Power considerations for generalized estimating equations analyses of four-level cluster randomized trials

Xueqi Wang[*,1,2], **Elizabeth L. Turner**[1,2], **John S. Preisser**[3], **Fan Li**[4,5]

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, 27707, USA

[2]Duke Global Health Institute, Durham, NC, 27707, USA

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

[4]Department of Biostatistics, Yale University School of Public Health, New Haven, CT, 06511, USA

[5]Center for Methods in Implementation and Prevention Science, Yale University, New Haven, CT, 06511, USA

Abstract

In this article, we develop methods for sample size and power calculations in four-level intervention studies when intervention assignment is carried out at any level, with a particular focus on cluster randomized trials (CRTs). CRTs involving four levels are becoming popular in health care research, where the effects are measured, for example, from evaluations (level 1) within participants (level 2) in divisions (level 3) that are nested in clusters (level 4). In such multi-level CRTs, we consider three types of intraclass correlations between different evaluations to account for such clustering: that of the same participant, that of different participants from the same division, and that of different participants from different divisions in the same cluster. Assuming arbitrary link and variance functions, with the proposed correlation structure as the true correlation structure, closed-form sample size formulas for randomization carried out at any level (including individually randomized trials within a four-level clustered structure) are derived based on the generalized estimating equations approach using the model-based variance and using the sandwich variance with an independence working correlation matrix. We demonstrate that empirical power corresponds well with that predicted by the proposed method for as few as 8 clusters, when data are analyzed using the matrix-adjusted estimating equations for the correlation parameters with a bias-corrected sandwich variance estimator, under both balanced and unbalanced designs.

[*]Corresponding author: xueqi.wang@duke.edu, Phone: +1-217-898-5630.

1 Introduction

Cluster randomized trials (CRTs) are commonly used to study the effectiveness of health care interventions within a pragmatic trials framework (Weinfurt et al., 2017). In CRTs, the unit of randomization is typically a group (or cluster) of individuals with outcome measurements taken on individuals themselves. Reasons to randomize clusters rather than individuals include administrative convenience and prevention of intervention contamination (Murray, 1998; Turner et al., 2017). CRTs typically include two levels, for example, where patients are nested within clinics that are randomized to intervention conditions. With an increase in the number of pragmatic trials embedded within health care systems, recent CRTs have been designed with multiple hierarchical levels involving, for example, patients, providers, and health care systems (Heo and Leon, 2008; Teerenstra et al., 2010; Liu and Colditz, 2020). Moreover, depending on the nature of the intervention and of the context, the unit of randomization may be at the highest level or one of the lower levels in the hierarchy. The inherent hierarchical structure of the health care delivery system demands rigorous multi-level methods that enable a precise evaluation of health care interventions.

While methods for designing three-level CRTs have been previously developed in Heo and Leon (2008), Teerenstra et al. (2010) and Cunningham and Johnson (2016), there has been little development on methods for designing CRTs with more than three levels and limited work on categorical outcomes for CRTs with more than two levels. In particular, we know of only one approach for four-level stepped wedge CRTs with methodology based on a linear mixed model (Teerenstra et al., 2019). Existing methods and their scope for designing trials with more than two levels are presented in Table 1. In particular, our two motivating examples pertain to CRTs with four levels, which necessitates new considerations on the within-cluster correlation structure and sample size determination. The first example is the Reducing Stigma among Healthcare Providers (RESHAPE) CRT conducted in Nepal, which evaluates a new intervention to improve accuracy of mental illness diagnosis in comparison to implementation as usual (IAU). In the RESHAPE CRT, binary diagnosis outcomes are observed for patients (level 1) nested in providers (level 2), who are nested in health facilities (level 3) within municipality (level 4), which is the unit of randomization. Figure 1 provides a hierarchical illustration of the RESHAPE trial. Our second example is the Health and Literacy Intervention (HALI) trial conducted in Kenya, which compared a literacy intervention to improve early literacy outcomes to usual practice (Jukes et al., 2017). In the HALI trial, repeated literacy continuous outcomes (level 1) are measured for children (level 2), who are nested within schools (level 3) of each Teacher Advisory Center (TAC) tutor zone (level 4). The TAC tutor zone is the unit of randomization which, like the RESHAPE trial, is at the highest level and which we refer to as the *cluster*.

To account for multiple levels of clustering, both marginal (population-averaged) and conditional (cluster-specific) models can be used to analyze the outcome data. In this article, we focus on developing new sample size procedures for four-level CRTs based on the marginal model due to its population-averaged interpretation which is of direct relevance to health-systems level research questions (Preisser et al., 2003). In particular, the intervention effect parameter in the marginal model describes the difference of average outcome between the source population included in the intervention and control clusters, and its interpretation is unaffected by the specification of the within-cluster correlation structure (Zeger et al., 1988). We characterize an *extended nested exchangeable* correlation structure appropriate for the four-level CRT, and develop general sample size equations assuming arbitrary link and variance functions. Special cases of continuous, binary and count outcomes with commonly-used link functions are presented, which generalize existing sample size formulas for two-level and three-level CRTs (Shih, 1997; Heo and Leon, 2008; Teerenstra et al., 2010). When the randomization is carried out at the highest level, we show that the variance inflation factor (VIF) due to the four-level clustered structure is equal to the largest eigenvalue of the extended nested exchangeable correlation structure, which depends on the number of units at each level and on three intraclass correlation coefficients (ICCs). We carried out an extensive simulation study to demonstrate the accuracy and robustness of the proposed sample size formula with a binary outcome, and applied our method to design the RESHAPE and HALI trials. Finally, although we focus on randomization at the highest level as in our motivating CRTs, we also develop more general results for sample size calculation to allow randomization at lower levels.

2   Generalized estimating equations and finite-sample adjustments

We first consider a four-level CRT, and use the RESHAPE trial (Figure 1) as a running example for illustration. Let Y_{ijkl} denote the outcome for patient $l = 1, \dots, L_{ijk}$ from provider $k = 1, \dots, K_{ij}$ nested in health facility $j = 1, \dots, M_i$ of municipality $i = 1, \dots, N$, and $\mathbf{X}_{ijkl} = (X_{ijkl1}, \dots, X_{ijklp})'$ denote a list of p covariates. Frequently in designing CRTs, $\mathbf{X}_{ijkl}$ only includes an intercept and a cluster-level intervention indicator, which equals one if the cluster is assigned to intervention and zero if the cluster is assigned to control. Let $\mu_{ijkl} = \mathrm{E}(Y_{ijkl}|\mathbf{X}_{ijkl})$ be the marginal mean outcome given $\mathbf{X}_{ijkl}$, which is specified via a generalized linear model

$$g(\mu_{ijkl}) = \mathbf{X}'_{ijkl}\boldsymbol{\beta}, \tag{1}$$

where g is a link function and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. The marginal variance function is specified as $\mathrm{var}(Y_{ijkl}|\mathbf{X}_{ijkl}) = \phi v_{ijkl}$, where ϕ is the common dispersion parameter and v_{ijkl} is an arbitrary variance function that depends on the marginal mean and possibly additional dispersion parameters. Denote $\kappa_{ijkl} = \sqrt{\phi v_{ijkl}}/\mu_{ijkl}$ as the coefficient of variation (CV) of the outcome. In addition to the marginal mean model, we propose to characterize the degree of similarity among the within-cluster outcomes through an extended nested exchangeable correlation structure with the following assumptions:

   i.   the correlation between different patients from the same provider is

$$\text{corr}\big(Y_{ijkl}, Y_{ijkl'}|\mathbf{X}_{ijkl}, \mathbf{X}_{ijkl'}\big) = \alpha_0, \text{ for } l \neq l';$$

ii. the correlation between patients from different providers but within the same health facility is

$$\text{corr}\big(Y_{ijkl}, Y_{ijk'l'}|\mathbf{X}_{ijkl}, \mathbf{X}_{ijk'l'}\big) = \alpha_1, \text{ for } k \neq k';$$

iii. the correlation between patients from different health facilities but within the same municipality is

$$\text{corr}\big(Y_{ijkl}, Y_{ij'k'l'}|\mathbf{X}_{ijkl}, \mathbf{X}_{ij'k'l'}\big) = \alpha_2, \text{ for } j \neq j'.$$

We term this correlation structure as the extended nested exchangeable structure because it accounts for an additional hierarchical nesting compared to the nested exchangeable structure developed for three-level CRTs in Teerenstra et al. (2010). Our three-correlation structure also differs from an existing three-parameter block exchangeable structure proposed for cohort stepped wedge designs (Li et al., 2018). To illustrate these differences in correlation structures, we provide specific matrix examples in Web Appendix A.

For simplicity, we assume a balanced design so that there is an equal number of units at each level across all clusters, i.e., $M_i = M \geq 2$ ($i = 1, \dots, N$), $K_{ij} = K \geq 2$ ($i = 1, \dots, N; j = 1, \dots, M_i$), and $L_{ijk} = L \geq 2$ ($i = 1, \dots, N; j = 1, \dots, M_i; k = 1, \dots, K_{ij}$). For each provider, let $\mathbf{Y}_{ijk} = \big(Y_{ijk1}, \dots, Y_{ijkL}\big)'$ and $\boldsymbol{\mu}_{ijk} = \big(\mu_{ijk1}, \dots, \mu_{ijkL}\big)'$ be the $L \times 1$ vector of outcomes and $L \times 1$ marginal mean vector, respectively. Furthermore, in each municipality, denote $\mathbf{Y}_i = (\mathbf{Y}'_{i11}, \mathbf{Y}'_{i12}, \dots \mathbf{Y}'_{iMK})'$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}'_{i11}, \boldsymbol{\mu}'_{i12}, \dots, \boldsymbol{\mu}'_{iMK})'$, where $\mathbf{Y}_i$ and $\boldsymbol{\mu}_i$ are of dimension MKL and include MK blocks of provider-specific outcome vectors (classified by combinations of different health facilities and providers); denote $\mathbf{X}_i$ as the $MKL \times p$ covariate matrix. We use the generalized estimating equations (GEE) approach (Liang and Zeger, 1986) to estimate the intervention effect in mean model (1). Define $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \beta'$, and let $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}$ be a working covariance matrix for $\mathbf{Y}_i$, where $\mathbf{A}_i$ is a MKL-dimensional diagonal matrix with elements of ϕv_{ijkl}, and $\mathbf{R}_i = \mathbf{R}_i(\alpha)$ is a working correlation matrix specified by the ICC vector $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$. The extended nested exchangeable correlation structure can be concisely represented as

$$\mathbf{R}_i = (1 - \alpha_0)\mathbf{I}_{MKL} + (\alpha_0 - \alpha_1)\mathbf{I}_{MK} \otimes \mathbf{J}_L + (\alpha_1 - \alpha_2)\mathbf{I}_M \otimes \mathbf{J}_{KL} + \alpha_2 \mathbf{J}_{MKL}, \quad (2)$$

where $\otimes$ is the Kronecker product, $\mathbf{J}_u = 1_u 1'_u$ is a $u \times u$ matrix of ones, and $\mathbf{I}_u$ is a $u \times u$ identity matrix. Of note, when we equate $\alpha_0 = \alpha_1$ or $\alpha_1 = \alpha_2$, $\mathbf{R}_i$ reduces to the two-parameter nested exchangeable correlation structure developed in Teerenstra et al. (2010). The following Theorem provides a closed-form characterization of the eigenvalues of $\mathbf{R}_i$. The proof of Theorem 2.1 can be found in Web Appendix B.

