a GEE analysis coupled with BC0 compared to the t-test. On the other hand, the two tests may have different implications for the class of bias-corrected sandwich estimators, which are known to provide different degrees of inflation relative to the uncorrected sandwich variance. In the subsequent power calculations, we assume for simplicity that each cluster recruits a cohort of the same size such that $N_i = N$, $\lambda_{i3} = \lambda_3$, and $\lambda_{i4} = \lambda_4$ for each cluster i.

3.1. Continuous Responses

When the response y_{ijk} is continuous and g is the identity link, a closed-form power formula can be obtained since we can derive an explicit expression for $\text{var}(\hat{\delta})$. We assume the covariance of $\boldsymbol{y}_i$ to be known and given by $\text{var}(\boldsymbol{y}_i) = \boldsymbol{V}_i$. Therefore, $\text{var}(\hat{\delta})$ is the $(T + 1, T + 1)$th element of the model-based variance $\boldsymbol{\Sigma}_1^{-1}$. In Web Appendix C, we show that the variance of the intervention effect estimator is

$$\text{var}(\hat{\delta}) = \frac{(\phi/N)IT\lambda_3\lambda_4}{\left(U^2 + ITU - TW - IV\right)\lambda_4 - \left(U^2 - IV\right)\lambda_3}, \quad (6)$$

where $U = \sum_{i=1}^{I}\sum_{j=1}^{T} X_{ij}$, $W = \sum_{j=1}^{T}(\sum_{i=1}^{I} X_{ij})^2$, and $V = \sum_{i=1}^{I}(\sum_{j=1}^{T} X_{ij})^2$ are design constants that only depend on the treatment sequence each cluster receives. This variance further depends on the correlation parameters through two eigenvalues of the block exchangeable correlation matrix, λ_3 and λ_4. Further, for fixed number of clusters I and periods T, as the cohort size $N \to \infty$, we could show that

$$\lim_{N\to\infty} \text{var}(\hat{\delta}) = \frac{\phi I(\alpha_0 - \alpha_1)\{\alpha_0 + (T-1)\alpha_1\}}{(IU - W)\{\alpha_0 + (T-1)\alpha_1\} + (U^2 - IV)\alpha_1}. \quad (7)$$

As a result, the limit of $\text{var}(\hat{\delta})$ is controlled by I, T, and two correlation parameters α_0, α_1 (see Web Appendix C for a detailed discussion); expression (7) may be useful as an approximation to variance (6) when N is large. We note that the GEE-based variance (6) is equivalent to the variance formula provided in Li et al. (2018) based on a linear mixed model with three random intercepts. Variance formula (6) generalizes the Hussey and Hughes (2007) formula based on a simple linear random intercept model to cohort designs. Specifically, if a single correlation parameter is used to characterize the working correlation $\boldsymbol{R}_i$ such that $\alpha_0 = \alpha_1 = \alpha_2$ in (2) and $\boldsymbol{R}_i$ is exchangeable, then (6) reduces to the variance in Hussey and Hughes (2007) from noting that ϕ is the total variance components of y_{ijk}. For a cross-sectional design, the within-period correlation, α_0, and the inter-period correlation, α_1, may be sufficient to represent the correlation structure, that is, one could equate $\alpha_2 = \alpha_1$ in (2) and adjust the values of λ_3 and λ_4 in (6) to obtain the appropriate variance. Plugging $\text{var}(\hat{\delta})$ in (4) and (5) gives the analytical power formula for a z-test and a t-test.

In general, directly solving equations (4) and (5) for the required number of clusters is difficult. However, in the following case we can derive the design effect of a stepped wedge CRT relative to an individually randomized study to facilitate sample size determination. Following Woertman et al. (2013), we assume that an equal number of clusters cross over to intervention at each step such that $m_s = m$, and further that an equal number of measurements are taken after each step such that $c_s = c$ for all $s = 1, \ldots, S$. Under such simplifications, we obtain a design with $I = Sm$ clusters, $T = b + Sc$ periods, and design constants $U = \frac{1}{2}S(S+1)mc$, $W = \left(\frac{1}{3}S^3 + \frac{1}{2}S^2 + \frac{1}{6}S\right)m^2c$, $V = \left(\frac{1}{3}S^3 + \frac{1}{2}S^2 + \frac{1}{6}S\right)mc^2$, which can be used to simplify $\text{var}(\hat{\delta})$. We compare the variance of $\hat{\delta}$ in a stepped wedge design versus its counterpart in an individually randomized design, where the sample mean difference is used to estimate δ. Given that the sample mean difference has variance $4\phi/(NSm)$, the design effect defined as the ratio of $\text{var}(\hat{\delta})$ under these two designs is given by,

$$\text{design effect} = \frac{3}{2c(S - 1/S)}\left\{\frac{(b + Sc)\lambda_3\lambda_4}{(Sc/2)\lambda_3 + (b + Sc/2)\lambda_4}\right\}. \quad (8)$$

