Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure

Fan Li[1,2]

[1]Department of Biostatistics, Yale University, New Haven, Connecticut

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

Abstract

A stepped wedge cluster randomized trial is a type of longitudinal cluster design that sequentially switches clusters to intervention over time until all clusters are treated. While the traditional posttest-only parallel design requires adjustment for a single intraclass correlation coefficient, the stepped wedge design allows multiple outcome measurements from the same cluster and so additional correlation parameters are necessary to characterize the within-cluster correlation structure. Although a number of studies have differentiated between the concepts of within-period and between-period correlations, only a few studies have allowed the between-period correlation to decay over time. In this article, we consider the proportional decay correlation structure for a cohort stepped wedge design, and provide a matrix-adjusted quasi-least squares approach to accurately estimate the correlation parameters along with the marginal intervention effect. We further develop the sample size and power procedures accounting for the correlation decay, and investigate the accuracy of the power procedure with continuous outcomes in a simulation study. We show that the empirical power agrees well with the prediction even with as few as nine clusters, when data are analyzed with matrix-adjusted quasi-least squares concurrently with a suitable bias-corrected sandwich variance. Two trial examples are provided to illustrate the new sample size procedure.

Keywords

finite-sample correction; generalized estimating equations (GEEs); group-randomized trial; proportional decay; quasi-least squares (QLS); sample size calculation

1 | INTRODUCTION

A unique feature of cluster randomized trials (CRTs) is that intact clusters, such as schools or clinics, are randomized to intervention arms.[1,2] Randomization at the cluster level often carries pragmatic considerations, eg, administrative convenience, political reasons and prevention of treatment contamination.[3] A stepped wedge CRT is a type of longitudinal design that sequentially switches clusters to intervention during the course of the study until all clusters are treated.[4] Such designs have become increasingly popular due to their logistical flexibility and perceived ethical benefits. Because individual outcomes within the same cluster tend be more similar than those in different clusters, the intraclass correlation coefficient (ICC) plays a central role in designing CRTs. While the traditional posttest-only parallel design (ie, parallel cluster randomized design without a baseline period as defined in the work of Murray[1]) requires adjustment for a single ICC, longitudinal cluster randomized design such as the crossover or stepped wedge design allows multiple outcome measurements from the same cluster and so naturally requires additional correlation parameters to characterize the within-cluster structure.[5,6] Correspondingly, sample size and power calculations for stepped wedge designs necessitate the specification of more than one correlation parameters. For example, Hemming et al[7] considered both the within-period and between-period ICCs in their sample size procedure for a cross-sectional design. Hooper et al[8] and Li et al[5] examined a three-correlation structure that additionally accounts for the within-individual repeated measurements in a closed-cohort design.

Despite existing development of multiparameter correlation structures for designing stepped wedge trials, most of them assumed a constant between-period ICC with a few exceptions. For example, in a cross-sectional design where outcome data are obtained from a different set of participants in each cluster period,[1,9] Hemming et al[7] allowed the between-period ICC to be different from the within-period ICC, but restricted the between-period ICC to be constant irrespective of the distance between periods. Relaxing the constant between-period ICC assumption for a cross-sectional design, Kasza et al[10,11] studied a nonuniform correlation structure with a decay parameter and proposed a sample size procedure that accounts for the exponential correlation decay. Grantham et al[12] further extended their sample size procedure to allow for continuous-time correlation decay in multiple-periods CRTs with continuous recruitment. From a trial planning standpoint, if correlation decay is present, Kasza et al[10] indicated that omitting the correlation decay in a cross-sectional design would either underestimate or overestimate the true variance of the intervention effect, which led to inaccurate sample size determination. As we demonstrate in Section 6.3, similar considerations carry over to closed-cohort designs, where outcome data are collected from the same set of participants in each cluster period.[1,9,13] Particularly, the constant between-period ICC assumption in the works of Hooper et al[8] and Li et al[5] may not always be realistic and it is therefore necessary to develop alternative design and analysis strategies accounting for correlation decay in cohort stepped wedge studies.

Two popular modeling approaches for stepped wedge designs are cluster-specific models (eg, random-effects models) and population-averaged models.[15] The parameter estimates from a population-averaged model can be interpreted as the marginal intervention effect for the participating individuals combined over all cluster periods, and may be preferred over

the cluster-specific models for trials conducted in the health policy or health services settings.[5] In this article, we consider a population-averaged model with a decay correlation structure. As indicated in Table 1, the first contribution of this article is to fill in the gap and characterize a proportional decay structure appropriate for cohort stepped wedge designs. Such a proportional decay structure has been previously introduced in analyzing clustered longitudinal data; see, eg, the works of Lefkopoulou et al,[16] Shults and Morrow[17] and Liu et al,[18] but has not been exemplified in CRTs with a staggered randomization. Based on the proportional decay structure, we derive a new closed-form variance expression to facilitate sample size and power determination. Based on the derived variance expression, we additionally obtain a simple-to-use design effect (DE) and study how the power depends on the correlation parameters. Since the sample size procedure requires input for the correlation parameters, accurate estimation of the correlations is instrumental for future planning of stepped wedge trials. Therefore, a second contribution of this article is to introduce a modified generalized estimating equations (GEEs) approach to accurately estimate the decay correlation structure along with the marginal mean parameters. The traditional GEE[19] is modified in the following two ways. (1) As simple moment estimators for the decay correlation structure may not be easy to obtain, we estimate the marginal mean and correlation parameters by quasi-least squares (QLS).[20,21] The QLS approach shares the same estimating equations with GEE regarding the marginal mean parameters, but is flexible enough to accommodate nonstandard correlation structures. Similar to the traditional GEE estimator, the QLS estimator is also robust to correlation misspecification. (2) Since stepped wedge cluster randomized design frequently includes a small number of clusters, we refine the QLS approach by incorporating appropriate finite-sample bias corrections to both the estimation of correlation parameters as well as the variance of the intervention effect.

The remainder of this article is organized as follows. The notation of cohort stepped wedge designs is introduced in Section 2. In Section 3, we provide the proportional decay correlation structure and discuss the QLS estimators to estimate the correlation structure. In Section 4, we develop closed-form procedures for sample size and power calculations based on the population-averaged model coupled with the proportional decay structure. We conduct a simulation study in Section 5 to examine the accuracy of the proposed power procedure when the trials are analyzed by the QLS approach. Section 6 provides two illustrative examples of cohort stepped wedge designs and Section 7 concludes.

2 | NOTATION AND BASIC SETUP

We consider a cohort stepped wedge design, where a closed cohort of individuals are enrolled at each of the I participating clusters at the start of the trial. We mainly focus on cohort designs to inform the applications in Section 6, and will defer the discussion of cross-sectional designs to Section 7. We assume the trial involves a total of T time periods. All clusters start from the control condition, and may be randomly chosen to switch to intervention during the course of the study, until all clusters are treated at the end of the Tth period. Individual participants will be scheduled for outcome measurement during each period, and so each individual has a total of T repeated measurements. Denote y_{ijt} as the outcome for individual j $(j = 1, …, N_i)$ from cluster i $(i = 1, …, I)$ at period t $(t = 1, …, T)$. A step is defined as the preplanned time point when at least one cluster crosses over from

control to intervention. We denote the total number of steps by S ($2 \leq S \leq T - 1$), and assume that m_s clusters cross over at step s such that $\sum_{s=1}^{S} m_s = I$. We assume a complete design in the terminology of the work of Hemming et al[7] such that outcome measurements are taken for all individuals during each period. Following the work of Woertman et al,[22] we define the baseline measurements as those taken before any cluster is randomized to intervention, and follow-up measurements as those taken after at least one cluster is randomized to intervention. We assume there are $b \geq 1$ baseline measurements planned under the control condition, and $c_s \geq 1$ follow-up measurements planned between step s and step $s + 1$ (or end of study). Each measurement time point is associated with a distinct time period and the total number of periods $T = b + \sum_{s=1}^{S} c_s$. A standard stepped wedge design is given by $b = c_s = 1$ for all s, and $T = S + 1$ ($T \geq 3$). A schematic illustration of a standard design with $I = 6$ clusters and $T = 4$ periods can be found in Figure 1. We also provide a schematic illustration of an alternative design with $I = 6$ clusters and $T = 6$ periods in Web Figure 1.

