Statistics in Medicine

Sample size calculation in hierarchical 2×2 factorial trials with unequal cluster sizes

Journal:	*Statistics in Medicine*
Manuscript ID	SIM-21-0104
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	04-Feb-2021
Complete List of Authors:	Tian, Zizhong; Yale University School of Public Health, Department of Biostatistics Esserman, Denise; Yale University Yale School of Public Health, Biostatistics Tong, Guangyu; Biostatistics Blaha, Ondrej; Yale University School of Public Health, Biostatistics Dziura, James; Yale University School of Medicine, Emergency Medicine; Yale University School of Public Health, Biostatistics Peduzzi, Peter; Yale University School of Public Health, Biostatistics Li, Fan; Yale University School of Public Health, Biostatistics
Keywords:	Coefficient of variation, interaction test, intersection-union test, linear mixed model, power analysis

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Sample size calculation in hierarchical 2×2 factorial trials with unequal cluster sizes

Zizhong Tian[1], Denise Esserman[1,2], Guangyu Tong[1,2], Ondrej Blaha[1,2], James Dziura[1,2], Peter Peduzzi[1,2], Fan Li[1,2,3,*]

[1]Department of Biostatistics, Yale University School of Public Health, Connecticut, USA
[2]Yale Center for Analytical Sciences, Yale University, Connecticut, USA
[3]Center for Methods in Implementation and Prevention Science, Yale University, Connecticut, USA

Correspondence
Fan Li*
Department of Biostatistics
Yale School of Public Health
New Haven CT, 06510, USA
Email: fan.f.li@yale.edu

Funding Information
This research was supported by the Clinical and Translational Science Awards (UL1TR001863) at Yale University.

Abstract

Motivated by a suicide prevention trial with hierarchical treatment allocation (cluster-level and individual-level treatments), we address the sample size requirements for testing the marginal treatment effects, both separately and jointly. Our development assumes a saturated linear mixed model, based on which null hypotheses that are of scientific interest are formalized. For each hypothesis, we derive closed-form sample size formulas based on a large-sample z-approximation, and provide finite-sample modifications based on a t-approximation. We relax the conventional equal-cluster-size assumption and express the sample size formulas as functions of the mean and coefficient of variation of cluster sizes. We find that the variance inflation for testing the marginal cluster-level treatment effect due to unequal cluster sizes resembles that derived for a two-arm parallel cluster randomized trial. In contrast, unequal cluster sizes have little impact on the sample size requirements for testing either the marginal individual-level treatment effect or the interaction effect between the two treatments. We conduct simulations to validate the proposed sample size formulas, and find the empirical power agrees well with the predicted power for each test. In addition, the t-approximation for testing the marginal cluster-level treatment effect often provides better control of type I error rate with a small number of clusters. The z-approximation for testing the individual-level marginal effect, however, has robust control of type I error rate even with a small number of clusters. Finally, we illustrate our sample size formulas to design the motivating suicide prevention factorial trial.

KEYWORDS:
Coefficient of variation, interaction test, intersection-union test, linear mixed model, power analysis, variable cluster sizes

1 | INTRODUCTION

Intervention programs with multiple components or treatments are common in health, behavioral and educational research. The factorial design is a rigorous framework to evaluate the effectiveness of different intervention components or treatments.[1] In a traditional 2×2 factorial trial with two treatments, T1 and T2, investigators could simultaneously randomize the two treatments and assign the individual participants to one of the four conditions: T1 only, T2 only, both T1 and T2, and double usual care. The cross-classification of participants into four conditions allows the identification of the marginal and the interaction effects

of different treatments.[2] While traditional factorial designs randomize treatments at the individual level, recent design variants have considered factorial randomization at the cluster level, where a cluster could be a school, clinic or hospital.[3] In a cluster randomized factorial trial, the intraclass correlation coefficient (ICC) of the outcome inflates the required sample size compared to an individually randomized factorial trial, and thus represents a key consideration for study planning.[3,4,5,6]

While sample size formulas for factorial designs with randomization carried out at the same level (cluster level or individual level) were previously studied,[2,3] sample size formulas when randomization is carried out at two different levels are less developed. Our motivating example is a hierarchical 2×2 factorial trial, which aims to assess the clinical effectiveness of a two-component intervention program for suicide prevention among community-dwelling transgender individuals. In this trial, participating clinics will be randomized to either the Caring Contacts (CC) arm or the usual care condition.[7] In addition, participants within each clinic will be individually randomized to either Cognitive Behavioral Therapy for Suicide Prevention (CBT-SP) arm or usual care.[8] Because participants are nested within clinics, the ICC of the outcome should be considered for study planning as in a cluster randomized factorial trial. However, unlike the cluster randomized factorial trial, individual-level randomization of the CBT-SP program necessitates additional design considerations for testing the treatment effects.

In the experimental design literature, the hierarchical 2×2 factorial design has also been named as the split-plot design.[9,10] A recent systematic review of split-plot trials suggested that rigorous methods for sample size calculation were lacking.[10] Shin et al.[11] developed a sample size procedure for testing (both separately or jointly) the main effects and interactions under a split-plot design with an arbitrary number of factor levels. However, their approach assumes moment-based estimators of the regression parameters, and therefore does not exploit the ICC during the analysis stage. Failure to account for the within-cluster correlation in the estimation of parameters is less statistically efficient and can lead to a larger sample size than necessary. In contrast, we consider generalized least square (GLS) estimators for the regression coefficients, and develop corresponding sample size formulas for testing a number of hypotheses of interest based on our motivating study. As listed in Table 1, our null hypotheses include (A1-A2) no marginal effect for each treatment, separately, (B) no interaction between the two treatments, (C) no marginal effect for both treatments and (D) no marginal effect for at least one treatment. Given that the two treatments may have an interaction effect, we define the marginal effects of interest as the average effect of one treatment across levels of the other treatment. This is akin to the "at the margin" idea for the analysis of individually randomized factorial trials.[12] Based on a linear mixed model, the marginal effect of each treatment can be expressed as a linear combination of regression parameters, and the null hypotheses (A1), (A2), (B), (C) are special cases of the general linear hypotheses. On the other hand, the null hypothesis (D) is a composite and will be addressed by an intersection-union test.[13,14]

TABLE 1 Types of hypotheses of interest in the motivating hierarchical 2×2 factorial trial. CC stands for the Caring Contact intervention, randomized at the clinic level, and CBT-SP stands for the Cognitive Behavioral Therapy for Suicide Prevention program, randomized at the participant level. Δ_x denotes the marginal effect of the CC program, Δ_z denotes the marginal effect of the CBT-SP program, and Δ_{xz} denotes the interaction effect of the CC and CBT-SP programs.

Label	Null hypothesis	Scientific interpretation of null
(A1)	H_0^{A1}: $\Delta_x = 0$	There is no effect due to the CC program compared with usual care among the trial population.
(A2)	H_0^{A2}: $\Delta_z = 0$	There is no effect due to the CBT-SP program compared with usual care among all clinics.
(B)	H_0^{B}: $\Delta_{xz} = 0$	There is no synergistic or antagonistic effect between the CC and CBT-SP intervention programs.
(C)	H_0^{C}: $\Delta_x = \Delta_z = 0$	There is no effect due to both the CC program and CBT-SP program among the trial population.
(D)	H_0^{D}: $\Delta_x = 0$ or $\Delta_z = 0$	There is no effect from at least one of the CC program and CBT-SP program among the trial population.

While Shin et al.[11] assumed equal cluster sizes in deriving the sample size formulas in a split-plot design, we will further relax the equal-cluster-size assumption to mimic more realistic scenarios observed in practice. Unequal cluster sizes arise frequently in pragmatic studies, in which participating providers or clinics naturally have different source population sizes or rates

of participation. While the implications of unequal cluster sizes have been studied in cluster randomized trials with a single intervention,[15,16,17,18] its implications for hierarchical factorial designs remain unclear. Assuming the cluster sizes are randomly sampled from an underlying distribution, we will derive, for each test, its approximate sample size formula that further depends on the mean cluster size as well as the coefficient of variation (CV) of cluster sizes. We show both analytically and numerically that the sample size requirement for testing the marginal cluster-level treatment effect tends to be more sensitive than that for testing the marginal individual-level treatment effect, and we further connect our results with previous results on unequal cluster sizes developed for cluster randomized trials.[18,19]

The remainder of this article is organized into the following sections. In Section 2, we introduce our linear mixed model and derive the large-sample covariance matrix for the GLS estimators assuming unequal cluster sizes. In Section 3, we present the closed-form sample size formulas for testing each hypothesis and discuss finite-sample considerations. In Section 4, we conduct a series of simulation studies to evaluate the accuracy of the proposed sample size formulas with unequal cluster sizes. We apply our proposed approach to calculate the required sample size for the suicide prevention factorial trial in Section 5, and Section 6 concludes with a brief discussion.

2 | STATISTICAL MODEL

2.1 | Linear mixed model for hierarchical 2×2 factorial trials

We consider a hierarchical 2×2 factorial trial with one treatment (T1) randomized at the cluster level and the second (T2) at the individual level. In the context of the suicidal prevention trial, T1 and T2 refer to the CC and CBT-SP programs, respectively. Let Y_{ij} be a continuous outcome measured from the jth individual ($j = 1, \ldots, m_i$) in the ith cluster ($i = 1, \ldots, n$). We assume $n\pi_x$ ($0 < \pi_x < 1$) clusters are randomized to T1, and $n(1-\pi_x)$ to usual care. Within each clusters, we further assume $m_i\pi_z$ ($0 < \pi_z < 1$) participants are randomized to T2, and the remaining $m_i(1-\pi_z)$ to usual care. We assume a "saturated" linear mixed model to characterize the treatment effects as

$$Y_{ij} = \beta_1 + \beta_2 X_i + \beta_3 Z_{ij} + \beta_4 X_i Z_{ij} + a_i + \epsilon_{ij} \tag{1}$$

where X_i is the indicator for the cluster-level treatment ($X_i = 1$ if cluster i is assigned to T1 and $X_i = 0$ otherwise), Z_{ij} is the indicator for the individual-level treatment ($Z_{ij} = 1$ if the jth individual in cluster i is assigned to T2 and $Z_{ij} = 0$ otherwise), $X_i Z_{ij}$ is the interaction between the two treatments, and β_4 describes the direction and magnitude of the interaction effect. To account for clustering, we assume $a_i \sim \mathcal{N}(0, \sigma_a^2)$ is a random intercept describing the unobserved between-cluster variability, and define $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ as the within-cluster random error. We further assume independence between a_i and ϵ_{ij}. Model (1) implies a common ICC, defined by $\rho = \sigma_a^2/\sigma_y^2$, where $\sigma_y^2 = \sigma_a^2 + \sigma_\epsilon^2$ is the total variance of Y_{ij}.[4,20]

Model (1) captures the average outcome for four types of patients based on their treatment status. The parameter β_1 represents the mean outcome for those assigned to double usual care, $\beta_1 + \beta_2$ represents the mean outcome for those receiving T1 only, $\beta_1 + \beta_3$ represents the mean outcome for those receiving T2 only, and $\beta_1 + \beta_2 + \beta_3 + \beta_4$ represents the mean outcome for those receiving both treatments. Because the main-effects parameters in model (1) may be less interpretable due to the existence of an interaction, we focus on the marginal effect of each treatment across levels of the other treatment for sample size consideration. From model (1)

$$E(Y_{ij}|X_i) = \beta_1 + \pi_z\beta_3 + (\beta_2 + \pi_z\beta_4)X_i = \beta_1 + \pi_z\beta_3 + \Delta_x X_i, \tag{2}$$

$$E(Y_{ij}|Z_{ij}) = \beta_1 + \pi_x\beta_2 + (\beta_3 + \pi_x\beta_4)Z_{ij} = \beta_1 + \pi_x\beta_2 + \Delta_z Z_{ij}. \tag{3}$$

These two expressions indicate that the marginal effect of each treatment can be represented as a linear contrast of model parameters. Namely, $\Delta_x = \beta_2 + \pi_z\beta_4$ and $\Delta_z = \beta_3 + \pi_x\beta_4$ represent the marginal effect of T1 and T2, respectively.

2.2 | Large-sample covariance matrix

To study the sample size requirements for the hierarchical 2×2 factorial trial, we first provide a closed-form characterization of the large-sample variance matrix of the regression parameter estimators. While we follow the general strategy considered in Jung and Ahn[21] and Yang et al.[22] to derive the 4×4 variance matrix, a major difference in our work is that we allow for unequal

cluster sizes. Specifically, we reparameterize model (1) by mean-centering the cluster-level treatment

$$Y_{ij} = b_1 + b_2(X_i - \pi_x) + b_3 Z_{ij} + b_4(X_i - \pi_x)Z_{ij} + a_i + \epsilon_{ij} \tag{4}$$

where $b_1 = \beta_1 + \beta_2\pi_x$, $b_2 = \beta_2$, $b_3 = \beta_3 + \beta_4\pi_x$, and $b_4 = \beta_4$. Define the design vector $D_{ij} = \left(1, (X_i - \pi_x), Z_{ij}, (X_i - \pi_x)Z_{ij}\right)^T$ and $D_i = (D_{i1}, \ldots, D_{im_i})^T$, then the Feasible Generalized Least Squares (FGLS) estimator for $b = (b_1, b_2, b_3, b_4)^T$ is given as $\hat{b} = \left(\sum_{i=1}^n D_i^T R_i^{-1} D_i\right)^{-1}\left(\sum_{i=1}^n D_i^T R_i^{-1} Y_i\right)$, where $R_i = (1-\rho)I_{m_i} + \rho J_{m_i}$ is the compound symmetric correlation matrix of the outcome, I_{m_i} is the $m_i \times m_i$ identity matrix, and J_{m_i} is the $m_i \times m_i$ matrix of ones. Assuming the cluster sizes come from a well-defined distribution $f(m_i)$ with finite first and second moments, as the number of clusters n becomes large, the root-n scaled FGLS estimator, $\sqrt{n}(\hat{b} - b)$, converges to a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \sigma_y^2\left(\lim_{n\to\infty} n^{-1}\sum_{i=1}^n D_i^T R_i^{-1} D_i\right)^{-1}$. In what follows, we provide an explicit form of the 4×4 matrix Σ to develop analytical sample size formulas based on the linear mixed model in equation (1).