Theorem 2.1 *The extended nested exchangeable correlation structure has the following eigenvalues:*

$$\lambda_1 = 1 - \alpha_0,$$

$$\lambda_2 = 1 + (L-1)\alpha_0 - L_{\alpha_1},$$

$$\lambda_3 = 1 + (L-1)\alpha_0 + L(K-1)\alpha_1 - LK\alpha_2,$$

$$\lambda_4 = 1 + (L-1)\alpha_0 + L(K-1)\alpha_1 + LK(M-1)\alpha_2,$$

with multiplicity $MK(L-1)$, $M(K-1)$, $M-1$ *and* 1, *respectively.*

Remark 2.2 When (a) $\alpha_0 \neq \alpha_1$, $\alpha_1 \neq \alpha_2$, and $\alpha_2 \neq 0$, and (b) $\alpha_0 \neq -(K-1)\alpha_1 + K\alpha_2$, $\alpha_0 \neq -(K-1)\alpha_1 - K(M-1)\alpha_2$, and $\alpha_1 \neq -(M-1)\alpha_2$, Theorem 2.1 implies that the extended nested exchangeable correlation structure has four distinct eigenvalues. Otherwise, at least two elements of $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are identical and the multiplicity of each distinct eigenvalue can be inferred from Theorem 2.1 by simple addition. For example, if $\alpha_0 = \alpha_1 (= 0$ or not), but the rest of the conditions in (a) and (b) hold, the extended nested exchangeable correlation structure has three distinct eigenvalues, with $\lambda_1 = \lambda_2$ whose multiplicity is $MK(L-1) + M(K-1) = M(KL-1)$.

These explicit forms of the eigenvalues developed in Theorem 2.1 facilitate an efficient determination of the joint validity of the correlation parameters. Specifically, valid values for $(\alpha_0, \alpha_1, \alpha_2)$ should ensure a positive definite $\mathbf{R}_i$ and are contained in the convex open set defined by $\min\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} > 0$. When the marginal mean model includes only an intercept and cluster-level intervention status, there is often an additional natural restriction on the range of α for binary outcomes given in Equation (8) of Qaqish (2003). When $\mathbf{X}_{ijkl}$ only includes an intercept and a cluster-level intervention indicator, it is straightforward to show that the upper bound of α is one, and the lower bound is negative (this lower bound depends on P_0 and P_1, which are the marginal prevalences of the outcome in the control and intervention arms, respectively). In practice, the ICC values are assumed to be positive and the natural constraints are satisfied. We will maintain the positive ICC assumption for the rest of the article.

In general, the GEE estimator $\hat{\beta}$ solves the β-estimating equations $\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i) = 0$. Because the ICC parameters are of interest when analyzing CRTs, we specify a second set of α-estimating equations to iteratively update the correlations that parameterize $\mathbf{V}_i$. In the presence of an unknown dispersion parameter, an additional estimating equation for ϕ needs to be specified beyond the β-estimating equations and α-estimating equations. To provide a correction to the small-sample bias in estimating α due to a limited number of clusters, we

adopt the matrix-adjusted estimating equations (MAEE) of Preisser et al. (2008); also see Appendix B of Li et al. (2018) for details of MAEE. As the number of clusters increases, $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a multivariate normal random vector with mean 0 and covariance estimated by the model-based variance $\mathbf{V}_R^{\text{model}} = \hat{\Sigma}_1^{-1} = \left\{ N^{-1} \sum_{i=1}^N \mathbf{D}_i'(\hat{\boldsymbol{\beta}}) \mathbf{V}_i^{-1}(\hat{\alpha}) \mathbf{D}_i(\hat{\boldsymbol{\beta}}) \right\}^{-1}$, or by the sandwich variance $\mathbf{V}_R^{\text{sandwich}} = \hat{\Sigma}_1^{-1} \hat{\Sigma}_0 \hat{\Sigma}_1^{-1}$ where

$$\hat{\Sigma}_0 = N^{-1} \sum_{i=1}^N \mathbf{C}_i \mathbf{D}_i'(\hat{\boldsymbol{\beta}}) \mathbf{V}_i^{-1}(\hat{\alpha}) (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)' \mathbf{V}_i^{-1}(\hat{\alpha}) \mathbf{D}_i(\hat{\boldsymbol{\beta}}) \mathbf{C}_i', \tag{3}$$

and $\mathbf{C}_i$ represents a multiplicative factor for small-sample bias correction in variance estimation. While both $\mathbf{V}_R^{\text{model}}$ and $\mathbf{V}_R^{\text{sandwich}}$ provide adequate quantification of the uncertainty in estimating $\hat{\boldsymbol{\beta}}$ when the extended nested exchangeable correlation structure is correctly specified, only the sandwich variance $\mathbf{V}_R^{\text{sandwich}}$ is asymptotically valid when the working correlation matrix is misspecified. For example, the independence working correlation matrix is misspecified when the true correlation structure follows the extended nested exchangeable structure. Furthermore, the extended nested exchangeable structure can also be misspecified if the true correlation structure is more complex with more than three ICC parameters, for instance, when the ICC parameters depend on covariates such as intervention arm or other cluster characteristics. In this article, we consider only a single example of misspecification, namely the first case with the independence working correlation matrix, with details discussed in Section 3.3.

In pragmatic CRTs, there is usually a limited number of units at the highest level. In the RESHAPE trial, it is only possible to randomize a total of 24 municipalities. Therefore, adjustments to the sandwich variance $\mathbf{V}_R^{\text{sandwich}}$ are required to reduce its potentially negative bias (Li and Redden, 2015). Setting $\mathbf{C}_i = \mathbf{I}_p$ in (3) provides the uncorrected sandwich estimator of Liang and Zeger (1986), denoted as BC0. Because BC0 tends to underestimate the variance when the number of clusters is small, we consider four types of small-sample adjustments. Define matrix $\mathbf{Q}_i = \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i (N \Sigma_1)^{-1}$. Setting $\mathbf{C}_i = (\mathbf{I}_p - \mathbf{Q}_i)^{-1/2}$ provides the bias-corrected variance of Kauermann and Carroll (2001), or BC1 (here we provide an equivalent representation based on the $\mathbf{Q}_i$ matrix instead of the cluster-leverage matrix; we provide additional details in Web Appendix C to clarify this subtle equivalence). Setting $\mathbf{C}_i = (\mathbf{I}_p - \mathbf{Q}_i)^{-1}$ provides the bias-corrected variance of Mancl and DeRouen (2001), or BC2. Setting $\mathbf{C}_i = \text{diag}\left[\left\{ 1 - \min(r, [\mathbf{Q}_i]_{jj}) \right\}^{-1/2} \right]$, where $r < 1$ is a user-defined bound with a default value of 0.75, provides the bias-corrected variance of Fay and Graubard (2001), or BC3. Based on the degree of multiplicative adjustments, we generally have BC0 < BC1 ≈ BC3 < BC2 (Preisser et al., 2008; Li et al., 2018). In addition to the three multiplicative adjustments, we also consider an additive bias correction of Morel et al. (2003), defined as BC4 = cBC0 + $\delta_N \phi \Sigma_1^{-1}$, where $c = \{(f - 1)/(f - p)\} \times \{N/(N - 1)\}$,

$f = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} L_{ijk}$ is the total number of observations, $\delta_N = \min\{0.5, p/(N-p)\}$ is the correction factor converging to 0 as N tends to infinity, and

$$\phi = \max\left[1, \text{trace}\left\{c\left(\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \hat{\mu}_i)(\mathbf{Y}_i - \hat{\mu}_i)' \mathbf{V}_i^{-1} \mathbf{D}_i\right)(N\Sigma_1)^{-1}\right\}/p\right].$$

One potential advantage of the additive bias correction BC4 over the multiplicative bias correction is that the former guarantees the positive definiteness of the estimated covariance (Morel et al., 2003). The general MAEE methods are implemented in the recent R package geeCRT (Yu et al., 2020). Source R code to implement MAEE in four-level CRTs (including BC4) are also available at https://github.com/XueqiWang/Four-Level_CRT.

3 Power and sample size considerations

At the design stage of a balanced four-level clustered trial, we consider an unadjusted marginal model (1) with $\mathbf{X}_{ijkl}$ including only an intercept and a binary cluster-level intervention indicator ($p = 2$). This simple structure is assumed for the rest of the article. Suppose we are interested in testing the null hypothesis of no intervention effect $H_0 : \beta_2 = 0$, using a two-sided t-test. Specifically, the asymptotic distribution of $\sqrt{N}(\hat{\beta}_2 - \beta_2)$ is normal with mean 0 and variance obtained as the lower-right element of $\text{cov}\{\sqrt{N}(\hat{\beta} - \beta)\}$. The Wald test statistic $\sqrt{N}\hat{\beta}_2/\sigma_\beta$, where $\sigma_\beta^2 = \text{var}(\sqrt{N}\hat{\beta}_2)$, will be compared to a t-distribution with N–2 degrees of freedom; the t-test is chosen because it often improves the test size in finite samples compared to normal approximations (Li and Redden, 2015; Teerenstra et al., 2010). The predicted power to detect an effect size b on the link function scale with a nominal type I error rate α is then

$$1 - \gamma = \Phi_{t, N-2}\left(t_{\alpha/2, N-2} + |b|\sqrt{\frac{N}{\sigma_\beta^2}}\right), \tag{4}$$

where $\Phi_{t,d}$ and $t_{\alpha,d}$ are the cumulative distribution function and $100\alpha\%$ percentile of the t-distribution with d degrees of freedom, respectively. Accordingly, the number of clusters or level-four units required to achieve $100(1 - \gamma)\%$ power must satisfy

$$N \geq \left(t_{\alpha/2, N-2} + t_{\gamma, N-2}\right)^2 \frac{\sigma_\beta^2}{b^2}. \tag{5}$$

Note that the above two equations are approximations, and Equation (5) must be solved iteratively. Equations (4) and (5) suggest that analytical power and sample size calculations depend on an explicit expression for the asymptotic variance σ_β^2. In what follows, we will assume the true correlation matrix of $\mathbf{Y}_i$ is extended nested exchangeable defined in Equation (2), and determine the explicit form of σ_β^2 when the working correlation is the true correlation structure or when the working correlation assumes independence.