3 | ANALYSIS CONSIDERATIONS: MODELS AND STATISTICAL INFERENCE

3.1 | Population-averaged models

The population-averaged model relates the marginal mean, μ_{ijt}, to the time trend and the intervention effect by

$$g\left(\mu_{ijt}\right) = \beta_t + X_{it}\delta, \tag{1}$$

where g is the link function and β_t is the tth period effect.[5] Furthermore, X_{it} is the intervention status, which is equal to 1 or 0 depending on whether cluster i receives intervention during period t, and δ describes the intervention effect on the link function scale. Model (1) can be regarded as the marginal counterpart of a number of existing random-effects models, such as those proposed in the works of Hemming et al,[7] Hooper et al[8] and Kasza et al.[10,11] Like these random-effects models, our marginal model (1) does not specify treatment by period interaction and so δ should be interpreted as the average intervention effect across periods. We write the collection of model parameters as $\theta = (\beta_1, \ldots, \beta_T, \delta)'$, the collection of intervention status for cluster i (a sequence of ones preceded by zeros) as $X_i = (X_{i1}, \ldots, X_{iT})'$, and define $v(\mu_{ijt})$ as a known variance function. To allow for potential correlation decay over time, we define the proportional decay correlation structure similar to the work of Lefkopoulou et al.[16] Specifically, we define the within-period correlation as the correlation between outcomes for two distinct individuals from the same cluster during the same period, ie, corr($y_{ijt}, y_{ij't}$) = τ for $j \neq j'$. The same definition is prevalently used in the posttest-only parallel designs.[1] We then assume a first-order autoregressive structure for the set of within-individual repeated measurements. Therefore, the within-individual correlation, which describes the association between outcomes measured at time t and t' of the same individual, is corr($y_{ijt}, y_{ijt'}$) = $\rho^{|t - t'|}$, $t \neq t'$. Finally, we define the between-period correlation as the correlation between outcome measured at time t

for individual j and outcome measured at time t' for individual j', and assumes a decay structure as corr$(y_{ijt}, y_{ij't'}) = \tau\rho^{|t-t'|}$ for $j \neq j'$, $t \neq t'$.

In a closed-cohort design, typically, the between-period correlation is much smaller than the within-individual correlation for any two fixed periods, since the former is defined for measurements from two distinct individuals, while the latter is defined for those from the same individual. This ordering of magnitude is reflected in the proportional decay structure because $\tau < 1$ and corr$(y_{ijt}, y_{ij't'}) = \tau\rho^{|t-t'|} <$ corr$(y_{ijt}, y_{ijt'}) = \rho^{|t-t'|}$. By comparison, the exponential decay structure by Kasza et al[10] is based on cross-sectional designs and therefore obviates the need for modeling the within-individual correlation from repeated measurements. In fact, the exponential decay structure assumes corr$(y_{ijt}, y_{ij't'}) = \tau\rho^{|t-t'|}$ irrespective of whether $j = j'$ since two different sets of participants are included in two different periods for a cross-sectional design. In this respect, parameters τ and ρ have similar interpretations in both structures. We provide a visual comparison of these two structures in Table 2. In summary, the proportional decay correlation structure is defined through two parameters, τ and ρ, with the former resembling the traditional ICC definition in a parallel design and the latter controlling for the degree of correlation decay. Of note, the works Shults and Morrow[17] and Liu et al[18] also adopted the same proportional decay structure in longitudinal CRTs with a parallel assignment, and we extend the discussion of this decay structure to CRTs with a staggered assignment.

3.2 | Quasi-least squares analysis

We use the QLS approach introduced by Shults and Morrow[17] to simultaneously estimate the intervention effect in model (1) and the correlation parameters in a cohort stepped wedge design. In particular, the QLS approach and the traditional GEE approach share the same estimating equations for the marginal mean parameters, whereas the former provides a flexible and convenient way to estimate nonstandard correlation structures. Furthermore, both the QLS and GEE approaches are robust to correlation misspecification, namely, estimators for the marginal intervention effect remain consistent even if the working correlation model is misspecified. In sufficiently large samples (usually $I \geq 30$), the robust sandwich variance could also be used to adequately quantify the uncertainty of the intervention effect estimate even under correlation misspecification. We refer the readers to the textbook of Shults and Hilbe[23] for a full exposition on the advantages of QLS over the traditional GEE.

Write $y_{ij} = (y_{ij1}, \ldots, y_{ijT})'$, $\mu_{ij} = (\mu_{ij1}, \ldots, \mu_{ijT})'$, $y_i = \left(y'_{i1}, \ldots, y'_{iN_i}\right)'$, and $\mu_i = \left(\mu'_{i1}, \ldots, \mu'_{iN_i}\right)'$.

Furthermore, define $D_i(\theta) = \partial\mu_i / \partial\theta'$, and let the working covariance of y_i be $V_i = \phi A_i^{1/2}(\theta)R_i(\alpha_0, \alpha_1)A_i^{1/2}(\theta)$, where $\phi > 0$ is the dispersion parameter,

$A_i(\theta) = \text{diag}\left\{A_{i1}(\theta), \ldots, A_{iN_i}(\theta)\right\}$, $A_{ij}(\theta) = \text{diag}\{v(\mu_{ij1}), \ldots, v(\mu_{ijT})\}$, and $R_i(\alpha_0, \alpha_1)$ is a

positive definite working correlation parameterized by α_0 and α_1. We assume the true correlation structure among elements of y_i is the proportional decay structure, denoted by $R_i(\tau, \rho)$. In matrix notation, we can verify that the proportional decay structure induces separability between τ and ρ in that $R_i(\tau, \rho) = G_i(\tau) \otimes R(\rho)$, where $G_i(\tau)$ is a $N_i \times N_i$

exchangeable correlation matrix (with respect to τ) and $F(\rho)$ is a $T \times T$ first-order autoregressive correlation matrix as follows:

$$
G_i(\tau) = \begin{pmatrix} 1 & \tau & \tau & \cdots & \tau \\ \tau & 1 & \tau & \cdots & \tau \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau & \tau & \tau & \cdots & 1 \end{pmatrix}, \quad F(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1. \end{pmatrix}.
$$

We could verify that the determinant

$$
\det\{R_i(\tau, \rho)\} = \det\{G_i(\tau)\}^T \det\{F(\rho)\}^{N_i} = (1 - \tau)^{T(N_i - 1)}\{1 + (N_i - 1)\tau\}^T(1 - \rho^2)^{(T-1)N_i}.
$$

Therefore, valid correlation values that ensure positive definite $R_i(\tau, \rho)$ should be contained in the triangular region

$$
\mathcal{S} = \left\{ (\tau, \rho): -\frac{1}{\max\{N_1, \ldots, N_I\} - 1} < \tau < 1, \ -1 < \rho < 1 \right\}. \tag{2}
$$

Finally, the inverse of the R_i also exists in closed form and is given by $R_i^{-1}(\tau, \rho) = G_i^{-1}(\tau) \otimes F_i^{-1}(\rho)$, where

$$
G_i^{-1}(\tau) = \frac{1}{1 - \tau} I_i - \frac{\tau}{(1 - \tau)\{(1 + (N_i - 1)\tau)\}} J_i, \quad F^{-1}(\rho) = \frac{1}{1 - \rho^2}\{I + \rho^2 C_2 - \rho C_1\},
$$

I_i is a $N_i \times N_i$ identity matrix, J_i is a $N_i \times N_i$ matrix of ones, $C_2 = \text{diag}(0, 1, \ldots, 1, 0)$, and C_1 is a $T \times T$ tridiagonal matrix with zeros on the main diagonal and ones on the two subdiagonals.

To introduce the QLS estimating equations, we further define $r_{ij}(\theta) = A_{ij}^{-1/2}(\theta)\left(y_{ij} - \mu_{ij}\right)$, and write $r_i(\theta) = \left(r'_{i1}(\theta), \ldots, r'_{iN_i}(\theta)\right)'$. The first-stage QLS estimates for θ, α_0, and α_1 are obtained by alternating between the following estimating equations until convergence:

$$
\sum_{i=1}^{I} D'_i(\theta) A_i^{-1/2}(\theta) R_i^{-1}(\alpha_0, \alpha_1) r_i(\theta) = 0, \tag{3}
$$

$$
\sum_{i=1}^{I} \frac{\partial}{\partial \alpha_0}\left\{ r'_i(\theta) R_i^{-1}(\alpha_0, \alpha_1) r_i(\theta) \right\} = 0, \tag{4}
$$

$$
\sum_{i=1}^{I} \frac{\partial}{\partial \alpha_1}\left\{ r'_i(\theta) R_i^{-1}(\alpha_0, \alpha_1) r_i(\theta) \right\} = 0. \tag{5}
$$

In particular, (3) is the usual GEE coupled with the proportional decay structure, and (4) and (5) are scalar equations for the first-stage correlation estimates. Further, closed-form solutions exist for $\hat{\alpha}_0$ and $\hat{\alpha}_1$ (within an iterative step) and are provided in Web Appendix A. Chaganty and Shults[21] showed that $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are asymptotically biased for τ and ρ. To eliminate the large-sample bias in the first-stage correlation estimates, Chaganty and Shults[21] provided the following second-stage estimating equations to obtain: $\hat{\tau}, \hat{\rho}$

$$\sum_{i=1}^{I} \text{tr}\left\{ \frac{\partial G_i^{-1}(\hat{\alpha}_0)}{\partial \alpha_0} G_i(\tau) \right\} = 0 \tag{6}$$

$$\text{tr}\left\{ \frac{\partial F^{-1}(\hat{\alpha}_1)}{\partial \alpha_1} F(\rho) \right\} = 0. \tag{7}$$

The closed-form solution for (6) and (7) are provided by the work of Shults and Morrow[17] as

$$\hat{\tau} = \sum_{i=1}^{I} \frac{N_i(N_i-1)\hat{\alpha}_0\{2 + (N_i-2)\hat{\alpha}_0\}}{\{1 + (N_i-1)\hat{\alpha}_0\}^2} \bigg/ \sum_{i=1}^{I} \frac{N_i(N_i-1)\{1 + (N_i-1)\hat{\alpha}_0^2\}}{\{1 + (N_i-1)\hat{\alpha}_0\}^2},$$

and $\hat{\rho} = 2\hat{\alpha}_1/(1 + \hat{\alpha}_1^2)$.