For each cluster i, the inverse of the compound symmetric correlation matrix can be obtained as[23]

$$R_i^{-1} = \frac{1}{1-\rho} I_{m_i} - \frac{\rho}{(1-\rho)[1+(m_i-1)\rho]} J_{m_i} = \frac{1}{1-\rho}(I_{m_i} + c_i J_{m_i}),$$

where $c_i = -\rho/[1+(m_i-1)\rho]$. Therefore, we can represent

$$\frac{1}{n}\sum_{i=1}^n D_i^T R_i^{-1} D_i = \frac{1}{n(1-\rho)}\sum_{i=1}^n D_i^T D_i + \frac{1}{n(1-\rho)}\sum_{i=1}^n c_i D_i^T J_{m_i} D_i. \tag{5}$$

Due to randomization, the cluster size distribution $f(m_i)$ is independent of both treatment indicators. Define $\bar{m} = E(m_i)$ as the mean cluster size, $\sigma_x^2 = \pi_x(1-\pi_x)$ is the Bernoulli variance of cluster-level treatment, we show in the Appendix that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\left(\sum_{j=1}^{m_i} Z_{ij}\right) = \bar{m}\pi_z.$$

This allows us to obtain

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n D_i^T D_i = \begin{bmatrix} \bar{m} & 0 & \bar{m}\pi_z & 0 \\ 0 & \bar{m}\sigma_x^2 & 0 & \bar{m}\pi_z\sigma_x^2 \\ \bar{m}\pi_z & 0 & \bar{m}\pi_z & 0 \\ 0 & \bar{m}\pi_z\sigma_x^2 & 0 & \bar{m}\pi_z\sigma_x^2 \end{bmatrix}.$$

We further define the following expectations for the functions of cluster sizes as

$$\bar{\eta}_r = E\left\{\frac{-m_i^r\rho}{1+(m_i-1)\rho}\right\},$$

for $r = 1, 2$ and write $\sigma_z^2 = \pi_z(1-\pi_z)$ as the Bernoulli variance of individual-level treatment. In the Appendix, we further show

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\left\{\frac{-\rho}{1+(m_i-1)\rho}\right\} m_i\left(\sum_{j=1}^{m_i} Z_{ij}\right) = \bar{\eta}_2\pi_z,$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\left\{\frac{-\rho}{1+(m_i-1)\rho}\right\}\left(\sum_{j=1}^{m_i} Z_{ij}\right)^2 = \bar{\eta}_2\pi_z^2 + \bar{\eta}_1\sigma_z^2.$$

These intermediate results allow us to obtain

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n c_i D_i^T J_{m_i} D_i = \begin{bmatrix} \bar{\eta}_2 & 0 & \bar{\eta}_2\pi_z & 0 \\ 0 & \bar{\eta}_2\sigma_x^2 & 0 & \bar{\eta}_2\pi_z\sigma_x^2 \\ \bar{\eta}_2\pi_z & 0 & \bar{\eta}_2\pi_z^2 + \bar{\eta}_1\sigma_z^2 & 0 \\ 0 & \bar{\eta}_2\pi_z\sigma_x^2 & 0 & \bar{\eta}_2\pi_z^2\sigma_x^2 + \bar{\eta}_1\sigma_z^2\sigma_x^2 \end{bmatrix}.$$

Therefore, based on (5), the large-sample variance of $\sqrt{n}(\hat{b} - b)$ can be obtained by block matrix inversion as

$$\Sigma = \sigma_y^2\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n D_i^T R_i^{-1} D_i\right)^{-1} = \sigma_y^2(1-\rho)\begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

where the component matrices can be derived explicitly as

$$\Sigma_{11} = \frac{(\bar{m}+\bar{\eta}_1)+(\bar{\eta}_2-\bar{\eta}_1)\pi_z}{(\bar{m}+\bar{\eta}_2)(\bar{m}+\bar{\eta}_1)(1-\pi_z)\sigma_x^2}\begin{bmatrix}\sigma_x^2 & 0\\ 0 & 1\end{bmatrix}, \quad \Sigma_{12}=\Sigma_{21}=-\pi_z\Sigma_{22}=\frac{-1}{(\bar{m}+\bar{\eta}_1)(1-\pi_z)\sigma_x^2}\begin{bmatrix}\sigma_x^2 & 0\\ 0 & 1\end{bmatrix}.$$

3 | SAMPLE SIZE ESTIMATION

3.1 | Test for a single marginal treatment effect

We first consider separately testing the null hypotheses concerning the marginal effect for each treatment. From equations (2) and (3), the marginal null hypotheses of interest are given by (A1) H_0^{A1}: $\Delta_x = 0$ and (A2) H_0^{A2}: $\Delta_z = 0$. Define δ_x and δ_z as the effect sizes for the marginal cluster-level and individual-level treatment effects, respectively. For testing H_0^{A1}, the total required number of clusters based on a two-sided Wald z-test is given by

$$n_{A1} = \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\omega_x}{\delta_x^2}, \tag{6}$$

where α and λ defines the prescribed type I and type II error rates, and $\omega_x = n\text{Var}(\hat{\Delta}_x) = n\text{Var}(\hat{\beta}_2+\pi_z\hat{\beta}_4)$. Likewise, for testing H_0^{A2}, the total required number of clusters with a nominal test size α and power $1-\lambda$ is given by

$$n_{A2} = \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\omega_z}{\delta_z^2}, \tag{7}$$

where $\omega_z = n\text{Var}(\hat{\Delta}_z) = n\text{Var}(\hat{\beta}_3+\pi_x\hat{\beta}_4)$. Therefore, sample size estimation for testing the marginal effect of a single treatment requires us to obtain explicit expressions for the variances ω_x and ω_z.

Based on model (4), we have

$$\omega_x = n\text{Var}(\hat{\beta}_2+\pi_z\hat{\beta}_4) = n\text{Var}(\hat{b}_2)+n\pi_z^2\text{Var}(\hat{b}_4)+2n\pi_z\text{Cov}(\hat{b}_2,\hat{b}_4) = \frac{\sigma_y^2(1-\rho)}{(\bar{m}+\bar{\eta}_2)\sigma_x^2}. \tag{8}$$

This expression depends on the cluster size distribution only through $\bar{m}+\bar{\eta}_2$, which we can further approximate using second-order Taylor series. Following the work of van Breukelen et al.[18,19] and the details in the Appendix, we can approximate

$$\bar{m}+\bar{\eta}_2 = \frac{1-\rho}{\rho}E\left\{\frac{m_i\rho}{1+(m_i-1)\rho}\right\} \approx \frac{\bar{m}(1-\rho)}{1+(\bar{m}-1)\rho}\left[1-\text{CV}^2\frac{\bar{m}\rho(1-\rho)}{\{1+(\bar{m}-1)\rho\}^2}\right],$$

which leads us to an approximate sample size formula for testing the marginal cluster-level treatment effect

$$n_{A1} \approx \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\sigma_y^2\{1+(\bar{m}-1)\rho\}}{\bar{m}\pi_x(1-\pi_x)\delta_x^2}\left[1-\text{CV}^2\frac{\bar{m}\rho(1-\rho)}{\{1+(\bar{m}-1)\rho\}^2}\right]^{-1}. \tag{9}$$

In the equations above, CV is defined as the ratio between the standard deviation and mean of the cluster sizes.

Two important implications emerge from sample size formula (9). First, under equal cluster sizes such that $m_i = \bar{m} = m$ for all i, equation (9) reduces to the usual sample size formula in a parallel cluster randomized trial.[4] We immediately recognize that the variance inflation factor (VIF) due to clustering in the hierarchical 2×2 factorial design, $1+(m-1)\rho$, has the same form as the usual VIF in a two-arm parallel cluster randomized trial. The caveat, however, is that the ICC in our model, ρ, is conditional on both X_i and Z_{ij}. This "conditioning" step may reduce the "marginal" ICC given X_i only and therefore lead to a smaller required sample size compared to a model adjusting for X_i only.[24,25] Second, the VIF due to unequal cluster sizes also takes the same form as the usual VIF derived in van Breukelen et al.[18] in a parallel cluster randomized trial, with the same caveat for interpreting ρ. When the CV of cluster sizes increases, the required number of clusters to detect effect size δ_x also increases as a nonlinear function. In addition, this VIF due to unequal cluster sizes has a parabolic relationship in ρ, and reaches its maximum when $\rho = 1/(\bar{m}+1)$.[18]

For testing the marginal effect of the individual-level treatment, the corresponding variance is given by

$$\omega_z = n\text{Var}(\hat{b}_3) = \frac{\sigma_y^2(1-\rho)}{(\bar{m}+\bar{\eta}_1)\sigma_z^2}, \tag{10}$$

which depends on the cluster size distribution only through $\bar{m} + \bar{\eta}_1$. By appealing to the same Taylor series technique, we have

$$\bar{\eta}_1 = -E\left\{\frac{m_i\rho}{1+(m_i-1)\rho}\right\} \approx -\frac{\bar{m}\rho}{1+(\bar{m}-1)\rho}\left[1-\mathrm{CV}^2\frac{\bar{m}\rho(1-\rho)}{\{1+(\bar{m}-1)\rho\}^2}\right],$$

which leads to

$$\bar{m}+\bar{\eta}_1 \approx \frac{\bar{m}\left[\{1+(\bar{m}-2)\rho\}\{1+(\bar{m}-1)\rho\}^2+\mathrm{CV}^2\bar{m}\rho^2(1-\rho)\right]}{\{1+(\bar{m}-1)\rho\}^3}. \tag{11}$$

Plugging this expression back into ω_z, we obtain the required number of clusters for testing the marginal individual-level treatment effect as

$$n_{A2} \approx \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\sigma_y^2(1-\rho)\{1+(\bar{m}-1)\rho\}^3}{\bar{m}\pi_z(1-\pi_z)\delta_z^2\left[\{1+(\bar{m}-2)\rho\}\{1+(\bar{m}-1)\rho\}^2+\mathrm{CV}^2\bar{m}\rho^2(1-\rho)\right]}. \tag{12}$$

In contrast to the sample size formula (9), sample size formula (12) implies that larger cluster size variability may reduce the required sample size for testing H_0^{A2} because n_{A2} is a decreasing function of the CV of cluster sizes. However, a closer examination of (12) also reveals that realistic degrees of cluster size variation (often with CV not exceeding 0.6) have a limited impact on the resulting sample size unless the mean cluster sizes $\bar{m}$ is extremely large (since the factor $\rho^2(1-\rho)$ is close to zero for common ICC values[4,25]). Furthermore, when the cluster sizes are all equal so that $m_i = \bar{m} = m$ for all i, sample size formula (12) reduces to

$$n_{A2} = \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\sigma_y^2(1-\rho)\{1+(m-1)\rho\}}{m\pi_z(1-\pi_z)\delta_z^2\{1+(m-2)\rho\}}. \tag{13}$$

Interestingly, this equation suggests that the design effect for testing H_0^{A2} due to clustering equals to $(1-\rho)\{1+(m-1)\rho\}/\{1+(m-2)\rho\}$, which is strictly smaller than one. In other words, the within-cluster correlation improves the efficiency for estimating the individual-level treatment effect. On the other hand, we also notice that our linear mixed model (1) could lead to a smaller variance for the marginal effect of the individual-level treatment. To see why, one can easily show that the required number of clusters (in a multi-center individually randomized trial) assuming a linear mixed model adjusting for Z_{ij} only is given by

$$n_{A2}^* = \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\sigma_y^2(1-\rho^*)}{m\pi_z(1-\pi_z)\delta_z^2}.$$

where ρ^* is the ICC conditional only on Z_{ij}. Because ρ^* is usually at least as large as ρ, we often have $n_{A2}^* > n_{A2}$. When the cluster size becomes large and $\rho^* \approx \rho$, the required sample sizes for testing the individual-level treatment effect are similar using model (1) versus the model only adjusting for Z_{ij}.