3.1 A general sample size expression with randomization at level four

Mimicking our motivating applications, we consider a total of N clusters or level-four units are randomized with $\pi_c N$ clusters assigned to the control arm and $(1 - \pi_c)N$ clusters to the intervention arm. Let $\mathbf{X}_c$ and $\mathbf{X}_t$ be the design matrix for clusters assigned to the control and intervention conditions, respectively, i.e., $\mathbf{X}_c = (\mathbf{1}_{MKL}, \mathbf{0}_{MKL})$ and $\mathbf{X}_t = (\mathbf{1}_{MKL}, \mathbf{1}_{MKL})$. Below, we consider a general sample size expression for outcomes with an arbitrary mean-variance relationship. To do so, we further define μ_c and μ_t to be the marginal mean for the control and intervention arms, respectively. Because the randomization is carried out at level four, the marginal means of all observations in each particular cluster (μ) are either all μ_c or all μ_t, depending on the intervention assignment. We also define κ_c and κ_t as the marginal CV of outcome in the control and intervention arms, respectively. Assuming an arbitrary link function g, we have $\mathbf{D}_i = (\partial g(\mu_c)/\partial \mu_c)^{-1}\mathbf{X}_c$ and $\mathbf{V}_i = \mu_c^2 \kappa_c^2 \mathbf{R}_i$ for all i in the control arm, $\mathbf{D}_i = (\partial g(\mu_t)/\partial \mu_t)^{-1}\mathbf{X}_t$ and $\mathbf{V}_i = \mu_t^2 \kappa_t^2 \mathbf{R}_i$ for all i in the intervention arm.

In this Section, we assume that the working correlation model is the extended exchangeable correlation structure, with correlation parameters estimated via the MAEE approach. Therefore σ_β^2 is the lower-right element of the model-based variance, given by the inverse of

$$\boldsymbol{\Sigma}_1 = N^{-1}\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i = \left(\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}\right)\begin{pmatrix} \pi_c \rho_c^{-2} + (1-\pi_c)\rho_t^{-2} & (1-\pi_c)\rho_t^{-2} \\ (1-\pi_c)\rho_t^{-2} & (1-\pi_c)\rho_t^{-2} \end{pmatrix},$$

where $\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}$ is the sum of all elements of the inverse correlation matrix $\mathbf{R}^{-1}$, $\rho_c = \mu_c \kappa_c \{\partial g(\mu_c)/\partial \mu_c\}$, and $\rho_t = \mu_t \kappa_t \{\partial g(\mu_t)/\partial \mu_t\}$. In Web Appendix D, we show that an explicit inverse of the extended nested exchangeable correlation matrix is

$$\mathbf{R}^{-1} = \frac{1}{\lambda_1}\mathbf{I}_{MKL} - \frac{\alpha_0 - \alpha_1}{\lambda_1 \lambda_2}\mathbf{I}_{MK}\otimes\mathbf{J}_L - \frac{\alpha_1 - \alpha_2}{\lambda_2 \lambda_3}\mathbf{I}_M\otimes\mathbf{J}_{KL} - \frac{\alpha_2}{\lambda_3 \lambda_4}\mathbf{J}_{MKL},$$

which gives $\mathbf{1}'\mathbf{R}^{-1}\mathbf{1} = MKL/\lambda_4$, where λ_1, λ_2, λ_3 and λ_4 are defined in Theorem 2.1. These intermediate results allow us to obtain the asymptotic variance expression as

$$\sigma_\beta^2 = \frac{\lambda_4}{MKL}\left(\frac{\rho_c^2}{\pi_c} + \frac{\rho_t^2}{1-\pi_c}\right). \tag{6}$$

Hence the required number of clusters must satisfy

$$N \geq \frac{\left(t_{\alpha/2, N-2} + t_{\gamma, N-2}\right)^2}{b^2} \times \frac{\lambda_4}{MKL}\left(\frac{\rho_c^2}{\pi_c} + \frac{\rho_t^2}{1-\pi_c}\right). \tag{7}$$

The variance expression (6) has several important implications for study planning, which we discuss in a series of remarks below.

Remark 3.1 In the absence of any clustering, one can easily derive the asymptotic variance of the intervention effect estimator in an individually randomized trial as

$$\tilde{\sigma}_\beta^2 = \frac{1}{MKL}\left(\frac{\rho_{\tilde{c}}^2}{\pi_c} + \frac{\rho_t^2}{1 - \pi_c}\right),$$

assuming a total of $N \times MKL$ individuals are recruited. As a consequence, the eigenvalue $\lambda_4 = \sigma_\beta^2/\tilde{\sigma}_\beta^2$ is the design effect or VIF due to the four-level clustered structure. This VIF generalizes the previous findings in two-level and three-level CRTs, where the corresponding VIFs have also been found to be the largest eigenvalues of the simple exchangeable and nested exchangeable correlation matrices, respectively (see Shih (1997), Liu et al. (2019) and discussions in Section 7 of Li et al. (2019) for VIF in two-level and three-level CRTs).

Remark 3.2 Because $\lambda_4 = 1 + (L-1)a_0 + L(K-1)a_1 + LK(M-1)a_2$ is monotonically increasing in all three ICC parameters, a_0, a_1 and a_2, larger values of any of these three ICC values would inflate the variance and decrease the design efficiency. Interestingly, the coefficient of each ICC parameter in the equation for λ_4 determines the relative importance of that ICC for variance inflation, and the sample size or variance inflation is most sensitive to changes in the correlation between patients from different health facilities within the same municipality (or the ICC at the highest hierarchy), a_2, which has the largest coefficient $LK(M-1)$. On the other hand, ignoring the fourth level of clustering by assuming $a_1 = a_2$ (as in a three-level CRT) could lead to a highly inaccurate power calculation if these two correlations are, in fact, different. That is, assuming the value of a_1 is fixed and known, when the true value of a_2 is larger than a_1, the required sample size would be underestimated; when the true value of a_2 is smaller than a_1, the required sample size would be overestimated. Ignoring the ICC at the lowest hierarchy, a_0, however, may have a relatively smaller impact on the power calculation, especially when the patient panel size L is not too large.

Remark 3.3 Expression (6) also suggests the optimal allocation of clusters does not depend on the ICC parameters. Specifically, the optimal randomization proportion that leads to the smallest σ_β^2 is obtained as $\pi_c^{\mathrm{opt}} = 1/2$ if $\rho_c = \rho_t$, and $\pi_c^{\mathrm{opt}} = \left(\rho_c^2 \pm |\rho_c\rho_t|\right)/\left(\rho_c^2 - \rho_t^2\right)$ (depending on which value is contained within the unit interval) otherwise. Details of the derivation for π_c^{opt} are provided in Web Appendix E.

Finally, for commonly-used link functions and different types of outcomes, we can elegantly specify the marginal variance v as a function of the marginal mean μ. In those cases, sample size Equation (7) can be recognized as more familiar expressions, examples of which are provided below.

Example 3.4 (**Sample size formula with a continuous outcome**) When Y_{ijkl} is continuous, and a Gaussian variance function is assumed, we have $v = 1$ for both trial arms, and the

marginal CV of the outcome $\kappa_c = \sqrt{\phi}/\mu_c$ and $\kappa_t = \sqrt{\phi}/\mu_t$. Under the identity link function, $\partial g(\mu_t)/\partial \mu_t = \partial g(\mu_c)/\partial \mu_c = 1$, and Equation (7) reduces to

$$N \geq \frac{\left(t_{\alpha/2,\, N-2} + t_{\gamma,\, N-2}\right)^2}{b^2} \times \frac{\phi \lambda_4}{\pi_c(1-\pi_c)MKL}. \tag{8}$$

In this case, ϕ is assumed as the homoscedastic marginal total variance of Y_{ijkl}, and the sample size expression (8) extends the results of Heo and Leon (2008) and Teerenstra et al. (2010) to account for an additional level of clustering. Additionally, the optimal randomization proportion π_c^{opt} is clearly 1/2 given $\rho_c = \rho_t$.

Example 3.5 (**Sample size formulas with a binary outcome**) When Y_{ijkl} is binary, the dispersion $\phi = 1$, and the variance function $v = \mu(1-\mu)$. This leads to the CV of the outcome in the control and intervention arms as $\kappa_c = \sqrt{(1-\mu_c)/\mu_c}$ and $\kappa_t = \sqrt{(1-\mu_t)/\mu_t}$. We consider the sample size formulas for three commonly-used link functions. Under the canonical logit link function, $g(\mu) = \log(\mu/(1-\mu))$, with $\partial g(\mu)/\partial \mu = \mu^{-1}(1-\mu)^{-1}$. Define $P_0 = (1 + \exp(-\beta_1))^{-1}$, $P_1 = (1 + \exp(-\beta_1 - \beta_2))^{-1}$ as the prevalence in the two arms, the required number of clusters to detect an effect size on the log odds ratio scale $b = \log(P_1/(1-P_1)) - \log(P_0/(1-P_0))$ must satisfy

$$N \geq \frac{\left(t_{\alpha/2,\, N-2} + t_{\gamma,\, N-2}\right)^2}{b^2} \times \frac{\lambda_4}{MKL}\left\{\frac{1}{\pi_c P_0(1-P_0)} + \frac{1}{(1-\pi_c)P_1(1-P_1)}\right\}. \tag{9}$$

Similarly, with an identity link function, we define the prevalence of outcome in the control and intervention clusters by $P_0 = \beta_1$, $P_1 = \beta_1 + \beta_2$, which should satisfy the boundary constraints $P_0, P_1 \in (0, 1)$. The general Equation (7) suggests that the required number of clusters to detect an effect size on the risk difference scale $b = P_1 - P_0$ satisfies

$$N \geq \frac{\left(t_{\alpha/2,\, N-2} + t_{\gamma,\, N-2}\right)^2}{b^2} \times \frac{\lambda_4}{MKL}\left\{\frac{P_0(1-P_0)}{\pi_c} + \frac{P_1(1-P_1)}{1-\pi_c}\right\}. \tag{10}$$

Finally, under the log link function, we write $P_0 = \exp(\beta_1)$ and $P_1 = \exp(\beta_1 + \beta_2)$, and define the effect size on the log relative risk scale to be $b = \log(P_1) - \log(P_0)$. The required sample size is provided by

$$N \geq \frac{\left(t_{\alpha/2,\, N-2} + t_{\gamma,\, N-2}\right)^2}{b^2} \times \frac{\lambda_4}{MKL}\left\{\frac{1-P_0}{\pi_c P_0} + \frac{1-P_1}{(1-\pi_c)P_1}\right\}. \tag{11}$$

Equations (9), (10) and (11) generalize the sample size requirements provided in Teerenstra et al. (2010) and Liu et al. (2019) from three-level CRTs to four-level CRTs.