This design effect generalizes the one given in Woertman et al. (2013) based on the Hussey and Hughes (2007) model and agrees with the design effect derived by Hooper et al. (2016) based on a linear mixed model. To estimate sample size, we could first compute the required

number of individuals in an individually randomized trial, and then multiply by (8) to obtain the required number of individuals in a cohort stepped wedge design (rounding up to the nearest integer or, for a balanced design, multiple of I). Given the required total number of individuals, the required number of clusters, and the required cohort size can be ascertained. The design effect could also be used to study how the correlations affect the required sample size. Since λ_3 and λ_4 increase as the within-period correlation α_0 increases, the required sample size will inflate given a larger value of α_0. Therefore, α_0 mimics the traditional intraclass correlation in a parallel design. However, the impact of the inter-period and within-individual correlations are less apparent from expression (8). We plot the design effect as a function of α_1 and α_2 for several scenarios in Web Figures 1 and 2. Both figures indicate that larger values of α_1 or α_2 reduce the required sample size when all the correlations are positive. Finally, we remark that power and sample size calculations with continuous responses do not depend on the period effect, as variance expression (6) is free of β_j.

3.2. Binary Responses

When y_{ijk} is binary and g is the canonical logit link, the desired variance $\text{var}(\hat{\delta})$ cannot be obtained in closed-form because the marginal variance $v(\mu_{ijk}) = \mu_{ijk}(1 - \mu_{ijk})$ depends on the marginal mean. However, power calculations can be performed by adapting the general methodology presented in Rochon (1998). To proceed, we specify a value for I and divide the participating clusters into S groups depending on the step at which they cross over to receive intervention, and so m_s clusters will be included in the sth group. Then, the expected longitudinal trajectory over T periods for the sth group of clusters will be assumed as $\boldsymbol{\mu}_s = (\mu_{s1}, \ldots, \mu_{sT})'$. This could be informed by previous trials with a similar endpoint or pilot data. Note that based on model (1), $\boldsymbol{\mu}_s$ is selected according to $\mu_{ijk} = \exp(\beta_j + X_{ij}\delta)/(1 + \exp(\beta_j + X_{ij}\delta))$. Since the design matrix of a cluster in group s based on mean model (1) is $\boldsymbol{Z}_s = (\boldsymbol{I}_T, \boldsymbol{X}_s) \otimes \mathbf{1}_N$ where $\boldsymbol{X}_s$ is the treatment sequence received by all clusters in group s, we can solve for $\boldsymbol{\theta}$ by generalized least squares, $\boldsymbol{\theta} = (\sum_{s=1}^{S} m_s \boldsymbol{Z}_s' \boldsymbol{W}_s \boldsymbol{Z}_s)^{-1} \sum_{s=1}^{S} m_s \boldsymbol{Z}_s' \boldsymbol{W}_s \boldsymbol{g}_s$, where $\boldsymbol{g}_s = (g(\mu_{s1}), \ldots, g(\mu_{sT}))'$, $\boldsymbol{W}_s = \boldsymbol{A}_s^{1/2} \boldsymbol{R}_s^{-1} \boldsymbol{A}_s^{1/2}$, $\boldsymbol{A}_s = \text{diag}\{v(\mu_{s1}), \ldots, v(\mu_{sT})\}$, and $\boldsymbol{R}_s = \boldsymbol{R}(\boldsymbol{\alpha})$ is the common block exchangeable correlation matrix. The detectable effect size expressed in the log odds ratio, δ, is the $(T + 1)$th element in $\boldsymbol{\theta}$. The variance $\text{var}(\hat{\delta})$ is the $(T + 1, T + 1)$th element of the model-based variance $\boldsymbol{\Sigma}_1^{-1} = (\sum_{s=1}^{S} m_s \boldsymbol{Z}_s' \boldsymbol{W}_s \boldsymbol{Z}_s)^{-1}$, and can be used in equations (4) or (5) for power calculations. Unlike with continuous responses, power calculation with binary responses implicitly depends on the assumed period effect specified in $\boldsymbol{\theta}$, and sensitivity analyses could be conducted to gauge how power changes due to different assumptions of the period effect.

