HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2023 March 25.

Published in final edited form as:

Biometrics. 2023 March ; 79(1): 98–112. doi:10.1111/biom.13596.

Sample size considerations for stepped wedge designs with subclusters

Kendra Davis-Plourde[1,2,3], **Monica Taljaard**[4,5], **Fan Li**[1,3]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA

[2]Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA

[3]Center for Methods in Implementation and Prevention Science, Yale University, New Haven, Connecticut, USA

[4]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[5]School of Epidemiology and Public Heath, University of Ottawa, Ottawa, Ontario, Canada

Abstract

The stepped wedge cluster randomized trial (SW-CRT) is an increasingly popular design for evaluating health service delivery or policy interventions. An essential consideration of this design is the need to account for both within-period and between-period correlations in sample size calculations. Especially when embedded in health care delivery systems, many SW-CRTs may have subclusters nested in clusters, within which outcomes are collected longitudinally. However, existing sample size methods that account for between-period correlations have not allowed for multiple levels of clustering. We present computationally efficient sample size procedures that properly differentiate within-period and between-period intracluster correlation coefficients in SW-CRTs in the presence of subclusters. We introduce an extended block exchangeable correlation matrix to characterize the complex dependencies of outcomes within clusters. For Gaussian outcomes, we derive a closed-form sample size expression that depends on the correlation structure only through two eigenvalues of the extended block exchangeable correlation structure. For non-Gaussian outcomes, we present a generic sample size algorithm based on linearization and elucidate simplifications under canonical link functions. For example, we show that the approximate sample size formula under a logistic linear mixed model depends on three eigenvalues of the extended block exchangeable correlation matrix. We provide an extension to accommodate unequal cluster sizes and validate the proposed methods via simulations. Finally, we illustrate our methods in two real SW-CRTs with subclusters.

Correspondence Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA. kendra.plourde@yale.edu.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://github.com/kldavisplourde/multilevelSWCRT.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 3–6 are available with this paper at the *Biometrics* website on Wiley Online Library. R code for predicting power and for conducting the simulation and application studies described in Sections 2–5 are available as supporting material and at https://github.com/kldavisplourde/multilevelSWCRT.

Keywords

cluster randomized trial; eigenvalues; extended block exchangeable correlation structure; generalized linear mixed models; power analysis

1 | INTRODUCTION

Stepped wedge cluster randomized trials (SW-CRTs) are an increasingly popular trial design for a variety of reasons including their potential benefits of increased power and statistical efficiency (Hemming and Taljaard, 2020). In an SW-CRT, one or more randomly allocated clusters are scheduled to transition from control to intervention at prespecified time points until all clusters are exposed to the intervention. Figure 1 provides a schematic of a typical SW-CRT in which eight clusters are allocated to four sequences and outcomes are observed within each of five time periods. SW-CRTs can employ either a cross-sectional or closed-cohort design, depending on whether repeated measurements are taken from the same or different individuals in each period.

A comprehensive methodological review of available statistical methods for SW-CRTs is presented in Li *et al.* (2021). While almost all current methods for designing SW-CRTs have assumed a two-level structure within periods (e.g., patients nested in practices), several recent trials have utilized data from patients nested within subclusters, giving rise to three-level structures within periods. For example, the Lumbar Imaging with Reporting of Epidemiology (LIRE) trial was an SW-CRT that evaluated the impact of inserting benchmark prevalence data in spinal imaging reports on subsequent health care utilization, and cross-sectionally sampled patients nested in primary care providers, who are nested in practices (Jarvik *et al.*, 2015). As another example, the Washington State Expedited Partner Therapy (EPT) SW-CRT randomized local health jurisdictions (LHJ) consisting of clinics and cross-sectionally measured chlamydia infection in patients (Golden *et al.*, 2015). In this paper, we present new methods for designing SW-CRTs that explicitly account for the three-level structure within each period while additionally accounting for the between-period correlation parameters.

For the standard parallel-arm CRT design (without repeated measures), Heo and Leon (2008) and Teerenstra *et al.* (2010) developed the design effect (DE) for three-level data with a Gaussian outcome. The DE is expressed as $1 + (N-1)\alpha_0 + N(K-1)\rho_0$, where α_0 and ρ_0 are the within-subcluster and between-subcluster intracluster correlation coefficients (ICCs), K and N are the number of subclusters and subcluster size, respectively. This DE represents the amount by which the sample size required for an individually randomized trial needs to be multiplied to obtain the sample size required for a three-level CRT and coincides with the leading eigenvalue of the *nested exchangeable* correlation structure (Li *et al.*, 2019). The same DE expression also applies to three-level CRTs with a binary outcome (Teerenstra *et al.*, 2010; Liu and Colditz, 2020). Frequently, it is assumed that the between-subcluster ICC does not exceed the within-subcluster ICC (Teerenstra *et al.*, 2010), which suggests, based on the DE, that $\rho_0 = \alpha_0$ would lead to conservative sample size estimates under parallel randomization.

The design and analysis of SW-CRTs have been conventionally based on linear mixed models (Li *et al.*, 2021). Assuming a Gaussian outcome, Hemming *et al.* (2015) proposed a generic framework for designing cross-sectional SW-CRTs and extended the Hussey and Hughes (2007) linear mixed model by including an additional random intercept at the subcluster level. The variance of the intervention effect estimator was obtained using the covariance matrix of the generalized least squares estimator, but no analytical formulas were provided. Teerenstra *et al.* (2019) extended the Hemming *et al.* (2015) approach to accommodate closed-cohort SW-CRTs and developed closed-form DE expressions. However, these approaches implicitly assume that the between-period ICC equals the within-period ICC, both within and between subclusters. Under this strong assumption, we show in Section 3.1 that the variance of the intervention effect vanishes as the subcluster size grows indefinitely, which may lead to an underestimated sample size when the between-period ICCs differ from the within-period ICCs. To the best of our knowledge, explicit sample size formulas that differentiate between-period and within-period ICCs in SW-CRTs with multiple levels of clustering have not been previously derived. Furthermore, with a binary outcome, Teerenstra *et al.* (2019) considered an approximation based on a linear mixed model, which in the case of a single level of clustering has already been shown to be inaccurate (Zhou *et al.*, 2020). To this end, it is necessary to develop more accurate sample size procedures that acknowledge the mean-variance relationship with a non-Gaussian outcome for multilevel SW-CRTs.

To address these issues, we consider a generalized linear mixed model (GLMM) that characterizes five possible sources of random variation in SW-CRTs with subclusters, while differentiating within-period and between-period ICCs. With a Gaussian outcome, we describe an *extended block exchangeable* correlation structure parameterized by at most five ICC parameters, depending on whether a cross-sectional or closed-cohort design is assumed at each level. We show in Section 3 that the variance of the intervention effect estimator depends on the ICCs only through two distinct eigenvalues of the extended block exchangeable correlation matrix. Although a scalar closed-form variance expression is unavailable for other exponential family outcomes, we propose a computationally efficient generic sample size algorithm based on cluster period averages. Under the canonical link functions, new strategies are provided to circumvent complex numerical integration. In addition, we also extend our approach to accommodate unequal cluster sizes. Our simulation studies to validate the proposed sample size methodology are presented in Section 4 and applications to two SW-CRTs with subclusters are presented in Section 5. Section 6 provides concluding remarks.

2 | MODELS AND GENERIC SAMPLE SIZE CONSIDERATIONS

2.1 | Three design variants

We consider a multilevel SW-CRT with I clusters and T periods, with subclusters nested in clusters and subjects nested in subclusters. We consider three possible variations of this multilevel stepped wedge design (Figure 2): (A) a closed-cohort design at both the subcluster and subject levels, in which case repeated measurements are taken on the same subjects in each subcluster over time; (B) a closed-cohort design on the subcluster level

but a cross-sectional design at the subject level, in which case different subjects in the same subcluster are sampled at each time period; (C) a cross-sectional design at both the subcluster and subject level, in which case different subjects within different subclusters are sampled in each cluster during each period. Our development will include all three variants.

2.2 | Statistical model

We consider a GLMM to represent the average secular trend, intervention effect, as well as the three-level structure within each period. Let Y_{ijkl} be the outcome of interest for individual $l = 1, \ldots, N_{ijk}$ nested in subcluster $k = 1, \ldots, K_{ij}$ nested in cluster $i = 1, \ldots, I$ during period $j = 1, \ldots, T$. For generality, we assume design (A), where a closed-cohort of subjects in the same set of subclusters are measured in the study during each period. Designs (B) and (C) will be cast as two special cases. For design (A), the conditional mean model for $\mu_{ijkl} = \mathbb{E}(Y_{ijkl} \mid b_i, c_{ik}, s_{ij}, \pi_{ijk}, \gamma_{ikl})$ is given by

$$g(\mu_{ijkl}) = \beta_j + \delta X_{ij} + b_i + c_{ik} + s_{ij} + \pi_{ijk} + \gamma_{ikl}, \quad (1)$$

where g is a link function, β_j represents the categorical secular trend, X_{ij} is the intervention status for cluster i at period j (equal to 1 if exposed under intervention and 0 otherwise), and δ is the intervention effect of interest on the link function scale. We also write $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_T, \delta)^\top$. Model (1) includes five random effects to reflect the multilevel structure of the data: $b_i \sim \mathcal{N}(0, \sigma_b^2)$ is the random cluster effect, $c_{ik} \sim \mathcal{N}(0, \sigma_c^2)$ is the random subcluster effect, $s_{ij} \sim \mathcal{N}(0, \sigma_s^2)$ is the random cluster-by-period interaction, and $\pi_{ijk} \sim \mathcal{N}(0, \sigma_\pi^2)$ is the random subcluster-by-period interaction. Furthermore, since design (A) involves a closed-cohort of subjects, $\gamma_{ikl} \sim \mathcal{N}(0, \sigma_\gamma^2)$ is the random subject-level effect arising from the repeated outcome measurements for each same subject. In particular, the random interactions, s_{ij} and $\boldsymbol{\pi}_{ijk}$, can represent variations within each cluster and subcluster due to time-varying characteristics of the cluster and each subcluster. The inclusion of these random interactions further allows the within-period ICCs to differ from the between-period ICCs at each level of clustering, which is considered critical for accurate power calculation even in SW-CRTs without subclusters (Taljaard *et al.*, 2016). Following convention in modeling for SW-CRTs (Li *et al.*, 2021), we assume all random effects are mutually independent and Y_{ijkl} is a random realization from a parametric distribution with mean μ_{ijkl} and higher order moments as potential functions of μ_{ijkl}. Model (1) includes several existing models for SW-CRTs as special cases. For example, when $\sigma_s^2 = \sigma_\pi^2 = \sigma_\gamma^2 = 0$, model (1) coincides with the model in Hemming *et al.* (2015) for cross-sectional SW-CRTs; when $\sigma_s^2 = \sigma_\pi^2 = 0$, model (1) becomes the model in Teerenstra *et al.* (2019) without cluster- or subcluster-by-time interactions; when $\sigma_c^2 = \sigma_\pi^2 = 0$, model (1) reduces to the model in Hooper *et al.* (2016) for closed-cohort SW-CRTs without subclusters. Finally, models for design (B) or (C) can be obtained by setting $\sigma_\gamma^2 = 0$ or $\sigma_\gamma^2 = \sigma_c^2 = 0$ in model (1), due to the absence of repeated assessments for the same subjects or the same subclusters.