Example 3.6 (**Sample size formula with a count outcome**) When Y_{ijkl} is a count outcome, a typical approach is to assume Y_{ijkl} follows the Poisson distribution. This assumption leads to $\varphi = 1$, $\nu = \mu$, and the CV of the Poisson outcome in each arm is $\kappa_c = \sqrt{1/\mu_c}$, $\kappa_t = \sqrt{1/\mu_t}$. Under the canonical log link function, $\partial g(\mu)/\partial\mu = 1/\mu$. Therefore, we define $\mu_c = \exp(\beta_1)$ and $\mu_t = \exp(\beta_1 + \beta_2)$ as the expected event rate in the two arms, and the required number of clusters to detect an effect size on the log rate ratio scale $b = \log(\mu_t/\mu_c)$ is given by

$$N \geq \frac{\left(t_{\alpha/2,\, N-2} + t_{\gamma,\, N-2}\right)^2}{b^2} \times \frac{\lambda_4}{MKL}\left\{\frac{1}{\pi_c\exp(\beta_1)} + \frac{1}{(1-\pi_c)\exp(\beta_1+\beta_2)}\right\}. \tag{12}$$

The sample size equation (12) generalizes the result in Amatya et al. (2013) from two-level CRTs to four-level CRTs.

3.2 Randomization at lower levels

Although our two motivating examples represent CRTs that randomize at the fourth level, there may sometimes be administrative or practical considerations for randomizing units at a lower level. For example, when such lower-level randomization would not lead to intervention contamination, there would be an increase in the number of units of randomization and hence in the statistical power. Another intuition for improved power is that the inclusion of within-cluster contrasts in the intervention effect contributes to the increase in effective sample size. For completeness, we provide some general results for randomizing at lower levels. In our running example, the RESHAPE trial, randomizing at level three refers to, within each of the N municipalities, randomizing $\pi_c M$ health facilities to the control arm and $(1 - \pi_c)M$ health facilities to the intervention arm. Likewise, randomizing at level two refers to, within each of NM health facilities, randomizing $\pi_c K$ providers in the control arm and $(1 - \pi_c)K$ providers in the intervention arm. Because patients are nested within each provider, a design that randomizes at level three or level two within the four-level clustered structure would still be classified as a CRT. In contrast, randomizing at level one would be an individually randomized trial for which $\pi_c L$ patients and $(1 - \pi_c)L$ patients are randomized to the control and intervention arms, respectively, within each of the NMK providers. Notice that the design vector $\mathbf{X}_{ijkl}$ can depend on the level of randomization, and we provide full details in Web Appendix F. To link these different contexts, we establish the following Theorem that applies to randomization at any level. The proof is in Web Appendix F.

Theorem 3.7 *Consider a four-level clustered trial, suppose the randomization is carried out at the rth level ($r = 1, 2, 3, 4$), and assume arbitrary link and variance functions. The asymptotic model-based variance for the GEE estimator $\hat{\beta}_2$ that uses the extended nested exchangeable working correlation is*

$$\sigma_{\hat{\beta}}^2 = \frac{\lambda_r}{MKL}\left(\frac{\rho_c^2}{\pi_c} + \frac{\rho_t^2}{1-\pi_c}\right) + \frac{(\lambda_4 - \lambda_r)(\rho_c - \rho_t)^2}{MKL}, \tag{13}$$

where $\rho_c = \mu_c \kappa_c \{\partial g(\mu_c)/\partial \mu_c\}$ and $\rho_t = \mu_t \kappa_t \{\partial g(\mu_t)/\partial \mu_t\}$, μ_c and μ_t are the marginal means for the control and intervention population, respectively, κ_c and κ_t are the marginal coefficients of variation of the outcome for the control and intervention population, respectively.

Clearly, Equation (13) includes Equation (6) as a special case, where the second term of (13) vanishes. Furthermore, Theorem 3.7 has several interesting implications, which we elaborate below in several remarks.

Remark 3.8 When randomizing at a lower level r 3, the design effect relative to an individually randomized trial without any clustering is given by

$$\text{design effect} = \lambda_r + (\lambda_4 - \lambda_r)(\rho_c - \rho_t)^2 \left(\frac{\rho_c^2}{\pi_c} + \frac{\rho_t^2}{1 - \pi_c}\right)^{-1}, \tag{14}$$

which generally depends on the distribution of the outcome only through the first two moments via ρ_c and ρ_t. With a continuous outcome, identity link function and a Gaussian variance function, we have $\rho_c = \rho_t$, and the design effect (14) when randomizing at level r simply reduces to the corresponding eigenvalue λ_r. This result indicates that as we randomize at a lower level, the design efficiency depends on the relative magnitude of the ICC parameters, because $\lambda_2 = \lambda_1 + L(\alpha_0 - \alpha_1)$, $\lambda_3 = \lambda_2 + LK(\alpha_1 - \alpha_2)$, and $\lambda_4 = \lambda_3 + LKM\alpha_2$. In our HALI trial context where we expect $\alpha_0 \geq \alpha_1 \geq \alpha_2$, randomization at a lower level leads to higher efficiency with a continuous outcome. These insights with a continuous outcome are in parallel to the design effect expressions provided in Cunningham and Johnson (2016) in a three-level design with a linear mixed model.

Remark 3.9 Expression (13) suggests that the optimal allocation proportion does not change according to the level of randomization. It is straightforward to see from Equation (13) that the optimal randomization probability π_c^{opt} minimizes $\left\{\rho_c^2/\pi_c + \rho_t^2/(1 - \pi_c)\right\}$ and equals to the result derived in Remark 3.3.

Finally, specific sample size formulas for randomization at lower levels can be derived based on Equation (5) and (13), under commonly-used link functions and variance functions as in Section 3.1. Details of these specific cases are provided in Web Appendix G.

3.3 Considerations on using an independence working correlation

In our derivations of the variance expression σ_β^2, we have assumed that the working correlation structure is correctly specified as the extended nested exchangeable structure, and the ICCs are estimated through MAEE. In the context of clustered trials, using MAEE has been recommended because reporting accurate ICC estimates can inform the design of future trials (Preisser et al., 2008; Teerenstra et al., 2010), and it is considered good practice per the CONSORT extension to CRTs (Campbell et al., 2012). Alternatively, one may pre-specify the primary analysis to be the GEE analysis with an independence working correlation and account for clustering simply via the sandwich variance. In a two-level design with equal cluster sizes, previous studies (Pan, 2001; Yu et al., 2020; Li and Tong, 2021) have found that GEE estimators with working exchangeable and working

independence have the same asymptotic efficiency when the randomization is carried out at the cluster level (i.e. the highest level in a two-level design). In Web Appendix H, we prove the following Theorem, which is a more general result that the same asymptotic equivalence holds in a four-level clustered design with equal "cluster" sizes at each level, regardless of the level of randomization.

Theorem 3.10 *Under balanced designs, suppose the randomization is carried out at the rth level* ($r = 1, 2, 3, 4$), *and assume arbitrary link and variance functions, with the extended nested exchangeable correlation structure as the true correlation structure. Using the sandwich variance with an independence working correlation matrix results in the same* σ_β^2 *as we obtain using the extended nested exchangeable working correlation in Theorem 3.7.*

Therefore, Theorem 3.7 still holds for GEE analysis assuming working independence and there is no difference in the derived sample size equations assuming equal "cluster" sizes at each level, when randomization is carried out at any one of the four levels.

4   Simulation study

Since we expect that the closed-form sample size formula for binary outcomes would be less accurate compared to those for continuous and count outcomes, we conducted a simulation study to evaluate the validity of our sample size formula for binary outcomes under the canonical logit link function for both balanced and unbalanced designs, where the latter has variable numbers of level-one units per level-two unit. Because we derived our sample size formulas assuming an equal number of patients within each provider, we consider the unbalanced design to assess the robustness of our formulas when the balance design assumption fails to hold in a specific and meaningful way (i.e., variable numbers of patients per provider in the RESHAPE trial). We focus on randomization at the fourth level with equal allocation to the two trial arms ($\pi_c = 1/2$), which mimics the motivating studies. Correlated binary data in each cluster were generated from a binomial model with marginal mean in (1) and extended nested exchangeable correlation structure using the methods of Qaqish (2003).

4.1   Balanced designs

We varied correlation values by choosing $a = (a_0, a_1, a_2) = \{(0.4, 0.1, 0.03), (0.15, 0.08, 0.02), (0.1, 0.02, 0.01), (0.05, 0.05, 0.02)\}$; these values resemble estimates from the HALI trial (Jukes et al., 2017) or assumptions for the RESHAPE trial. Marginal mean parameters β_1 and β_2 were induced from the marginal means P_0 in the control arm and P_1 in the intervention arm for assessing the empirical power; β_2 was fixed at 0 for assessing the empirical type I error rate. The nominal type I error rate was fixed at 5%. For illustration, we fixed $M \in \{2, 3\}$, $K \in \{3, 4\}$ and $L \in \{5, 10\}$. The total number of clusters N was determined as the smallest number ensuring that the predicted power was 80%, and ranged from 8 to 30 across 30 simulation scenarios. For each scenario, 1000 data replications were generated and analyzed using GEE for the mean model and MAEE for the extended nested exchangeable working correlation structure. We consider 7 variance estimators for the intervention effect: the model-based variance (MB), BC0, BC1, BC2, the variance estimator

of Ford and Westgate (2017) with standard error obtained as the average of those from BC1 and BC2 (denoted as AVG), BC3, and BC4. The convergence rate exceeded 99% for most scenarios except for a few cases (a summary of convergence rates for all scenarios under balanced designs is presented in Web Table 1). Since the nominal type I error rate was 5%, according to the margin of error from a binomial model with 1000 replications, we considered an empirical type I error rate from 3.6% to 6.4% as acceptable. Similarly, since the predicted power was at least 80% for each scenario, we considered an empirical power differing at most 2.6% from the predicted power as acceptable. These acceptable bounds will be labeled with gray dashed lines in all related Figures summarizing the simulation results.

Table 2 summarizes the results for empirical type I error rates for t-tests using different variance estimators. The type I error rates with BC1 were valid across almost all scenarios, except for only one scenario where the test became slightly liberal (0.068). BC0 often gave inflated type I error rates, while MB, BC2, AVG, BC3, and BC4 sometimes led to overly conservative type I error rates. Table 3 summarizes the results for predicted power and empirical power using different variance estimators. The empirical power with BC1 corresponded well with the predicted power throughout. BC0 provided higher empirical power than predicted in most scenarios, and MB sometimes led to higher empirical power than prediction; AVG, BC3, and BC4 sometimes gave lower empirical power than predicted, while BC2 almost always led to lower empirical power than predicted. Overall, the t-test with BC1 performed best; the performance among AVG, BC3, and BC4 were very similar.