3.2 | Interaction test

One potential advantage of model (1) is that it introduces a formal test for potential interaction between the two treatments. Specifically, the null hypothesis of no interaction is given as (B) H_0^B: $\Delta_{xz} = \beta_4 = 0$. The required number of clusters for testing H_0^B based on a two-sided Wald z-test is given by

$$n_B = \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\omega_{xz}}{\delta_{xz}^2}, \tag{14}$$

where δ_{xz} is the target interaction effect size, and results in Section 2.2 suggest

$$\omega_{xz} = n\mathrm{Var}(\hat{\Delta}_{xz}) = n\mathrm{Var}(\hat{\beta}_4) = n\mathrm{Var}(\hat{b}_4) = \frac{\sigma_y^2(1-\rho)}{(\bar{m}+\bar{\eta}_1)\pi_x\pi_z(1-\pi_x)(1-\pi_z)}.$$

Based on expression (11), we obtain the approximate sample size formula

$$n_B \approx \frac{(z_{1-\alpha/2}+z_{1-\lambda})^2\sigma_y^2(1-\rho)\{1+(\bar{m}-1)\rho\}^3}{\bar{m}\pi_x\pi_z(1-\pi_x)(1-\pi_z)\delta_{xz}^2\left[\{1+(\bar{m}-2)\rho\}\{1+(\bar{m}-1)\rho\}^2+\mathrm{CV}^2\bar{m}\rho^2(1-\rho)\right]}. \tag{15}$$

Importantly, the required sample size only depends on the interaction effect size δ_{xz} regardless of the magnitude of the main effect parameters in model (1). Furthermore, the required sample size $n_B = n_{A2}(\delta_z/\delta_{xz})^2\{\pi_x(1-\pi_x)\}^{-1}$, which, depending on the relative effect size δ_z/δ_{xz}, may be larger or smaller than the required sample size for testing the marginal individual-level

treatment effect. Similar to the sample size requirement for testing the marginal individual-level treatment effect, n_B is insensitive to realistic degrees of cluster size variability. Finally, when the cluster sizes are all equal, we obtain

$$n_B = \frac{(z_{1-\alpha/2} + z_{1-\lambda})^2 \sigma_y^2 (1-\rho)\{1+(m-1)\rho\}}{m\pi_x\pi_z(1-\pi_x)(1-\pi_z)\delta_{xz}^2\{1+(m-2)\rho\}}, \tag{16}$$

which is a special case of the formula derived in Yang et al.[22] for testing the interaction between a cluster-level treatment and an individual-level binary covariate with zero covariate ICC.[24]

3.3 | Joint test

The derivation of the approximate variance formulas allows us to further develop a sample size procedure for simultaneously testing the two marginal treatment effects. In this case, we may be interested in the null hypothesis of no effect for both treatments, H_0^C: $\Delta_x = \Delta_z = 0$. To obtain the sample size requirement for such a joint test, we need to obtain the covariance between the two marginal treatment effect estimators. Based upon the results in Section 2.2, we can show

$$n\text{Cov}(\hat{\Delta}_x, \hat{\Delta}_z) = n\text{Cov}(\hat{\beta}_2 + \pi_x\hat{\beta}_4, \hat{\beta}_3 + \pi_z\hat{\beta}_4) = n\text{Cov}(\hat{b}_2 + \pi_x\hat{b}_4, \hat{b}_3) = 0,$$

which indicates that the two marginal treatment effect estimators are asymptotically orthogonal. From the property of the FGLS estimator, the scaled vector of the marginal treatment effect estimators converges to a bivariate normal distribution,

$$\sqrt{n}\begin{bmatrix}\hat{\Delta}_x - \delta_x \\ \hat{\Delta}_z - \delta_z\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \Omega = \begin{bmatrix}\omega_x & 0\\ 0 & \omega_z\end{bmatrix}\right). \tag{17}$$

This motivates a simple Wald test statistic $J = n\left(\hat{\omega}_x^{-1}\hat{\Delta}_x^2 + \hat{\omega}_z^{-1}\hat{\Delta}_z^2\right)$, which asymptotically follows a Chi-square distribution with 2 degrees of freedom and a non-centrality parameter $n\left(\omega_x^{-1}\delta_x^2 + \omega_z^{-1}\delta_z^2\right)$. Therefore, given the target effect sizes, the power equation of the joint test is

$$1 - \lambda = \int_{\chi^2_{1-\alpha}(2)}^{\infty} f\left(x; 2, n\left(\omega_x^{-1}\delta_x^2 + \omega_z^{-1}\delta_z^2\right)\right) dx, \tag{18}$$

where $\chi^2_{1-\alpha}(2)$ is the upper-α quantile of the Chi-square distribution with 2 degrees of freedom and $f(x; 2, \theta)$ is the probability density function of the non-central Chi-square distribution with non-centrality parameter θ. To estimate the sample size for the joint test, one could first fix the values of ICC, mean cluster sizes, CV, and the two effect sizes, and then specify a series of increasing integers n. The required sample size n_C can then be obtained by searching the minimum among the integers that provides $(1-\lambda)$ power according to (18).

3.4 | Intersection-union test

While the joint test rejects the null when at least one treatment has an effect on the outcome, investigators may conclude the "success" of a trial only when both treatments are effective. In this particular case, the alternative hypothesis is formulated as H_1^D: $\Delta_x \neq 0$ and $\Delta_z \neq 0$, while the composite null hypothesis holds when at most one treatment has an effect on the outcome. The intersection-union (I-U) test is often used to test the composite null hypothesis, and has been previously applied in trials with multiple co-primary endpoints; see, for example, Chuang et al.,[13] Sozu et al.[26] and Li et al.[27] While these previous applications focused on one-sided alternatives, we expand this approach to test a two-sided alternative H_1^D. Compared to the joint test in Section 3.3 which answers whether there exists at least one treatment that is effective, the I-U test examines whether both treatments are effective.

Specifically, the I-U test considers a bivariate test statistic, $W = (W_x, W_z)^T$, where $W_x = \hat{\Delta}_x/\sqrt{\hat{\omega}_x/n}$ and $W_z = \hat{\Delta}_z/\sqrt{\hat{\omega}_z/n}$. From our previous result (17), it follows that $W = (W_x, W_z)^T$ approximately follows a bivariate normal distribution with mean $\left(\delta_x/\sqrt{\omega_x/n}, \delta_z/\sqrt{\omega_z/n}\right)^T$ and a covariance matrix equal to the 2×2 identity matrix, and therefore the I-U test rejects H_0^D when both $|W_x| > z_{1-\alpha/2}$ and $|W_z| > z_{1-\alpha/2}$.[13,26,27] Given the effect sizes δ_x and δ_z, the power formula for the two-sided I-U

test can be written as

$$\begin{aligned}1-\lambda =&P\left[\left\{|W_x|>z_{1-\alpha/2}\right\}\cap\left\{|W_z|>z_{1-\alpha/2}\right\}\right]\\ =&\Phi\left(z_{\alpha/2};\delta_x/\sqrt{\omega_x/n},1\right)\Phi\left(z_{\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)\\ &+\Phi\left(z_{\alpha/2};\delta_x/\sqrt{\omega_x/n},1\right)\left\{1-\Phi\left(z_{1-\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)\right\}\\ &+\left\{1-\Phi\left(z_{1-\alpha/2};\delta_x/\sqrt{\omega_x/n},1\right)\right\}\Phi\left(z_{\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)\\ &+\left\{1-\Phi\left(z_{1-\alpha/2};\delta_x/\sqrt{\omega_x/n},1\right)\right\}\left\{1-\Phi\left(z_{1-\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)\right\}\\ =&\left\{1+\Phi\left(z_{\alpha/2};\delta_x/\sqrt{\omega_x/n},1\right)-\Phi\left(z_{1-\alpha/2};\delta_x/\sqrt{\omega_x/n},1\right)\right\}\\ &\times\left\{1+\Phi\left(z_{\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)-\Phi\left(z_{1-\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)\right\},\end{aligned}\tag{19}$$

where $\Phi(\bullet;\mu,\sigma^2)$ is the cumulative distribution function corresponding to a normal distribution with mean μ and variance σ^2. Equation (19) can be used to numerically estimate the required sample size. Specifically, the investigators need to specify the values of ICC, mean and CV of cluster sizes, and the effect sizes. Then, a series of increasing integers n can be plugged into equation (19) to compute the power. The smallest integer n_D that corresponds to no smaller than $(1-\lambda)$ power is then given as the estimated number of clusters to power the I-U test with the composite null H_0^D. Finally, because the I-U test rejects H_0^D only when W_x and W_z both fall beyond the critical value, it is straightforward to see that the I-U test requires a sample size at least as large as that required by the test for marginal effect of a single treatment in Section 3.1. In other words, $n_D \geq \max\{n_{A1}, n_{A2}\}$.

3.5 | Finite-sample considerations

Due to financial and human resource constraints, a frequent limitation of research designs using clusters (such as health centers or clinics) is that a small number of clusters are available, even though the clusters may have moderate to large sizes. For example, recent systematic reviews of published cluster randomized trials found that more than half of the studies reviewed included 24 or fewer clusters.[28,29] With a limited number of clusters available, the Wald z-test may carry an inflated type I error rate when studying the marginal cluster-level treatment effect, and a t-test coupled with the between-within degrees of freedom (df $= n-2$) has been suggested to preserve the nominal test size.[30] This finite-sample consideration necessitates modifications to the sample size procedures concerning the test for marginal cluster-level treatment effect, the joint test, and the I-U test. On the other hand, because the total sample size is usually much larger than the number of clusters, the Wald-test for the marginal individual-level treatment effect or the interaction effect has sufficient within-cluster degrees of freedom such that the z-approximation of the null distribution is adequate.

For testing the marginal cluster-level treatment effect, we still proceed with the test statistic $W_x = \hat{\Delta}_x/\sqrt{\hat{\omega}_x/n}$, which approximately follows a t-distribution under H_0^{A1}. Under the alternative, W_x follows the noncentral t-distribution with noncentrality parameter $\delta_x/\sqrt{\omega_x/n}$. The corresponding power formula is given by

$$1-\lambda = 1-\Psi_{n-2}\left(t_{1-\alpha/2,n-2};\delta_x/\sqrt{\omega_x/n}\right)+\Psi_{n-2}\left(t_{\alpha/2,n-2};\delta_x/\sqrt{\omega_x/n}\right)\tag{20}$$

where $t_{1-\alpha/2,n-2}$ and $t_{\alpha/2,n-2}$ are the upper- and lower- $\alpha/2$ quantile of the central t-distribution with $n-2$ degrees of freedom, and $\Psi_{n-2}(\bullet;\theta)$ is the cumulative distribution function of the noncentral t-distribution with $n-2$ degrees of freedom and noncentrality parameter θ. Although equation (20) should in principal be solved iteratively, a non-iterative approximation could be made by computing the required sample size n_{A1} through (9) and then multiplying by $(n_{A1}+1)/(n_{A1}-1)$.[31]

For testing H_0^C based on the omnibus statistics J defined in Section 3.3, due to asymptotic independence between $\hat{\Delta}_x$ and $\hat{\Delta}_z$, the null distribution can now be approximated by $F(1,n-2)+\chi^2(1)$, which is the mixed central F-χ^2 distribution (i.e., distribution for the sum of an independent central F-random variable and an independent central Chi-square random variable). Because the critical value of this null distribution is not directly available, we draw 10,000 simulations from $F(1,n-2)$ and $\chi^2(1)$, and numerically identify the upper-α quantile to form the associated rejection region. Under the alternative, the omnibus test statistic J approximately follows $F(1,n-2,n\omega_x^{-1}\delta_x^2)+\chi^2(1,n\omega_z^{-1}\delta_z^2)$, which is the mixed noncentral F-χ^2 distribution (i.e., distribution for the sum of an independent noncentral F-random variable and an independent noncentral Chi-square random variable). For each candidate n, we then draw 10,000 simulations from $F(1,n-2,n\omega_x^{-1}\delta_x^2)+\chi^2(1,n\omega_z^{-1}\delta_z^2)$ and compute the

power as the proportion of draws that fall beyond the critical value of the mixed central F-χ^2 distribution. The required sample size n_C is identified as the smallest number of clusters such that the power is at least $1-\lambda$.

Finally, for testing H_0^D, we can replace the z-based I-U test in Section 3.4 with a mixed t- and z-based I-U test to improve the validity for testing the marginal cluster-level treatment effect. The power formula for the mixed t- and z-based I-U test could be written as

$$1-\lambda = \left\{1+\Psi_{n-2}\left(t_{\alpha/2,n-2};\delta_x/\sqrt{\omega_x/n}\right)-\Psi_{n-2}\left(t_{1-\alpha/2,n-2};\delta_x/\sqrt{\omega_x/n}\right)\right\} \times\left\{1+\Phi\left(z_{\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)-\Phi\left(z_{1-\alpha/2};\delta_z/\sqrt{\omega_z/n},1\right)\right\}, \quad (21)$$

where $\Phi(\bullet;\mu,\sigma^2)$ and $\Psi(\bullet;\theta)$ are defined earlier. The same iterative algorithm can be used to search for the smallest value, n_D, that satisfies the power equation. The use of the t-approximation, as we shall see in Section 4, can help maintain the correct type I error rate with a limited number of clusters and therefore greatly improve the validity for designing hierarchical factorial trials.