4. Simulation Study

We conducted a simulation study to assess the empirical size of the GEE Wald-tests in the context of a cohort stepped wedge CRT. Further, we compared the accuracy of the predicted power to the empirical power of the tests that maintained the nominal size. We generated correlated continuous responses in each cluster from a multivariate normal distribution with

mean specified by model (1) and variance $\boldsymbol{R}(\boldsymbol{\alpha})$ (g is the identity link and the total variance $\phi = 1$). For illustration purposes, a gently increasing period effect was assumed such that $\beta_1 = 0$ and $\beta_{j+1} - \beta_j = 0.1 \times (0.5)^{j-1}$ for $j \geq 1$. We noted that the conclusions were insensitive to the choice of the period effect since these effects were accounted for in the GEE analyses. The effect size $\delta/\phi^{1/2}$ was fixed at zero for studying empirical test size and varied from {0.65, 0.40, 0.35, 0.25} for studying power. Additionally, correlated binary responses within clusters were generated from a binomial model with marginal mean specified by (1) with a logit link and correlation $\boldsymbol{R}(\boldsymbol{\alpha})$ using the method of Qaqish (2003). Baseline prevalence $e^{\beta_1}/\left(1 + e^{\beta_1}\right)$ for all clusters were chosen from {0.75, 0.70, 0.65}. We assumed a gently decreasing period effect on the logit scale such that $\beta_j - \beta_{j+1} = 0.1 \times 0.5^{j-1}$ for $j \geq 1$. The effect size in odds ratio $\exp(\delta)$ was fixed at 1 for studying empirical size and varied from {0.25, 0.30, 0.45, 0.60} for studying power. For both types of responses, we chose $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2) = \{(0.03, 0.015, 0.2), (0.1, 0.05, 0.2), (0.01, 0.005, 0.4)\}$ to represent a range of different correlation values. In particular, the values of α_0 are representative of small correlations commonly reported in parallel CRTs (Murray, 1998), and the inter-period correlation α_1 was assumed smaller than α_0, as observed in Martin et al. (2016). The values of α_2 were chosen to reflect small to moderate within-individual correlations in longitudinal studies, and were assumed to be larger than α_0 and α_1. We have also studied scenarios with larger values of α_2; findings remained similar and the details are omitted for brevity.

We varied the number of clusters I from 8 to 25 since CRTs usually involve a limited number of clusters. We also varied the cohort size N from 4 to 25, and the number of steps S from 2 to 6 to ensure the predicted power was at least 80%. For simplicity, we focused on the standard design with $b = 1$ baseline measurement and $c_s = 1$ follow-up measurement after each step. We assumed that an equal number of $m_s = I/S$ clusters crossed over to intervention at a randomly-assigned step and so I is a multiple of $S = T - 1$. For each scenario, we generated 1000 data replicates and fit GEE for the mean model and MAEE for the block exchangeable correlation structure (the bias-corrected moment-based estimator for ϕ given in Web Appendix B was used with continuous responses; ϕ was set as 1 with binary responses). We considered both two-sided t-tests and z-tests for testing $H_0 : \delta = 0$, constructed from the use of five different variance estimators for $\hat{\delta}$, the model-based variance, BC0, BC1, BC2, and BC3. The convergence rate exceeded 98% for the majority of simulation scenarios except for a few cases (a summary of convergence rates along with the corresponding simulation scenarios is provided in Web Tables 1 and 2). The nominal test size was fixed at 5%, and we considered an empirical size between 3.6% and 6.4% to be acceptable according to the margin of error with 1000 replicates from a binomial model. Similarly, given the predicted power for each scenario was at least 80%, we considered an empirical power that differs at most 2.6% from the nominal value to be in agreement with the predicted power.

Figure 2 summarizes the empirical type I error rates of the z-test and the t-test with different variance estimators for continuous responses. Overall, the z-test tended to be more liberal compared with the corresponding t-test. The empirical size of a z-test was close to nominal level with the use of model-based variance or BC2, when there are at least 18 clusters

(tended to be liberal otherwise). The use of BC1 and BC3 with a *t*-test carried valid type I error rates across all simulation scenarios (only occasionally conservative), while the use of model-based variance or BC2 with a *t*-test was often conservative. The use of BC0 frequently led to inflated type I error rates, especially when it was coupled with a *z*-test. The findings for binary responses are similar and presented in Web Figure 3. Using a *z*-test, the empirical power based on the model-based variance was close to the prediction, while the empirical power based on BC2 was lower than predicted (Figure 3 and Table 1). Although *z*-tests based on the other variance estimators closely matched the analytical power, they carried an inflated size throughout. Using a *t*-test, the empirical power based on BC1 and BC3 corresponded well with the predicted power, while the empirical power based on the model-based variance may slightly exceed the prediction. The *t*-test with BC2 had lower power than predicted in most scenarios. Simulations with binary responses yielded qualitatively similar results, which are presented in Web Figure 4 and Web Table 3.