The above simulations were repeated with data fit using GEE under working independence. The results for empirical type I error rates are summarized in Web Table 2, analogous to Table 2; the results for power are summarized in Web Table 3, analogous to Table 3. With working independence, MB is invalid and gave inflated type I error rates. All other variance estimators yielded very similar results to those estimating the extended nested exchangeable working correlation structure with MAEE (Table 2 and Table 3). In fact, we found out that under balanced designs, the variance estimators BC0, BC1, BC2, AVG, and BC3 are numerically equivalent under either working independence or working extended nested exchangeable correlation matrix (minor differences in empirical type I error and power due to non-convergence of MAEE in a few iterations), when the true correlation structure is extended nested exchangeable. We formally state this result in the following Remark, with the proof presented in Web Appendix I.

Remark 4.1 Under balanced designs, suppose the randomization is carried out at the 4th level, and assume arbitrary link and variance functions, with the extended nested exchangeable correlation structure as the true correlation structure. Then, GEE analysis using the extended nested exchangeable working correlation matrix or using an independence working correlation matrix result in the same estimators $\hat{\beta}$, BC0, BC1, BC2, AVG, and BC3.

4.2 Unbalanced designs

Although we derive our sample size formula assuming an equal number of patients within each provider, in practice, providers may have variable numbers of patients. We assessed the robustness of the proposed sample size formula under this specific type of variable

panel size, and further illustrate the comparisons between using the true working correlation model versus the independence working correlation model. For each scenario in Section 4.1, keeping other parameters unchanged, we generated numbers of patients per provider L_{ijk} from a gamma distribution with mean equal to $\bar{L} \in \{5, 10\}$ and CV ranging from $\{0.25, 0.50, 0.75, 1\}$. We computed the required sample size ignoring the cluster size variability, but estimate the empirical power under variable cluster sizes for GEE estimators using MAEE with the extended nested exchangeable working correlation structure versus using working independence. The full results for empirical type I error rates and empirical power using different variance estimators are presented in Web Tables 4–11 for GEE analyses using the extended nested exchangeable working correlation structure with MAEE, in Web Tables 12–19 for GEE analyses using working independence, and in Web Figures 1–8 for comparisons of GEE analyses between using the extended nested exchangeable working correlation structure with MAEE and using working independence. As BC1 performed best under balanced designs, here we focus discussion of the results on BC1. Figure 2 summarizes the empirical type I error rates for t-tests with BC1 across the range of estimated sample sizes, with gray dashed lines indicating the acceptable bounds calculated from a binomial sampling model (detailed calculation of the acceptable bounds is provided in Section 4.1). Using the extended nested exchangeable working correlation structure with MAEE, most scenarios had valid type I error rates, with liberal results in only five cases across all CV values of cluster sizes (Scenario 13 for CV = 0; Scenarios 17 and 30 for CV = 0.25; Scenario 26 for CV = 0.75; Scenario 25 for CV = 1.00). On the other hand, as the CV of cluster sizes increases, the GEE estimator with working independence started to exhibit inflated type I error rates. Figure 3 summarizes the power results for t-tests with BC1, where gray dashed lines indicate the acceptable bounds for difference in empirical versus predicted power (detailed calculation of the acceptable bounds is provided in Section 4.1). Surprisingly, the empirical power corresponded well with that predicted for most scenarios when the data are analyzed by GEE and MAEE with the extended nested exchangeable working correlation structure, regardless of the cluster size variability; across all CV values of cluster sizes, there were only five cases where the empirical power appeared slightly lower than the predicted (Scenarios 9, 24, and 25 for CV = 0.75; Scenarios 9 and 23 for CV = 1.00). In sharp contrast, the empirical power for independence GEE had unacceptable lower empirical power than that predicted when the CV of cluster sizes increased from 0.25 to 1.

5 Applications

5.1 The RESHAPE trial

We apply our sample size formula to determine the required sample size in the RESHAPE trial, which is briefly described in Section 1. The RESHAPE trial compares two strategies – RESHAPE and IAU – for the primary outcome of accuracy of diagnosis of mental illness, defined as the proportion of all patients seen by providers who are accurately diagnosed as determined by psychiatrists. The unit of randomization is the municipality with equal allocation to the two trial arms ($\pi_c = 1/2$), and patient-level binary outcomes are collected to measure diagnostic accuracy. The design question focuses on calculating the required number of municipalities to achieve at least 80% power at the 5% nominal test size.

According to the local conditions in Nepal, we anticipate 3 health facilities per municipality, 3 providers per health facility, and roughly 36 patients per provider. The researchers expect IAU and RESHAPE will result in an accurate diagnosis of $P_0 = 78.5\%$ and $P_1 = 88\%$, respectively. From extensive discussions with the study team and preliminary simulations for the study protocol, the anticipated ICC between different patients of the same provider is $a_0 = 0.05$, the ICC between different patients of different providers from the same health facility is $a_1 = 0.04$, and the ICC between different patients of different providers from different health facilities in the same municipality is $a_2 = 0.03$. Given $M = 3$, $K = 3$, $L = 36$, and $a = (0.05, 0.04, 0.03)$, under the canonical logit link function, Equation (9) and (4) suggested that the required number of municipalities is $N = 22$ with power of 82.65%. The same result can also be obtained by using the design effect formula (14). Specifically, suppose this is an individually randomized trial without any clustering, then the required number of patients is 562 (with the working degrees of freedom of a t-test adjusted to match the four-level CRT design). By multiplying the design effect $\lambda_4 = 12.11$, the required number of patients is 6806 and the required number of municipalities is 6806/(3×3×36) ≈ 22. Although the research team has good faith in a_0, there may be uncertainty in a_1 and a_2. We conducted a sensitivity analysis by varying the values of these two correlations. Figure 4 shows the sensitivity of power as a function of a_1 and a_2 at $a_0 = \{0.025, 0.05, 0.1\}$, assuming $N = 22$ under the logit link function. As expected, the predicted power decreases as a_0, a_1, or a_2 increases. Specifically for $a_0 = 0.05$, power remains above 70% for $a_1 \leq 0.07$ and $a_2 \leq 0.04$. In Web Appendix L, we provide additional illustrative calculations of the required sample size under the identity and log link functions, as well as when randomization is hypothetically carried out in lower levels under the RESHAPE trial context. Noticeably, if randomization is carried out at the patient level (level 1), the required number of municipalities can dramatically reduce to as few as 6.

5.2 The HALI trial

We also apply our method to design the HALI trial introduced in Section 1. The HALI trial compared two strategies – the HALI literacy intervention and the usual instruction – in terms of children's literacy, assessed by various tests such as for spelling and English letter knowledge (Jukes et al., 2017). The randomization and implementation were conducted at the TAC tutor zone level with equal randomization ($\pi_c = 1/2$), and the first primary outcome of spelling score (ranging from 0 to 20) was measured at the student level. The design focused on calculating the required number of TAC tutor zones to achieve at least 80% power at the 5% nominal test size. In the trial sample, there were between 3 and 6 schools in each tutor zone, with 25 children per school, and 2 follow-up visits for literacy tests. For simplicity, we assumed there were 4 schools per TAC tutor zone. According to the results reported in Table 5 of Juke's paper (Jukes et al., 2017), the correlation between two spelling scores of each child was $a_0 = 0.445$, the correlation between two spelling scores of different children from the same school was $a_1 = 0.104$, and the correlation between two spelling scores of different children from different schools in the same TAC tutor zone was $a_2 = 0.008$. The researchers initially expected an effect size of 0.19 standard deviation (SD) for spelling scores. Given $M = 4$, $K = 25$, $L = 2$, $b = 0.19\sqrt{\varphi}$, and $a = (0.445, 0.104, 0.008)$, Equation (8) and (4) suggested that the required number of TAC tutor zones was $N = 36$ with power of 80.87%. Figure 5 shows the sensitivity of power as a function of a_1 and a_2 at $a_0 = $

{0.4, 0.445, 0.5}, assuming $N = 36$. Predicted power decreases as a_0, a_1, or a_2 increases. In Web Appendix L, we also provide illustrative calculations of the required sample size under a different effect size, and when randomization is hypothetically carried out in lower levels under the HALI trial context. As expected, the required number of TAC tutor zones reduces to as few as 8 if randomization is conducted at the children level (level 2).

6 Discussion

In this article, we provide a comprehensive investigation on the sample size and power considerations for four-level intervention studies assuming arbitrary link and variance functions, when intervention assignment is carried out at any level, with a particular focus on CRTs. We develop the extended nested exchangeable correlation structure, which is a generalization of the nested exchangeable correlation structure in three-level CRTs (Teerenstra et al., 2010) and the simple exchangeable correlation structure commonly used in two-level CRTs. That is, the proposed method provides a very general sample size formula that could be applied to continuous, binary and count outcomes in designs with up to four levels. The results suggest that sample size and power calculations using the proposed method are valid under plausible values of the three correlations (a_0, a_1, a_2) for studies with 8 or more clusters. In practice, sensitivity analyses of sample size and power should be performed by varying correlation values within possible ranges, as demonstrated in Section 5. It is worth noting that the combination of the correlation parameters is valid only if the resulting correlation matrix is positive definite, which can be checked analytically by linear constraints presented in Section 2. Finally, although we have primarily focused on four level clustered designs, it is possible to extend our approach to accommodate more than four levels. However, this extension depends on a successful generalization of (2) to a more complex correlation matrix and the derivation of its analytical inverse. This interesting extension, however, is beyond the scope of this article and will be pursued in future work.

In the proposed sample size formula, we assume a balanced design with equal "cluster" sizes at each level and assume that the analytic model included only an intercept and a binary cluster-level intervention indicator. At the design stage, assuming the true correlation is extended nested exchangeable, using the model-based variance and using the sandwich variance with an independence working correlation matrix result in the same required sample size for the balanced case. At the analysis stage, using the extended nested exchangeable working correlation structure and using an independence working correlation matrix for GEE analyses result in the same estimators $\hat{\beta}$, BC0, BC1, BC2, AVG, and BC3 for the balanced case (Theorem 3.10 and Remark 4.1). However, using the extended nested exchangeable working correlation structure for GEE analyses can protect the study from losing power under unbalanced designs in real-world studies, and allows us to report values of correlations to adhere to the CONSORT Statement (Schulz et al., 2010; Campbell et al., 2012). For such reasons, we recommend modeling the underlying correlation structures and caution the use of independence working correlation matrix for GEE-based design and analysis of multi-level CRTs.

We compared using the extended nested exchangeable working correlation structure with MAEE and using an independence working correlation matrix for GEE analyses, in our

simulation study with t-tests. Under balanced designs, BC1 among 7 variance estimators performed best with a near nominal type I error rate and adequate power relative to the proposed sample size formula for as few as 8 clusters using either of these two analysis methods. Similarly, the use of BC1 with a *t*-test was recommended for three-level CRTs with as few as 10 clusters by Teerenstra et al. (2010) and for stepped wedge CRTs with as few as 8 clusters by Li et al. (2018). We showed that based on BC1, using the extended nested exchangeable working correlation structure with MAEE, the type I error rate and power maintained acceptable levels when there were variable numbers of level-one units per level-two unit, while using an independence working correlation matrix could not ensure an acceptable type I error rate or power as the CV of numbers of level-one units increased. Therefore, under unbalanced designs, the proposed method of sample size calculations can also be used with the mean number of level-one units provided, and the GEE *t*-test with the use of BC1 can protect the type I error rate and maintain adequate power for four-level CRTs with as few as 8 clusters.