2.3 | Generic sample size requirement

In the design phase, the power to detect an effect size $\delta \neq 0$ with a two-sided α-level Wald test is approximately (Harrison and Brady, 2004)

$$\text{power} \approx 1 - \Phi_t\left(t_{\alpha/2,\text{DoF}}; \text{DoF}, |\delta|/\sqrt{\text{var}(\hat{\delta})}\right), \quad (2)$$

where $t_{\alpha/2,\text{DoF}}$ is the upper $\alpha/2$th quantile of the central t-distribution with specified degrees of freedom (DoF) and $\Phi_t(t; \text{DoF}, \Lambda)$ is the cumulative t-distribution function with DoF and noncentrality parameter Λ. The variance of the intervention effect, $\text{var}(\hat{\delta})$, is an implicit function of the number of clusters (I) and other design parameters. While power expression (2) is asymptotically equivalent to its counterpart with a standard normal distribution as the DoF approach infinity, we consider the t-distribution with DoF = $I - 2$ since it has been found to maintain a valid empirical type I error rate with a limited number of clusters in SW-CRTs (Ford and Westgate, 2020; Li *et al.*, 2021). By Equation (2), the sample size requirement for the multilevel SW-CRT requires us to characterize an expression of $\text{var}(\hat{\delta})$ at the design phase. To do so, we first assume equal cluster and subcluster sizes such that $K_{ij} = K$ and $N_{ijk} = N$. We relax this assumption in Section 3.3.

3 | VARIANCE OF THE INTERVENTION EFFECT ESTIMATOR

3.1 | Gaussian outcomes

With a Gaussian outcome, we consider g in model (1) to be the identity function, and assume $Y_{ijkl} = \mu_{ijkl} + \epsilon_{ijkl}$, where $\epsilon_{ijkl} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is an independent residual error. In this linear mixed model, we define $\sigma^2 = \sigma_b^2 + \sigma_c^2 + \sigma_s^2 + \sigma_\pi^2 + \sigma_\gamma^2 + \sigma_\epsilon^2$ as the total variance. Under design (A), this linear mixed model induces five different ICCs: (i) the within-period within-subcluster ICC, α_0; (ii) the between-period within-subcluster ICC, α_1; (iii) the within-period between-subcluster ICC, ρ_0; (iv) the between-period between-subcluster ICC, ρ_1; and (v) the within-subject autocorrelation, α_2. We explicitly define the ICCs as functions of variance components under each design variant in Table 1.

As the variance components are nonnegative, we implicitly have $\alpha_0 \geq \alpha_1 \geq \rho_1$, $\alpha_0 \geq \rho_0 \geq \rho_1$, and $\alpha_2 \geq \alpha_1 \geq \rho_1$ without further restrictions. To aid in the interpretation of ICCs, an investigator could also parameterize the between-period ICCs based on the within-period ICCs by assuming a particular cluster autocorrelation coefficient (CAC), defined as the ratio of between-period to within-period ICCs (Hooper *et al.*, 2016; Martin *et al.*, 2016). Because ICCs are conventionally used in designing CRTs and intuitive to understand, we will characterize the variance of the intervention effect with these five ICC parameters and obtain the variances for design (B) or (C) by setting $\alpha_2 = \alpha_1$, or $\alpha_2 = \alpha_1 = \rho_1$. In addition, the variance components or ICCs should ensure a positive definite correlation structure for all outcomes within each cluster. Specifically, if we define $\boldsymbol{Y}_i = \left(\boldsymbol{Y}_{i1}^\top, \ldots \boldsymbol{Y}_{iT}^\top\right)^\top$, where $\boldsymbol{Y}_{ij} = \left(\boldsymbol{Y}_{ij1}^\top, \ldots, \boldsymbol{Y}_{ijK}^\top\right)^\top$, and $\boldsymbol{Y}_{ijk} = (Y_{ijk1}, \ldots, Y_{ijkN})^\top$, then the induced correlation structure, $\boldsymbol{R}_i = \text{corr}(\boldsymbol{Y}_i)$, follows an *extended block exchangeable* structure and can be written as a linear combination of six basis matrices,

$$\begin{aligned} \boldsymbol{R}_i = {} & (1 - \alpha_0 - \alpha_2 + \alpha_1)\boldsymbol{I}_{TKN} + (\alpha_0 - \rho_0 - \alpha_1 + \rho_1)\boldsymbol{I}_{TK} \\ & \otimes \boldsymbol{J}_N + (\rho_0 - \rho_1)\boldsymbol{I}_T \otimes \boldsymbol{J}_{KN} \\ & + (\alpha_2 - \alpha_1)\boldsymbol{J}_T \otimes \boldsymbol{I}_{KN} + (\alpha_1 - \rho_1)\boldsymbol{J}_T \\ & \otimes \boldsymbol{I}_K \otimes \boldsymbol{J}_N + \rho_1 \boldsymbol{J}_{TKN}, \end{aligned} \quad (3)$$

where $\boldsymbol{I}_u$ is a $u \times u$ identity matrix and $\boldsymbol{J}_u = \mathbf{1}_u\mathbf{1}_u^\top$ is a $u \times u$ matrix of ones. Evidently, the diagonal and off-diagonal $KN \times KN$ blocks of $\boldsymbol{R}_i$ are block exchangeable given by $\boldsymbol{I}_K \otimes \{(1-\alpha_0)\boldsymbol{I}_N + (\alpha_0 - \rho_0)\boldsymbol{J}_N\} + \rho_0\boldsymbol{J}_{KN}$ and $\boldsymbol{I}_K \otimes \{(\alpha_2 - \alpha_1)\boldsymbol{I}_N + (\alpha_1 - \rho_1)\boldsymbol{J}_N\} + \rho_1\boldsymbol{J}_{KN}$, respectively, and, therefore, the structure (3) resembles the block exchangeable structure studied in Li *et al.* (2018). For deriving the variance of the intervention effect, we first provide the two key intermediate results on the induced correlation structure.

Lemma 1. *The induced extended block exchangeable correlation matrix has at most six unique eigenvalues, which can be expressed as linear functions of the five ICC parameters:*

$$\begin{aligned}
\lambda_1 &= 1 - \alpha_0 - \alpha_2 + \alpha_1 \\
\lambda_2 &= 1 - \alpha_0 - \alpha_2 + \alpha_1 + N(\alpha_0 - \alpha_1 - \rho_0 + \rho_1) \\
\lambda_3 &= 1 - \alpha_0 - \alpha_2 + \alpha_1 + N\{\alpha_0 - \alpha_1 + (K-1)(\rho_0 - \rho_1)\} \\
\lambda_4 &= 1 - \alpha_0 + (T-1)(\alpha_2 - \alpha_1) \\
\lambda_5 &= 1 - \alpha_0 + (T-1)(\alpha_2 - \alpha_1) \\
&\quad + N\{\alpha_0 - \rho_0 + (T-1)(\alpha_1 - \rho_1)\} \\
\lambda_6 &= 1 - \alpha_0 + (T-1)(\alpha_2 - \alpha_1) \\
&\quad + N[\alpha_0 + (T-1)\alpha_1 + (K-1)\{\rho_0 + (T-1)\rho_1\}]
\end{aligned} \tag{4}$$

with algebraic multiplicities $(T-1)K(N-1)$, $(T-1)(K-1)$, $T-1$, $K(N-1)$, $K-1$, *and 1, respectively. The set of ICC parameters* $\{\alpha_0, \alpha_1, \alpha_2, \rho_0, \rho_1\}$ *for which* $\mathbf{R}_i$ *is positive definite corresponds to the convex open subset characterized by* $\min\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\} > 0$.

Lemma 2. *The extended block exchangeable correlation matrix has a closed-form inverse, which shares the same set of basis matrices with* $\mathbf{R}_i$*, with coefficients determined by the set of eigenvalues. The inverse is represented as*

$$\begin{aligned}
\boldsymbol{R}_i^{-1} &= \frac{1}{\lambda_1}\boldsymbol{I}_{TKN} - \frac{\lambda_2 - \lambda_1}{N\lambda_1\lambda_2}\boldsymbol{I}_{TK} \otimes \boldsymbol{J}_N + \frac{\lambda_2 - \lambda_3}{KN\lambda_2\lambda_3}\boldsymbol{I}_T \otimes \boldsymbol{J}_{KN} \\
&+ \frac{1}{T}\left(\frac{1}{\lambda_4} - \frac{1}{\lambda_1}\right)\boldsymbol{J}_T \otimes \boldsymbol{I}_{KN} \\
&+ \frac{1}{T}\left(\frac{\lambda_2 - \lambda_1}{N\lambda_1\lambda_2} - \frac{\lambda_5 - \lambda_4}{N\lambda_4\lambda_5}\right)\boldsymbol{J}_T \otimes \boldsymbol{I}_K \otimes \boldsymbol{J}_N \\
&+ \frac{1}{TK}\left(\frac{\lambda_5 - \lambda_6}{N\lambda_5\lambda_6} - \frac{\lambda_2 - \lambda_3}{N\lambda_2\lambda_3}\right)\boldsymbol{J}_{TKN}.
\end{aligned}$$

Lemmas 1 and 2 derive the eigenvalues as well as an explicit inverse of the extended block exchangeable correlation structure (derivation details and eigenvalue expressions under each design variant are provided in Web Appendix A and Web Table 1). By parameterizing $\boldsymbol{R}_i^{-1}$ as a function of the unique eigenvalues, the cumbersome expressions on individual ICCs are avoided in deriving an explicit variance expression of the intervention effect.