5. Application

We illustrate our approach to determine the required sample size of a cohort CRT evaluating the effect of an intervention on physical function of end-stage renal disease patients in Australia (Bennett et al., 2013). The intervention was an accredited exercise physiologist coordinated resistance exercise program, offered to improve the health-related quality of life for dialysis patients. Each hemodialysis clinic was a cluster, within which patients were recruited and followed over time. Since there was prior evidence signaling benefit on physical quality of life from the resistance exercise, a stepped wedge design was considered appropriate. The randomization was conducted at the clinic level, and responses were measured at the patient level. The duration of the study was 48 weeks, with evenly spaced $T = 4$ periods. There were $I = 15$ participating clinics, and $S = 3$ steps were considered (a schematic illustration is in Figure 1). No exercise programs were offered during the first period. The clinics were randomly split into three groups, each containing five clinics, crossing over to receive intervention at week 12, 24, or 36. For illustration, we assumed a standard design with one measurement at the end of each period ($b = c_s = 1$) and thus a total of 4 measurements will be recorded for each patient. Since the number of participating clinics was pre-planned, we focused on calculating the required number of patients per clinic to achieve at least 80% power at the 5% nominal test size.

The primary outcome was the 30-second sit-to-stand (STS) test, which measured the number of times a participant could rise from and return to a seated position in a 30-second time frame. A standardized effect size of 0.65 was estimated from a previous study on STS test for hemodialysis patients with end-stage renal disease (Bennett et al., 2013). The within-period correlation was provided as $\alpha_0 = 0.03$ from a previous study (Littenberg and MacLean, 2006); we assumed the inter-period correlation as half of the within-period correlation, $\alpha_1 = 0.015$, and assigned a conservative value, $\alpha_2 = 0.2$, to the within-individual correlation. Since $I = 15$, we planned to use a *t*-statistic instead of a *z*-statistic to avoid potential type I error rate inflation from the GEE analysis. Given that an equal number of clusters switched to intervention at each step, the design constants are $U = IT/2$, $W = I^2T(2T - 1)/\{6(T - 1)\}$, $V = IT(2T - 1)/6$, and variance (6) is

$$\text{var}(\hat{\delta}) = \frac{12(\phi/N)(T-1)\lambda_3\lambda_4}{I(T-2)\{(T-1)\lambda_3 + (T+1)\lambda_4\}}. \quad (9)$$

Power was estimated using equations (5) and (9) to be 72.9% if $N = 3$ and 83.7% if $N = 4$, therefore four patients would be recruited in each clinic. Because an equal number of clusters crossed over to the intervention at each step, we could alternatively use design effect (8) directly for sample size estimation. Suppose this was an individually randomized trial, 92 patients would be required (the working degrees of freedom of a t-test is 10 to match the stepped wedge design). Assuming three patients would be recruited in each clinic, the design effect is 0.58, indicating a total of 54 patients would be needed and 18 clinics would be required. Because we could only afford 15 clinics, we adjusted $N = 4$, and re-computed the design effect to be 0.60. Hence 55 patients would be required for a total of 55/4 ≈ 14 clinics. Therefore, including four patients in each of the 15 clinics guaranteed 80% power.

We investigated the sensitivity of power calculation by varying the values of a_1 and a_2, which are less commonly reported than a_0. The power prediction was plotted as function of a_1 and a_2 at $a_0 = \{0.03, 0.06, 0.1\}$ assuming $I = 15$, $T = 4$ and $N = 4$. Notably, power decreases as a_0 increases but increases as either a_1 or a_2 increases. When $a_0 = 0.03$, the power remains above 80% regardless of the values of $a_1 \in (0, 0.1)$ and $\alpha_2 \in (0, 0.5)$. When a_0 takes the upper bound, 0.06, reported in Littenberg and MacLean (2006), small values of a_1 and a_2 may result in slight loss in power (≈ 78% at the lower left corner of Figure 4(b)). Additional power loss was observed with a moderate within-period correlation, $a_0 = 0.1$.