When designing four-level clustered trials in practice, one potential challenge is to identify accurate ICC values for the extended nested exchangeable correlation structure. This challenge, in fact, is not unique to parallel-arm four-level designs, but also equally applies to other CRTs with complex correlation structures (e.g., cluster randomized crossover design and stepped wedge design). Ideally, one could conduct a four-level pilot CRT (e.g., Kim et al. (2006)) to generate ICC estimates. In general, the study planners should search the literature for published ICCs for their outcomes of interest. Depending on the level of clustering in the published trials, only certain components of α may be available. In other cases, study planners may also consider using historical or routinely collected data from clusters to estimate α_0, α_1, α_2 and guide the planning of study. When there is a large uncertainty for components of α, we strongly recommend conducting a sensitivity analysis by considering a range of values for ICCs in power analysis, as we already demonstrated in Section 5. Finally, we wish to note that publishing ICCs has long been advocated; for example, see Murray et al. (2004) for a table listing 14 research articles presenting ICCs for a variety of endpoints, groups and populations; Preisser et al. (2007) for published ICCs for nine binary youth alcohol use measures from the Youth Survey of the Enforcing Underage Drinking Laws Program; and Korevaar et al. (2021) for published ICC estimates from the CLustered OUtcome Dataset bank to inform the design of longitudinal CRTs for a wide range of outcomes. We encourage more such efforts to assist the planning for complex multi-level CRTs and to strengthen the connection between the methodological development on study design and statistical practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

collaboration with us (XW and ELT). We thank the Editor and two anonymous referees for their constructive comments and suggestions, which have improved the exposition of this work.

References

Amatya A, Bhaumik D, and Gibbons RD (2013). Sample size determination for clustered count data. Statistics in Medicine 32, 4162–4179. [PubMed: 23589228]

Campbell MK, Piaggio G, Elbourne DR, and Altman DG (2012). Consort 2010 statement: extension to cluster randomised trials. Bmj 345, e5661. [PubMed: 22951546]

Cunningham TD and Johnson RE (2016). Design effects for sample size computation in three-level designs. Statistical Methods in Medical Research 25, 505–519. [PubMed: 23070588]

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

Ford WP and Westgate PM (2017). Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. Biometrical Journal 59, 478–495. [PubMed: 28128854]

Heo M and Leon AC (2008). Statistical power and sample size requirements for three level hierarchical cluster randomized trials. Biometrics 64, 1256–1262. [PubMed: 18266889]

Jukes MC, Turner EL, Dubeck MM, Halliday KE, Inyega HN, Wolf S, et al. (2017). Improving literacy instruction in kenya through teacher professional development and text messages support: A cluster randomized trial. Journal of Research on Educational Effectiveness 10,449–481.

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

Kim H-Y, Preisser JS, Gary Rozier R, and Valiyaparambil JV (2006). Multilevel analysis of group-randomized trials with binary outcomes. Community Dentistry and Oral Epidemiology 34, 241–251. [PubMed: 16856945]

Korevaar E, Kasza J, Taljaard M, Hemming K, Haines T, Turner EL, Thompson JA, Hughes JP, and Forbes AB (2021). Intra-cluster correlations from the CLustered Outcome Dataset bank to inform the design of longitudinal cluster trials. Clinical Trials page 17407745211020852.

Li F, Forbes AB, Turner EL, and Preisser JS (2019). Power and sample size requirements for gee analyses of cluster randomized crossover trials. Statistics in Medicine 38, 636–649. [PubMed: 30298551]

Li F and Tong G (2021). Sample size and power considerations for cluster randomized trials with count outcomes subject to right truncation. Biometrical Journal 63, 1052–1071. [PubMed: 33751620]

Li F, Turner EL, and Preisser JS (2018). Sample size determination for gee analyses of stepped wedge cluster randomized trials. Biometrics 74,1450–1458. [PubMed: 29921006]

Li P and Redden DT (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. Statistics in medicine 34, 281–296. [PubMed: 25345738]

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

Liu J and Colditz GA (2020). Sample size calculation in three-level cluster randomized trials using generalized estimating equation models. Statistics in medicine 39,3347–3372. [PubMed: 32720717]

Liu J, Liu L, and Colditz GA (2019). Optimal designs in three-level cluster randomized trials with a binary outcome. Statistics in Medicine 38, 3733–3746. [PubMed: 31162709]

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

Morel JG, Bokossa M, and Neerchal NK (2003). Small sample correction for the variance of gee estimators. Biometrical Journal: Journal of Mathematical Methods in Biosciences 45, 395–409.

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

Murray DM, Varnell SP, and Blitstein JL (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. American Journal of Public Health 94,423–432. [PubMed: 14998806]

Pan W (2001). Sample size and power calculations with correlated binary data. Controlled clinical trials 22,211–227. [PubMed: 11384786]

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 E-Y, and Wolfson M (2007). The importance and role of intracluster correlations in planning cluster trials. Epidemiology 18, 552. [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 with specified marginal means and correlations. Biometrika 90, 455–463.

Schulz KF, Altman DG, Moher D, and Consort Group (2010). Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. Trials 11, 32. [PubMed: 20334632]

Shih WJ (1997). Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. Biometrical Journal 39, 899–908.

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]

Teerenstra S, Taljaard M, Haenen A, Huis A, Atsma F, Rodwell L, et al. (2019). Sample size calculation for stepped-wedge cluster-randomized trials with more than two levels of clustering. Clinical Trials 16, 225–236. [PubMed: 31018678]

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]

Weinfurt KP, Hernandez AF, Coronado GD, DeBar LL, Dember LM, Green BB, et al. (2017). Pragmatic clinical trials embedded in healthcare systems: Generalizable lessons from the NIH Collaboratory. BMC Medical Research Methodology 17, 1–10. [PubMed: 28056835]

Yu H, Li F, Rathouz P, Turner EL, and Preisser J (2020). geeCRT: Bias-Corrected GEE for Cluster Randomized Trials. R package version 0.0.1, https://CRAN.R-project.org/package=geeCRT.

Yu H, Li F, and Turner EL (2020). An evaluation of quadratic inference functions for estimating intervention effects in cluster randomized trials. Contemporary Clinical Trials Communications 19, 100605. [PubMed: 32728648]