Assuming the variance components are known, the feasible generalized least squares estimator for $\boldsymbol{\theta}$ is given by $\hat{\theta} = \left(\sum_{i=1}^{I} \boldsymbol{Z}_i^\top \boldsymbol{R}_i^{-1} \boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{I} \boldsymbol{Z}_i^\top \boldsymbol{R}_i^{-1} \boldsymbol{Y}_i\right)$, where $\boldsymbol{Z}_i = (\boldsymbol{I}_T, \boldsymbol{X}_i) \otimes \mathbf{1}_{KN}$ is the design matrix for fixed effects and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iT})^\top$ is the vector of cluster-level intervention indicators. As the number of clusters becomes large, $\hat{\theta}$ follows a multivariate normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\sigma^2\left(\sum_{i=1}^{I} \boldsymbol{Z}_i^\top \boldsymbol{R}_i^{-1} \boldsymbol{Z}_i\right)^{-1}$, whose $(T+1, T+1)$th element corresponds to the variance of the intervention effect estimator. Based on Lemma 2, we show in Web Appendix A that a closed-form expression of $\mathrm{var}(\hat{\delta})$ exists, which

reveals the impact of various ICC parameters on sample size and power. We summarize the results in Theorem 1.

Theorem 1. *Assuming known variance components, the closed-form variance of the intervention effect estimator based on a linear mixed model with a Gaussian outcome is*

$$var(\hat{\delta}) = \frac{\sigma^2}{IKNtr(\mathbf{\Omega})} \times \frac{T\lambda_6\lambda_3}{T\lambda_6 - \{1 + (T-1)\tau_X\}(\lambda_6 - \lambda_3)}, \tag{5}$$

where

$$\mathbf{\Omega} = I^{-1}\sum_{i=1}^{I} \mathbf{X}_i\mathbf{X}_i^{\top} - \left(I^{-1}\sum_{i=1}^{I}\mathbf{X}_i\right)\left(I^{-1}\sum_{i=1}^{I}\mathbf{X}_i^{\top}\right)$$

is the covariance matrix of the intervention vector under a specific design and $\tau_X = \{(T-1)tr(\mathbf{\Omega})\}^{-1}\{\mathbf{1}_T^{\top}\mathbf{\Omega}\mathbf{1}_T - tr(\mathbf{\Omega})\} \in [-1, 1]$ *is the generalized ICC of the intervention, which is the ratio of average covariance over the average variance and measures the similarity between the intervention status for each cluster in different periods* (Kistner and Muller, 2004). *With all other design parameters fixed, larger values of the within-period ICCs,* $\{\alpha_0, \rho_0\}$*, are always associated with larger required sample size, whereas larger values of the between-period ICCs,* $\{\alpha_1, \rho_1, \alpha_2\}$*, are associated with smaller required sample size when* $\tau_X < (\lambda_6^2 - \lambda_3^2)/\{\lambda_6^2 + (T-1)\lambda_3^2\}$.

Theorem 1 reveals that the variance of the intervention effect in a linear mixed model is free of the secular trend, as long as it is adjusted for in the model. The variance (5) also depends on the five ICCs only through two eigenvalues of the extended block exchangeable correlation matrix, λ_3 and λ_6. Our results also confirm the different roles of the within-period ICCs and the between-period ICCs. Specifically in the design phase, assuming larger values of α_0 and ρ_0 will only lead to a conservative sample size estimate. Likewise, assuming smaller values of between-period ICCs or even ignoring them by considering $\alpha_2 = \alpha_1 = \rho_1 = 0$ will lead to a conservative sample size estimate if the constraint on τ_X holds. These directional results can guide decisions on design parameters to avoid an underpowered trial in the absence of accurate ICC estimates for an SW-CRT with subclusters. Finally, while we assume design (A), variance expressions for design (B) or (C) can be easily obtained by enforcing $\alpha_2 = \alpha_1$, or $\alpha_2 = \alpha_1 = \rho_1$ in computing the two eigenvalues.

As expected from the generality of model (1), our variance expression (5) includes a number of variances previously derived as special cases. First, we observe that the variance (5) can be alternatively represented by

$$\text{var}(\hat{\delta}) = \frac{(\sigma^2/KN)IT\lambda_6\lambda_3}{(U^2 + ITU - TW - IV)\lambda_6 - (U^2 - IV)\lambda_3}, \tag{6}$$

where

$$U = \sum_{i=1}^{I} \sum_{j=1}^{T} X_{ij}, V = \sum_{i=1}^{I} \left(\sum_{j=1}^{T} X_{ij} \right)^2,$$

and

$$W = \sum_{j=1}^{T} \left(\sum_{i=1}^{I} X_{ij} \right)^2$$

are typical design constants that only depend on the sequence of treatment indicators for each cluster. The connection between the two variance expressions is made clear through the observance that $\mathbf{1}_T^\top \boldsymbol{\Omega} \mathbf{1}_T = I^{-2}(IV - U^2)$ and $t(\boldsymbol{\Omega}) = I^{-2}(IU - W)$. Using this equivalent variance expression (6), we see that in the absence of subclusters, the variance derived in Li *et al.* (2018) under a closed-cohort SW-CRT is obtained by setting $\rho_0 = \alpha_0$ and $\rho_1 = \alpha_1$; the variance of the Hooper *et al.* (2016) linear mixed model under a cross-sectional SW-CRT can be obtained by additionally requiring $\alpha_2 = \alpha_1$; whereas the basic Hussey and Hughes (2007) linear mixed model is obtained by equating all five ICCs. In the presence of subclusters, our variance expression also includes that derived in Teerenstra *et al.* (2019) as a special case when we assume $\rho_1 = \rho_0$ and $\alpha_1 = \alpha_0$. However, we caution against this simplification assumption, because based on (6), we observe that $\lim_{N \to \infty}(\lambda_3/N) = (\alpha_0 - \alpha_1) + (K-1)(\rho_0 - \rho_1)$, and $\lim_{N \to \infty}(\lambda_6/N) = \alpha_0 + (T-1)\alpha_1 + (K-1)\{\rho_0 + (T-1)\rho_1\} = \kappa(\alpha_0, \alpha_1, \rho_0, \rho_1)$, which leads to a limiting variance

$$\lim_{N \to \infty} \text{var}(\hat{\delta}) = \frac{(\sigma^2/K)\{(\alpha_0 - \alpha_1) + (K-1)(\rho_0 - \rho_1)\} I \kappa(\alpha_0, \alpha_1, \rho_0, \rho_1)}{(IU - W)\kappa(\alpha_0, \alpha_1, \rho_0, \rho_1) + (U^2 - IV)\{\alpha_1 + (K-1)\rho_1\}}.$$

Therefore, assuming the between-period ICCs are equal to the within-period ICCs, the limiting variance approaches zero and the required number of clusters approaches one as the subcluster size increases indefinitely, similar to the discussion in Taljaard *et al.* (2016) without subclusters. From this perspective, it is reasonable to differentiate the between-period ICCs from the within-period ICCs and avoid the risks associated with too few clusters in the design phase. Furthermore, based on the proposed analytical variance (6), we show in Web Appendix B that the most efficient SW-CRT with subclusters allocates more clusters to the intervention during the second and last period.

Finally, while we have thus far focused on stepped wedge designs with a staggered allocation of intervention to clusters, the derivations of Theorem 1 do not involve restrictions on $\boldsymbol{X}_i$, and therefore expression (5) applies more generally to any multiperiod CRT, including the longitudinal parallel-arm and repeated crossover designs. Clearly, variance (5) reveals that higher efficiency is achieved with a larger total variance of the intervention status, $tt(\boldsymbol{\Omega})$, and a smaller generalized ICC of the intervention, τ_X, both of which a specific design will implicitly characterize. For example, with $T = 5$ periods and I as a multiple of $(T-1)$, a standard stepped wedge design corresponds to $t(\boldsymbol{\Omega}) = 0.625$ and $\tau_X = 0.25$, a parallel longitudinal design corresponds to $tt(\boldsymbol{\Omega}) = 1.25$ and $\tau_X = 1$, whereas a repeated

crossover design engenders $tr(\boldsymbol{\Omega}) = 1.25$ and $\tau_X = -0.2$. In Web Table 2, we show that a repeated crossover design and a parallel longitudinal design have the same values of $t(\boldsymbol{\Omega})$, which is generally larger than that under a standard stepped wedge design. Furthermore, the generalized ICC of the intervention indicator is maximal under a longitudinal parallel-arm design, whereas $\tau_X \in (0, 1)$ in a standard stepped wedge design. However, the generalized ICC, τ_X, is often negative under a repeated crossover design. Such insights could facilitate the comparison of relative efficiency for alternative designs with subclusters.

3.2 | Non-Gaussian exponential family outcomes

When the outcome is binary or count, Theorem 1 is not directly applicable because the residual variance function of the outcome is no longer a constant. Specifically, with a link function g, we can rewrite the conditional mean of the outcome Y_{ijkl} from (1) as

$$\mu_{ijkl} = g^{-1}(\eta_{ijkl}) = g^{-1}(\beta_j + \delta X_{ij} + b_i + c_{ik} + s_{ij} + \pi_{ijk} + \gamma_{ikl}), \tag{7}$$

where the fixed and random effects are defined in Section 2.2. We further define the conditional variance of the outcome as $\phi\zeta(\mu_{ijkl})$, where ϕ is a common dispersion. Without loss of generality, we assume $\phi = 1$ but the following procedure applies to arbitrary $\phi > 0$. For example, the variance function of a binary outcome is parameterized as $\zeta(\mu_{ijkl}) = \mu_{ijkl}(1 - \mu_{ijkl})$. To approximate the large-sample variance of the maximum likelihood estimator for $\hat{\delta}$, we linearize model (7) by a first-order Taylor expansion about the estimated fixed- and random-effects (Breslow and Clayton, 1993; Amatya and Bhaumik, 2018) such that