6. Discussion

Since a cohort stepped wedge design involves repeated measurements for fixed sets of individuals, we consider a block exchangeable correlation structure that models three distinct types of correlations for power calculations. The within-period correlation, a_0, is similar to the conventional intraclass correlation in a parallel CRT. Larger values of a_0 inflate the required sample size for a given level of power. By contrast, larger (positive) values of the inter-period correlation, a_1, and the within-individual correlation, a_2, appear to reduce the required sample size. In practice, power calculations should be guided by reasonable estimates of these correlations. The within-period correlation is usually estimated by the intraclass correlation reported in previous trials with a similar endpoint. However, the inter-period correlation is less commonly reported, although such reporting practice is advocated (Preisser et al., 2007; Martin et al., 2016). This type of correlation was also discussed in designing crossover CRTs, and a default value of half the within-period correlation has been recommended in the absence of external information (Giraudeau et al., 2008). The within-individual correlation could perhaps be obtained from published longitudinal studies with a similar endpoint. In any case, the sensitivity of sample size and power should be investigated for a range of correlation values, as illustrated in Section 5. Notably, the combination of (a_0, a_1, a_2) is valid only if the resulting correlation matrix is positive definite; this condition could be efficiently checked by the set of linear constraints provided in Section 2.

In finite samples, we found that the normality-based *z*-test coupled with the model-based variance with correlation estimated by MAEE performed well with a correct size and adequate power when there were at least 18 clusters. In this case, the *z*-test has higher power than the *t*-test and is preferred. This finding agrees with the simulations of Lu et al. (2007) based on a two-correlation model. They also reported that for a within-cluster covariate, the normality-based confidence interval with BC2 produced better coverage than BC0 and BC1. Since the treatment varies within each cluster over time, we confirmed that the *z*-test using BC2 produced closer to nominal size compared to BC0 and BC1. However, this test was underpowered. With as few as eight clusters, a *t*-test coupled with BC1 or BC3 might be favored. Similarly, the use of BC1 with a *t*-test was recommended by Teerenstra et al. (2010) in a three-level CRT involving as few as 10 clusters. We noticed that the *t*-test with the model-based variance was conservative under the null. Even though its empirical power remained unaffected across the effect sizes we investigated, the model-based variance may produce lower power than the use of either BC1 or BC3 for smaller effect sizes (Web Figure 5). Finally, the *t*-test might not be universally recommended. In extreme cases where a single cluster crosses over at each step, the number of mean model parameters in (1) exceeds the number of clusters and the *t*-test degrees of freedom become inappropriate. To overcome this issue, one could assume a linear time trend in model (1) and use this parsimonious parameterization to estimate sample size and conduct subsequent analysis. However, such a design may be discouraged for practical reasons since it takes longer to finish, and will likely increase the burden to collect repeated measurements (Hussey and Hughes, 2007).

In summary, the block exchangeable correlation structure represents a three-correlation parameter model and applies to a cohort stepped wedge CRT. For a cross-sectional design, the nested exchangeable structure can be obtained as a special case of the block exchangeable structure, and our results still apply. We have assumed that the block exchangeable structure is the correctly specified correlation model for the responses. However, the GEE analyses based on sandwich variance estimators are robust to correlation misspecification in that they provide consistent estimation for the intervention effect. If it is anticipated at the design phase that the working correlation model is misspecified, power calculations require supplementation of a hypothesized true correlation structure and the robust sandwich variance should be used to develop a modified sample size procedure (Rochon, 1998).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The research of Dr. John Preisser in this article was partially supported by the North Carolina Translational Research and Clinical Sciences Institute, CTSA grant number UL1TR001111. Li is grateful to the International *Biometric* Society, Eastern North American Region 2018 Student Paper Award Committee for receiving a distinguished paper award in the student paper competition based on an earlier version of this article. The authors thank the Editor, Associate Editor, and two anonymous referees for their critical reading and constructive comments which greatly improved an earlier version of this article.

References

Bennett PN, Daly RM, Fraser SF, Haines T, Barnard R, Ockerby C, et al., (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, 204–210. [PubMed: 24070232]

Fay MP and Graubard BI (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics 57, 1198–1206. [PubMed: 11764261]

Giraudeau B, Ravaud P, and Donner A (2008). Sample size calculation for cluster randomized cross-over trials. Statistics in Medicine 27, 5578–5585. [PubMed: 18646266]

Hemming K, Lilford R, and Girling AJ (2015). Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Statistics in Medicine 34, 181–196. [PubMed: 25346484]

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, 4718–4728. [PubMed: 27350420]

Hughes JP, Granston TS, and Heagerty PJ (2015). Current issues in the design and analysis of stepped wedge trials. Contemporary Clinical Trials 45, 55–60. [PubMed: 26247569]

Hussey MA and Hughes JP (2007). Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials 28, 182–191. [PubMed: 16829207]

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

Li F, Turner EL, and Preisser JS (2018). Optimal allocation of clusters in cohort stepped wedge designs. Statistics and Probability Letters 137, 257–263.