Zeger SL, Liang K-Y, and Albert PS (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics pages 1049–1060. [PubMed: 3233245]
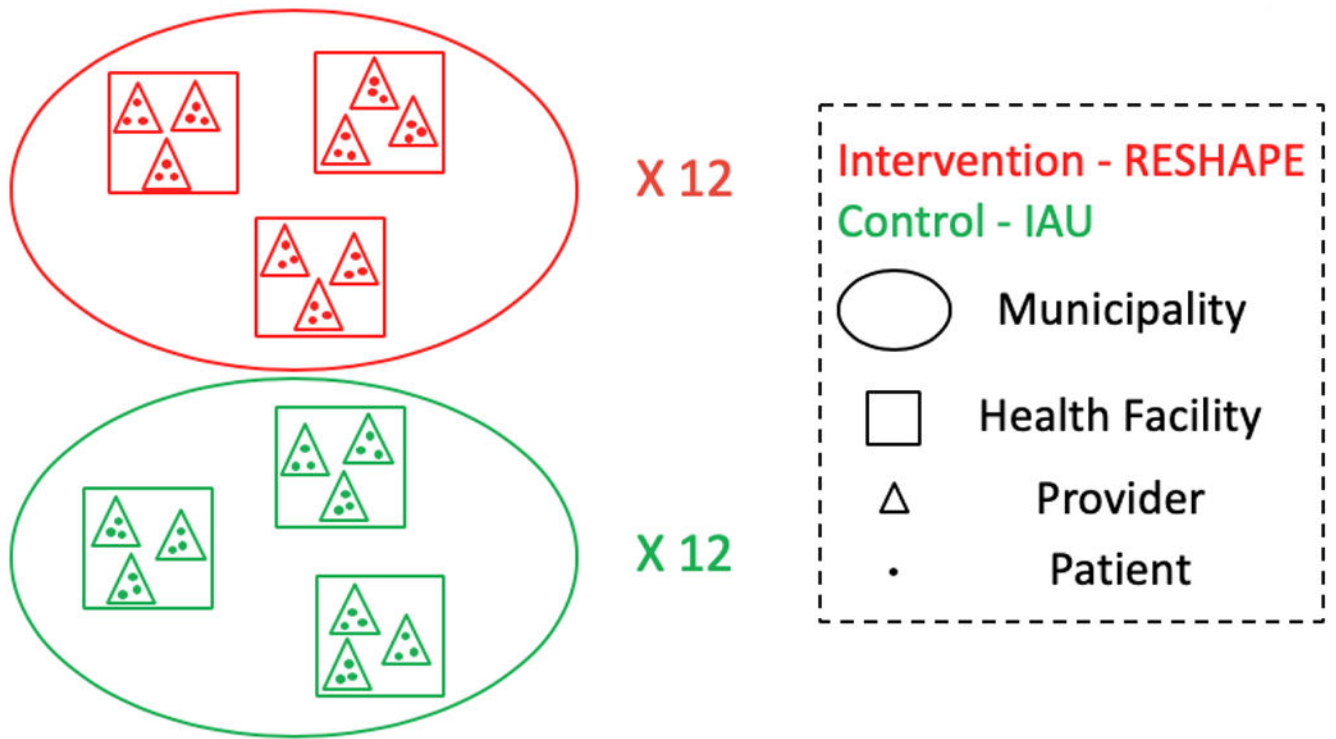
Figure 1.
An illustration of the RESHAPE trial with 12 municipalities per arm; the outcome of interest is whether patients are accurately diagnosed by provider (as determined by psychiatrist diagnosis). A motivating design question is: Given 3 health facilities per municipality, 3 providers per health facilities, and 3 patients per provider, how many municipalities are needed for a two-sided test to identify a difference in diagnostic accuracy of 88% vs. 78% at 90% power given the 5% significance level?
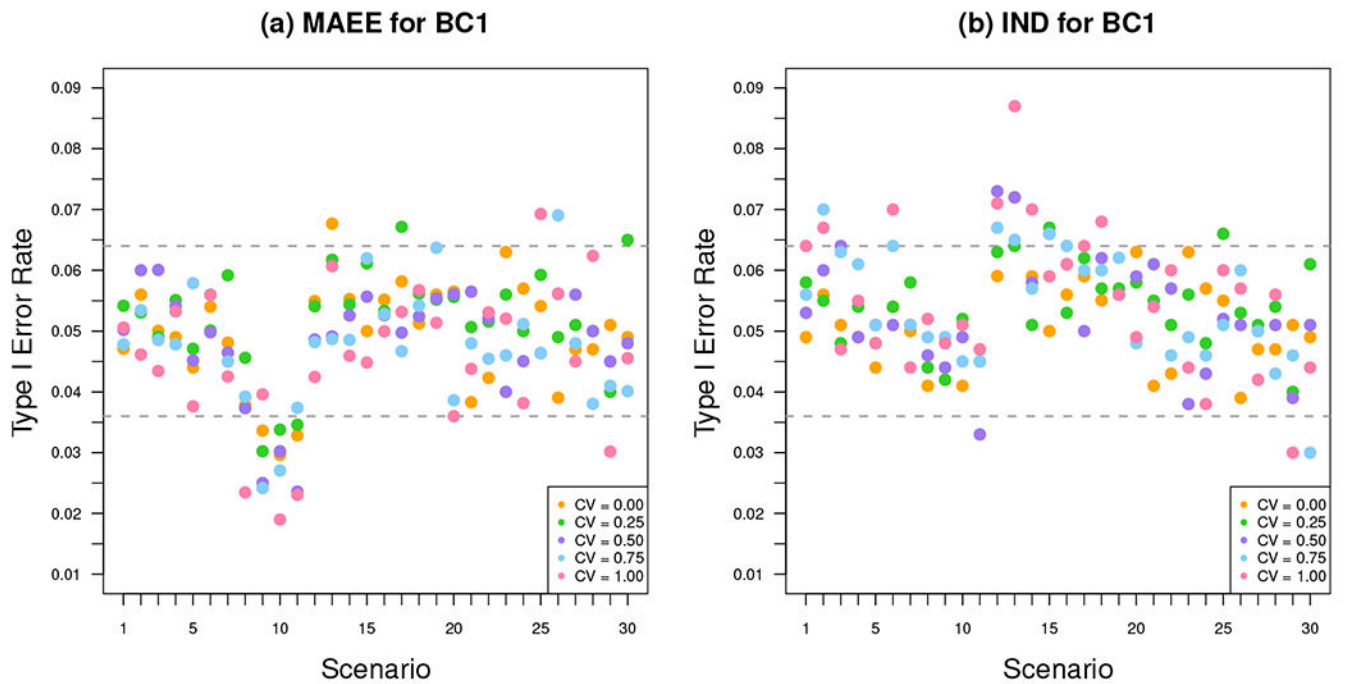
Figure 2.
Empirical type I error rates of GEE analyses using (a) the extended nested exchangeable
working correlation structure with MAEE and (b) an independence working correlation
matrix, based on BC1: bias-corrected sandwich estimator of Kauermann and Carroll (2001),
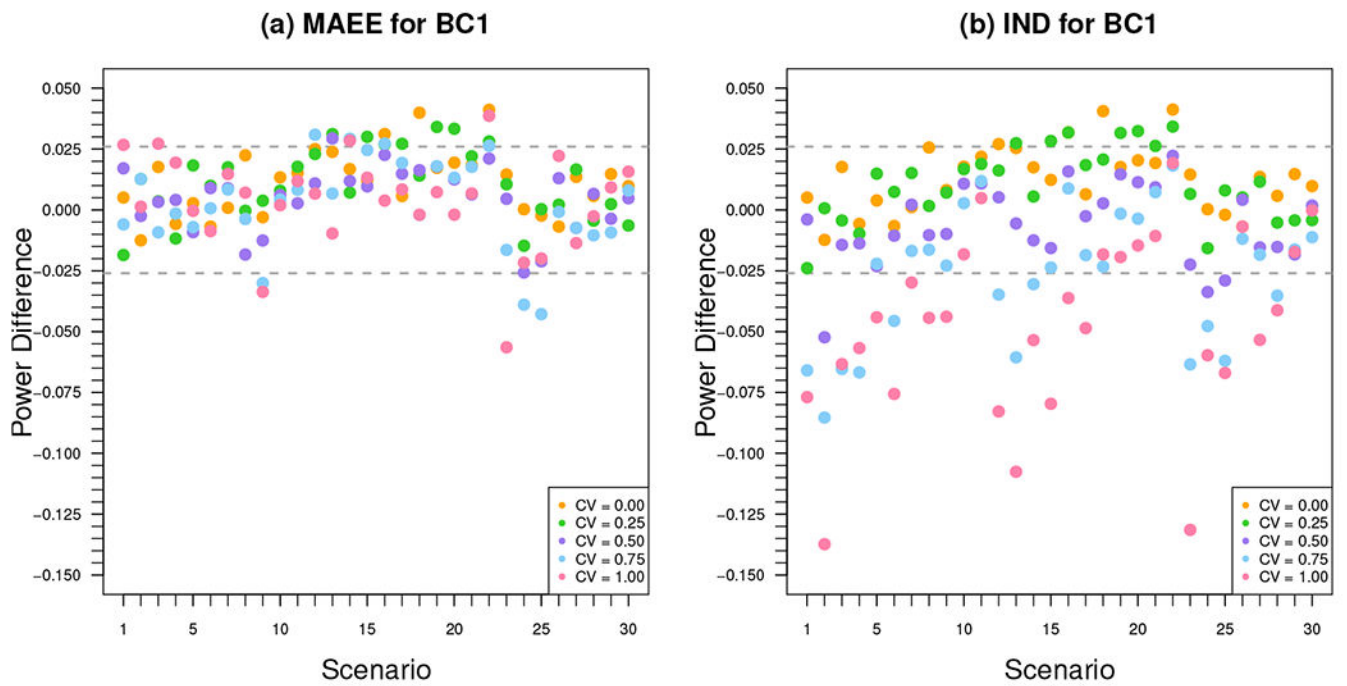under four-level CRTs.

Figure 3.
Difference between the empirical power and the predicted power of GEE analyses using
(a) the extended nested exchangeable working correlation structure with MAEE and (b) an
independence working correlation matrix, based on BC1: bias-corrected sandwich estimator
of Kauermann and Carroll (2001), under four-level CRTs.
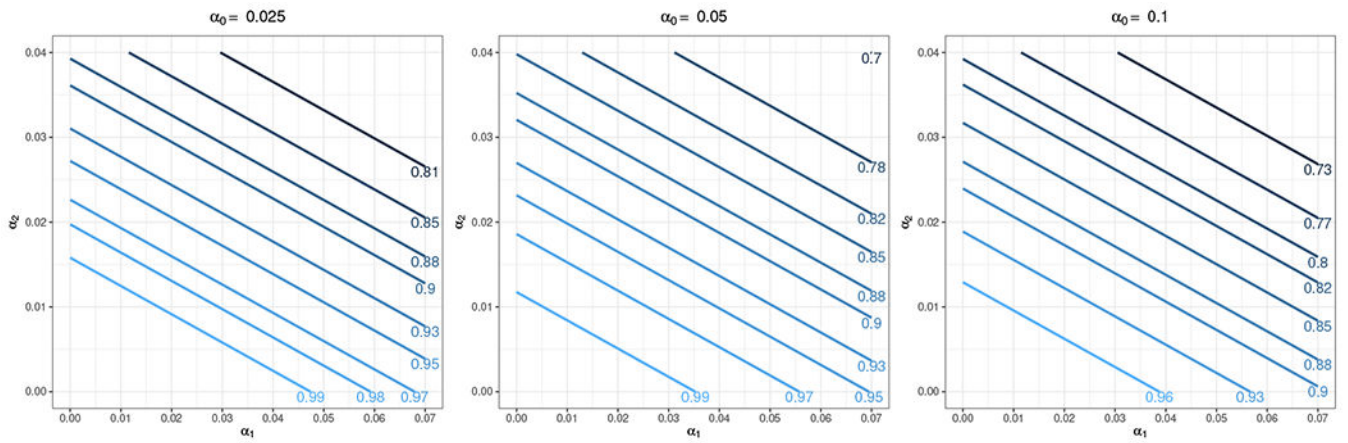
Figure 4.
Predicted power contours as a function of a_1 and a_2 at $a_0 = \{0.025, 0.05, 0.1\}$, with $N = 22$, $M = 3$, $K = 3$, $L = 36$, $P_0 = 78.5\%$, $P_1 = 88\%$ for the RESHAPE trial, under the logit link function.
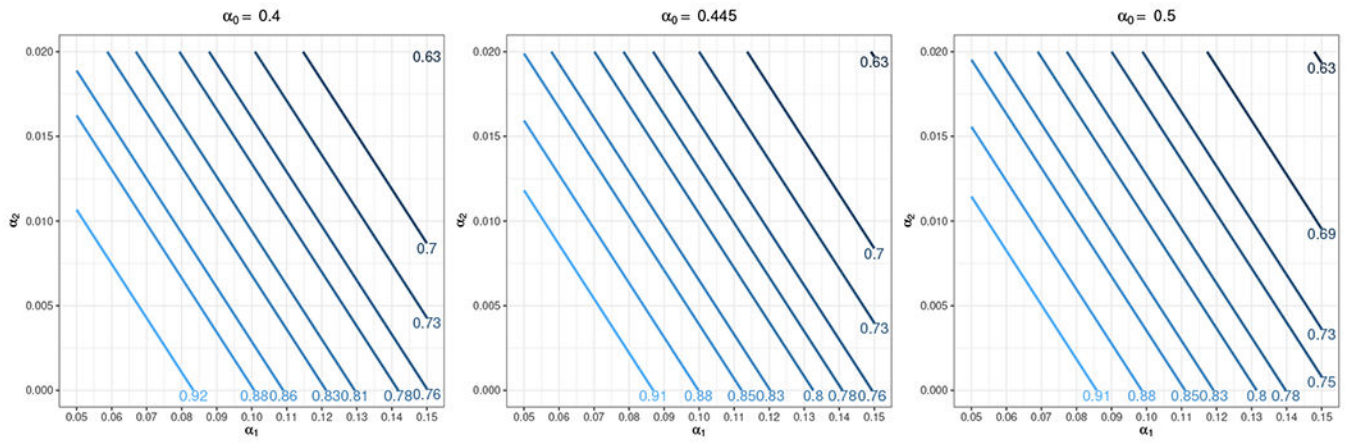
Figure 5.
Predicted power contours as a function of a_1 and a_2 at $a_0 = \{0.4, 0.445, 0.5\}$, with $N = 36$, $M = 4$, $K = 25$, $L = 2$, $b = 0.19\sqrt{\varphi}$ for the HALI trial.

Table 1

Brief summary of existing methods for designing trials with more than two levels.

Reference	Design	Model[a]	Outcome	Link
Heo and Leon (2008)	Three-level CRT	GLMM	Continuous	Identity
Teerenstra et al. (2010)	Three-level CRT	GEE	Continuous	Identity
			Binary	Logit
Cunningham and Johnson (2016)	Three-level Clustered Trial	GLMM	Continuous	Identity
Teerenstra et al. (2019)	Four-level stepped-wedge CRT	GLMM	Continuous	Identity
			Binary	Identity
			Count	Identity
Liu and Colditz (2020)	Three-level CRT	GEE	Continuous	Identity
			Binary	Logit
			Count	Log

[a]GLMM: generalized linear mixed model. GEE: generalized estimating equations.