$$\begin{aligned} \boldsymbol{Y}_{ij} &= \hat{\boldsymbol{\mu}}_{ij} + \widehat{\boldsymbol{\Delta}}_{ij}\boldsymbol{Z}_{ij}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \widehat{\boldsymbol{\Delta}}_{ij}\mathbf{1}_{KN}\left(b_i - \hat{b}_i\right) \\ &+ \widehat{\boldsymbol{\Delta}}_{ij}(\boldsymbol{I}_K \otimes \mathbf{1}_N)(\boldsymbol{c}_{ij} - \hat{\boldsymbol{c}}_{ij}) + \widehat{\boldsymbol{\Delta}}_{ij}\mathbf{1}_{KN}(s_{ij} - \hat{s}_{ij}) \\ &+ \widehat{\boldsymbol{\Delta}}_{ij}(\boldsymbol{I}_K \otimes \mathbf{1}_N)(\boldsymbol{\pi}_{ij} - \hat{\boldsymbol{\pi}}_{ij}) + \widehat{\boldsymbol{\Delta}}_{ij}(\boldsymbol{\gamma}_{ij} - \hat{\boldsymbol{\gamma}}_{ij}) + \boldsymbol{\epsilon}_{ij}, \end{aligned} \tag{8}$$

where $\boldsymbol{\Delta}_{ij} = \mathrm{diag}(\Delta_{ij11}, \ldots, \Delta_{ijKN}) = \left\{\partial g^{-1}(\boldsymbol{\eta}_{ij})/\partial \boldsymbol{\eta}_{ij}^{\top}\right\}^{-1}$ is a diagonal matrix of derivatives, $\boldsymbol{\eta}_{ij} = (\eta_{ij11}, \ldots, \eta_{ijKN})^{\top}$, $\boldsymbol{\mu}_{ij} = (\mu_{ij11}, \ldots, \mu_{ijKN})^{\top}$, $\boldsymbol{Z}_{ij} = (\boldsymbol{e}_j, X_{ij}) \otimes \mathbf{1}_{KN}$ ($\boldsymbol{e}_j$ is the jth row of $\boldsymbol{I}_T$), $\boldsymbol{c}_{ij} = (c_{i1}, \ldots, c_{iK})^{\top}$, $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \ldots, \pi_{ijK})^{\top}$, $\boldsymbol{\gamma}_{ij} = (\gamma_{i11}, \ldots, \gamma_{iKN})^{\top}$, $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij11}, \ldots, \epsilon_{ijKN})^{\top}$, and $\mathrm{var}(\epsilon_{ijkl}) = \zeta(\mu_{ijkl})$. Therefore, if we define the vector of pseudo-outcomes as $\boldsymbol{Y}_{ij}^{*} = \widehat{\boldsymbol{\Delta}}_{ij}^{-1}(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_{ij}) + \hat{\boldsymbol{\eta}}_{ij}$ and rearrange the terms in (8), we obtain an approximate linear mixed model with $\boldsymbol{Y}_{ij}^{*} = \boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij}^{*}$, with a modified random residual error $\boldsymbol{\epsilon}_{ij}^{*} = \widehat{\boldsymbol{\Delta}}_{ij}^{-1}\boldsymbol{\epsilon}_{ij}$. Define the collection of all pseudo-outcomes in cluster i as $\boldsymbol{Y}_i^{*} = \left(\boldsymbol{Y}_{i1}^{*\top}, \ldots, \boldsymbol{Y}_{iT}^{*\top}\right)^{\top}$, we show in Web Appendix C that

$$\begin{aligned} \boldsymbol{V}_i &= \mathrm{cov}\left(\boldsymbol{Y}_i^{*}\right) \approx \mathbb{E}\left\{\boldsymbol{\Delta}_i^{-1}\zeta(\boldsymbol{\mu}_i)\boldsymbol{\Delta}_i^{-1}\right\} \\ &+ \sigma_{\pi}^2(\boldsymbol{I}_{TK} \otimes \boldsymbol{J}_N) + \sigma_S^2(\boldsymbol{I}_T \otimes \boldsymbol{J}_{KN}) + \sigma_{\gamma}^2(\boldsymbol{J}_T \otimes \boldsymbol{I}_{KN}) \\ &+ \sigma_c^2(\boldsymbol{J}_T \otimes \boldsymbol{I}_K \otimes \boldsymbol{J}_N) + \sigma_b^2\boldsymbol{J}_{TKN}, \end{aligned} \tag{9}$$

with $\boldsymbol{\Delta}_i = \oplus_{j=1}^{T}\boldsymbol{\Delta}_{ij}$ where "$\oplus$" is a block diagonal operator with nonzero matrices along the diagonal and zero values elsewhere, and the expectation is taken over the distribution of all the random effects. In general, $\mathbb{E}\left\{\boldsymbol{\Delta}_i^{-1}\zeta(\boldsymbol{\mu}_i)\boldsymbol{\Delta}_i^{-1}\right\}$ can be computed via numerical integration and depends on the conditional mean of outcomes through the secular trend and intervention status, thus $\boldsymbol{V}_i$ will be cluster specific. The approximate covariance matrix of the estimator

for fixed-effects parameters based on the pseudo-outcomes is then $\left(\sum_{i=1}^{I} \boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i\right)^{-1}$, whose $(T+1, T+1)$th element is

$$\begin{aligned}&\mathrm{var}(\hat{\delta}) \\ &= \left[\sum_{i=1}^{I} (\boldsymbol{X}_i^\top \otimes \mathbf{1}_{KN}^\top)\boldsymbol{V}_i^{-1}(\boldsymbol{X}_i \otimes \mathbf{1}_{KN}) - \left\{\sum_{i=1}^{I} (\boldsymbol{X}_i^\top \otimes \mathbf{1}_{KN}^\top)\boldsymbol{V}_i^{-1}(\boldsymbol{I}_T \otimes \mathbf{1}_{KN})\right\} \times \left\{\sum_{i=1}^{I} (\boldsymbol{I}_T^\top \otimes \mathbf{1}_{KN}^\top)\boldsymbol{V}_i^{-1}(\boldsymbol{I}_T \otimes \mathbf{1}_{KN})\right\}^{-1} \times \left\{(\boldsymbol{I}_T^\top \otimes \mathbf{1}_{KN}^\top)\sum_{i=1}^{I} \boldsymbol{V}_i^{-1}(\boldsymbol{X}_i \otimes \mathbf{1}_{KN})\right\}\right]^{-1}\end{aligned}$$

.

In the design stage, computation of the above variance requires us to invert the $TKN \times TKN$ variance matrix, $\boldsymbol{V}_i$, for each cluster, which could require substantial computational time as the subcluster size or the number of subclusters become large. However, because the fixed effects in (7) only depends on each cluster period, we provide an equivalent but computationally more efficient variance expression in the following lemma.

Lemma 3. *The variance of the intervention effect estimator in the GLMM* (7) *is equivalently written as*

$$var(\hat{\delta}) = \left\{\sum_{i=1}^{I} \boldsymbol{X}_i^\top \widetilde{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i - \left(\sum_{i=1}^{I} \boldsymbol{X}_i^\top \widetilde{\boldsymbol{V}}_i^{-1}\right) \times \left(\sum_{i=1}^{I} \widetilde{\boldsymbol{V}}_i^{-1}\right)^{-1} \left(\sum_{i=1}^{I} \widetilde{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i\right)\right\}^{-1}, \tag{10}$$

where $\widetilde{\boldsymbol{V}}_i = (KN)^{-1}\boldsymbol{E}_i + \left(K^{-1}\sigma_\pi^2 + \sigma_s^2\right)\boldsymbol{I}_T + \left\{\sigma_b^2 + K^{-1}\sigma_c^2 + (KN)^{-1}\sigma_\gamma^2\right\}\boldsymbol{J}_T$ *is the* $T \times T$ *matrix characterizing the covariance of the cluster period means of the pseudo-outcomes* $\overline{\boldsymbol{Y}}_i^* = (KN)^{-1}\left(\boldsymbol{I}_T \otimes \mathbf{1}_{KN}^\top\right)\boldsymbol{Y}_i^*$ *, and* $\boldsymbol{E}_i = \mathrm{diag}\left[\mathbb{E}\left\{\Delta_{i111}^{-1}\zeta(\mu_{i111})\Delta_{i111}^{-1}\right\}, \ldots, \mathbb{E}\left\{\Delta_{iT11}^{-1}\zeta(\mu_{iT11})\Delta_{iT11}^{-1}\right\}\right]$.

By providing an equivalent variance expression, Lemma 3 indicates that one only needs to invert a set of $T \times T$ matrices to obtain the variance for power calculation for general exponential family outcomes with a certain mean-variance structure and alleviates the computational burden associated with inverting a series of $TKN \times TKN$ matrices. While Lemma 3 is easy to verify when $\boldsymbol{V}_i$ has a closed-form inverse (as when the variance function $\zeta(\mu_{ijkl}) \propto 1$), it is not trivial for general variance functions and requires us to exploit the block structure of $\mathbb{E}\left\{\boldsymbol{\Delta}_i^{-1}\zeta(\boldsymbol{\mu}_i)\boldsymbol{\Delta}_i^{-1}\right\}$ under a stepped wedge design; the detailed proof is provided in Web Appendix C. Finally, Lemma 3 implies that $\mathrm{var}(\hat{\delta})$ can be equivalently considered as that obtained from a linear mixed model for the cluster period mean pseudo-outcomes $\overline{\boldsymbol{Y}}_i^*$ and is therefore an extension of the results in Li *et al.* (2021) to more complex correlation structures, under equal subcluster sizes.