4 | A SIMULATION STUDY

4.1 | Simulation design

We carried out a simulation study to assess the performance of the proposed sample size formulas in a hierarchical 2×2 factorial trial with equal randomization ($\pi_x=\pi_z=1/2$). Based on the sample size equations we derived in Section 3, the number of clusters is determined by the following parameters: nominal type I error rate (α), power ($1-\lambda$), total variance (σ_y^2), ICC (ρ), mean cluster size ($\bar{m}$), CV of cluster sizes, and the effect sizes for different hypotheses (δ_x, δ_z, or δ_{xz}). Throughout, we fixed the total variance σ_y^2 at 1, nominal type I error α at 0.05 and the desired power level $1-\lambda$ at 0.8, and varied the remaining parameters. We considered two levels of mean cluster sizes $\bar{m}\in\{50,100\}$, and three levels of ICC $\rho\in\{0.02,0.05,0.1\}$, based on the range commonly reported in the cluster randomized design literature.[4,25] The CV of cluster sizes were chosen from CV $\in\{0,0.3,0.6,0.9\}$ with CV = 0 representing equal cluster sizes. Our experiences suggest that most cluster randomized trials have CV no larger than 0.6, and therefore the scenario with CV = 0.9 corresponds to an extreme case for illustration.[32,33] To ensure a realistic range of predicted sample sizes, we separately specified effect sizes for each type of hypothesis. We chose $\delta_x\in\{0.2,0.4\}$ for testing H_0^{A1}, $\delta_z\in\{0.1,0.15\}$ for testing H_0^{A2}, $\delta_{xz}\in\{0.2,0.3\}$ for testing H_0^B, $(\delta_x,\delta_z)\in\{(0.2,0.1),(0.25,0.1)\}$ for testing H_0^C and $(\delta_x,\delta_z)\in\{(0.2,0.1),(0.4,0.2)\}$ for testing H_0^D. In summary, we considered $2\times3\times4\times2=48$ parameter combinations for each one of the five null hypotheses. For each parameter combination, we estimated the number of clusters n that gives at least 80% power and rounded to the nearest even integer above to ensure an exactly equal randomization. We used the predicted cluster number n to simulate correlated outcomes and obtain the empirical power to validate the accuracy of the formula-based power prediction.

We generated correlated outcome data based on model (1). We fixed $\beta_1=1$ for simplicity. As described in Section 3, while the interaction effect size δ_{xz} directly corresponds to β_4, the marginal effect sizes δ_x, δ_z are linear combinations of regression parameters and only imply constraints for β_2, β_3 and β_4. For testing null hypotheses H_0^{A1}, H_0^{A2} and H_0^B, we fixed $\beta_2=0.15$ and $\beta_3=0.05$ and computed β_4 based on the assumed effect sizes. Specifically, we have $\beta_4=2(\delta_x-\beta_2)$, $\beta_4=2(\delta_z-\beta_3)$, and $\beta_4=\delta_{xz}$. When designing the simulations for testing H_0^C and H_0^D, we fixed $\beta_2=0.15$ and solved β_3 and β_4 based on the corresponding linear constraints. Table 2 summarized the specification of regression parameters in each simulation scenario to match the assumed marginal effect sizes. Finally, given values of $\bar{m}$ and CV, we simulate varying cluster sizes using $m_i\sim\text{Gamma}(g,h)$, where the shape parameter $g=\text{CV}^{-2}$ and the rate parameter $h=\bar{m}^{-1}\text{CV}^{-2}$. The simulated cluster size m_i was rounded to the nearest integer and ensured to be at least 2 for computational stability. Finally, the cluster-specific random intercept a_i was randomly generated from $\mathcal{N}(0,\rho)$, and the random error ϵ_{ij} was independently generated from $\mathcal{N}(0,1-\rho)$. For each parameter combination, we generated 5,000 hypothetical factorial trials for the evaluation of empirical type I error under the null and power under the alternative.

For each simulated hypothetical factorial trial, we fitted the linear mixed model (1) using the restricted maximum likelihood estimation (REML) and carried out the corresponding test for inference. Under each null, the empirical type I error rate was computed as the proportion of false rejections among the 5,000 trials. Under the alternative, the empirical power was calculated as the proportion of correct rejections among the 5,000 trials, and was compared with the power prediction based on our proposed formulas. Finally, for the tests involving the marginal effect of the cluster-level treatment, i.e., those associated with H_0^{A1}, H_0^C and H_0^D, we replicated the simulations based on the modified sample size methods discussed in Section 3.5, using

TABLE 2 Specification of regression parameters for generating correlated outcome data in different simulation scenarios. (A1) represents the test for marginal cluster-level treatment effect, (A2) represents the test for marginal individual-level treatment effect, (B) represents the interaction test, (C) represents the joint test, and (D) represents the intersection-union test. β_2, β_3, and β_4 stands for the true regression parameters corresponding to the cluster-level treatment effect, the individual-level treatment effect, and the interaction effect, respectively.

Test label	Hypothesis	β_2	β_3	β_4	Effect size
(A1)	Null (H_0^{A1})	0.15	0.05	−0.3	$\Delta_x = 0$
	Alternative (H_1^{A1})	0.15	0.05	$2(\delta_x - 0.15)$	$\Delta_x = \delta_x$
(A2)	Null (H_0^{A2})	0.15	0.05	−0.1	$\Delta_z = 0$
	Alternative (H_1^{A2})	0.15	0.05	$2(\delta_z - 0.05)$	$\Delta_z = \delta_z$
(B)	Null (H_0^{B})	0.15	0.05	0	$\Delta_{xz} = 0$
	Alternative (H_1^{B})	0.15	0.05	δ_{xz}	$\Delta_{xz} = \delta_{xz}$
(C)	Null (H_0^{C})	0.15	0.15	−0.3	$\Delta_x = \Delta_z = 0$
	Alternative (H_1^{C})	0.15	$\delta_z - \delta_x + 0.15$	$2(\delta_x - 0.15)$	$\Delta_x = \delta_x, \Delta_z = \delta_z$
(D)	Null (H_0^{D})	0.15	$\delta_z + 0.15$	−0.3	$\Delta_x = 0, \Delta_z = \delta_z$
	Alternative (H_1^{D})	0.15	$\delta_z - \delta_x + 0.15$	$2(\delta_x - 0.15)$	$\Delta_x = \delta_x, \Delta_z = \delta_z$

the same parameter constellations, to assess the potential improvement of type I error rate. Our simulations were carried out in R (version 3.6.2) and the linear mixed model was fitted using the `nlme` package.[34]

4.2 | Simulation results

Web Table 1 and 2 present the estimated required number of clusters (n_{A1}), empirical type I error (ψ), empirical power (ϕ) and predicted power ($\hat{\phi}$) corresponding to testing the marginal effect of the cluster-level treatment based on two levels of effect sizes. The t-test with the between-within degrees of freedom can require more clusters to achieve a similar level of power compared to the z-test. However, compared to the z-test, the t-test has more robust control of the empirical type I error rate, especially with a larger effect size δ_x and a smaller number of clusters. For example, with as few as 6 clusters, the z-test carries a 10% type I error rate. On the other hand, the empirical power of both the z-test and t-test agree well with the prediction even if the CV of cluster sizes is as extreme as 0.9, which confirms the accuracy of the proposed formulas. In addition, we also observe that the required sample size can be sensitive to the CV of cluster sizes when the effect size is relatively small (Web Table 1). Overall, the findings for testing the marginal cluster-level treatment effect in our hierarchical factorial trial are consistent with the previous findings in parallel cluster randomized trials.[18]

When testing the marginal individual-level treatment effect (Web Table 3) and the interaction effect (Table 3), the z-test provides close to nominal test size and dispenses the need for any finite-sample corrections. In our simulation design, we set $\delta_z/\delta_{xz} = 1/2 = \sqrt{\pi_x(1-\pi_x)}$, and therefore the estimated sample size n_B and n_{A2} are identical in Web Table 3 and Table 3. Confirming our analytical discussion in Section 3.1, the estimated sample size is insensitive to the CV of cluster sizes as the ICC in clustered designs is usually small.[4,25] In general, the empirical power of the z-test for H_0^{A2} and H_0^{B} is close to the formula prediction, and confirms the accuracy of our sample size formulas. However, the empirical power of the test appears to be slightly lower than the prediction when the mean cluster size $\bar{m} = 100$ and the CV of cluster sizes becomes 0.9. In this case, the estimated number of clusters is often smaller than 15 and the large-sample approximation under unequal cluster sizes may be less accurate. With a larger cluster size, the empirical power of the z-test for testing the individual-level treatment effect and the interaction effect matches the formula prediction even when the CV of cluster sizes is equal to 0.9.

Web Tables 4 and 5 present the estimated required number of clusters (n_C), empirical type I error (ψ), empirical power (ϕ) and predicted power ($\hat{\phi}$) corresponding to the joint test with two levels of effect sizes. For the simulation parameters we considered, the estimated sample sizes are generally similar between the large-sample Chi-square test and the mixed F-χ^2 test. However, the mixed F-χ^2 test corrects for the type I error rate inflation when the estimated sample size is smaller than 20, and is favored for validity considerations. The empirical power of the mixed F-χ^2 test also matches well with the prediction. Finally, Web Table 6 and Table 4 present the estimated required number of clusters (n_D), empirical type I error (ψ), empirical power (ϕ) and

TABLE 3 Required number of clusters n_B, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the interaction test. Notation: δ_{xz} is the interaction effect size, $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			$\delta_{xz} = 0.2$				$\delta_{xz} = 0.3$			
		CV	n_B	ψ	ϕ	$\hat{\phi}$	n_B	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	64	0.05	0.82	0.81	28	0.05	0.80	0.81
		0.3	64	0.05	0.81	0.81	28	0.05	0.81	0.81
		0.6	64	0.05	0.82	0.81	28	0.05	0.81	0.81
		0.9	64	0.05	0.80	0.81	28	0.05	0.79	0.81
	$\rho = 0.05$	0	62	0.05	0.82	0.81	28	0.05	0.81	0.82
		0.3	62	0.06	0.81	0.81	28	0.05	0.82	0.82
		0.6	62	0.05	0.80	0.81	28	0.05	0.82	0.82
		0.9	62	0.05	0.82	0.81	28	0.05	0.80	0.82
	$\rho = 0.10$	0	58	0.06	0.81	0.80	26	0.05	0.81	0.81
		0.3	58	0.05	0.81	0.80	26	0.05	0.80	0.81
		0.6	58	0.05	0.80	0.80	26	0.05	0.80	0.81
		0.9	58	0.05	0.81	0.80	26	0.05	0.79	0.81
$\bar{m} = 100$	$\rho = 0.02$	0	32	0.04	0.82	0.81	14	0.05	0.81	0.81
		0.3	32	0.05	0.82	0.81	14	0.05	0.81	0.81
		0.6	32	0.05	0.81	0.81	14	0.05	0.79	0.81
		0.9	32	0.05	0.80	0.81	14	0.05	0.76	0.81
	$\rho = 0.05$	0	32	0.04	0.83	0.82	14	0.05	0.83	0.82
		0.3	32	0.05	0.83	0.82	14	0.05	0.82	0.82
		0.6	32	0.05	0.82	0.82	14	0.05	0.81	0.82
		0.9	32	0.05	0.81	0.82	14	0.05	0.77	0.82
	$\rho = 0.10$	0	30	0.05	0.82	0.82	14	0.05	0.85	0.84
		0.3	30	0.05	0.81	0.82	14	0.05	0.84	0.84
		0.6	30	0.05	0.82	0.82	14	0.05	0.82	0.84
		0.9	30	0.05	0.80	0.82	14	0.05	0.79	0.84

predicted power ($\hat{\phi}$) corresponding to the I-U test with two levels of effect sizes. The general patterns are similar to the joint test. Specifically, our sample size procedure accurately predicted the power for both the z-based and the mixed z- and t-based I-U test, with the latter carrying close to the nominal type I error rate with fewer than 30 clusters.

5 | APPLICATION TO THE SUICIDE PREVENTION FACTORIAL TRIAL

We illustrate the proposed sample size formulas in the context of the motivating suicide prevention trial. The suicide prevention trial considers a hierarchical 2×2 factorial design, and aims to study the clinical effectiveness of two treatment strategies, CC delivered at the cluster level and CBT-SP delivered at the individual level, for suicide prevention among community-dwelling transgender individuals. Clinic will be randomized in a 1:1 ratio to usual care or to deliver CC, an efficacious suicide prevention approach that involves sending participants brief, non-demanding expressions of care and concern at specified time intervals.[7] Participants within a clinic be randomized in a 1:1 ratio to receive the CBT-SP program or usual care. CBT-SP consists of acute and continuation phases, each lasting about 12 individual sessions, and includes a chain analysis of the suicidal event, safety plan development, skill building, psychoeducation, family intervention, and relapse (recurrence of suicidal behavior) prevention.[8] Since the depression is an outcome along the causal pathway to suicide attempt or suicide death,[35] we consider it as an intermediate outcome to evaluate the clinical effectiveness of our two interventions. The level of depression will be