Table 2

Simulation scenarios and empirical type I error rates[a] of GEE/MAEE analyses based on different variance estimators, using the extended nested exchangeable working correlation structure under balanced four-level CRTs. MB: Model-based variance estimator; BC0: Uncorrected sandwich estimator of Liang and Zeger (1986); BC1: Bias-corrected sandwich estimator of Kauermann and Carroll (2001); BC2: Bias-corrected sandwich estimator of Mancl and DeRouen (2001); AVG: Bias-corrected sandwich estimator with standard error as the average of those from BC1 and BC2; BC3: Bias-corrected sandwich estimator of Fay and Graubard (2001); BC4: Bias-corrected sandwich estimator of Morel et al. (2003).

P_0	P_1	a[b]	n	m	k	l	MB	BC0	BC1	BC2	AVG	BC3	BC4
0.2	0.5	A1	14	2	3	5	**0.045**	0.068	**0.047**	**0.042**	**0.045**	**0.045**	**0.043**
0.2	0.5	A1	14	2	3	10	**0.046**	0.078	**0.056**	**0.041**	**0.047**	**0.049**	**0.046**
0.2	0.5	A1	14	2	4	5	**0.044**	0.069	**0.050**	0.036	**0.044**	**0.047**	**0.044**
0.2	0.5	A1	12	3	3	5	**0.046**	0.070	**0.049**	0.033	**0.042**	**0.038**	**0.039**
0.2	0.5	A2	10	2	3	5	0.033	**0.060**	**0.044**	0.024	0.032	0.033	0.029
0.2	0.5	A2	10	2	3	10	**0.045**	0.077	**0.054**	0.032	**0.041**	**0.040**	**0.037**
0.2	0.5	A2	10	2	4	5	**0.045**	0.067	**0.048**	0.032	**0.039**	**0.043**	**0.037**
0.2	0.5	A2	8	3	3	5	**0.036**	0.071	**0.038**	0.021	0.029	0.030	0.028
0.2	0.5	A3	8	2	3	5	0.033	**0.051**	0.034	0.014	0.021	0.027	0.021
0.2	0.5	A3	8	3	3	5	0.031	**0.060**	0.030	0.016	0.021	0.023	0.020
0.2	0.5	A4	8	3	3	5	0.032	**0.052**	0.033	0.014	0.025	0.024	0.024
0.1	0.3	A1	22	2	3	5	**0.037**	0.070	**0.055**	**0.047**	**0.048**	**0.050**	**0.047**
0.1	0.3	A1	20	2	3	10	**0.037**	0.079	0.068	**0.046**	**0.057**	**0.059**	**0.051**
0.1	0.3	A1	20	2	4	5	**0.039**	0.073	**0.055**	**0.045**	**0.049**	**0.050**	**0.046**
0.1	0.3	A1	16	3	3	5	0.033	0.073	**0.050**	0.036	**0.043**	**0.040**	**0.041**
0.1	0.3	A2	16	2	3	5	**0.038**	0.074	**0.055**	**0.040**	**0.046**	**0.047**	**0.044**
0.1	0.3	A2	14	2	3	10	**0.041**	0.072	**0.058**	**0.040**	**0.049**	**0.044**	**0.042**
0.1	0.3	A2	14	2	4	5	**0.042**	0.076	**0.051**	0.037	**0.047**	**0.044**	**0.040**
0.1	0.3	A2	12	3	3	5	**0.040**	0.083	**0.056**	**0.040**	**0.049**	**0.048**	**0.044**
0.1	0.3	A3	12	2	3	5	**0.042**	0.070	**0.057**	**0.041**	**0.048**	**0.046**	**0.045**
0.1	0.3	A3	10	3	3	5	**0.037**	**0.059**	**0.038**	0.028	0.029	0.029	0.028
0.1	0.3	A4	10	3	3	5	0.033	0.066	**0.042**	0.025	0.034	0.032	0.029
0.5	0.7	A1	26	2	4	5	**0.063**	0.066	**0.063**	**0.049**	**0.055**	**0.057**	**0.054**
0.5	0.7	A2	16	3	3	5	**0.059**	0.065	**0.057**	**0.041**	**0.049**	**0.049**	**0.049**
0.5	0.7	A3	12	2	4	5	**0.054**	0.069	**0.054**	0.037	**0.041**	**0.040**	**0.040**
0.5	0.7	A4	14	3	3	5	**0.048**	**0.061**	**0.039**	0.033	0.035	0.035	0.034
0.8	0.9	A2	30	3	3	5	**0.047**	**0.053**	**0.047**	**0.041**	**0.044**	**0.046**	**0.044**
0.8	0.9	A3	22	2	4	5	**0.046**	**0.056**	**0.047**	0.037	**0.039**	**0.039**	**0.040**
0.8	0.9	A4	28	2	4	5	**0.051**	**0.057**	**0.051**	0.037	**0.042**	**0.043**	**0.041**
0.8	0.9	A4	24	3	3	5	**0.050**	**0.056**	**0.049**	**0.041**	**0.046**	**0.046**	**0.046**

[a]Bold text indicates acceptable empirical type I error rate (from 3.6% to 6.4%).

[b]A1: $a = (0.4, 0.1, 0.03)$; A2: $a = (0.15, 0.08, 0.02)$; A3: $a = (0.1, 0.02, 0.01)$; A4: $a = (0.05, 0.05, 0.02)$.

Table 3

Simulation scenarios, predicted power, and empirical power[a] of GEE/MAEE analyses based on different variance estimators, using the extended nested exchangeable working correlation structure under balanced four-level CRTs. MB: Model-based variance estimator; BC0: Uncorrected sandwich estimator of Liang and Zeger (1986); BC1: Bias-corrected sandwich estimator of Kauermann and Carroll (2001); BC2: Bias-corrected sandwich estimator of Mancl and DeRouen (2001); AVG: Bias-corrected sandwich estimator with standard error as the average of those from BC1 and BC2; BC3: Bias-corrected sandwich estimator of Fay and Graubard (2001); BC4: Bias-corrected sandwich estimator of Morel et al. (2003).

P_0	P_1	a[b]	n	m	k	l	Pred[c]	MB	BC0	BC1	BC2	AVG	BC3	BC4
0.2	0.5	A1	14	2	3	5	0.817	**0.821**	0.857	**0.822**	0.767	**0.798**	**0.795**	**0.797**
0.2	0.5	A1	14	2	3	10	0.845	**0.829**	0.876	**0.833**	0.790	0.811	0.806	0.808
0.2	0.5	A1	14	2	4	5	0.866	**0.891**	0.909	**0.884**	0.854	**0.871**	**0.869**	**0.870**
0.2	0.5	A1	12	3	3	5	0.857	**0.856**	0.891	**0.851**	0.793	0.824	0.821	0.822
0.2	0.5	A2	10	2	3	5	0.808	**0.814**	0.881	**0.811**	0.743	**0.782**	0.776	0.776
0.2	0.5	A2	10	2	3	10	0.870	**0.868**	0.907	**0.863**	0.792	0.829	0.827	0.829
0.2	0.5	A2	10	2	4	5	0.852	**0.860**	0.897	**0.853**	0.784	0.823	0.820	0.820
0.2	0.5	A2	8	3	3	5	0.800	**0.822**	0.895	**0.823**	0.712	**0.776**	0.772	0.768
0.2	0.5	A3	8	2	3	5	0.851	**0.859**	0.916	**0.848**	0.752	0.793	0.794	0.785
0.2	0.5	A3	8	3	3	5	0.936	**0.955**	0.981	**0.950**	0.904	**0.927**	**0.930**	**0.926**
0.2	0.5	A4	8	3	3	5	0.892	**0.914**	0.957	**0.907**	0.830	**0.869**	0.865	0.860
0.1	0.3	A1	22	2	3	5	0.829	0.860	0.877	**0.854**	0.828	0.843	0.838	0.840
0.1	0.3	A1	20	2	3	10	0.818	0.849	0.865	**0.841**	0.813	0.830	0.827	0.829
0.1	0.3	A1	20	2	4	5	0.841	0.875	0.885	**0.857**	0.829	0.843	0.839	0.841
0.1	0.3	A1	16	3	3	5	0.805	**0.831**	0.851	**0.817**	0.783	**0.804**	**0.803**	**0.805**
0.1	0.3	A2	16	2	3	5	0.844	0.885	0.907	0.875	**0.833**	0.857	0.852	0.856
0.1	0.3	A2	14	2	3	10	0.849	**0.862**	0.894	**0.854**	0.807	**0.827**	0.826	0.829
0.1	0.3	A2	14	2	4	5	0.829	0.880	0.897	0.869	**0.822**	0.849	0.841	0.846
0.1	0.3	A2	12	3	3	5	0.826	0.858	0.892	**0.844**	0.798	**0.828**	0.825	0.825
0.1	0.3	A3	12	2	3	5	0.873	0.903	0.930	**0.892**	0.854	**0.877**	**0.874**	**0.876**
0.1	0.3	A3	10	3	3	5	0.898	0.928	0.959	**0.916**	0.871	**0.892**	**0.890**	**0.891**
0.1	0.3	A4	10	3	3	5	0.837	0.887	0.919	0.878	**0.799**	**0.840**	0.829	0.832
0.5	0.7	A1	26	2	4	5	0.823	**0.841**	0.854	**0.838**	0.814	**0.825**	**0.825**	**0.824**
0.5	0.7	A2	16	3	3	5	0.831	**0.838**	0.864	**0.831**	0.789	**0.810**	**0.813**	**0.811**
0.5	0.7	A3	12	2	4	5	0.827	**0.829**	0.860	**0.825**	0.778	0.800	**0.806**	0.797
0.5	0.7	A4	14	3	3	5	0.868	**0.857**	0.894	**0.861**	0.814	0.835	0.835	0.832
0.8	0.9	A2	30	3	3	5	0.804	**0.825**	0.838	**0.818**	**0.800**	**0.810**	**0.810**	**0.811**
0.8	0.9	A3	22	2	4	5	0.804	**0.815**	0.829	**0.810**	0.788	**0.801**	**0.805**	**0.801**
0.8	0.9	A4	28	2	4	5	0.824	**0.844**	0.850	**0.839**	0.826	**0.836**	**0.837**	**0.836**
0.8	0.9	A4	24	3	3	5	0.813	**0.830**	0.847	**0.823**	0.799	**0.810**	**0.811**	**0.811**

[a] Bold text indicates acceptable empirical power (differing at most 2.6% from the predicted power).

[b] A1: a = (0.4, 0.1, 0.03); A2: a = (0.15, 0.08, 0.02); A3: a = (0.1, 0.02, 0.01); A4: a = (0.05, 0.05, 0.02).

[c]Pred: Predicted power based on *t*-test.