From variance expression (10), it is critical to compute the nonzero diagonal elements of $\boldsymbol{E}_i$, where the expectation involves integration over all the random effects. While numerical or Monte Carlo integration can be used as a general solution, the computation of these terms can be considerably simplified under canonical link functions. For example, with a binary outcome and a canonical logit link, we can use the property of the Gaussian moment generating function to obtain the jth diagonal element of $\boldsymbol{E}_i$ as

$$\mathbb{E}\left\{\Delta_{ij11}^{-1}\zeta(\mu_{ij11})\Delta_{ij11}^{-1}\right\} = 2 + 2\exp\left(\frac{\sigma_b^2+\sigma_c^2+\sigma_s^2+\sigma_\pi^2+\sigma_\gamma^2}{2}\right) \times \cosh(\beta_j + X_{ij}\delta),$$

where $\cosh(t) = (e^t + e^{-t})/2$ is the hyperbolic cosine function. In cases when the ICCs are more intuitive parameters than variance components to consider in the design phase, we could use the latent response formulation and define the total variance as $\sigma^2 = \sigma_b^2 + \sigma_c^2 + \sigma_s^2 + \sigma_\pi^2 + \sigma_\gamma^2 + \pi^2/3$, where $\pi^2/3$ is the variance of the standard logistic distribution (Eldridge *et al.*, 2009). This formulation allows us to reparameterize $\text{var}(\hat{\delta})$ with σ^2 and the five ICCs (Table 1). Specifically, using Lemma 3 the variance matrix under a binary outcome with a logit link can be represented by

$$\widetilde{\boldsymbol{V}}_i = \frac{1}{KN}\boldsymbol{E}_i + \frac{(\lambda_3-\lambda_1)\sigma^2}{KN}\boldsymbol{I}_T + \frac{(\lambda_6-\lambda_3)\sigma^2}{TKN}\boldsymbol{J}_T, \tag{11}$$

with

$$\boldsymbol{E}_i = 2\boldsymbol{I}_T + 2\exp\left\{\frac{(1-\lambda_1)\sigma^2}{2}\right\}\text{diag}\{\cosh(\beta_1 + X_{i1}\delta), \ldots, \cosh(\beta_T + X_{iT}\delta)\}, \tag{12}$$

and λ_1, λ_3, and λ_6 are three eigenvalues of the extended block exchangeable correlation matrix defined in Lemma 1. Unlike Theorem 1, Equations (11) and (12) suggest the variance of the intervention effect in a logistic linear mixed model depends on the ICCs through an additional eigenvalue λ_1. In Web Figure 1, we study the relationship between the five ICCs and $\text{var}(\hat{\delta})$ and confirm that the observations in Theorem 1 remain valid under a logistic linear mixed model with the exception of the within-subject autocorrelation, α_2, for which larger values will lead to more conservative sample size estimates. In Web Appendix D, we provide the expression of $\boldsymbol{E}_i$ for a count outcome with canonical log link and gamma outcome with a canonical inverse link which can be easily computed without complex numerical integration.

3.3 | Extension to unequal cluster sizes

In multilevel SW-CRTs, it may not always be realistic to assume equal cluster sizes for power calculation. To extend our procedure to unequal cluster sizes, we assume the number of subclusters and the subcluster size only vary across clusters but remain constant within each cluster over time; namely, $K_{ij} = K_i$ and $N_{ijk} = N_i$. A similar assumption was made in previous studies incorporating unequal cluster sizes in sample size calculations for SW-CRTs (Harrison *et al.*, 2019; Matthews, 2020). This assumption is considered appropriate because the efficiency of SW-CRTs is primarily driven by the between-cluster imbalance rather than the within-cluster imbalance over time (Tian *et al.*, 2021). For given I and T and assuming $K_i \sim f_K(K_i; \overline{K}, \text{CV}_K)$ and $N_i \sim f_N(N_i; \overline{N}, \text{CV}_N)$, where $f(\bullet; a, b)$ represents a valid density or mass function with mean a and coefficient of variation (CV) b, the expected variance of $\hat{\delta}$ can be used for sample size determination and is given by

$$\operatorname{var}\left(\hat{\delta}\right) = \int \cdots \int \operatorname{var}(\hat{\delta} \mid \mathcal{O}) \prod_{i=1}^{I} \times \left\{ f_K\left(K_i; \overline{K}, \mathrm{CV}_K\right) f_N\left(N_i; \overline{N}, \mathrm{CV}_N\right) dK_i dN_i \right\}, \tag{13}$$

where $\mathcal{O} = \{(K_i, N_i), i = 1, \ldots, I\}$ represents a specific design with unequal cluster sizes and $\operatorname{var}(\hat{\delta} \mid \mathcal{O})$ can be efficiently computed based on a modified version of Lemma 3 provided in Web Appendix E. To obtain $\operatorname{var}(\hat{\delta})$, one can impose any sensible parametric distributions for f_K and f_N and approximate (13) via Monte Carlo integration. For example, one can assume f_K and f_N as gamma distributions with specified mean and CV. A total of R sets of cluster size configurations, $\left\{\mathcal{O}^{(r)}, r = 1, \ldots, R\right\}$, are then drawn from the specified gamma distributions (rounded to nearest integer) to obtain $\operatorname{var}(\hat{\delta}) \approx R^{-1} \sum_{r=1}^{R} \operatorname{var}\left(\hat{\delta} \mid \mathcal{O}^{(r)}\right)$. By averaging over R designs (e.g., $R = 1000$), $\operatorname{var}(\hat{\delta})$ accounts for the anticipated variations in cluster sizes and can be used in (2) for power calculation. Finally, when $\mathrm{CV}_K = \mathrm{CV}_N = 0$, the above procedure coincides with that in Sections 3.1 and 3.2 as both f_K and f_N degenerate to a point mass. Additional details are provided in Web Appendix E.

4 | SIMULATION STUDY

We conducted a simulation study to assess the accuracy of our sample size procedures for a Gaussian outcome and a binary outcome. For illustration, we focused on a closed-cohort design at the subcluster level and a cross-sectional design at the subject level (design (B)). We generated correlated Gaussian outcomes based on the linear mixed model, $Y_{ijkl} = \beta_j + \delta X_{ij} + b_i + c_{ik} + s_{ij} + \pi_{ijk} + \epsilon_{ijkl}$, by constraining the total variance components $\sigma^2 = 1$. Given σ^2, we consider three sets of ICCs (recall that $\alpha_2 = \alpha_1$ under design (B)), $(\alpha_0, \alpha_1, \rho_0, \rho_1) = \{(0.1, 0.05, 0.025, 0.0125), (0.03, 0.015, 0.0075, 0.00375), (0.01, 0.005, 0.0025, 0.00125)\}$, which determines the value of each variance component. The within-period ICCs were chosen to mimic commonly reported values in parallel-arm CRTs, and the between-period ICCs correspond to a CAC of 0.5. By Theorem 1, $\operatorname{var}(\hat{\delta})$ is invariant to the period effects, and therefore we only considered a gently increasing secular trend with $\beta_1 = 0$ and $\beta_{j+1} - \beta_j = 0.1 \times (0.5)^{j-1}$ for $j \geq 1$. On the other hand, we generated correlated binary outcomes from a Bernoulli distribution with $Y_{ijkl} \sim \mathrm{Bern}(\mu_{ijkl})$, with $\mu_{ijkl} = 1/\{1 + \exp(-\beta_j - \delta X_{ij} - b_i - c_{ik} - s_{ij} - \pi_{ijk})\}$. Following Li *et al.* (2018), we assumed a slightly decreasing secular trend on the logit scale with a baseline prevalence of 0.7 such that $\beta_1 = 1/\{1 + \exp(-0.7)\}$ and $\beta_j - \beta_{j+1} = 0.1 \times (0.5)^{j-1}$ for $j \geq 1$. To specify the variance components, we employ the latent response formulation in Section 3.2 by setting the residual variance as $\pi^2/3$ and mapping the three sets of ICCs considered for Gaussian outcomes to the variance components on the logit scale (Web Appendix F).

We simulated standard stepped wedge designs such that an equal number of clusters crossed over to treatment at a randomly assigned step. Motivated by a recent systematic review of SW-CRTs (Grayling *et al.*, 2017) and assuming equal cluster sizes, we varied the number of clusters (I) from eight to 30, the number of subclusters per cluster (K) from two to six, the number of periods (T) from four to seven, and allowed a maximum of 15 subjects per subcluster (N). For Gaussian outcomes, we assumed standardized effect sizes, $\delta/\sigma \in \{0.1,$

0.2, 0.25, 0.35, 0.4, 0.5}, and for binary outcomes, we considered effect sizes on the odds ratio scale, $\exp(\delta) \in \{0.8, 0.75, 0.7, 0.65, 0.6, 0.5\}$. Each specific parameter combination was selected to ensure the predicted power is at least 80% based on a two-sided 5% level Wald test. For a Gaussian outcome, the predicted power is based on Equation (2) and Theorem 1; for a binary outcome, the predicted power is based on Equation (2), Lemma 3, and Equations (11) and (12). The empirical power of the test is obtained as the proportion correctly rejecting H_0 over 1000 simulated SW-CRTs, and the agreement between the predicted and empirical power was used to confirm the accuracy of the proposed method. Finally, we assessed the empirical type I error rate to confirm the validity of the Wald test. For this purpose, we simply set $\delta/\sigma = 0$ with a Gaussian outcome and $\exp(\delta) = 1$ with a binary outcome. Further, in Web Tables 3 and 4, we compare the predicted power assuming correctly CAC = 0.5 with the predicted power assuming equal within- and between-period ICCs (incorrectly assuming CAC = 1). Our results show that incorrectly assuming equal within- and between-period ICCs typically leads to overly confident power predictions. Finally, to assess the accuracy of our approach in Section 3.3 with unequal cluster sizes, we chose four typical scenarios and draw K_i and N_i from gamma distributions with $\text{CV}_K \in \{0, 0.25, 0.5\}$ and $\text{CV}_N \in \{0, 0.25, 0.5, 0.75, 1.0\}$.

4.1 | Simulation results for Gaussian outcomes

We present the empirical type I error rate and empirical and predicted power of the Wald test under each scenario with a Gaussian outcome (Table 2). We used the restricted maximum likelihood estimator for model fitting, based on which the Wald test is carried out. The empirical type I error rate was generally conservative, and the empirical and predicted power were in agreement with the largest difference within 3%. Similar results were found when the Wald test was computed from (unrestricted) maximum likelihood estimation (Web Table 5). Lastly, in Web Tables 6 and 7 we present the simulation results under unequal cluster sizes. The empirical type I error rate was conservative overall, and differences between empirical and predicted power were within −4.4% and 3.5%.

4.2 | Simulation results for binary outcomes

We present the empirical type I error rate, and empirical and predicted power of the Wald test under each scenario using the logistic linear mixed model fitted via the Laplace approximation (Table 3). Similarly, the empirical type I error rate was well controlled under the nominal level, and the empirical power was in agreement with the predicted power, with the differences ranging from −0.9% to 5.3%. This suggests that our sample size procedure based on first-order Taylor expansion at most results in slightly conservative power prediction. Results were similar when penalized quasi-likelihood (which is computationally more efficient) was used to fit the model (Web Table 8). Lastly, in Web Tables 9 and 10, we present the simulation results under unequal cluster sizes. Overall, the empirical type I error rate was adequately controlled and differences between empirical and predicted power were within −6.1% and 8.2%. Overall, these results confirm that the procedure in Section 3.3 sufficiently captures the trend of the empirical power under unequal cluster sizes.