3.3 | Bias-corrected correlation estimation

Although the correlation estimates obtained from the second-stage QLS estimating equations are unbiased in large samples, they could be subject to finite-sample bias. The finite-sample bias is a typical consideration in stepped wedge CRTs,[5] as they usually include a small number of clusters ($I \leq 30$). We refine the QLS approach with finite-sample bias corrections to the correlation estimating equations by utilizing the cluster leverage,[24] defined as $H_i = D_i(\theta)\left(\sum_{i=1}^{I} D_i'(\theta)V_i D_i(\theta)\right)^{-1} D_i'(\theta)V_i$. Specifically, when I is small, the estimated residual $y_i - \hat{\mu}_i$ tends to be biased toward zero, and following the work of Preisser et al,[25] we have $E[(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'] \approx (I - H_i)\text{cov}(y_i) = \phi(I - H_i)A_i^{1/2}\text{corr}(y_i)A_i^{1/2}$ and, therefore, $E[r_i(\hat{\theta})r_i'(\hat{\theta})] \approx \phi A_i^{-1/2}(I - H_i)A_i^{1/2}\text{corr}(y_i)$. This last equation suggests that $\phi^{-1}A_i^{-1/2}(I - H_i)^{-1}A_i^{1/2}r_i(\hat{\theta})r_i'(\hat{\theta})$ is a better estimator for $\text{corr}(y_i)$ compared to the simple cross-product $\phi^{-1}r_i(\hat{\theta})r_i'(\hat{\theta})$, since the former accounts for finite-sample bias in a multiplicative fashion. Furthermore, observe that

$$r_i'(\theta)R_i^{-1}(\alpha_0, \alpha_1)r_i(\theta) = \text{tr}\left\{R_i^{-1}(\alpha_0, \alpha_1)r_i(\theta)r_i'(\theta)\right\} \propto \phi^{-1}\text{tr}\left\{R_i^{-1}(\alpha_0, \alpha_1)r_i(\theta)r_i'(\theta)\right\}$$

for all values of α_0, α_1, and θ; we propose to replace the first-stage estimating equations (4) and (5) by

$$\sum_{i=1}^{I} \frac{\partial}{\partial \alpha_0} \text{tr}\left\{ R_i^{-1}(\alpha_0, \alpha_1) \tilde{R}_i(\theta) \right\} = \sum_{i=1}^{I} \frac{\partial}{\partial \alpha_0} \text{tr}\left[\left\{ G_i^{-1}(\alpha_0) \otimes F^{-1}(\alpha_1) \right\} \tilde{R}_i(\theta) \right] = 0 \qquad (8)$$

$$\sum_{i=1}^{I} \frac{\partial}{\partial \alpha_1} \text{tr}\left\{ R_i^{-1}(\alpha_0, \alpha_1) \tilde{R}_i(\theta) \right\} = \sum_{i=1}^{I} \frac{\partial}{\partial \alpha_1} \text{tr}\left[\left\{ G_i^{-1}(\alpha_0) \otimes F^{-1}(\alpha_1) \right\} \tilde{R}_i(\theta) \right] = 0, \qquad (9)$$

where $\tilde{R}_i(\theta) = \phi^{-1} A_i^{-1/2} (I - H_i)^{-1} A_i^{1/2} r_i(\theta) r_i'(\theta)$ represents the matrix-adjusted estimator for the correlation structure. The solutions obtained from (8) and (9) could effectively reduce the finite-sample bias in $\hat{\alpha}_0$ and $\hat{\alpha}_1$, which would in turn decrease the finite-sample bias in the QLS estimators for τ and ρ. Of note, similar finite-sample matrix adjustment was developed by Preisser et al[25] for the Prentice-type GEE,[26] and we extend this finite-sample bias-correction approach to the QLS estimating equations. Accurately estimating the correlation parameters in the analysis stage has practical implications since these estimates could be used to guide the planning of future trials.[1] Additional details of the matrix-adjusted estimating equations, (8) and (9), along with the closed-form updates are provided in Web Appendix B.

3.4 | Bias-corrected covariance estimation

The availability of a small number of clusters may also have implications for estimating the variance using GEE-based approaches.[27] In general, the variance of the marginal mean model parameter $\hat{\theta}$ can be estimated using the model-based variance $\Omega_1^{-1} = \left(\sum_{i=1}^{I} D_i' V_i D_i \right)^{-1}$ or the sandwich variance $\Omega_1^{-1} \Omega_0 \Omega_1^{-1}$, where

$$\Omega_0 = \sum_{i=1}^{I} C_i D_i' V_i^{-1} B_i (y_i - \mu_i)(y_i - \mu_i)' B_i' V_i^{-1} D_i C_i, \qquad (10)$$

and both Ω_0 and Ω_1 are evaluated at $(\hat{\theta}, \hat{\tau}, \hat{\rho})$. When both C_i and B_i are identity matrices, Equation (10) reduces to the uncorrected sandwich estimator of Liang and Zeger,[19] which we denote as BC0. BC0 provides valid inference regardless of the correct specification of the working correlation R_i, as long as the number of clusters is sufficiently large (I 30), while the consistency of the model-based variance requires the correct specification of the correlation structure. As $r_i(\hat{\theta})$ is biased toward zero with a limited number of clusters, BC0 is likely to underestimate the variance and alternative choices of matrices C_i and B_i maybe necessary to provide a partial correction to the finite-sample bias.[27] We consider three popular approaches for bias corrections summarized as follows and also in Table 3: the finite-sample correction due to the work of Kauermann and Carroll,[28] or BC1, is given by $C_i = I$ and $B_i = (I - H_i)^{-1/2}$; the finite-sample correction due to the work of Mancl and DeRouen,[29] or BC2, is given by $C_i = I$ and $B_i = (I - H_i)^{-1}$; and the finite-sample correction due to the work of Fay and Graubard,[30] or BC3, given by $C_i = \text{diag}\left\{ \left(1 - \min\left\{ \zeta, \left[D_i' V_i^{-1} D_i \Omega_1^{-1} \right]_{jj} \right\} \right)^{-1/2} \right\}$ and $B_i = I$, where the bound parameter ζ is a

user-defined constant (< 1) with a default value 0.75. Because the matrix elements of the cluster leverage are between 0 and 1, we generally have BC0 < BC1 < BC2 in terms of the amount of correction.[25] Furthermore, Scott et al[31] have shown that BC3 tends to be close to BC1. Of note, the estimation of dispersion parameter should only affect the model-based variance. Similar to the work of Liang and Zeger,[19] we propose to consistently update the dispersion parameter from iteration s to $s + 1$ by

$$\hat{\phi}^{(s+1)} = \hat{\phi}^{(s)} \sum_{i=1}^{I} \operatorname{tr}(\tilde{R}_i) / \left\{ \sum_{i=1}^{I} TN_i - (T+1) \right\}.$$