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

Littenberg B and MacLean CD (2006). Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. BMC Medical Research Methodology 6, 20–30. [PubMed: 16672056]

Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala SI, and Wolfson M (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics 63, 935–941. [PubMed: 17825023]

Mancl LA and DeRouen TA (2001). A covariance estimator for GEE with improved small-sample properties. Biometrics 57, 126–134. [PubMed: 11252587]

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, 402–413. [PubMed: 27524396]

Murray DM (1998). Design and Analysis of Group-Randomized Trials. New York, NY: Oxford University Press.

Preisser JS, Lu B, and Qaqish BF (2008). Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. Statistics in Medicine 27, 5764–5785. [PubMed: 18680122]

Preisser JS, Reboussin BA, Song EY, and Wolfson M (2007). The importance and role of intracluster correlations in planning cluster trials. Epidemiology 18, 552–560. [PubMed: 17879427]

Preisser JS, Young ML, Zaccaro DJ, 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, 1235–1254. [PubMed: 12687653]

Qaqish BF (2003). A family of multivariate binary distributions for simulating correlated binary variables. Biometrika 90, 455–463.

Rochon J (1998). Application of GEE procedures for sample size calculations in repeated measures experiments. Statistics in Medicine 17, 1643–1658. [PubMed: 9699236]

Scott JM, DeCamp A, Juraska M, Fay MP, and Gilbert PB (2017). Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Statistical Methods in Medical Research 26, 583–597. [PubMed: 25267551]

Teerenstra S, Lu B, Preisser JS, Van Achterberg T, and Borm GF (2010). Sample size considerations for GEE analyses of three-level cluster randomized trials. Biometrics 66, 1230–1237. [PubMed: 20070297]

Turner EL, Li F, Gallis JA, Prague M, and Murray DM (2017). Review of recent methodological developments in group-randomized trials: part 1–design. American Journal of Public Health 107, 907–915. [PubMed: 28426295]

Woertman W, De Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, and Teerenstra S (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. Journal of Clinical Epidemiology 66, 752–758. [PubMed: 23523551]

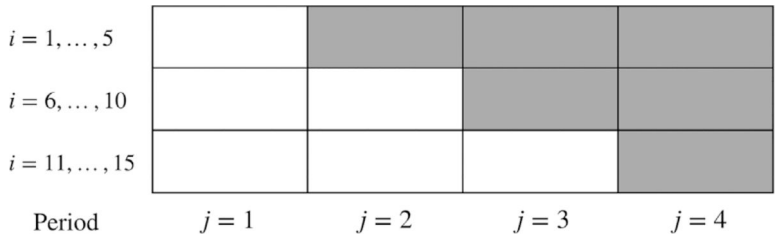

Figure 1.

A schematic illustration of a standard, complete stepped wedge design with $I = 15$ clusters and $T = 4$ periods. There is $b = 1$ baseline measurement and $S = 3$ steps. Only $c_s = 1$ follow-up measurement is taken after each step. Each row represents five randomly selected clusters that cross over to receive intervention at a pre-determined step. A blank cell indicates the control condition and a gray cell indicates the intervention condition.