5 | APPLICATIONS TO SW-CRTs WITH SUBCLUSTERS

5.1 | Lumbar Imaging with Reporting of Epidemiology trial

The LIRE trial aimed to evaluate the effect of adding prevalence data to spine imaging reports on subsequent spine-related health care utilization (Jarvik *et al.*, 2015). The study planned to randomize $I = 100$ practices consisting of a total of 1700 primary care providers (PCPs) over $T = 6$ periods; each practice is a cluster and each PCP represents a subcluster. This is a closed-cohort design on the subcluster level but a cross-sectional design at the subject level (design (B)). While the number of PCPs per practice varied, we assume $K = 17$ PCPs per practice for illustration. The primary outcome was log-transformed spine-related relative value units (RVUs), a continuous composite measure of back pain. Assuming the median and total variance of RVU is approximately 3.56 and 2.5 (Jarvik *et al.*, 2020), a 5% reduction in median due to treatment corresponds to a standardized effect size of around −0.1. Based on preliminary data, an overall ICC was estimated to be 0.013 with a 95% confidence interval of (0.00, 0.046). We therefore assume the within-period within-PCP ICC to be the upper bound of the preliminary estimates, $a_0 = 0.046$, and a slightly smaller within-period between-PCP ICC of $\rho_0 = 0.04$. Assuming a CAC of 0.5 further gives us $a_1 = 0.023$ and $\rho_1 = 0.02$. Based on (2) and our variance expression (5), we found having $N = 77$ subjects per PCP leads to 87.5% power for a two-sided 5% test.

To assess the sensitivity of our power calculation to ICC specifications, we looked at power trends for $a_0 \in (0, 0.1)$ with various ratios of ρ_0/a_0 across varying levels of CAC (0.2, 0.5, 0.8). In concordance with our findings in Theorem 1, larger within-period ICCs (a_0, ρ_0) and smaller between-period ICCs (a_1, ρ_1) correspond to more conservative power predictions (Figure 3), thus we are confident that our ICC specifications likely produced a conservative power estimate. Alternatively, we have also derived a DE based on the closed-form variance expression (5) for sample size determination. The details and illustrative calculations that reach the same power results in the LIRE trial are provided in Web Appendix G. Further, in Web Appendix H we provide sample size determinations for the LIRE trial under alternative cluster level designs. Finally, in Web Appendix I we provide an application to the LIRE trial assuming unequal cluster sizes using the expected variance expression (13).

5.2 | Washington State EPT trial

In the Washington State EPT study, investigators were interested in evaluating whether an expedited partner therapy, the treatment of sex partners of people with sexually transmitted infections without medical evaluation, would decrease the risk of chlamydia reinfection (Golden *et al.*, 2015). Since the chlamydia infection status was binary, we illustrate the following calculations with a logistic linear mixed model. The study included $I = 24$ LHJ that were randomly assigned to intervention at one of four steps ($T = 5$). Each LHJ includes clinics that provide subject-level outcomes over time. Of the clinics with repeated measures, over half were sampled at each period; thus for simplicity, we assume a closed-cohort design at the clinic level and a cross-sectional design at the subject level (design (B)). A total of 219 clinics participated in chlamydia testing, but to be conservative we assume the number of clinics per LHJ is $K = 5$. In the design of this study, investigators aimed to detect a prevalence ratio of 0.7 and assumed a baseline prevalence of 0.05. Because the

outcome is rare, we assume the effect size expressed as an odds ratio can be approximated by the prevalence ratio and obtain the required number of subjects per clinic (N) to achieve at least 80% power for a two-sided 5% test. Without distinguishing between clusters and subclusters, Li *et al.* (2021) estimated the within-period ICC to be 0.007 and the between-period ICC to be 0.004 based on marginal models. We consider these values to be the within-period between-clinic and between-period within-clinic ICCs (defined based on the latent response formulation in Section 3.2) such that $\rho_0 = 0.007$ and $a_1 = 0.004$, and set the remaining ICCs to be $a_0 = 0.008$ and $\rho_1 = 0.0035$ (corresponding to a CAC of 0.5). We assume a slightly decreasing time effect as in our simulations and find based on Equation (2), Lemma 3, and Equations (11) and (12) that including $N = 42$ subjects per clinic gives us 89.5% power. As a sensitivity analysis, we considered a larger decreasing period effect such that $\beta_j - \beta_{j+1} = 1 \times (0.5)^{j-1}$ for $j \geq 1$, which increased $N = 139$ to attain 89.5% power. On the other hand, using a smaller decreasing period effect, $\beta_j - \beta_{j+1} = 0.01 \times (0.5)^{j-1}$ for $j \geq 1$, reduced our required number of subjects per clinic to $N = 37$ to achieve 89.3% power.

We assessed the sensitivity of our power calculation to ICC specifications by examining the power trends for varying $a_0 \in (0, 0.05)$ with various ratios of ρ_0/a_0 across varying levels of CAC (0.2, 0.5, 0.8) and time trends ($\beta_j - \beta_{j+1} = \{0.01, 0.1, 1\} \times (0.5)^{j-1}$ for $j \geq 1$) (Web Figure 2). We found that larger within-period ICCs (a_0, ρ_0) and smaller between-period ICCs (a_1, ρ_1) correspond to more conservative power predictions, thus our current ICC specifications likely produced a conservative power estimate. Furthermore, in Web Appendix J we provide sample size determinations for the EPT study under alternative cluster level designs. Finally, in Web Appendix K we provide an application to the EPT study assuming unequal cluster sizes using the expected variance expression (13).

6 | CONCLUDING REMARKS

In this study, we presented new sample size procedures for SW-CRTs with subclusters. With a Gaussian outcome, we characterized an extended block exchangeable correlation structure induced by the random-effects assumption of the linear mixed model. The extended block exchangeable correlation structure is parameterized by at most five ICC parameters and intentionally differentiate between the within-period and between-period ICCs to avoid unrealistically small sample size estimates (Taljaard *et al.*, 2016). We derived the variance of the intervention effect estimator, which depends on the ICC parameters only through two eigenvalues of the extended block exchangeable correlation matrix. Assuming a GLMM with non-Gaussian outcomes, we further presented a generic framework for approximating the variance of the intervention effect, and specific examples are provided under the canonical link functions. Of note, while our primary focus is GLMMs, our methods can be extended to accommodate generalized estimating equations; the details are presented in Web Appendix L.

While our sample size and variance expressions are derived based on large sample approximation, many SW-CRTs may have a limited number of clusters. Our simulation results suggest that with as few as eight clusters, a Wald t-test can preserve adequate type I error rate and maintain sufficient power, thus validating our methods. To sum up, we reiterate two key contributions from this work. First, our variance expressions are

computationally convenient, thus obviating the need for simulation-based power calculation; the latter approach can quickly become impractical in SW-CRTs with subclusters due to the need for searching across many design parameters, as well as the associated computational cost for repeatedly fitting complex multilevel models. Second, we have characterized the relationship between various ICCs and power in the presence of subclusters and confirmed that assuming smaller between-period ICCs typically leads to larger and conservative sample sizes. In the presence of limited external data to inform sample size calculations in SW-CRTs with subclusters, it may therefore be prudent to assume smaller between-period ICC values.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work is supported by the National Institute of Aging (NIA) of the National Institutes of Health (NIH) under Award Number U54AG063546, which funds NIA Imbedded Pragmatic Alzheimer's Disease and AD-Related Dementias Clinical Trials Collaboratory (NIA IMPACT Collaboratory). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors are grateful to Professor Jim Hughes for providing data and information from the Washington State EPT trial. We also thank the associate editor and an anonymous referee for their valuable suggestions, which greatly improved the exposition of this work.

Funding information

National Institute on Aging, Grant/Award Number: U54AG063546

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this paper.

REFERENCES

Amatya A and Bhaumik DK (2018) Sample size determination for multilevel hierarchical designs using generalized linear mixed models. Biometrics, 74, 673–684. [PubMed: 28901009]

Breslow NE and Clayton DG (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88, 9–25.

Eldridge SM, Ukoumunne OC and Carlin JB (2009) The intracluster correlation coefficient in cluster randomized trials: a review of definitions. International Statistical Review, 77, 378–394.

Ford WP and Westgate PM (2020) Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. Statistics in Medicine, 39, 2779–2792. [PubMed: 32578264]

Golden MR, Kerani RP, Stenger M, Hughes JP, Aubin M, Malinski C, et al. (2015) Uptake and population-level impact of expedited partner therapy (EPT) on *Chlamydia trachomatis* and *Neisseria gonorrhoeae*: the Washington state community-level randomized trial of EPT. PLoS Medicine, 12, e1001777. [PubMed: 25590331]

Grayling MJ, Wason JMS and Mander AP (2017) Stepped wedge cluster randomized controlled trial designs : a review of reporting quality and design features. Trials, 18, 1–13. [PubMed: 28049491]

Harrison DA and Brady AR (2004) Sample size and power calculations using the noncentral t-distribution. The Stata Journal, 4, 142–153.

Harrison LJ, Chen T and Wang R (2019) Power calculation for cross-sectional stepped wedge cluster randomized trials with variable cluster sizes. Biometrics, 76, 951–962. [PubMed: 31625596]

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

Hemming K and Taljaard M (2020) Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? International Journal of Epidemiology, 49, 1043–1052. [PubMed: 32386407]

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

Hooper R, Teerenstra S, de Hoop E and Eldridge S (2016) Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in Medicine, 35, 4718–4728. [PubMed: 27350420]

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

Jarvik JG, Comstock BA, James KT, Avins AL, Bresnahan BW, Deyo RA, et al. (2015) Lumbar imaging with reporting of epidemiology (LIRE)-protocol for a pragmatic cluster randomized trial. Contemporary Clinical Trials, 45, 157–163. [PubMed: 26493088]

Jarvik JG, Meier EN, James KT, Gold LS, Tan KW, Kessler LG, et al. (2020) The effect of including benchmark prevalence data of common imaging findings in spine image reports on health care utilization among adults undergoing spine imaging: a stepped-wedge randomized clinical trial. JAMA Network Open, 3, e2015713–e2015713. [PubMed: 32886121]

Kistner EO and Muller KE (2004) Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. Psychometrika, 69, 459–474. [PubMed: 25152541]

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

Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER and Heagerty PJ (2021) Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. Statistical Methods in Medical Research, 30, 612–639. [PubMed: 32631142]

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

Li F, Yu H, Rathouz PJ, Turner EL and Preisser JS (2021) Marginal modeling of cluster period means and intraclass correlations in stepped wedge designs with binary outcomes. Biostatistics, 39, 3347–3372.