4 | DESIGN CONSIDERATIONS: SAMPLE SIZE AND POWER CALCULATIONS

4.1 | Closed-form variance of the intervention effect

Under the null hypothesis H_0: $\delta = \delta_0$, the large-sample variance of $\sqrt{n}(\hat{\delta} - \delta_0)$ is provided by the ($T + 1$, $T + 1$) element of the large-sample covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$. Since the QLS estimator $\hat{\delta}$ is asymptotically normal, we could use the z-test statistic $\hat{\delta}/\sqrt{\operatorname{var}(\hat{\delta})}$ to test the null of no intervention effect H_0: $\delta = 0$, and the power to detect an intervention effect of size $\delta \neq 0$ with a prescribed type I error rate α is approximately

$$\text{power} = \Phi\left(z_{\alpha/2} + \frac{|\delta|}{\sqrt{\operatorname{var}(\hat{\delta})}}\right), \tag{11}$$

where Φ is the standard normal cumulative distribution function and $z_{\alpha/2}$ is the normal quantile such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Because there is uncertainty in estimating the asymptotic variance $\operatorname{var}(\hat{\delta})$, an alternative two-sided test uses the same statistic but refers to the t-distribution. We consider two choices of degrees of freedom (DoF). The first DoF dates back to the work of Mancl and DeRouen[29] and equal to the number of clusters minus the number of regression parameters; this DoF has been previously used in the GEE analyses of parallel CRTs,[32] three-level CRTs,[33] crossover CRTs,[6] stepped wedge CRTs,[5,34] and shown to have test size not exceeding the nominal level. The second DoF was suggested in the PhD dissertation of Ford[35] and specifies DoF $= I - 2$. This choice of DoF was found to provide excellent control of type I error rate for GEE analyses of both parallel CRTs and stepped wedge CRTs in the work of Ford and Westgate.[35,36] With the same effect size δ and prescribed type I error rate α, the power of the t-test is approximately

$$\text{power} = \Psi_{\text{DoF}}\left(t_{\alpha/2, \text{DoF}} + \frac{|\delta|}{\sqrt{\operatorname{var}(\hat{\delta})}}\right), \tag{12}$$

where Ψ_{DoF} is the t distribution function and the quantile $t_{\alpha/2}$ is chosen such that $\Psi_{\text{DoF}}(t_{\alpha/2}) = 1 - \alpha/2$. We notice that, because the t-distribution has a heavier tail compared with the standard normal distribution, the QLS z-test is more likely to result in an inflated type I error rate with the use of BC0 than the corresponding QLS t-test. As the bias-corrected sandwich variance estimators (BC1, BC2, and BC3) provide different degrees of inflation relative to the uncorrected variance BC0, an investigation of both the z- and t-tests coupled with the

collection of alternative variance estimators could help inform the practical choice among the analytical options for stepped wedge CRTs.

To assist the design of stepped wedge trials allowing for correlation decay, we derive a new closed-form variance expression for $\hat{\delta}$ assuming the outcome is continuous and g is the identity link function. We will return to categorical outcomes and nonlinear link functions in Section 7. To do so, we follow the work of Shih[37] and assume the covariance of Y_i to be known as $\text{var}(Y_i) = Vi$. Therefore, $\text{var}(\hat{\delta})$ is the $(T+1, T+1)$ element of the model-based variance Ω_1^{-1}. We further assume a balanced design such that an equal number of participants will be recruited in each cluster prior to the first period, so that $N_i = N$. Such a simplification assumption is routinely made in designing stepped wedge trials.

Under a balanced design, we could write the design matrix corresponding to cluster i as $Zi = 1N \otimes (I_T, X_i)$, where 1_N is a N-vector of ones. Then, the variance of the intervention effect $\hat{\delta}$ is equal to the lower-right element of $\phi\left\{\sum_{i=1}^I Z_i' R_i^{-1}(\tau, \rho) Z_i\right\}^{-1}$, where ϕ is the marginal variance. We show in Web Appendix C that a closed-form variance expression for $\hat{\delta}$ is

$$\text{var}(\hat{\delta}) = \frac{(\phi I/N)\left(1 - \rho^2\right)\{1 + (N-1)\tau\}}{(IU - W)\left(1 + \rho^2\right) - 2(IV - Q)\rho}, \tag{13}$$

where the design constant $U = \sum_{i=1}^I \sum_{t=1}^T X_{it}$ is the total number of cluster periods exposed under the intervention condition, $W = \sum_{t=1}^T \left(\sum_{i=1}^I X_{it}\right)^2$ is the squared number of clusters receiving the intervention summed across periods, and $V = \sum_{i=1}^I \sum_{t=1}^T X_{it} X_{i,t+1}$ and $Q = \sum_{t=1}^{T-1} \left(\sum_{i=1}^I X_{it}\right)\left(\sum_{i=1}^I X_{i,t+1}\right)$ are cross-product terms resulting from the decay correlation structure. It is interesting to see that this variance expression does not depend on the magnitude of the period effect β_t as long as they are controlled for in the marginal mean model. Noticeably, the QLS-based variance (13) extends the formula due to the work of Liu et al[18] to longitudinal cluster designs with staggered randomization. Furthermore, as the cohort size N becomes large, the variance expression converges to

$$\lim_{N \to \infty} \text{var}(\hat{\delta}) = \frac{\phi I\left(1 - \rho^2\right)\tau}{(IU - W)\left(1 + \rho^2\right) - 2(IV - Q)\rho}, \tag{14}$$

which is a finite constant since $|\rho| < 1$ and $\tau > 0$ for large N, according to (2). Therefore, the limit of the variance is a positive constant determined by available design resources I, T, and two correlation values, τ and ρ, and cannot be made arbitrarily small. In other words, the power of the stepped wedge design may not be increased to one by solely increasing the cohort size, which is consistent with the known results for parallel cluster randomized designs.[1] For this reason, when N is large, variance (14) could be used in the design stage to approximate the variance (13). Finally, given hypothesized values for I, N, T, and correlation parameters, variance expression (13) or (14) can be used in Equations (11) and (12) to obtain the predicted power.

4.2 | The design effect

For determining the required sample size based on Equations (11) and (12), it is straightforward to solve N by fixing the required number of clusters I but not the other way around. However, with a predetermined cohort size N for each cluster, we could postulate a series of values for I and find the smallest value such that the estimated power is at least equal to the prescribed level. Additionally, in the following case studied by Woertman et al,[22] we could derive a simple expression for the design effect (DE) relative to an individually randomized trial to simplify sample size calculation. Specifically, we assume that an equal number of clusters switch to intervention at each step so that $m_s = m$, and, furthermore, an equal number of measurements are taken between steps such that $c_s = s$ for all $s = 1, \ldots, S$. We then write the total number of clusters $I = Sm$ and total number of periods $T = b + Sc$, and the design constants become

$$U = \frac{1}{2}S(S+1)mc, \quad W = \left(\frac{1}{3}S^3 + \frac{1}{2}S^2 + \frac{1}{6}S\right)m^2c,$$
$$V = U - Sm, \quad Q = W - \frac{1}{2}S(S+1)m^2.$$

Plugging the design constants back into the variance formula (13) and dividing by the variance of the two-sample mean difference $4\phi/(NSm)$, we obtain

$$\text{DE} = \frac{3S}{2(S-1)} \frac{\left(1 - \rho^2\right)}{\left\{(S+1)c(1-\rho)^2 + 6\rho\right\}} \{1 + (N-1)\tau\}. \tag{15}$$

The aforementioned DE allows us to easily study how the design resources affect the statistical efficiency relative to individual randomization and how the correlation parameters affect the statistical power. For example, since the DE is free of b, the relative design efficiency does not change according to the number of baseline periods. However, for fixed values of the correlation parameters, increasing the number of steps S and number of measurements between steps c decreases the DE and increases the efficiency. On the other hand, for fixed design resources, larger values of the within-period correlation τ increases the DE, confirming that τ functions as the traditional ICC of a parallel CRT. By contrast, the role of correlation parameter ρ is characterized by $f(\rho) = (1 - \rho^2)/[(S+1)c(1-\rho)^2 + 6\rho]$, which is monotonically increasing on $(-1, r)$ and decreasing on $(r, 1)$, where

$$r = 1 + \frac{\sqrt{3}\{\sqrt{3} - \sqrt{2(S+1)c - 3}\}}{(S+1)c - 3} \in (0, 1).$$

For convenience, we can also define the decay parameter $d = 1 - \rho$ so that $d = 0$ and $d = 1$ correspond to no decay and total decay, respectively. Since it is more plausible that $\rho \in (0, 1)$, the aforementioned result suggests that, with an increasing level of correlation decay, the DE first increases to its largest value and then decreases, with the maximum DE obtained at $\rho = r$. A numerical illustration of the DE as a function of the decay parameter is provided in Web Figure 2.

5 | A SIMULATION STUDY

5.1 | Simulation design

We carry out a simulation study (1) to compare the correlation estimates from the uncorrected QLS and the proposed matrix-adjusted QLS (MAQLS), and (2) to evaluate the utility of the proposed power formula for QLS-based analyses of stepped wedge CRTs. For the second objective, we first determine the empirical type I error rates for the QLS-based tests coupled with alternative variance estimators, and then identify valid tests (those with a close-to-nominal type I error rate) whose empirical power corresponds well with the predicted power from the proposed formula. Findings specific to the second objective are informative for practical data analysis since we prefer tests that maintain a valid size and meanwhile demonstrate empirical power that is at least the magnitude of the analytical prediction.