TABLE 4 Required number of clusters n_D, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the intersection-union test with and without finite-sample correction. The marginal cluster-level treatment effect size is $\delta_x = 0.4$ and the marginal individual-level treatment effect size is $\delta_z = 0.2$. Notation: $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			z-based I-U test				t- and z-based I-U test			
		CV	n_D	ψ	ϕ	$\hat{\phi}$	n_D	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	18	0.06	0.84	0.85	18	0.04	0.83	0.84
		0.3	18	0.05	0.84	0.84	18	0.04	0.83	0.83
		0.6	18	0.06	0.83	0.84	18	0.04	0.82	0.83
		0.9	18	0.05	0.80	0.83	18	0.04	0.79	0.81
	$\rho = 0.05$	0	20	0.06	0.82	0.83	20	0.04	0.80	0.80
		0.3	20	0.06	0.83	0.83	22	0.05	0.85	0.85
		0.6	20	0.06	0.79	0.81	22	0.05	0.83	0.83
		0.9	22	0.07	0.81	0.83	22	0.05	0.78	0.81
	$\rho = 0.10$	0	26	0.07	0.81	0.81	28	0.05	0.83	0.82
		0.3	26	0.06	0.81	0.81	28	0.05	0.82	0.82
		0.6	28	0.06	0.82	0.83	28	0.05	0.79	0.80
		0.9	28	0.06	0.79	0.81	30	0.05	0.81	0.81
$\bar{m} = 100$	$\rho = 0.02$	0	10	0.08	0.84	0.85	12	0.05	0.89	0.89
		0.3	10	0.07	0.82	0.85	12	0.04	0.88	0.89
		0.6	10	0.07	0.80	0.84	12	0.05	0.87	0.87
		0.9	10	0.08	0.74	0.81	12	0.05	0.81	0.85
	$\rho = 0.05$	0	14	0.07	0.83	0.84	16	0.05	0.85	0.85
		0.3	14	0.07	0.84	0.84	16	0.05	0.83	0.84
		0.6	14	0.07	0.81	0.82	16	0.05	0.82	0.83
		0.9	16	0.07	0.82	0.86	16	0.05	0.78	0.81
	$\rho = 0.10$	0	22	0.06	0.80	0.81	24	0.05	0.81	0.81
		0.3	22	0.07	0.81	0.81	24	0.05	0.81	0.81
		0.6	24	0.07	0.83	0.83	26	0.04	0.83	0.83
		0.9	24	0.06	0.80	0.82	26	0.05	0.80	0.82

measured using the nine-item Patient Health Questionnaire (PHQ-9), a 9 item scale with a total score ranging from 0 to 27. We treat the score as a continuous outcome with larger values indicating ascending symptoms of depression.[35]

Figure 1 presents the sample size requirements for five different tests that can be relevant for planning the trial (also see Table 1 for these hypotheses of potential interest). Each panel plots the combinations of mean cluster size and the number of clusters for a two-sided test with 0.05 significance level to achieve 80% power, given a fixed set of CV of cluster sizes. We interpret each test separately and therefore do not further consider corrections for multiple testing. Because the t-approximation could substantially improve the empirical type I error rate, our calculation considers the t-based finite-sample corrections, whenever applicable. For producing each panel, we hypothesize that the standardized effect size for the marginal effect of the CC program is $\delta_x/\sigma_y = 0.25$, and that for the marginal effect of the CBT-SP program is $\delta_z/\sigma_y = 0.33$. We also assume the standardized effect size of the interaction effect to be $\delta_{xz}/\sigma_y = 0.3$. The ICC characterizing the within-cluster similarity is assumed to be 0.01, and the allocation ratio $\pi_x = \pi_z = 1/2$ under equal randomization. Because each clinic on average is likely to recruit no more than 100 participants, we vary the mean cluster size from 10 to 100 for each test.

Panel (A1) in Figure 1 presents the sample size requirement for testing the marginal effect of the CC program across four levels of cluster size variations measured by CV. Under equal cluster sizes (CV = 0), as the mean cluster size increases, the required number of clusters decreases from 58 to 14. At the same level of mean cluster size, a larger CV will slightly inflate n_{A1}. This observation is consistent with the findings in our simulation study and the prior results studied for a two-arm parallel

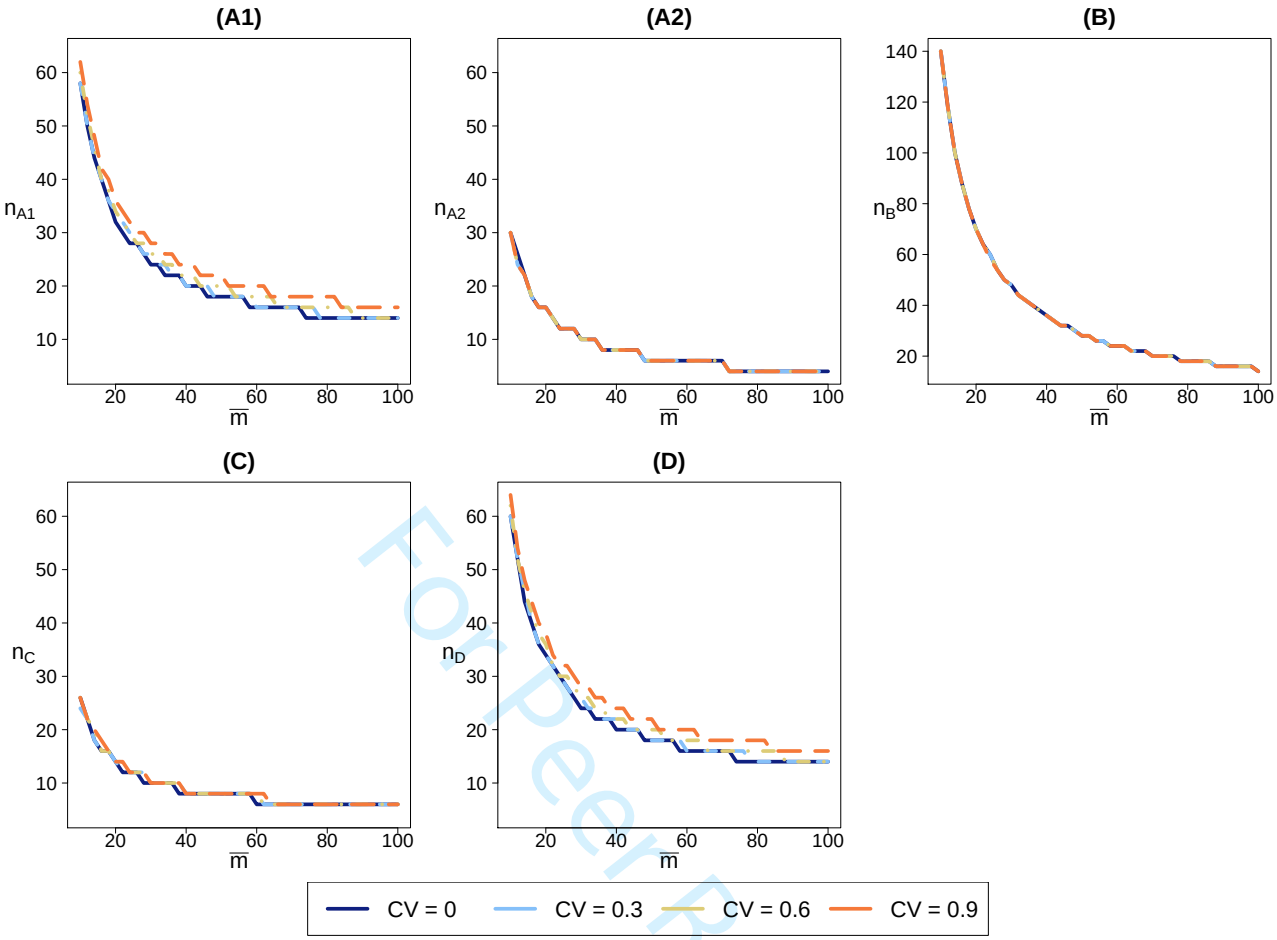

FIGURE 1 Required number of clusters n and mean cluster sizes $\bar{m}$ to achieve 80% power across four levels of cluster size variability for five types of hypothesis tests for the marginal effect of Caring Contacts (CC) program and the Cognitive Behavioral Therapy for Suicide Prevention (CBT-SP) program in the motivating trial. (A1) stands for the test for marginal treatment effect of the CC program, (A2) stands for the test for marginal treatment effect of the CBT-SP program, (B) stands for the interaction test, (C) stands for the joint test, and (D) stands for the intersection-union test.

cluster randomized trial.[18] In contrast, Panel (A2) and (B) indicate that the CV of cluster sizes has negligible influence on the sample size requirements for testing the marginal effect of the CBT-SP program or the interaction effect. This is expected because $\rho^2(1-\rho) = 9.9 \times 10^{-5} \approx 0$, and therefore the term involving CV in equation (12) and (16) is negligible. In addition, we observe that, under a similar level of effect size, the interaction test can require a substantially larger number of clusters compared to the tests for marginal effect associated with the CC program or the CBT-SP program. For example, to ensure an 80% power for the test for marginal effect of the CBT-SP program, the required number of clusters decreases from 30 to 4 as the mean cluster size increases from 10 to 100. However, for the interaction test, the required number of clusters decreases from 140 to 14, even though the effect sizes for the CBT-SP program and the interaction are somewhat similar.

Panels (C) and (D) present the sample size requirement for the joint test and the I-U test for the effects of the CC and CBT-SP programs. With our design parameters, the cluster size variability has only a minimal impact on the required sample size corresponding to the joint test, but the I-U test is more susceptible to cluster size variability, similar to the test for marginal effect of the CC program. Interestingly, with the same level of sample size, the joint test is always more powerful than the I-U test with the same level of effect sizes. Furthermore, the required sample size for the joint test approximates that for the marginal treatment effect test of the CBT-SP program, while the required sample size for the I-U test approximates that for the marginal treatment effect test of the CC program. Overall, the panels in Figure 1 provide different possibilities for the investigators to decide on recruitment parameters according to their resources and choices of hypothesis tests. For example, if each clinic can on

average recruit 20 participants due to the effort required to engage community-dwelling transgender individuals and assuming CV = 0.3, recruiting 35 clinics would have 80% power for the two separate tests of the marginal treatment effects, the joint test and the I-U test to detect our assumed effect sizes; at least 70 clinics may be required to detect the assumed interaction effect size.

6 | DISCUSSION

In this article, we developed a set of sample size and power formulas in a hierarchical 2×2 factorial trial with a cluster-level treatment and an individual-level treatment. Based on a continuous outcome, we considered different types of statistical tests for the analysis of the factorial trial, including (A1) the test for marginal cluster-level treatment effect, (A2) the test for marginal individual-level treatment effect, (B) the test for the interaction effect, (C) the joint test for the two marginal effects, and (D) the I-U test for the two marginal effects. For tests (A1), (A2), (C) and (D), we focused on the marginal effect of each treatment defined across the levels of the other treatment for a population-averaged interpretation. In addition, we showed that the estimators for the two marginal treatment effects are asymptotically independent, facilitating a simple derivation of the sample size requirements for the joint test and I-U test. Our simulations indicate that the proposed sample size formulas can accurately track the empirical power of each test under a wide range of parameter constellations. We applied our formulas to study the sample size requirements for each test in the motivating suicide prevention factorial trial, and illustrated different possibilities on the number of clusters and average cluster sizes to achieve the desired level of power under a fixed set of design parameters (i.e., effect sizes, type I error, power).

While sample size formulas for planning research designs with clusters often assume an equal cluster size, our development relaxed this condition and provided approximations under unequal cluster sizes. Importantly, even in the presence of an individual-level treatment, the VIF for testing the marginal cluster-level treatment effect due to unequal cluster sizes has the same form as the VIF developed for a two-arm parallel cluster randomized trial.[18] Therefore, the design implications from cluster randomized trials can be extrapolated to the marginal cluster-level treatment effect test in our hierarchical 2×2 factorial designs. For example, van Breukelen[18] pointed out that the loss of efficiency (or the inflation of sample size) rarely exceeds 10% for cluster randomized trials analyzed by linear mixed models, which should apply to the marginal analysis of cluster-level treatment based on our linear mixed model (1) with two treatments. In contrast, unequal cluster sizes have negligible effect on the sample size requirement for testing the marginal individual-level treatment effect. This is mainly because the factor involving CV is multiplied by the average cluster size and $\rho^2(1-\rho)$, the latter of which is usually fairly small in research designs with clusters. Furthermore, because the sample size formula for the interaction test is proportional to that for the marginal individual-level treatment effect test, it is similarly insensitive to unequal cluster sizes. Finally, our simulations and data application suggest that unequal cluster sizes may have a larger impact on the I-U test compared to the joint test, because the power of the I-U test directly depends on the power of the marginal cluster-level treatment effect test.

Because research designs with clusters may not enlist a large number of clusters, and the z-test for the marginal cluster-level treatment effect may carry an inflated type I error rate in small samples, our sample size development also considers a t-approximation with the between-within degrees of freedom[30] as a finite-sample consideration. In contrast, because the individual treatment variable changes within each cluster and exploits within-cluster comparisons for estimation, the tests for the marginal individual-level treatment effect and the interaction effect between two treatments have sufficient within-cluster degrees of freedom, which ensures the accuracy of the z-test even when the number of clusters is limited. These considerations were included in developing the sample size formulas for the joint test and the I-U test. In particular, with the t-approximation applied at the cluster level, we show that the joint test has a mixed F-χ^2 distribution under the null and a noncentral F-χ^2 distribution under the alternative. We numerically determined the rejection region and computed the power for the assumed effect sizes because the F-χ^2 distribution is not as standard as the usual Chi-square approximation for the joint test. Similarly, a mixed t- and z-based I-U test was considered in our sample size method. Our simulations demonstrated that the t-approximation not only improves the test size (compared to the z-approximation) for the marginal cluster-level treatment effect test, but also for the joint test and the I-U test, suggesting its necessity in developing a valid sample size procedure. Such finite-sample considerations were also included in determining the required sample size for the motivating suicide prevention trial.