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

Martin J, Girling A, Nirantharakumar K, Ryan R, Marshall T and Hemming K (2016) Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. Trials, 17, 1–12. [PubMed: 26725476]

Matthews JN (2020) Highly efficient stepped wedge designs for clusters of unequal size. Biometrics, 76, 1167–1176. [PubMed: 31961447]

Taljaard M, Teerenstra S, Ivers NM and Fergusson DA (2016) Substantial risks associated with few clusters in cluster randomized and stepped wedge designs. Clinical Trials, 13, 459–463. [PubMed: 26940696]

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

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

Tian Z, Preisser J, Esserman D, Turner E, Rathouz P and Li F (2021) Impact of unequal cluster sizes for GEE analyses of stepped wedge cluster randomized trials with binary outcomes. Biometrical Journal 10.1002/bimj.202100112.

Zhou X, Liao X, Kunz LM, Normand S-LT, Wang M and Spiegelman D (2020) A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. Biostatistics, 1, 102–121.

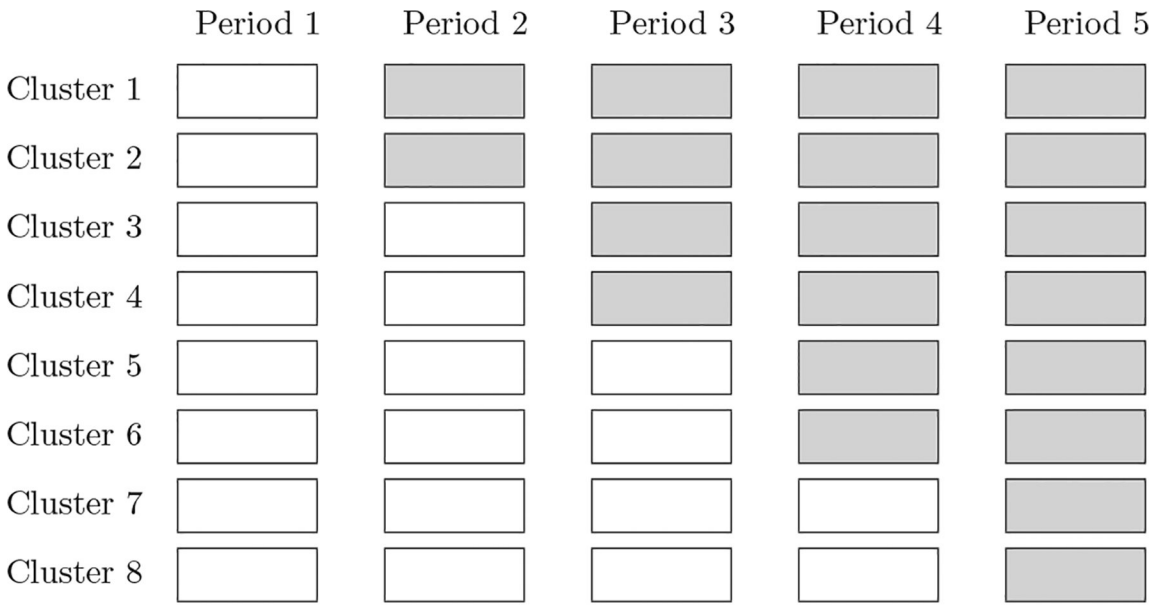
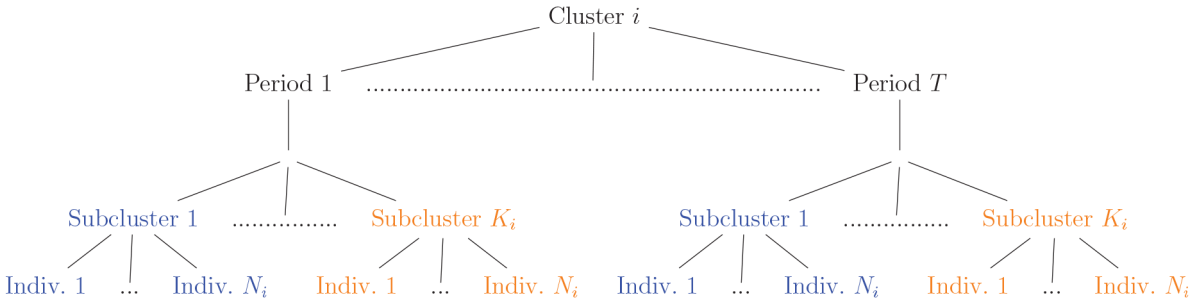


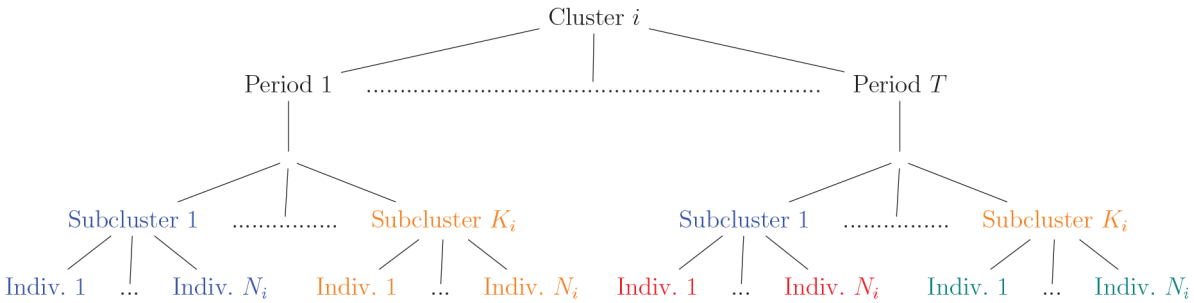
FIGURE 1.

A schematic illustration of a SW-CRT with eight clusters and five periods. Each white cell indicates a cluster period under the control condition and each gray cell indicates a cluster period under the intervention condition. There are in total $S = 4$ distinct intervention sequences

(A) Closed-cohort design at both the subcluster and subject levels: repeated measurements are taken on the same subjects in each subcluster over time.



(B) Closed-cohort design on the subcluster level but a cross-sectional design at the subject level: different subjects in the same subcluster are sampled at each time period.



(C) Cross-sectional design at both the subcluster and subject level: different subjects within different subclusters are sampled in each cluster during each period.

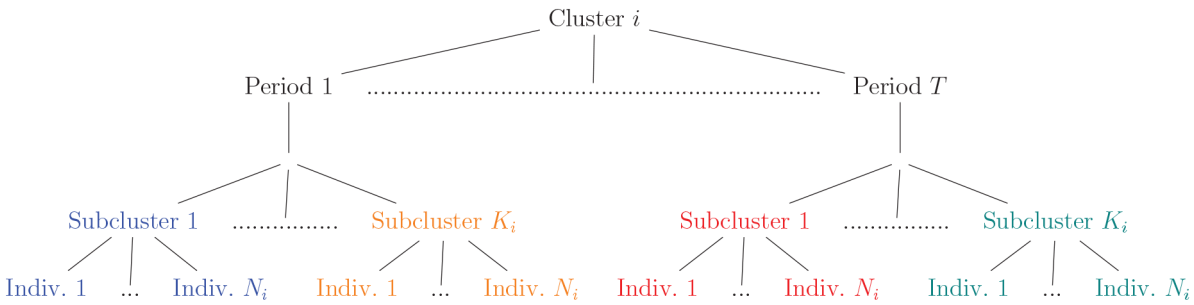


FIGURE 2.

Three design variants for a multilevel SW-CRT with T periods, K_i subclusters per cluster, and N_i individuals (Indiv.) per subcluster. Colors denote unique subclusters and individuals

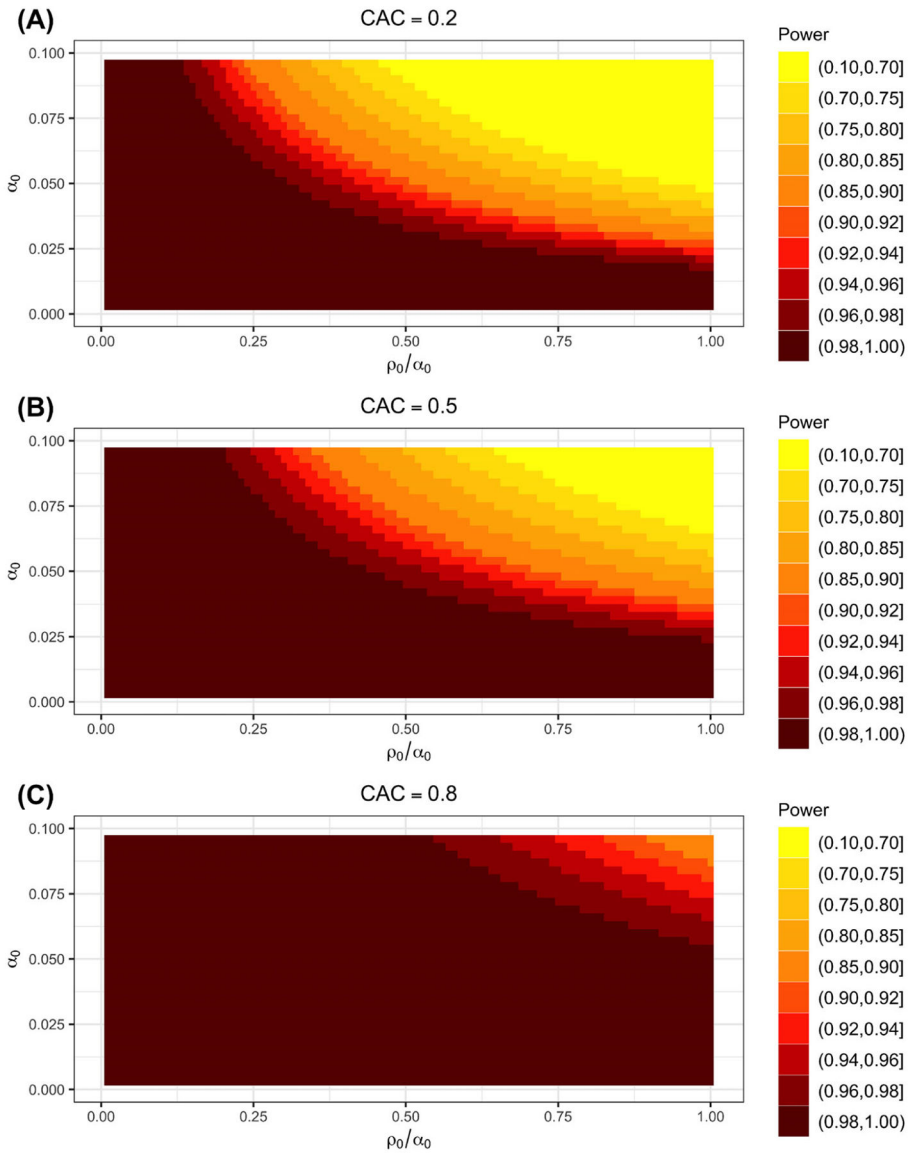


FIGURE 3.