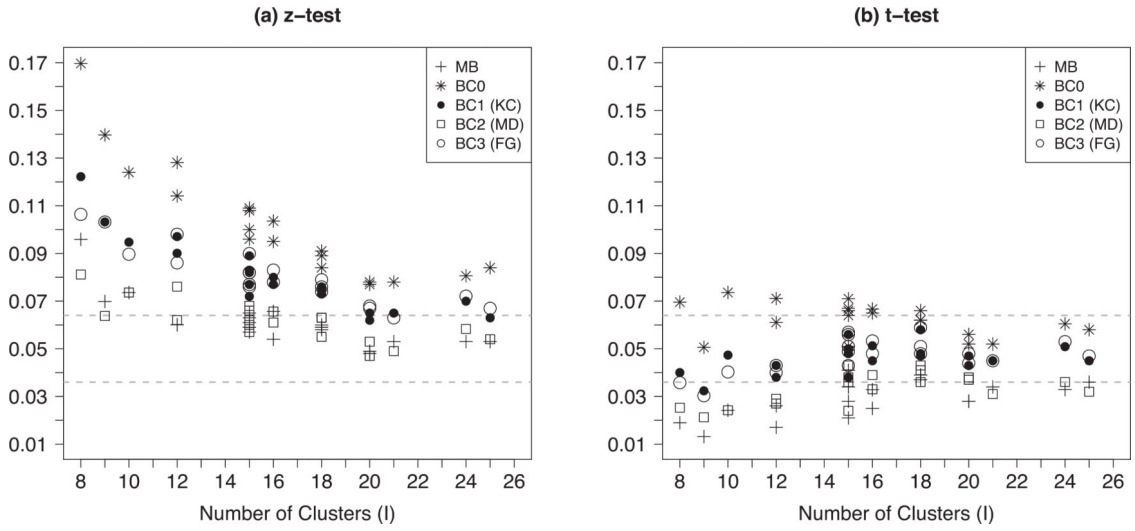

Figure 2.

Empirical type I error rates for GEE-based (a) *z*-tests and (b) *t*-tests for continuous responses. MB: model-based variance; BC0: uncorrected sandwich variance; BC1: KC-corrected sandwich variance; BC2: MD-corrected sandwich variance; BC3: FG-corrected sandwich variance.

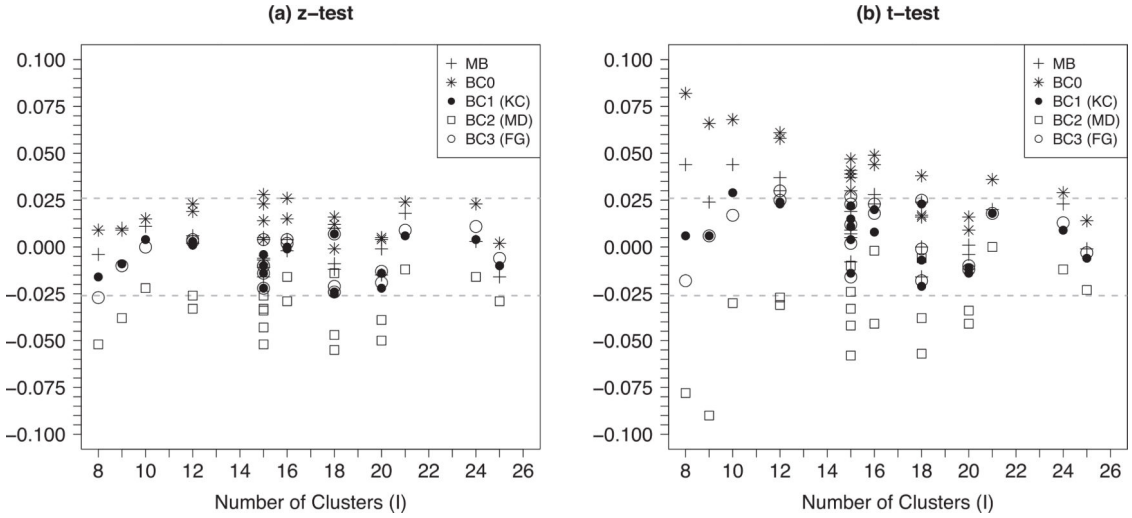

Figure 3.

Differences between the empirical power and the predicted power of GEE-based (a) *z*-tests and (b) *t*-tests for continuous responses. MB: model-based variance; BC0: uncorrected sandwich variance; BC1: KC-corrected sandwich variance; BC2: MD-corrected sandwich variance; BC3: FG-corrected sandwich variance.