There are some limitations that we plan to address in our future work. First, we have only considered a hierarchical 2×2 factorial design as in our motivating example, while other factorial clustered designs can have more than two arms at each level. For example, Shin et al.[11] developed a sample size procedure for the split-plot design with K factors (treatment arms), but they have not considered testing the marginal effect for each factor (they have only considered testing the main effects and interaction

effects) and only assumed an equal cluster size. It would be interesting to extend our work to accommodate an arbitrary level of treatments at each level, while allowing for unequal cluster sizes. Second, we only assumed a single level of clustering as in our motivating trial. However, research designs with clusters may have multiple levels of clustering; for example, sample size formulas have been developed to accommodate a linear mixed model with two random intercepts in three-level cluster randomized trials.[36,37] We anticipate it would be possible to extend our sample size methodology to accommodate a factorial trial with an additional level of clustering, such as when the community-dwelling transgender participants are nested in professional healthcare providers, who are nested in clinics. In this case, the cluster-level treatment can be delivered to either the clinic or the professional healthcare provider, which require different sample size considerations. Third, while we considered several test statistics for different null hypotheses, we interpreted each test separately and have not addressed multiple testing when more than one test is used for the primary analysis. Nevertheless, our formulas can be easily combined with Bonferroni correction as a conservative approach to control for family-wise error rate. Finally, our development assumes a linear mixed model with a continuous outcome, and we plan to carry out future work to extend our methods to accommodate a hierarchical factorial trial with a binary outcome.

SUPPORTING INFORMATION

Web Tables 1-6 references in the article can be found at the online supplementary materials available at *Wiley Library Online*. R code for reproducing the simulation results and the Figure in Section 5 is available at the author's GitHub page https://github.com/BillyTian/code_Hierarchical2x2Factorial. The proposed sample size formulas are also implemented in an open-source R package **H2x2Factorial** that will be available on the Comprehensive R Archive Network (CRAN).

ACKNOWLEDGEMENTS

This work is supported by CTSA Grant Number UL1 TR000142 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICT OF INTEREST

The authors have no conflict of interest.

APPENDIX

We provide full details for deriving the explicit form of the 4×4 covariance matrix $\Sigma = \sigma_y^2 \left(\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} D_i^T R_i^{-1} D_i\right)^{-1}$. Recall that we have defined the design vector $D_{ij} = \left(1, (X_i - \pi_x), Z_{ij}, (X_i - \pi_x)Z_{ij}\right)^T$ and $D_i = (D_{i1}, \ldots, D_{im_i})^T$. For each cluster i, the inverse of the compound symmetric correlation matrix can be obtained as

$$R_i^{-1} = \frac{1}{1-\rho} I_{m_i} - \frac{\rho}{(1-\rho)[1+(m_i-1)\rho]} J_{m_i} = \frac{1}{1-\rho}(I_{m_i} + c_i J_{m_i}),$$

where $c_i = -\rho/[1+(m_i-1)\rho]$. Therefore, we can write

$$\frac{1}{n}\sum_{i=1}^{n} D_i^T R_i^{-1} D_i = \frac{1}{n(1-\rho)}\sum_{i=1}^{n} D_i^T D_i + \frac{1}{n(1-\rho)}\sum_{i=1}^{n} c_i D_i^T J_{m_i} D_i. \tag{22}$$

Using the fact that $Z_{ij}^2 = Z_{ij}$, we can expand one of the main components in expression (22),

$$\frac{1}{n}\sum_{i=1}^{n} D_i^T D_i = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} m_i & m_i(X_i-\pi_x) & \sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij} \\ m_i(X_i-\pi_x) & m_i(X_i-\pi_x)^2 & (X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)^2\sum_{j=1}^{m_i} Z_{ij} \\ \sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij} & \sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij} \\ (X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)^2\sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_x)^2\sum_{j=1}^{m_i} Z_{ij} \end{bmatrix}.$$

Assuming the cluster sizes are non-informative, then the cluster size distribution $f(m_i)$ is independent of the assignment of randomized interventions. Define $\bar{m} = E(m_i)$ as the mean cluster size, $\sigma_x^2 = \pi_x(1-\pi_x)$ is the variance of cluster-level intervention indicator. Treating $\sum_{i=1}^{m_i} Z_{ij}$ as a random variable, then it has mean $m_i\pi_z$ and variance $m_i\sigma_z^2 = m_i\pi_z(1-\pi_z)$ due to binomial sampling. Invoking the Weak Law of Large Numbers (WLLN) for the independent but non-identically distributed random variable, we can write

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} Z_{ij}\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E\left(\sum_{j=1}^{m_i} Z_{ij}\right) = \pi_z \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_i = \bar{m}\pi_z. \tag{23}$$

For the first entry of the matrix above, it follows that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_i = E(m_i) = \bar{m}.$$

By the independence among the two interventions and m_i, the WLLN, and what is proved in equation (23), we can obtain the other entries in the matrix above

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_i(X_i-\pi_x) = E(m_i)E(X_i-\pi_x) = 0, \quad \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_i(X_i-\pi_x)^2 = E(m_i)\mathrm{Var}(X_i) = \bar{m}\sigma_x^2,$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{(X_i-\pi_x)\sum_{j=1}^{m_i} Z_{ij}\right\} = E(X_i-\pi_x)\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} Z_{ij}\right) = 0,$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{(X_i-\pi_x)^2\sum_{j=1}^{m_i} Z_{ij}\right\} = \mathrm{Var}(X_i)\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} Z_{ij}\right) = \sigma_x^2\bar{m}\pi_z.$$

This will allow us to write

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} D_i^T D_i = \bar{m}\begin{bmatrix} 1 & 0 & \pi_z & 0 \\ 0 & \sigma_x^2 & 0 & \pi_z\sigma_x^2 \\ \pi_z & 0 & \pi_z & 0 \\ 0 & \pi_x\sigma_x^2 & 0 & \pi_z\sigma_x^2 \end{bmatrix}.$$

Similarly, we can expand the other component in equation (22),

$$\frac{1}{n}\sum_{i=1}^{n} c_i D_i^T J_{m_i} D_i = \frac{1}{n}\sum_{i=1}^{n} c_i\left(\sum_{j=1}^{m_i} D_{ij}\right)\left(\sum_{j=1}^{m_i} D_{ij}^T\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-\rho}{1+(m_i-1)\rho}\right\}\begin{bmatrix} m_i^2 & m_i^2(X_i-\pi_X) & m_i\sum_{j=1}^{m_i} Z_{ij} & m_i(X_i-\pi_X)\sum_{j=1}^{m_i} Z_{ij} \\ m_i^2(X_i-\pi_X) & m_i^2(X_i-\pi_X)^2 & m_i(X_i-\pi_X)\sum_{j=1}^{m_i} Z_{ij} & m_i(X_i-\pi_X)^2\sum_{j=1}^{m_i} Z_{ij} \\ m_i\sum_{j=1}^{m_i} Z_{ij} & m_i(X_i-\pi_X)\sum_{j=1}^{m_i} Z_{ij} & \left(\sum_{j=1}^{m_i} Z_{ij}\right)^2 & (X_i-\pi_X)\left(\sum_{j=1}^{m_i} Z_{ij}\right)^2 \\ m_i(X_i-\pi_X)\sum_{j=1}^{m_i} Z_{ij} & m_i(X_i-\pi_X)^2\sum_{j=1}^{m_i} Z_{ij} & (X_i-\pi_X)\left(\sum_{j=1}^{m_i} Z_{ij}\right)^2 & (X_i-\pi_X)^2\left(\sum_{j=1}^{m_i} Z_{ij}\right)^2 \end{bmatrix}.$$

Define the following two expectations of the functions of cluster sizes,

$$\bar{\eta}_1 = E\left\{\frac{-m_i\rho}{1+(m_i-1)\rho}\right\}, \qquad \bar{\eta}_2 = E\left\{\frac{-m_i^2\rho}{1+(m_i-1)\rho}\right\}.$$

We show the derivations of some representative entries in the matrix above, by independence and WLLN,

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-\rho}{1+(m_i-1)\rho}m_i\sum_{j=1}^{m_i}Z_{ij}\right\} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-m_i\rho}{1+(m_i-1)\rho}E\left(\sum_{j=1}^{m_i}Z_{ij}\right)\right\}\\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-m_i^2\rho}{1+(m_i-1)\rho}\right\}\pi_z\\
&= \bar{\eta}_2\pi_z,\\
\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-\rho}{1+(m_i-1)\rho}\left(\sum_{j=1}^{m_i}Z_{ij}\right)^2\right\} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-\rho}{1+(m_i-1)\rho}\mathrm{E}\left[\left(\sum_{j=1}^{m_i}Z_{ij}\right)^2\right]\right\}\\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-\rho}{1+(m_i-1)\rho}\left[\mathrm{Var}\left(\sum_{j=1}^{m_i}Z_{ij}\right)+\left\{\mathrm{E}\left(\sum_{j=1}^{m_i}Z_{ij}\right)\right\}^2\right]\right\}\\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{-\rho}{1+(m_i-1)\rho}\left[m_i\sigma_z^2+m_i^2\pi_z^2\right]\right\}\\
&= \bar{\eta}_2\pi_z^2+\bar{\eta}_1\sigma_z^2.
\end{aligned}$$

These allow us to write

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}c_iD_i^TJ_{m_i}D_i=\begin{bmatrix}\bar{\eta}_2 & 0 & \bar{\eta}_2\pi_z & 0\\ 0 & \bar{\eta}_2\sigma_x^2 & 0 & \bar{\eta}_2\pi_z\sigma_x^2\\ \bar{\eta}_2\pi_z & 0 & \bar{\eta}_2\pi_z^2+\bar{\eta}_1\sigma_z^2 & 0\\ 0 & \bar{\eta}_2\pi_z\sigma_x^2 & 0 & \bar{\eta}_2\pi_z^2\sigma_x^2+\bar{\eta}_1\sigma_z^2\sigma_x^2\end{bmatrix}.$$

Combining the two components based on equation (22), we can write

$$\lim_{n\to\infty}\frac{1-\rho}{n}\sum_{i=1}^{n}D_i^TR_i^{-1}D_i=\begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}=\begin{bmatrix}\bar{m}+\bar{\eta}_2 & 0 & (\bar{m}+\bar{\eta}_2)\pi_z & 0\\ 0 & (\bar{m}+\bar{\eta}_2)\sigma_x^2 & 0 & (\bar{m}+\bar{\eta}_2)\pi_z\sigma_x^2\\ (\bar{m}+\bar{\eta}_2)\pi_z & 0 & \bar{m}\pi_z+\bar{\eta}_2\pi_z^2+\bar{\eta}_1\sigma_z^2 & 0\\ 0 & (\bar{m}+\bar{\eta}_2)\pi_z\sigma_x^2 & 0 & \bar{m}\pi_z\sigma_x^2+\bar{\eta}_2\pi_z^2\sigma_x^2+\bar{\eta}_1\sigma_z^2\sigma_x^2\end{bmatrix}.$$

We can observe that $\Omega_{12}=\Omega_{21}=\pi_z\Omega_{11}$. Using block matrix inversion, we can obtain

$$\begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1}=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}=\begin{bmatrix}(\Omega_{11}-\Omega_{12}\Omega_{22}^{-1}\Omega_{21})^{-1} & -\Omega_{11}^{-1}\Omega_{12}(\Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12})^{-1}\\ -\Omega_{22}^{-1}\Omega_{21}(\Omega_{11}-\Omega_{12}\Omega_{22}^{-1}\Omega_{21})^{-1} & (\Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12})^{-1}\end{bmatrix}.$$

We can compute that

$$\begin{aligned}
\Sigma_{11}&=(\Omega_{11}-\Omega_{12}\Omega_{22}^{-1}\Omega_{21})^{-1}=\begin{bmatrix}\frac{(\bar{m}+\bar{\eta}_1)+(\bar{\eta}_2-\bar{\eta}_1)\pi_z}{(\bar{m}+\bar{\eta}_2)(\bar{m}+\bar{\eta}_1)(1-\pi_z)} & 0\\ 0 & \frac{(\bar{m}+\bar{\eta}_1)+(\bar{\eta}_2-\bar{\eta}_1)\pi_z}{(\bar{m}+\bar{\eta}_2)(\bar{m}+\bar{\eta}_1)(1-\pi_z)\sigma_x^2}\end{bmatrix},\\
\Sigma_{22}&=(\Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12})^{-1}=(\Omega_{22}-\pi_z\Omega_{12})^{-1}=\begin{bmatrix}\frac{1}{(\bar{m}+\bar{\eta}_1)\pi_z(1-\pi_z)} & 0\\ 0 & \frac{1}{(\bar{m}+\bar{\eta}_1)\pi_z(1-\pi_z)\sigma_x^2}\end{bmatrix},\\
\Sigma_{12}&=\Sigma_{21}=-\Omega_{11}^{-1}\Omega_{12}(\Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12})^{-1}=\begin{bmatrix}\frac{1}{(\bar{m}+\bar{\eta}_1)(1-\pi_z)} & 0\\ 0 & \frac{1}{(\bar{m}+\bar{\eta}_1)(1-\pi_z)\sigma_x^2}\end{bmatrix}.
\end{aligned}$$

Therefore, we can obtain the expression of Σ or $n\mathrm{Var}(\hat{b})$ as

$$\Sigma=\sigma_y^2\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}D_i^TR_i^{-1}D_i\right)^{-1}=\sigma_y^2(1-\rho)\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}.$$

Next, we show the full details for the approximations accounting for the varying cluster sizes. Define CV to be the coefficient of variation of cluster sizes, then we have $\mathrm{CV}=\sigma_m/\bar{m}$, where σ_m is the standard deviation of the cluster sizes. Following the

power series strategy applied in van Breukelen et al.,[18] we define $d = m_i - \bar{m}$, and it follows that

$$E\left\{\frac{m_i\rho}{1+(m_i-1)\rho}\right\} = E\left\{\frac{\bar{m}+d}{\bar{m}+d+(1-\rho)/\rho}\right\}$$
$$= E\left\{\left(\frac{\bar{m}+d}{\bar{m}+(1-\rho)/\rho}\right)\left(\frac{1}{1+\frac{d}{\bar{m}+(1-\rho)/\rho}}\right)\right\}$$
$$= E\left\{\left(\frac{\bar{m}+d}{\bar{m}+(1-\rho)/\rho}\right)\sum_{q=0}^{\infty}\left(\frac{-d}{\bar{m}+(1-\rho)/\rho}\right)^q\right\}.$$