Within-cluster correlated continuous outcomes were generated from a multivariate normal distribution with mean given by $\mu_{ijt} = \beta_t + X_{it}\delta$ and covariance $\phi R(\tau, \rho)$, where $R(\tau, \rho)$ is the proportional decay structure defined in Section 3.1. We set the marginal variance $\phi = 1$ and assumed a gently increasing period effect such that $\beta_1 = 0$ and $\beta_{t+1} - \beta_t = 0.1 \times (0.5)^{t-1}$ for t 1. As discussed before, the predicted power should be insensitive to the magnitude of the period effects as long as they are accounted for in the QLS analyses. We fix the effect size $\delta/\phi^{1/2}$ at zero for studying test size and choose $\delta/\phi^{1/2}$ from {0.2, 0.3, 0.4, 0.5} for studying power. We choose the within-period correlation $\tau \in \{0.03, 0.1\}$, which represent typical ICC values reported in the parallel CRTs.[1] We further chose $\rho \in \{0.2, 0.8\}$, representing large and moderate degree of correlation decay over time. The number of clusters is varied from 9 to 24 as stepped wedge CRTs usually include a limited number of clusters. We specify the number of periods 3 T 8 as these values are frequently reported in practice, according to recent reviews by the works of Martin et al[38] and Grayling et al.[39] Finally, the cohort size is chosen as 5 N 24 to ensure that the predicted power is at least 80%. For illustration, we focus on standard stepped wedge designs so that there is only one baseline period and the number of steps $S = T - 1$. In other words, an equal number of I/S clusters cross over to intervention during each step, and the outcome is measured only once for each individual between consecutive steps. For each scenario, 10 000 data replications were generated and analyzed using both QLS and MAQLS. For the first objective, we report the percent relative bias in estimating τ and ρ. In general, an unbiased approach for estimating the correlation parameters is preferred since accurate reporting of correlations is critical for planning future trials. Web Tables 1 and 2 provide a summary of the simulation scenarios along with the convergence rates. The convergence rates are similar between QLS and MAQLS, and all exceed 97%. For the second objective, we consider both the z-tests and the t-tests for testing the null hypothesis of no intervention effect, coupled with five different variance estimators for $\hat{\delta}$, namely, the model-based variance, BC0, BC1, BC2, and BC3. The nominal type I error rate is held fixed at 5%, and we consider an empirical type I error rate between 4.5% and 5.5% to be acceptable based on the margin of error from a binomial model with 10 000 replications. By a similar reasoning, since the predicted power in each scenario is at least 80%, we consider an empirical power that differs by no more than 0.8% from the predicted value to be acceptable.

5.2 | Results

For the first objective, we summarize in Table 4 and Web Table 3 the percent relative bias in estimating the correlations with QLS and MAQLS. It is evident that the percent bias in estimating the within-period correlation τ is much larger than that in estimating the correlation parameter ρ, without respect to the incorporation of matrix adjustment to the first-stage estimating equations (4) and (5). However, the QLS estimator for τ exhibits noticeable downward bias, especially when the number of clusters is not large. By contrast, MAQLS substantially reduces such finite-sample bias and improves the estimation of τ. On the other hand, the QLS estimator for the parameter ρ seems more accurate in that the absolute percent bias only occasionally exceeds one. Nevertheless, MAQLS still mildly improves the estimation of ρ in that the absolute percent bias is always maintained under one. The comparative findings between QLS and MAQLS are consistent regardless of the magnitude of intervention effect δ. Therefore, MAQLS is the preferred approach because it provides much less biased estimates for the correlation parameters; these more accurate correlation estimates will eventually facilitate accurate estimation of sample size and power for future cohort stepped wedge trials.

For the second objective, we present the empirical type I error rates of the z-tests and t-tests for the QLS and MAQLS analyses in Web Tables 4 to 6 and Web Figure 3. Overall, we observe that the matrix adjustment to the correlation estimation mildly affects the tests with the model-based variance but has little impact on the tests with the sandwich variance. This is in accordance with the work of Lu et al,[40] who observed the same results for the GEE analyses of pretest-posttest CRTs. Since MAQLS provides more accurate estimation of the correlations, we will focus on this approach. Table 5 summarizes the empirical type I error rates of the MAQLS z-tests and t-tests with DoF $= I - 2$. We leave the results for t-tests with DoF $= I - (T + 1)$ to Web Table 6 as these tests are conservative in many cases. From Table 5, we observe that MAQLS z-tests are more liberal than the corresponding MAQLS t-tests. The type I error rates of the MAQLS z-tests coupled with the model-based variance or BC2 are close to nominal when I 20, while the MAQLS z-tests coupled with BC0, BC1, or BC3 are always liberal. By contrast, only the MAQLS t-tests with BC0 remain liberal, while MAQLS t-tests with BC1 or BC3 maintain close-to-nominal size and MAQLS t-tests with model-based variance or BC2 are conservative. Overall, the t-tests with DoF $= I - 2$ and BC1 demonstrate test sizes that consistently agree with the nominal level.

Web Tables 7 to 9 and Web Figure 4 present the predicted and empirical power for each simulation scenario. Because we are only interested in tests that maintain close-to-nominal sizes, we summarize in Table 6 the differences between the empirical and the predicted power only for the MAQLS z-tests and t-tests with DoF $= I - 2$. Among the z-tests, only the choice of BC2 provides substantially lower power than predicted. While the choices of model-based variance BC1 and BC3 provide adequate power for the z-tests in a number of scenarios, one should be cautious in adopting these tests with a small number of clusters since they may carry an inflated test size. On the other hand, the empirical power for MAQLS t-tests coupled with the model-based variance, BC1 or BC3 corresponds reasonably well with the analytical prediction from the proposed formula even for as few as nine clusters, while the empirical power for MAQLS t-tests with BC2 still tends to be

substantially lower than predicted. Interestingly, the MAQLS *t*-test with DoF = $I - (T + 1)$ coupled with model-based variance, BC1 or BC3 also demonstrates empirical power fairly close to prediction, even though these tests are more conservative under the null. These tests may not be preferred over the *t*-tests with DoF = $I - 2$ since they are less powerful.

6 | NUMERICAL ILLUSTRATIONS

6.1 | The AEP study

We illustrate the proposed sample size procedure to design a cohort stepped wedge CRT that aims to study the effect of an exercise intervention on the physical function of patients with end-stage renal disease.[41] The intervention was an accredited exercise physiologist (AEP) coordinated resistance exercise program, offered at hemodialysis clinics to improve the quality of life for dialysis patients. During the planning phase, it was determined that $I = 15$ clinics (clusters) were available, and would be randomized over $T = 4$ periods evenly spaced across 48 weeks. At baseline, no exercise programs were offered to any clinic. At week 12, 36 and 48, a random subset of 5 clinics cross over from control to intervention. A closed cohort of patients was recruited at baseline, and would be followed up during the study period. The primary patient-level outcome was the 30-second sit-to-stand test, recording the number of times a patient could rise from and return to a seated position in a 30-second time frame. The 30-second sit-to-stand test was conducted at the end of each period, resulting in four outcome measurements per patient. Based on a prior study within a similar context, a conservative estimate of the effect size was given by $\delta/\phi = 0.325$,[42] and the within-period correlation was estimated to be $\tau = 0.03$.[43] With $I = 15$ clusters, the simulations suggest the MAQLS *t*-test with DoF = $I - 2 = 13$ could maintain nominal size and adequate power; we illustrate the power calculation based on the *t*-test statistic.

Given this is a standard stepped wedge design where an equal number of clinics switch to intervention at each step, we can show that $U = IT/2$, $W = I^2T(2T - 1)/\{6(T - 1)\}$, $V = I(T - 2)/2$ and $Q = I^2T(T - 2)/\{3(T - 1)\}$. The variance expression (13) is then simplified to

$$\text{var}(\hat{\delta}) = \frac{6(\phi/N)(T - 1)\left(1 - \rho^2\right)\{1 + (N - 1)\tau\}}{I(T - 2)\left\{T(1 - \rho)^2 + 6\rho\right\}}. \tag{16}$$

If we anticipate large correlation decay so that $\rho = 0.2$, the power is estimated using Equations (11) and (16) to be 79.4% if $N = 21$ and 80.5% if $N = 22$. Therefore, at least $N = 22$ patients should be recruited in each clinic to achieve 80% power under the proportional decay structure. On the other hand, we could arrive at the same results by using the DE formula (15). For example, in an individual randomized study, 348 patients would be required for the hypothesized effect size. Assuming 21 patients will be included in each clinic, the DE is approximately 0.92, indicating a total of 320 patients in approximately 15.2 clinics would be required. Since the study affords to randomize only 15 clinics, we increase the cohort size to $N = 22$, resulting in a DE 0.94. Therefore, 326 patients are required for a total of $326/22 \approx 14.8$ clinics, and we conclude that 22 patients in 15 clinics ensured 80% power.