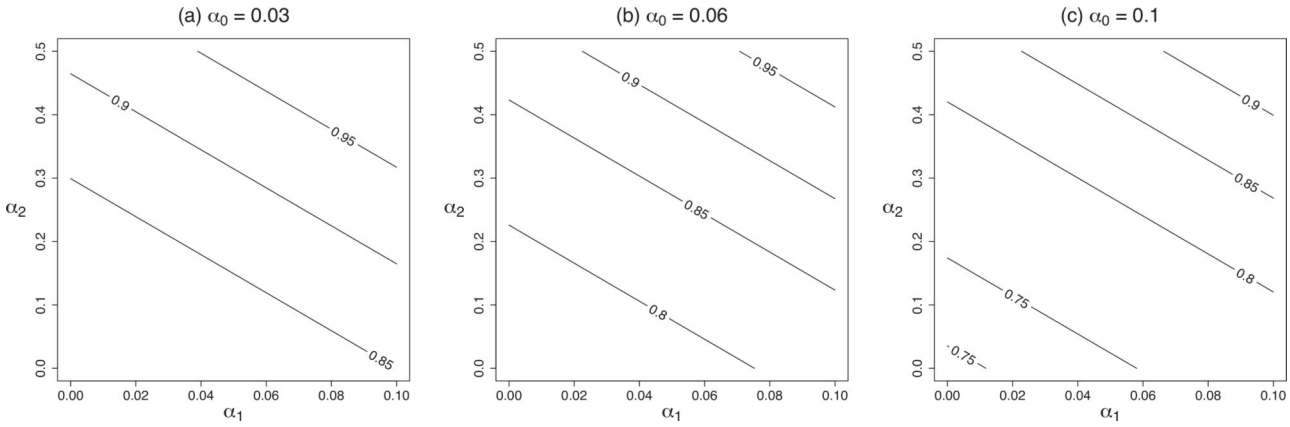

Figure 4.

Predicted power contours as a function of inter-period correlation, α_1, and within-individual correlation, α_2, holding the within-period correlation α_0 = {0.03, 0.06, 0.1}, I = 15, T = 4, and N = 4. The block exchangeable correlation matrix is positive definite across the range of values for all displayed combinations of α_0, α_1, and α_2.

Table 1

Simulation scenarios, predicted power based on z-test and t-test, along with the corresponding empirical power of GEE analyses using different variance estimators for continuous responses. Bias-corrected estimation of correlation parameters uses MAEE.

					z-test			t-test			
Effect Size[a]	α[b]	I	N	T	Pred[c]	MB[d]	BC2[f]	Pred[d]	MB[d]	BC1[e]	BC3[g]
0.65	A1	9	11	4	0.974	0.984	0.936	0.838	0.862	0.844	0.844
0.65	A1	8	24	3	0.965	0.961	0.913	0.812	0.856	0.818	0.794
0.65	A1	15	4	4	0.904	0.888	0.852	0.837	0.829	0.823	0.821
0.65	A1	10	16	3	0.956	0.967	0.934	0.866	0.910	0.895	0.883
0.65	A1	12	6	4	0.938	0.944	0.912	0.852	0.882	0.875	0.877
0.40	A1	12	14	5	0.941	0.943	0.908	0.838	0.875	0.862	0.868
0.40	A1	20	6	5	0.896	0.895	0.857	0.850	0.851	0.839	0.840
0.40	A1	15	12	4	0.893	0.898	0.860	0.824	0.843	0.835	0.836
0.40	A1	21	8	4	0.896	0.914	0.884	0.856	0.876	0.874	0.874
0.40	A1	15	8	6	0.938	0.927	0.912	0.866	0.871	0.888	0.893
0.40	A2	16	12	5	0.887	0.891	0.858	0.815	0.838	0.823	0.833
0.40	A2	18	15	4	0.866	0.854	0.811	0.810	0.794	0.789	0.792
0.40	A2	15	10	6	0.900	0.894	0.866	0.812	0.819	0.827	0.835
0.40	A2	18	6	7	0.909	0.900	0.862	0.844	0.844	0.837	0.843
0.40	A2	15	25	4	0.874	0.878	0.831	0.800	0.809	0.804	0.802
0.35	A3	18	10	4	0.913	0.923	0.899	0.867	0.891	0.890	0.892
0.35	A3	16	9	5	0.932	0.930	0.916	0.875	0.903	0.895	0.898
0.35	A3	20	6	5	0.890	0.875	0.840	0.842	0.838	0.828	0.830
0.25	A3	25	8	6	0.899	0.883	0.870	0.864	0.863	0.858	0.861
0.25	A3	24	7	7	0.899	0.902	0.883	0.859	0.882	0.868	0.872

[a] Effect size $\delta/\phi^{1/2}$.

[b] A1: $\boldsymbol{\alpha}$ = (0.03, 0.015, 0.2); A2: $\boldsymbol{\alpha}$ = (0.1, 0.05, 0.2); A3: $\boldsymbol{\alpha}$ = (0.01, 0.005, 0.4).

[c] Pred: Predicted power.

[d] MB: Model-based variance.

[e] BC1: Bias-corrected sandwich variance of Kauermann and Carroll (2001).

[f] BC2: Bias-corrected sandwich variance of Mancl and DeRouen (2001).

[g] BC3: Bias-corrected sandwich variance of Fay and Graubard (2001).