Expanding the series and discarding all terms d^q with $q > 2$ in the equation above, we have the approximation that

$$E\left\{\frac{m_i\rho}{1+(m_i-1)\rho}\right\} \approx \frac{\bar{m}}{\bar{m}+(1-\rho)/\rho} + \frac{E(d)}{\bar{m}+(1-\rho)/\rho} - \frac{\bar{m}E(d)}{(\bar{m}+(1-\rho)/\rho)^2} - \frac{E(d^2)}{(\bar{m}+(1-\rho)/\rho)^2} + \frac{\bar{m}E(d^2)}{(\bar{m}+(1-\rho)/\rho)^3}.$$

Since it also follows that $E(d) = 0$ and $E(d^2) = \sigma_m^2 = \mathrm{CV}^2\bar{m}^2$, we can write

$$E\left\{\frac{m_i\rho}{1+(m_i-1)\rho}\right\} \approx \frac{\bar{m}}{\bar{m}+(1-\rho)/\rho} - \frac{\mathrm{CV}^2\bar{m}^2}{(\bar{m}+(1-\rho)/\rho)^2} + \frac{\mathrm{CV}^2\bar{m}^3}{(\bar{m}+(1-\rho)/\rho)^3} = \frac{\bar{m}}{\bar{m}+(1-\rho)/\rho}\left\{1-\mathrm{CV}^2\frac{\bar{m}(1-\rho)/\rho}{\{\bar{m}+(1-\rho)/\rho\}^2}\right\}.$$

Therefore, we can use this key approximation to further derive $\bar{\eta}_1$, $\bar{\eta}_2$ and other required expressions in Σ:

$$\bar{\eta}_1 \approx -\frac{\bar{m}}{\bar{m}+(1-\rho)/\rho}\left\{1-\mathrm{CV}^2\frac{\bar{m}(1-\rho)/\rho}{\{\bar{m}+(1-\rho)/\rho\}^2}\right\} = -\frac{\bar{m}\rho}{1+(\bar{m}-1)\rho}\left\{1-\mathrm{CV}^2\frac{\bar{m}\rho(1-\rho)}{\{1+(\bar{m}-1)\rho\}^2}\right\},$$

then, it follows that

$$\bar{m}+\bar{\eta}_1 = \bar{m} - \frac{\bar{m}\rho}{1+(\bar{m}-1)\rho} + \mathrm{CV}^2\frac{\bar{m}^2\rho^2(1-\rho)}{\{1+(\bar{m}-1)\rho\}^3} = \frac{\bar{m}\{1+(\bar{m}-2)\rho\}\{1+(\bar{m}-1)\rho\}^2 + \mathrm{CV}\bar{m}^2\rho^2(1-\rho)}{\{1+(\bar{m}-1)\rho\}^3}.$$

We continue to derive the expression of $\bar{\eta}_2$ by simple mathematical manipulations,

$$\bar{\eta}_2 = E\left\{\frac{-m_i^2\rho}{1+(m_i-1)\rho}\right\} = -\bar{m} + E\left\{\frac{-m_i^2\rho}{1+(m_i-1)\rho} + m_i\right\} = -\bar{m} + E\left\{\frac{m_i(1-\rho)}{1+(m_i-1)\rho}\right\} = -\bar{m} - \frac{1-\rho}{\rho}\bar{\eta}_1,$$

and we can then use the approximation above to write

$$\bar{m}+\bar{\eta}_2 = -\frac{1-\rho}{\rho}\bar{\eta}_1 \approx \frac{\bar{m}(1-\rho)}{1+(\bar{m}-1)\rho}\left\{1-\mathrm{CV}^2\frac{\bar{m}\rho(1-\rho)}{\{1+(\bar{m}-1)\rho\}^2}\right\}.$$

After deriving the covariance matrix corresponding to the reparameterized coefficient parameters and incorporating the considerations of variable cluster size, we can derive the covariance matrix of the actual beta parameters. Specifically, for all types of the proposed tests, we need the variances and covariance of $\hat{\beta}_4$, $\hat{\beta}_2 + \pi_z\hat{\beta}_4$ and $\hat{\beta}_3 + \pi_z\hat{\beta}_4$. Recall that $b_1 = \beta_1 + \pi_x\beta_2$, $b_2 = \beta_2$, $b_3 = \beta_3 + \pi_x\beta_4$, $b_4 = \beta_4$, we can then write

$$n\mathrm{Var}(\hat{\beta}_4) = n\mathrm{Var}(\hat{b}_4) = \frac{\rho_y^2(1-\rho)}{(\bar{m}+\bar{\eta}_1)\pi_z(1-\pi_z)\sigma_x^2} = \frac{\rho_y^2(1-\rho)\{1+(\bar{m}-1)\rho\}^3}{\pi_z(1-\pi_z)\pi_x(1-\pi_x)\bar{m}[\{1+(\bar{m}-2)\rho\}\{1+(\bar{m}-1)\rho\}^2 + \mathrm{CV}^2\bar{m}\rho^2(1-\rho)]},$$

$$n\mathrm{Var}(\hat{\beta}_2 + \pi_z\hat{\beta}_4) = n\{\mathrm{Var}(\hat{b}_2) + \pi_z^2\mathrm{Var}(\hat{b}_4) + 2\pi_z\mathrm{Cov}(\hat{b}_2, \hat{b}_4)\} = \frac{\sigma_y^2(1-\rho)}{(\bar{m}+\bar{\eta}_2)\sigma_x^2} \approx \frac{\sigma_y^2(1+(\bar{m}-1)\rho)}{\pi_x(1-\pi_x)\bar{m}}\left\{1-\mathrm{CV}^2\frac{\bar{m}\rho(1-\rho)}{\{1+(\bar{m}-1)\rho\}^2}\right\}^{-1},$$

$$n\mathrm{Var}(\hat{\beta}_3 + \pi_x\hat{\beta}_4) = n\mathrm{Var}(\hat{b}_3) = \frac{\sigma_y^2(1-\rho)}{(\bar{m}+\bar{\eta}_1)\pi_z(1-\pi_z)} \approx \frac{\sigma_y^2(1-\rho)\{1+(\bar{m}-1)\rho\}^3}{\pi_z(1-\pi_z)\bar{m}[\{1+(\bar{m}-2)\rho\}\{1+(\bar{m}-1)\rho\}^2 + \mathrm{CV}^2\bar{m}\rho^2(1-\rho)]}.$$

Further, we can verify that

$$n\mathrm{Cov}(\hat{\beta}_2 + \pi_z\hat{\beta}_4, \hat{\beta}_3 + \pi_x\hat{\beta}_4) = n\mathrm{Cov}(\hat{b}_2 + \pi_z\hat{b}_4, \hat{b}_3) = n\mathrm{Cov}(\hat{b}_2, \hat{b}_3) + n\mathrm{Cov}(\hat{b}_3, \hat{b}_4)\pi_z = 0.$$

References

1. Collins LM, Dziak JJ, Kugler KC, Trail JB. Factorial experiments: Efficient tools for evaluation of intervention components. *American Journal of Preventive Medicine* 2014; 47(4): 498–504.

2. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. *Medical Research Methodology* 2003; 3(1): 1–5.

3. Dziak JJ, Nahum-Shani I, Collins LM. Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods* 2012; 17(2): 153.

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

5. Donner A, Klar N. *Design and Analysis of Group-Randomized Trials in Health Research*. New York, NY: Oxford University Press . 2000.

6. Mdege ND, Brabyn S, Hewitt C, Richardson R, Torgerson DJ. The 2× 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: A systematic review. *Journal of Clinical Epidemiology* 2014; 67(10): 1083–1092.

7. Landes SJ, Kirchner JE, Areno JP, et al. Adapting and implementing Caring Contacts in a Department of Veterans Affairs emergency department: A pilot study protocol. *Pilot and Feasibility Studies* 2019; 5(1): 115.

8. Stanley B, Brown G, Brent DA, et al. Cognitive-behavioral therapy for suicide prevention (CBT-SP): treatment model, feasibility, and acceptability. *Journal of the American Academy of Child & Adolescent Psychiatry* 2009; 48(10): 1005–1013.

9. Montgomery DC. *Design and Analysis of Experiments*. John Wiley & Sons . 2017.

10. Goulão B, MacLennan G, Ramsay C. The split-plot design was useful for evaluating complex, multilevel interventions, but there is need for improvement in its design and report. *Journal of Clinical Epidemiology* 2018; 96: 120–125.

11. Shin Y, Lafata JE, Cao Y. Statistical power in two-level hierarchical linear models with arbitrary number of factor levels. *Journal of Statistical Planning and Inference* 2018; 194: 106–121.

12. Kahan BC, Tsui M, Jairath V, et al. Reporting of randomized factorial trials was frequently inadequate. *Journal of Clinical Epidemiology* 2020; 117: 52–59.

13. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine* 2007; 26(6): 1181–1192.

14. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal* 2012; 54(5): 716–729.

15. Kerry SM, Martin Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in medicine* 2001; 20(3): 377–390.

16. Manatunga AK, Hudgens MG, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal* 2001; 43(1): 75–86.

17. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 2006; 35(5): 1292–1300.

18. van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine* 2007; 26(13): 2589–2603.

19. Candel MJ, van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Statistics in Medicine* 2010; 29(14): 1488–1501.

20. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of recent methodological developments in group-randomized trials: Part 2—analysis. *American Journal of Public Health* 2017; 107(7): 1078–1086.

21. Jung SH, Ahn CW. Sample size for a two-group comparison of repeated binary measurements using GEE. *Statistics in Medicine* 2005; 24(17): 2583–2596.

22. Yang S, Li F, Starks MA, F. HA, Mentz RJ, Choudhury KR. Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. *Statistics in Medicine* 2020; 00(00): 0000–0000.

23. Li F, Turner EL, Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* 2018; 74(4): 1450–1458.

24. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials.. *Psychological methods* 1997; 2(2): 173.

25. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review* 2003; 27(1): 79–103.

26. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine* 2010; 29(21): 2169–2179.

27. Li D, Cao J, Zhang S. Power analysis for cluster randomized trials with multiple binary co-primary endpoints. *Biometrics* 2020; 76(4): 1064–1074.

28. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 2016; 17(1): 72.

29. Ivers N, Taljaard M, Dixon S, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: Review of random sample of 300 trials, 2000-8. *British Medical Journal* 2011; 343.

30. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *Medical Research Methodology* 2015; 15(1): 1–12.

31. Steel dRG, Torrie JH. *Principles and Procedures of Statistics: A Biometrical Approach*. McGraw-Hill . 1986.

32. Bhasin S, Gill TM, Reuben DB, et al. A randomized trial of a multifactorial strategy to prevent serious fall injuries. *New England Journal of Medicine* 2020; 383(2): 129–140.

33. Coronado GD, Petrik AF, Vollmer WM, et al. Effectiveness of a mailed colorectal cancer screening outreach program in community health clinics: the STOP CRC cluster randomized clinical trial. *JAMA Internal Medicine* 2018; 178(9): 1174–1181.

34. Pinheiro J, Bates D. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media . 2006.

35. Simon GE, Rutter CM, Peterson D, et al. Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death?. *Psychiatric Services* 2013; 64(12): 1195–1202.

36. Heo M, Leon AC. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics* 2008; 64(4): 1256–1262.

37. Teerenstra S, Lu B, Preisser JS, Van Achterberg T, Borm GF. Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics* 2010; 66(4): 1230–1237.

Supplementary materials for "Sample size calculation in hierarchical 2×2 factorial trials with unequal cluster sizes" by Tian et al.