While the within-period correlation estimate was available from prior studies, published estimates of the correlation parameter ρ (or decay parameter $d = 1 - \rho$) are currently rare. For this reason, we carry out a sensitivity analysis on the power and present the results in panel (A) of Figure 2, where we fix the design resources but vary $\tau \in (0.03, 0.06)$ and $\rho \in (0, 1)$. Note that the upper bound of the within-period correlation τ was reported by the work of Littenberg and MacLean[43] and is used in this assessment. As expected, larger values of the within-period correlation reduce the study power, and furthermore, given a certain value of the within-period correlation, a greater magnitude of decay (smaller ρ or larger d) generally reduces the study power unless $\rho \approx 0$ (or $d \approx 1$). For the hypothesized $\tau = 0.03$, the study power remains at least close to 80% regardless of the correlation decay. On the other hand, the amount of correlation decay could result in further power loss if the within-period correlation τ increases. Nevertheless, the power loss is at most around 10% even if the within-period correlation τ approximates the upper bound 0.06.

6.2 | The CORE study

We next illustrate the proposed sample size procedure by designing the CORE stepped wedge trial. The CORE study aims to evaluate the patient-centered service design in health providers to improve the psychosocial recovery outcomes for people with severe mental illness in Australia.[44] The new service design intervention adopted the experience-based co-design to identify users' positive and negative experiences of the service, and involved patients' participation to co-design solutions to the negative experiences. A total of $I = 11$ teams from four health service providers would be participating the study; each team involved a number of service users who will be affected by the intervention. A stepped wedge design was considered appropriate for the study due to logistical constraint in simultaneously introducing the intervention to more than a few teams. The experience-based co-design intervention will be delivered to the clusters in three waves, each with a duration of 9 months. Four teams will start the intervention in wave 1 and wave 2, respectively, while the remaining three teams receive the intervention in the final wave. In other words, the study includes four periods, with a baseline period lasting about 6 months.

The outcome of interest is the improvement in psychosocial recovery measured by the recovery assessment scale revised,[45] and was measured for each user at the end of baseline period and each of the three follow-up period. The standardized effect size on the psychosocial recovery outcome was estimated to be 0.35, and the within-period correlation was assumed to be $\tau = 0.1$.[44] Since the study affords to randomize only 11 clusters, there may be a risk of inflated type I error rate with a z-test. As the t-test with DoF $= I - 2 = 9$ performs best with respect to empirical size and power in the simulations, we determine the required cohort size based on a test using expressions (12) and (13). Assuming the correlation decay is only moderate so that $\rho = 0.8$, power is estimated to be 0.79 for $N = 8$ and 0.81 for $N = 9$, barring drop out. Therefore, $N = 9$ is required to ensure 80% power given a 5% test size. We further conducted a sensitivity analysis to see how power changes according to the degree of correlation decay, and presented the power contour in panel (B) of Figure 2. Due to the small sample size and the heavy tail of the t distribution, the study is sensitive to correlation decay when $\tau = 0.1$, and remains so even if τ approaches zero. On the other hand, the actual study planned to recruit $N = 30$ users in each team. With this larger

cohort size, panel (C) of Figure 2 suggests that the power becomes less sensitive to the correlation decay, especially as the within-period correlation approaches zero. For example, if τ 0.02, the study power remains at least around 80% regardless of the amount of correlation decay.

6.3 | Consequences of specifying a nondecay correlation structure

For GEE analyses of cohort stepped wedge designs, Li et al[5] recently developed the block exchangeable correlation structure for estimating the sample size and power. With a continuous outcome and identity link function, the block exchangeable correlation structure is also implied by the linear mixed effects model for cohort studies discussed in the works of Hooper et al,[8] and Girling and Hemming.[46] To understand the implications of alternative correlation models, we compare the variance of the intervention effect estimator obtained under the proportional decay correlation structure to that obtained under the block exchangeable correlation structure. Specifically, both the proportional decay structure and the block exchangeable correlation structure assume a constant within-period correlation, ie, $\mathrm{corr}(y_{ijt}, y_{ij't}) = \tau$ for j j'. However, the latter correlation model also assumes constant between-period and within-individual correlations such that $\mathrm{corr}\left(y_{ijt}, y_{ij't'}\right) = \alpha_1^{\mathrm{BE}}$ for j j', t t' and $\mathrm{corr}\left(y_{ijt}, y_{ijt'}\right) = \alpha_2^{\mathrm{BE}}$, t t'. To focus ideas, we carry out the comparisons based on the standard stepped wedge designs with a single baseline period and an equal number of clinics switching to intervention at each step. We refer to the variance of $\hat{\delta}$ obtained under the proportional decay structure as $\mathrm{var}^{\mathrm{PD}}(\hat{\delta})$, whose expression is provided in (16). We refer to the variance of $\hat{\delta}$ obtained under the block exchangeable structure as $\mathrm{var}^{\mathrm{BE}}(\hat{\delta})$, whose expression is derived in the work of Li et al[5] as

$$\mathrm{var}^{\mathrm{BE}}(\hat{\delta}) = \frac{12(\phi/N)(T-1)\lambda_3\lambda_4}{I(T-2)\{(T-1)\lambda_3 + (T+1)\lambda_4\}}, \tag{17}$$

with $\lambda_3 = 1 + (N-1)\left(\tau - \alpha_1^{\mathrm{BE}}\right) - \alpha_2^{\mathrm{BE}}$ and

$\lambda_4 = 1 + (N-1)\tau + (T-1)(N-1)\alpha_1^{\mathrm{BE}} + (T-1)\alpha_2^{\mathrm{BE}}$ as the two distinct eigenvalues of the block exchangeable matrix. If we define function $h\left(\alpha_1^{\mathrm{BE}}, \alpha_2^{\mathrm{BE}}\right) = (N-1)\alpha_1^{\mathrm{BE}} + \alpha_2^{\mathrm{BF}}$, we can write the relative variance as

$$\frac{\mathrm{var}^{\mathrm{PD}}(\hat{\delta})}{\mathrm{var}^{\mathrm{BE}}(\hat{\delta})} = \frac{\left(1 - \rho^2\right)}{2\left\{T(1-\rho)^2 + 6\rho\right\}}\left[\frac{(T-1)\{1 + (N-1)\tau\}}{1 + (N-1)\tau + (T-1)h\left(\alpha_1^{\mathrm{BE}}, \alpha_2^{\mathrm{BE}}\right)} + \frac{(T+1)\{1 + (N-1)\tau\}}{1 + (N-1)\tau - h\left(\alpha_1^{\mathrm{BE}}, \alpha_2^{\mathrm{BE}}\right)}\right]. \tag{18}$$

For each value of the within-period correlation τ and each value of the decay parameter $d = 1 - \rho$ in the proportional decay model, there may exist pairs of values for $(\alpha_1^{\mathrm{BE}}, \alpha_2^{\mathrm{BE}})$ that result in the same variance of the intervention effect. Obtaining the equal variance correspondence is equivalent to finding the straight line $h\left(\alpha_1^{\mathrm{BE}}, \alpha_2^{\mathrm{BE}}\right) = \eta$ that solves