Contour plots illustrating the relationship between ICCs and power in our application study of the LIRE trial. ICCs include: within-period within-subcluster a_0; between-period within-subcluster a_1; within-period between-subcluster ρ_0; and between-period between-subcluster ρ_1. Various a_0 specifications are shown on the *y*-axis and various ρ_0 specifications are shown on the *x*-axis as a ratio of a_0. Between-period specifications are denoted by the CAC. Darker colors correspond to higher values of power

TABLE 1

Definition of ICCs: within-period within-subcluster (a_0), between-period within-subcluster (a_1), within-subject auto-correlation (a_2), within-period between-subcluster (ρ_0), and between-period between-subcluster (ρ_1) under each design variant: (A) Closed-cohort design at both the subcluster and subject levels (total variance $\sigma_A^2 = \sigma_b^2 + \sigma_c^2 + \sigma_s^2 + \sigma_\pi^2 + \sigma_\gamma^2 + \sigma_\epsilon^2$), (B) Closed-cohort design on the subcluster level but a cross-sectional design at the subject level (total variance $\sigma_B^2 = \sigma_b^2 + \sigma_c^2 + \sigma_S^2 + \sigma_\pi^2 + \sigma_\epsilon^2$), and (C) Cross-sectional design at both the subcluster and subject level (total variance $\sigma_C^2 = \sigma_b^2 + \sigma_s^2 + \sigma_\pi^2 + \sigma_\epsilon^2$)

ICC	Definition	Design (A)	Design (B)	Design (C)
a_0	corr (Y_{ijkl}, $Y_{ijkl'}$)	$\left(\sigma_b^2 + \sigma_c^2 + \sigma_S^2 + \sigma_\pi^2\right)/\sigma_A^2$	$\left(\sigma_b^2 + \sigma_c^2 + \sigma_s^2 + \sigma_\pi^2\right)/\sigma_B^2$	$\left(\sigma_b^2 + \sigma_s^2 + \sigma_\pi^2\right)/\sigma_C^2$
a_1	corr (Y_{ijkl}, $Y_{ij'kl'}$)	$\left(\sigma_b^2 + \sigma_c^2\right)/\sigma_A^2$	$\left(\sigma_b^2 + \sigma_c^2\right)/\sigma_B^2$	ρ_1
a_2	corr (Y_{ijkl}, $Y_{ij'kl}$)	$\left(\sigma_b^2 + \sigma_c^2 + \sigma_\gamma^2\right)/\sigma_A^2$	a_1	ρ_1
ρ_0	corr (Y_{ijkl}, $Y_{ijk'l'}$)	$\left(\sigma_b^2 + \sigma_s^2\right)/\sigma_A^2$	$\left(\sigma_b^2 + \sigma_S^2\right)/\sigma_B^2$	$\left(\sigma_b^2 + \sigma_S^2\right)/\sigma_C^2$
ρ_1	corr (Y_{ijkl}, $Y_{ij'k'l'}$)	σ_b^2/σ_A^2	σ_b^2/σ_B^2	σ_b^2/σ_C^2

TABLE 2

Estimated required number of clusters *I*, subclusters per cluster *K*, participants per subcluster *N*, periods *T*, empirical type I error (Test Size), empirical power (Empirical), and predicted power (Predicted) obtained from sample size formula for given effect size δ/σ, within-period ICCs for within- and between-subcluster (α_0, ρ_0), and between-period ICCs for within- and between-subcluster (α_1, ρ_1) assuming a cluster autocorrelation of 0.5, when outcome is Gaussian ($n = 1000$)

δ/σ	(α_0, ρ_0)	(α_1, ρ_1)	*I*	*K*	*N*	*T*	Test Size	Empirical	Predicted
0.1	(0.03, 0.0075)	(0.015, 0.00375)	24	6	15	7	3.6	88.2	85.3
	(0.01, 0.0025)	(0.005, 0.00125)	30	6	15	4	4.3	82.5	82.2
			24	5	10	7	4.4	82.9	81.4
0.2	(0.1, 0.025)	(0.05, 0.0125)	24	6	10	4	4.5	85.4	83.3
			18	3	12	7	3.9	84.0	81.8
	(0.03, 0.0075)	(0.015, 0.00375)	18	3	15	4	3.2	81.3	80.0
			15	3	10	6	4.0	81.1	80.8
	(0.01, 0.0025)	(0.005, 0.00125)	12	6	10	4	2.8	82.2	82.6
			10	4	10	6	2.7	79.4	80.0
0.25	(0.1, 0.025)	(0.05, 0.0125)	21	4	10	4	5.2	83.7	84.6
			18	2	10	7	4.6	83.2	83.5
	(0.03, 0.0075)	(0.015, 0.00375)	15	4	8	4	2.7	82.9	81.4
			12	2	10	7	3.5	81.0	80.2
	(0.01, 0.0025)	(0.005, 0.00125)	24	2	8	4	4.1	83.7	84.3
			10	3	9	6	1.8	83.6	83.6
0.35	(0.1, 0.025)	(0.05, 0.0125)	12	4	9	4	2.8	84.5	83.2
			10	3	8	6	1.9	83.6	82.9
	(0.03, 0.0075)	(0.015, 0.00375)	9	3	12	4	2.4	84.3	83.5
			16	2	5	5	3.0	86.9	84.0
	(0.01, 0.0025)	(0.005, 0.00125)	9	3	9	4	2.5	82.1	82.9
			8	3	7	5	1.7	77.7	80.0
0.4	(0.1, 0.025)	(0.05, 0.0125)	18	2	7	4	3.2	87.9	86.2
			12	2	8	5	3.5	80.8	82.0
	(0.03, 0.0075)	(0.015, 0.00375)	9	3	8	4	1.4	80.8	82.5
			8	3	7	5	1.1	84.3	83.5

δ/σ	(α_0, ρ_0)	(α_1, ρ_1)	I	K	N	T	Test Size	Empirical	Predicted
0.5	(0.01, 0.0025)	(0.005, 0.00125)	15	2	5	4	3.3	81.2	83.3
			12	2	5	5	1.8	85.4	85.1
	(0.1, 0.025)	(0.05, 0.0125)	12	2	7	4	3.3	82.6	84.7
			12	2	4	5	3.2	84.3	82.5
	(0.03, 0.0075)	(0.015, 0.00375)	9	2	8	4	1.8	87.4	85.4

TABLE 3

Estimated required number of clusters *I*, subclusters per cluster *K*, participants per subcluster *N*, periods *T*, empirical type I error (Test Size), empirical power (Empirical), and predicted power (Predicted) obtained from sample size formula for given effect size exp(δ), within-period ICCs for within- and between-subcluster (α_0, ρ_0), and between-period ICCs for within- and between-subcluster (α_1, ρ_1) assuming a cluster autocorrelation of 0.5, when outcome is binary with canonical logit link (n = 1000)

exp(δ)	(α_0, ρ_0)	(α_1, ρ_1)	*I*	*K*	*N*	*T*	Test Size	Empirical	Predicted
0.8	(0.03, 0.0075)	(0.015, 0.00375)	18	6	15	7	4.2	82.2	80.7
	(0.01, 0.0025)	(0.005, 0.00125)	27	6	15	4	5.2	85.3	84.2
			25	4	12	6	2.7	81.2	81.0
0.75	(0.1, 0.025)	(0.05, 0.0125)	25	6	15	6	5.2	85.8	82.8
			24	5	15	7	4.6	88.1	83.1
	(0.03, 0.0075)	(0.015, 0.00375)	27	5	12	4	5.1	83.4	80.6
			30	3	10	6	5.2	84.0	83.3
	(0.01, 0.0025)	(0.005, 0.00125)	21	6	10	4	2.8	82.0	80.5
			12	4	15	7	2.4	84.6	81.5
0.7	(0.1, 0.025)	(0.05, 0.0125)	30	5	14	4	4.4	82.9	82.3
			18	4	15	7	4.2	86.7	81.7
	(0.03, 0.0075)	(0.015, 0.00375)	18	6	10	4	4.8	82.7	80.6
			15	3	15	6	4.4	82.7	81.2
	(0.01, 0.0025)	(0.005, 0.00125)	18	4	12	4	3.6	83.2	82.3
			20	2	15	5	3.9	81.5	81.8
0.65	(0.1, 0.025)	(0.05, 0.0125)	21	6	12	4	4.1	88.3	83.6
			18	3	12	7	3.9	87.6	84.1
	(0.03, 0.0075)	(0.015, 0.00375)	24	3	10	4	4.6	86.4	85.0
			20	2	10	6	3.3	85.7	83.7
	(0.01, 0.0025)	(0.005, 0.00125)	15	4	10	4	2.5	84.8	82.7
			12	3	14	5	3.1	85.4	85.2
0.6	(0.1, 0.025)	(0.05, 0.0125)	18	5	10	4	3.6	85.9	82.3
			12	3	15	7	4.6	88.1	82.8
	(0.03, 0.0075)	(0.015, 0.00375)	16	2	12	5	4.6	85.3	83.9
			15	2	10	6	3.7	87.1	84.0

exp(δ)	(α_0, ρ_0)	(α_1, ρ_1)	I	K	N	T	Test Size	Empirical	Predicted
	(0.01, 0.0025)	(0.005, 0.00125)	21	2	10	4	3.7	84.6	85.5
			12	3	8	5	2.4	81.6	80.0
0.5	(0.1, 0.025)	(0.05, 0.0125)	15	3	10	4	3.0	86.9	83.2
			16	2	9	5	3.3	84.7	82.5
	(0.03, 0.0075)	(0.015, 0.00375)	15	2	9	4	2.6	88.2	84.1