1 | WEB TABLES

WEB TABLE 1 Required number of clusters n_{A1}, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the test for marginal cluster-level treatment effect with and without finite-sample correction. The marginal cluster-level treatment effect size is $\delta_x = 0.2$. Notation: $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			z-test				t-test			
		CV	n_{A1}	ψ	ϕ	$\hat{\phi}$	n_{A1}	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	32	0.06	0.80	0.81	34	0.05	0.80	0.81
		0.3	32	0.06	0.80	0.80	34	0.05	0.81	0.80
		0.6	36	0.06	0.82	0.82	38	0.05	0.83	0.82
		0.9	40	0.06	0.82	0.81	42	0.05	0.83	0.81
	$\rho = 0.05$	0	56	0.06	0.81	0.81	58	0.05	0.81	0.81
		0.3	56	0.05	0.80	0.81	58	0.05	0.81	0.81
		0.6	60	0.05	0.80	0.81	62	0.05	0.82	0.81
		0.9	66	0.06	0.81	0.81	68	0.05	0.81	0.81
	$\rho = 0.10$	0	94	0.05	0.80	0.81	96	0.05	0.81	0.81
		0.3	94	0.05	0.81	0.80	96	0.05	0.80	0.80
		0.6	98	0.05	0.80	0.80	100	0.05	0.81	0.80
		0.9	104	0.05	0.80	0.80	106	0.05	0.80	0.80
$\bar{m} = 100$	$\rho = 0.02$	0	24	0.06	0.80	0.81	26	0.05	0.82	0.81
		0.3	24	0.06	0.81	0.80	26	0.05	0.80	0.80
		0.6	26	0.06	0.80	0.81	28	0.05	0.80	0.81
		0.9	30	0.06	0.81	0.82	32	0.05	0.83	0.82
	$\rho = 0.05$	0	48	0.05	0.81	0.81	50	0.05	0.82	0.81
		0.3	48	0.06	0.80	0.81	50	0.05	0.80	0.81
		0.6	50	0.05	0.81	0.81	52	0.05	0.80	0.81
		0.9	54	0.06	0.81	0.81	56	0.05	0.79	0.81
	$\rho = 0.10$	0	86	0.05	0.80	0.80	88	0.05	0.80	0.80
		0.3	88	0.05	0.81	0.81	90	0.05	0.81	0.81
		0.6	88	0.06	0.80	0.80	90	0.05	0.81	0.80
		0.9	92	0.05	0.79	0.80	94	0.06	0.79	0.80

WEB TABLE 2 Required number of clusters n_{A1}, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the test for marginal cluster-level treatment effect with and without finite-sample correction. The marginal cluster-level treatment effect size is $\delta_x = 0.4$. Notation: $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			z-test				t-test			
		CV	n_{A1}	ψ	ϕ	$\hat{\phi}$	n_{A1}	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	8	0.07	0.80	0.81	12	0.04	0.88	0.88
		0.3	8	0.08	0.78	0.80	12	0.05	0.87	0.87
		0.6	10	0.08	0.84	0.86	12	0.05	0.84	0.85
		0.9	10	0.09	0.79	0.81	14	0.06	0.86	0.87
	$\rho = 0.05$	0	14	0.07	0.82	0.81	16	0.05	0.81	0.81
		0.3	14	0.08	0.79	0.81	16	0.05	0.80	0.80
		0.6	16	0.08	0.83	0.84	18	0.05	0.83	0.83
		0.9	18	0.07	0.84	0.84	20	0.06	0.84	0.84
	$\rho = 0.10$	0	24	0.06	0.80	0.81	26	0.05	0.81	0.81
		0.3	24	0.06	0.81	0.81	26	0.05	0.81	0.81
		0.6	26	0.06	0.82	0.83	28	0.06	0.82	0.83
		0.9	26	0.06	0.79	0.80	28	0.05	0.80	0.80
$\bar{m} = 100$	$\rho = 0.02$	0	6	0.10	0.80	0.81	10	0.05	0.89	0.89
		0.3	6	0.10	0.78	0.80	10	0.05	0.89	0.89
		0.6	8	0.10	0.87	0.88	10	0.05	0.86	0.87
		0.9	8	0.10	0.82	0.84	10	0.06	0.83	0.83
	$\rho = 0.05$	0	12	0.09	0.79	0.81	14	0.05	0.81	0.80
		0.3	12	0.08	0.80	0.81	16	0.05	0.85	0.86
		0.6	14	0.07	0.84	0.85	16	0.05	0.83	0.84
		0.9	14	0.08	0.81	0.83	16	0.05	0.81	0.82
	$\rho = 0.10$	0	22	0.06	0.80	0.81	24	0.05	0.81	0.81
		0.3	22	0.07	0.81	0.81	24	0.05	0.81	0.81
		0.6	22	0.06	0.78	0.80	26	0.04	0.83	0.83
		0.9	24	0.06	0.80	0.82	26	0.05	0.80	0.82

WEB TABLE 3 Required number of clusters n_{A2}, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the test for marginal individual-level treatment effect. Notation: δ_z is the marginal individual-level treatment effect size, $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			$\delta_z = 0.1$				$\delta_z = 0.15$			
		CV	n_{A2}	ψ	ϕ	$\hat{\phi}$	n_{A2}	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	64	0.05	0.82	0.81	28	0.05	0.81	0.81
		0.3	64	0.05	0.82	0.81	28	0.05	0.81	0.81
		0.6	64	0.05	0.80	0.81	28	0.05	0.80	0.81
		0.9	64	0.05	0.81	0.81	28	0.05	0.79	0.81
	$\rho = 0.05$	0	62	0.05	0.81	0.81	28	0.05	0.82	0.82
		0.3	62	0.06	0.82	0.81	28	0.06	0.82	0.82
		0.6	62	0.05	0.80	0.81	28	0.05	0.81	0.82
		0.9	62	0.06	0.79	0.81	28	0.05	0.80	0.82
	$\rho = 0.10$	0	58	0.05	0.80	0.80	26	0.05	0.81	0.81
		0.3	58	0.06	0.80	0.80	26	0.05	0.81	0.81
		0.6	58	0.05	0.80	0.80	26	0.05	0.80	0.81
		0.9	58	0.05	0.80	0.80	26	0.05	0.79	0.81
$\bar{m} = 100$	$\rho = 0.02$	0	32	0.05	0.82	0.81	14	0.05	0.81	0.81
		0.3	32	0.05	0.82	0.81	14	0.05	0.81	0.81
		0.6	32	0.05	0.82	0.81	14	0.05	0.79	0.81
		0.9	32	0.05	0.80	0.81	14	0.05	0.76	0.81
	$\rho = 0.05$	0	32	0.05	0.83	0.82	14	0.05	0.82	0.82
		0.3	32	0.05	0.83	0.82	14	0.05	0.82	0.82
		0.6	32	0.05	0.83	0.82	14	0.05	0.80	0.82
		0.9	32	0.05	0.81	0.82	14	0.05	0.78	0.82
	$\rho = 0.10$	0	30	0.05	0.82	0.82	14	0.05	0.84	0.84
		0.3	30	0.06	0.82	0.82	14	0.05	0.84	0.84
		0.6	30	0.05	0.81	0.82	14	0.05	0.82	0.84
		0.9	30	0.05	0.81	0.82	14	0.05	0.80	0.84

WEB TABLE 4 Required number of clusters n_C, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the joint test with and without finite-sample correction. The marginal cluster-level treatment effect size is $\delta_x = 0.2$ and the marginal individual-level treatment effect size is $\delta_z = 0.1$. Notation: $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			χ^2 test				mixed F-χ^2 test			
		CV	n_C	ψ_0	ϕ_0	$\hat{\phi}$	n_C	ψ_0	ϕ_0	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	26	0.05	0.81	0.81	28	0.05	0.84	0.82
		0.3	26	0.06	0.81	0.80	28	0.05	0.82	0.81
		0.6	28	0.07	0.81	0.81	30	0.05	0.82	0.83
		0.9	30	0.06	0.82	0.81	32	0.05	0.81	0.82
	$\rho = 0.05$	0	36	0.06	0.81	0.81	38	0.05	0.82	0.82
		0.3	36	0.06	0.81	0.81	38	0.05	0.81	0.81
		0.6	38	0.06	0.82	0.82	38	0.05	0.81	0.80
		0.9	40	0.05	0.83	0.82	40	0.05	0.82	0.80
	$\rho = 0.10$	0	44	0.05	0.81	0.80	46	0.05	0.80	0.81
		0.3	44	0.05	0.81	0.80	46	0.04	0.81	0.81
		0.6	46	0.05	0.82	0.82	46	0.04	0.80	0.80
		0.9	46	0.06	0.80	0.81	48	0.04	0.80	0.81
$\bar{m} = 100$	$\rho = 0.02$	0	18	0.06	0.84	0.84	18	0.05	0.80	0.80
		0.3	18	0.06	0.83	0.83	18	0.05	0.79	0.80
		0.6	18	0.06	0.83	0.82	20	0.05	0.82	0.83
		0.9	20	0.07	0.84	0.84	20	0.06	0.80	0.81
	$\rho = 0.05$	0	24	0.06	0.83	0.83	24	0.05	0.81	0.81
		0.3	24	0.06	0.82	0.83	24	0.05	0.80	0.80
		0.6	24	0.06	0.82	0.82	24	0.05	0.80	0.80
		0.9	24	0.06	0.79	0.81	26	0.05	0.81	0.82
	$\rho = 0.10$	0	28	0.05	0.83	0.83	28	0.05	0.81	0.81
		0.3	28	0.06	0.83	0.83	28	0.05	0.80	0.81
		0.6	28	0.06	0.83	0.82	28	0.05	0.81	0.81
		0.9	28	0.06	0.80	0.82	28	0.05	0.78	0.81

WEB TABLE 5 Required number of clusters n_C, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the joint test with and without finite-sample correction. The marginal cluster-level treatment effect size is $\delta_x = 0.25$ and the marginal individual-level treatment effect size is $\delta_z = 0.15$. Notation: $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			χ^2 test				mixed F-χ^2 test			
		CV	n_C	ψ	ϕ	$\hat{\phi}$	n_C	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	16	0.06	0.85	0.85	16	0.05	0.81	0.81
		0.3	16	0.06	0.85	0.84	16	0.04	0.81	0.81
		0.6	16	0.08	0.82	0.83	18	0.05	0.83	0.85
		0.9	18	0.06	0.84	0.85	18	0.04	0.80	0.81
	$\rho = 0.05$	0	20	0.06	0.84	0.83	20	0.05	0.80	0.80
		0.3	20	0.06	0.83	0.83	22	0.05	0.85	0.84
		0.6	20	0.07	0.81	0.82	22	0.05	0.83	0.83
		0.9	20	0.07	0.79	0.80	22	0.05	0.80	0.81
	$\rho = 0.10$	0	22	0.06	0.81	0.80	24	0.05	0.81	0.81
		0.3	22	0.06	0.81	0.80	24	0.05	0.81	0.82
		0.6	24	0.06	0.83	0.83	24	0.05	0.81	0.81
		0.9	24	0.06	0.81	0.82	24	0.05	0.79	0.80
$\bar{m} = 100$	$\rho = 0.02$	0	10	0.08	0.86	0.85	12	0.05	0.88	0.87
		0.3	10	0.08	0.85	0.85	12	0.05	0.87	0.88
		0.6	10	0.08	0.83	0.84	12	0.05	0.86	0.86
		0.9	10	0.09	0.80	0.82	12	0.05	0.82	0.85
	$\rho = 0.05$	0	12	0.08	0.83	0.82	14	0.05	0.84	0.84
		0.3	12	0.07	0.83	0.82	14	0.04	0.84	0.84
		0.6	12	0.07	0.82	0.82	14	0.04	0.82	0.83
		0.9	12	0.07	0.78	0.81	14	0.05	0.79	0.83
	$\rho = 0.10$	0	14	0.07	0.85	0.84	16	0.04	0.86	0.86
		0.3	14	0.06	0.85	0.84	16	0.05	0.86	0.86
		0.6	14	0.07	0.83	0.84	16	0.05	0.84	0.86
		0.9	14	0.07	0.81	0.84	16	0.04	0.82	0.85

WEB TABLE 6 Required number of clusters n_D, empirical type I error ψ, empirical power ϕ, and predicted power $\hat{\phi}$ corresponding to the intersection-union test with and without finite-sample correction. The marginal cluster-level treatment effect size is $\delta_x = 0.2$ and the marginal individual-level treatment effect size is $\delta_z = 0.1$. Notation: $\bar{m}$ is the mean cluster size, ρ is the ICC, CV is the coefficient of variation of cluster sizes. The results were based on 5,000 simulations.

			z-based I-U test				t- and z-based I-U test			
		CV	n_D	ψ	ϕ	$\hat{\phi}$	n_D	ψ	ϕ	$\hat{\phi}$
$\bar{m} = 50$	$\rho = 0.02$	0	66	0.04	0.81	0.81	66	0.04	0.81	0.81
		0.3	66	0.04	0.81	0.81	66	0.04	0.81	0.81
		0.6	66	0.05	0.81	0.80	68	0.05	0.81	0.81
		0.9	68	0.05	0.79	0.80	70	0.04	0.81	0.81
	$\rho = 0.05$	0	76	0.05	0.81	0.80	78	0.04	0.81	0.81
		0.3	78	0.05	0.82	0.81	78	0.04	0.81	0.81
		0.6	80	0.05	0.81	0.81	80	0.04	0.80	0.81
		0.9	84	0.05	0.81	0.81	84	0.05	0.81	0.81
	$\rho = 0.10$	0	102	0.05	0.80	0.80	104	0.05	0.81	0.81
		0.3	102	0.05	0.81	0.80	104	0.05	0.82	0.80
		0.6	106	0.05	0.79	0.81	108	0.05	0.81	0.81
		0.9	110	0.05	0.80	0.80	112	0.05	0.79	0.80
$\bar{m} = 100$	$\rho = 0.02$	0	36	0.05	0.80	0.80	38	0.04	0.83	0.82
		0.3	38	0.05	0.82	0.82	38	0.05	0.81	0.81
		0.6	38	0.05	0.81	0.81	40	0.05	0.83	0.83
		0.9	40	0.05	0.80	0.81	42	0.05	0.82	0.83
	$\rho = 0.05$	0	52	0.05	0.80	0.80	54	0.05	0.81	0.81
		0.3	52	0.05	0.80	0.80	54	0.05	0.80	0.81
		0.6	54	0.05	0.80	0.81	56	0.05	0.81	0.81
		0.9	58	0.06	0.80	0.82	58	0.05	0.79	0.80
	$\rho = 0.10$	0	86	0.05	0.80	0.80	88	0.05	0.80	0.80
		0.3	88	0.05	0.81	0.81	90	0.05	0.81	0.81
		0.6	90	0.05	0.82	0.81	92	0.05	0.81	0.81
		0.9	92	0.05	0.79	0.80	94	0.06	0.79	0.80