$\text{var}^{\text{PD}}(\hat{\delta})/\text{var}^{\text{BE}}(\hat{\delta}) = 1$. Because the relative variance is quadratic in $h\left(\alpha_1^{\text{BE}}, \alpha_2^{\text{BE}}\right)$, such a line may not always exist within the plausible range of $(\alpha_1^{\text{BE}}, \alpha_2^{\text{BE}})$ that ensures a positive definite block exchangeable correlation matrix. We confirm this observation by plotting the relative variance for varying correlation parameters. Fixing $\tau = 0.03$, $T = 4$ and $N = 20$ similar to the accredited exercise physiologist study, we present in Figure 3 the contour of $\text{var}^{\text{PD}}(\hat{\delta})/\text{var}^{\text{BE}}(\hat{\delta})$ over the $(\alpha_1^{\text{BE}}, \alpha_2^{\text{BE}})$ plane for each level of decay $d \in \{0, 1, \ldots, 0.9\}$. The dashed thick line indicates the equal variance correspondence. We also present the contour plots by specifying $N = 100$, $T = 8$, and $\tau = 0.1$ in Web Figures 5 to 7. It is evident that the existence and location of the equal variance line on the $(\alpha_1^{\text{BE}}, \alpha_2^{\text{BE}})$ plane depends on the value of correlation decay, cohort size and number of periods. If the equal variance line exists, the value of τ and number of periods only affects its location while cohort size N further affects its slope, as expected from inspecting expression (18). Apart from the equal variance correspondence, the two variances are generally different. Depending on the gray shades (colored shades in the online version), either the proportional decay or the block exchangeable structure will lead to a larger variance of the intervention effect and require a larger sample size. As a result, using the block exchangeable correlation structure in the presence of correlation decay could lead to an overestimation or underestimation of the true intervention effect variance, and vice versa. Particularly, the variances returned from these two correlation models become close when $\alpha_1^{\text{BE}}, \alpha_2^{\text{BE}}$ approximate zero and when the decay parameter, $d = 1 - \rho$, approximates one. These two restrictions result in many nearly-zero entries in the block exchangeable and proportional decay correlation matrices, increasing the dependence of sample size estimation on the within-period correlation τ. Therefore, it is anticipated that the variances from the two models become similar in this particular scenario, even though in general there could be large differences between $\text{var}^{\text{PD}}(\hat{\delta})$ and $\text{var}^{\text{BE}}(\hat{\delta})$. To summarize, our key message for cohort stepped wedge designs echo the principal findings in the work of Kasza et al[10] for cross-sectional designs. It is possible to grossly overestimate or underestimate the variance of the intervention effect if the correlation model is misspecified, except in restrictive scenarios. In practice, researchers could investigate the sensitivity of sample size estimates to misspecification of the correlation structure when there is limited preliminary data at the design stage of the trial.

7 | DISCUSSION

This article expanded on the design and analysis considerations for cohort stepped wedge CRTs in the presence of correlation decay. Since a cohort design involves repeated outcome assessments for fixed sets of individuals, we adopted the proportional decay structure of Lefkopoulou et al[16] to characterize the within-cluster correlations among the outcome measurements. Based on a marginal mean model accounting for the treatment and period effects, we developed a new sample size and power procedure to design stepped wedge CRTs accounting for such correlation decay. To apply this procedure, a key step is to obtain reasonable values for the correlation parameters. The within-period correlation, τ, is similar to the traditional ICC in a parallel CRT, and may often be found in previous studies with a similar endpoint. By contrast, the correlation parameter, ρ (or decay parameter $d = 1 - \rho$), is

not as commonly reported in the literature, and therefore the sensitivity of power should be investigated across a range of values for ρ, as illustrated in Section 6. Given that accurate reporting of correlations is vitally important for designing future stepped wedge trials, we also provided an improved MAQLS approach to estimate the correlation parameters along with the marginal mean parameters. The MAQLS has little impact on the estimation of the marginal mean parameters and the associated statistical tests coupled with the sandwich variance, but it substantially reduces the bias in estimating the within-period correlation τ and mildly improves the estimation of ρ, as confirmed in our simulation study.

In our simulation study with a small number of clusters, we found that, regardless of choice of variance estimators, the t-tests provide better control of the type I error rates compared to the z-tests, which are often liberal. Regardless of the two choices of DoF, the t-tests coupled with the model-based variance, BC1 and BC3 preserve the nominal size and demonstrate empirical power that agrees well with analytical prediction. Since the t distribution with a smaller DoF has a heavier tail, which implies a smaller power under the alternative, we prefer the t-tests with DoF $= I - 2$ for the design and analysis of cohort stepped wedge CRTs under the proportional correlation decay structure. An additional piece of evidence that favors DoF $= I - 2$ is found in the recent simulation study by Ford,[35] who showed that the GEE t-tests with DoF $= I - 2$ provide satisfactory control of type I error rates even for $I = 6$ but in the absence of correlation decay. Additional work is needed to investigate whether these extremely small sample sizes could provide adequate power under the proportional decay structure. On the other hand, although the t-tests with DoF $= I - (T + 1)$ could provide adequate power using the model-based variance, BC1 or BC3, they are frequently conservative under the null with a small number of clusters. In fact, one needs a minimum of $T+2$ clusters to provide at least one DoF, which further limits the applications of such a t-test in the design and analysis of small stepped wedge trials.

Recent reviews of stepped wedge CRTs[9,38,39] indicated that both the cross-sectional and cohort designs were common in practice. Although we have developed the design and analysis strategies specifically for cohort stepped wedge CRTs, a parallel discussion for cross-sectional stepped wedge CRTs could be equally informative. As discussed in Section 3.1, the exponential decay structure is often used to model correlation decay in multiperiod CRTs with repeated cross-sectional samples.[10] Assuming there are N_i observations in each period for each cluster, we could write the exponential decay structure as $L_i(\tau, \rho) = (1 - \tau)I_{TNi} + \tau J_N \otimes R(\rho)$ without changing the interpretation of τ and ρ from the cohort setting. To estimate the intervention effect and correlation parameters, the MAQLS approach could still be applied once we replace the second-stage estimating equations 6 and (7) by

$$\sum_{i=1}^{I} \text{tr}\left\{\frac{\partial L_i^{-1}(\hat{\alpha}_0, \hat{\alpha}_1)}{\partial \alpha_0}L_i(\tau, \rho)\right\} = 0,$$

$$\sum_{i=1}^{I} \text{tr}\left\{\frac{\partial L_i^{-1}(\hat{\alpha}_0, \hat{\alpha}_1)}{\partial \alpha_1}L_i(\tau, \rho)\right\} = 0.$$

Such modifications are necessary because τ and ρ are no longer separable in $L_i(\tau, \rho)$, and hence, updates for τ and ρ do not come in closed forms. Correspondingly, the inseparability between τ and ρ also precludes the derivation of an analytical inverse $L_i^{-1}(\tau, \rho)$, and therefore, one may not be able to obtain a simple algebraic expression for $\text{var}(\hat{\delta})$. As a result, sample size and power calculation requires numerically inverting the correlation matrix $L_i(\tau, \rho)$. In fact, with a continuous outcome and the identity link, it is straightforward to show that the QLS-based sample size procedure with $L_i(\tau, \rho)$ coincides with the mixed-effects model-based sample size procedure developed in the work of Kasza et al[10] with exponential correlation decay.

We have assumed that each individual has complete follow-up during the study. In reality, individual dropout may be anticipated and could be accounted for in the design phase. Given an expected attrition rate γ, a simple and commonly used strategy is to inflate the required total sample size by $1/(1 - \gamma)$, so that a complete-trajectory GEE analysis may provide adequate power if the drop-out or missingness is completely at random.[47] More sophisticated approaches that deal with monotone missingness have been studied for repeated-measure randomized trials and may be adapted to the stepped wedge context by considering staggered treatment assignments and appropriate levels of clustering.[48,49] In any event, trial implementation methodologies to prevent attrition bias in longitudinal CRTs merit further investigation.[50]

One simplification we made in the sample size and power calculations was to assume equal cluster (cohort) sizes. It has been shown that cluster size imbalance leads to reduced power in parallel CRTs and therefore may be accounted for in the design phase.[51] For a stepped wedge trial, Girling[52] computed the relative efficiency of unequal versus equal cluster sizes by assuming a linear mixed-effects model without correlation decay. It was concluded that the efficiency loss due to unequal cluster sizes is unlikely to exceed 12% across a wide range design of resources and correlation values. Nevertheless, a corresponding expression for the relative efficiency accounting for correlation decay is currently not available and should merit additional study. The availability of such expressions for relative efficiency could inform the amount of additional design resources required to compensate the efficiency loss due to unequal cluster sizes. Another limitation of our design strategy is that we have assumed the proportional decay correlation is the correctly specified within-cluster dependency structure. However, both the QLS or MAQLS estimators for the intervention effect remain consistent even if the correlation structure is misspecified. If it is anticipated in the design phase that the working correlation may be misspecified, one could follow the general idea of Rochon[48] to develop a modified sample size procedure based on the sandwich variance.

Finally, we have assumed a continuous outcome and an identity link function, corresponding to the scenarios of the illustrative examples. In practice, categorical outcomes could be collected in cohort stepped wedge designs. Under correlation decay, we could extend the QLS-based sample size procedure to accommodate binary and count outcomes by following the steps outlined in section 3.2 of the work of Li et al.[5] In those cases, a further complication is that the variance is an explicit function of the marginal mean, and so the

magnitude of the period effects necessarily affects the variance for the intervention effect. Future work is required to investigate the operating characteristics of such extensions for calculating sample size and power with binary and count outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

REFERENCES

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

2. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Hoboken, NJ: Wiley; 2000.

3. Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of recent methodological developments in group-randomized trials: part 1–design. Am J Public Health. 2017;107(6):907–915. [PubMed: 28426295]

4. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28(2):182–191. [PubMed: 16829207]

5. Li F, Turner EL, Preisser JS. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. Biometrics. 2018;74(4):1450–1458. [PubMed: 29921006]

6. Li F, Forbes AB, Turner EL, Preisser JS. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. Statist Med. 2019;38(4):636–649.

7. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Statist Med. 2015;34(2):181–196.

8. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statist Med. 2016;35(26):4718–4728.

9. Barker D, McElduff P, D'Este C, Campbell MJ. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med Res Methodol. 2016;16(69):1–19. [PubMed: 26728979]

10. Kasza J, Hemming K, Hooper R, Matthews JNS, Forbes AB. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Methods Med Res. 2017;28(3):703–716. 10.1177/0962280217734981 [PubMed: 29027505]

11. Kasza J, Forbes AB. Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. Stat Methods Med Res. 2018;28(10–11): 3112–3122. 10.1177/0962280218797151 [PubMed: 30189794]

12. Grantham KL, Kasza J, Heritier S, Hemming K, Forbes AB. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. Statist Med. 2019;38(11):1918–1934. 10.1002/sim.8089

13. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. Designing a stepped wedge trial: Three main designs, carry-over effects and randomisation approaches. Trials. 2015;16(1).

14. Li F, Turner EL, Preisser JS. Optimal allocation of clusters in cohort stepped wedge designs. Stat Probab Lett. 2018;137:257–263.

15. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Statist Med. 2003;22(8): 1235–1254.

16. Lefkopoulou M, Moore D, Ryan L. The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. J Am Stat Assoc. 1989;84(407):810–815.

17. Shults J, Morrow AL. Use of quasi-least squares to adjust for two levels of correlation. Biometrics. 2002;58(3):521–530. [PubMed: 12229986]

18. Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. Statist Med. 2002;21(12):1787–1801.

19. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.

20. Chaganty NR. An alternative approach to the analysis of longitudinal data via generalized estimating equations. J Stat Plan Inference. 1997;63(1):39–54.

21. Chaganty NR, Shults J. On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. J Stat Plan Inference. 1999;76(1):145–161.

22. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol. 2013;66(7):752–758. [PubMed: 23523551]

23. Shults J, Hilbe JM. Quasi-Least Squares Regression. Boca Raton, FL: Chapman and Hall/CRC; 2014.

24. Preisser JS, Qaqish BF. Deletion diagnostics for generalised estimating equations. Biometrika. 1996;83(3):551–562.

25. Preisser JS, Lu B, Qaqish BF. Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. Statist Med. 2008;27(27):5764–5785.

26. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics. 1988;44(4):1033–1048. [PubMed: 3233244]

27. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of recent methodological developments in group-randomized trials: part 2–analysis. Am J Public Health. 2017;107(7):1078–1086. [PubMed: 28520480]

28. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc. 2001;96(456):1387–1396.

29. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57(1):126–134. [PubMed: 11252587]

30. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics. 2001;57(4):1198–1206. [PubMed: 11764261]

31. Scott JM, DeCamp A, Juraska M, Fay MP, Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Stat Methods Med Res. 2017;26(2):583–597. [PubMed: 25267551]

32. Li F, Turner EL, Heagerty PJ, Murray DM, Vollmer WM, Delong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. Statist Med. 2017;36:3791–3806.

33. 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. [PubMed: 20070297]

34. Grayling MJ, Wason JMS, Mander AP. Group sequential designs for stepped-wedge cluster randomised trials. Clinical Trials. 2017;14(5):507–517. [PubMed: 28653550]

35. Ford WP. Improved Standard Error Estimation for Maintaining the Validities of Inference in Small-Sample Cluster Randomized Trials and Longitudinal Studies [dissertation]. Lexington, KY: Theses and Dissertations–Epidemiology and Biostatistics, University of Kentucky; 2018.

36. Ford WP, Westgate PM. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. Biometrical Journal. 2017;59(3):478–495. [PubMed: 28128854]

37. Shih WJ. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. Biometrical Journal. 1997;39(8):899–908.

38. Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. BMJ Open. 2016;6:e010166.

39. Grayling MJ, Wason JMS, Mander AP. Stepped wedge cluster randomized controlled trial designs : a review of reporting quality and design features. Trials. 2017;18(33):1–13. [PubMed: 28049491]

40. Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala SI, Wolfson M. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics. 2007;63(3):935–941. [PubMed: 17825023]

41. Bennett PN, Daly RM, Fraser SF, et al. The impact of an exercise physiologist coordinated resistance exercise program on the physical function of people receiving hemodialysis: a stepped wedge randomised control study. BMC Nephrology. 2013;14(1):204–210. [PubMed: 24070232]

42. Cappy CS, Jablonka J, Schroeder ET. The effects of exercise during hemodialysis on physical performance and nutrition assessment. J Ren Nutr. 1999;9(2):63–70. [PubMed: 10089261]

43. Littenberg B, MacLean CD. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. BMC Med Res Methodol. 2006;6(1):20–30. [PubMed: 16672056]

44. Palmer VJ, Chondros P, Piper D, et al. The CORE study protocol: a stepped wedge cluster randomised controlled trial to test a co-design technique to optimise psychosocial recovery outcomes for people affected by mental illness in the community mental health setting. BMJ Open. 2015;5:e006688.

45. Lusczakoski KD, Olmos-Gallo PA, McKinney CJ, Starks R, Huff S. Measuring recovery related outcomes: a psychometric investigation of the recovery markers inventory. Community Ment Health J. 2014;50:896–902. [PubMed: 24838373]

46. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Statist Med. 2016;35(13):2149–2166.

47. Little RJ, Rubin DB. Statistical Analysis with Missing Data. 2nd ed.Hoboken, NJ: Wiley; 2002 Hardcover.

48. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Statist Med. 1998;17(14): 1643–1658.

49. Jung S-H, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. Statist Med. 2003;22(8): 1305–1315.

50. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials. 2015;16(1):351–361. [PubMed: 26278521]

51. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol. 2006;35(5):1292–1300. [PubMed: 16943232]

52. Girling AJ. Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. Statist Med. 2018;37(30):4652–4664.

	$t = 1$	$t = 2$ $s = 1$	$t = 3$ $s = 2$	$t = 4$ $s = 3$
cluster 1	0	1	1	1
cluster 2	0	1	1	1
cluster 3	0	0	1	1
cluster 4	0	0	1	1
cluster 5	0	0	0	1
cluster 6	0	0	0	1
	$b = 1$	$c_1 = 1$	$c_2 = 1$	$c_3 = 1$

FIGURE 1.
A schematic illustration of a standard stepped wedge design with $I = 6$ clusters and $T = 4$ periods. Each cell with a zero entry indicates a control cluster period and each cell with a one entry indicates an intervention cluster period
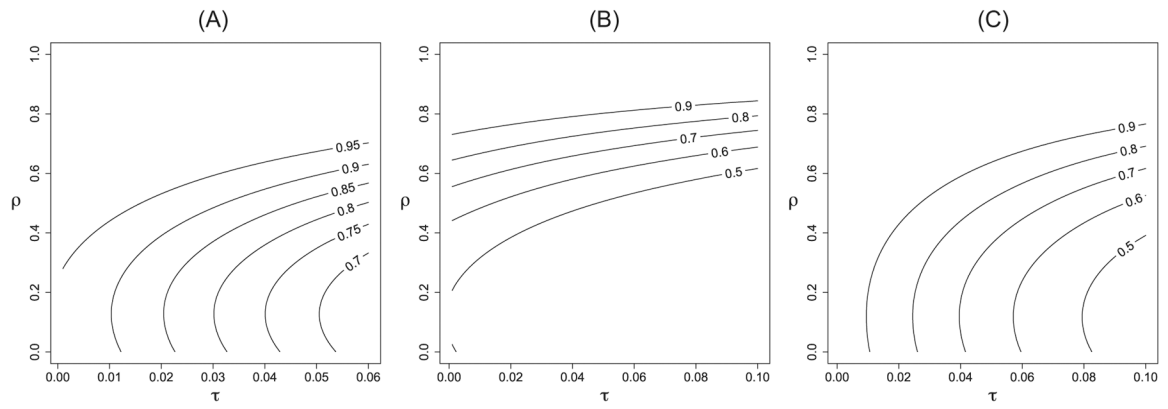
FIGURE 2.

Sensitivity analysis of study power for (A) the accredited exercise physiologist (AEP) study with $I = 15$ clusters and $N = 22$ individuals within each cluster, (B) the CORE study with $I = 11$ clusters and $N = 9$ individuals within each cluster, and (C) the CORE study with $I = 11$ clusters and $N = 30$ individuals within each cluster
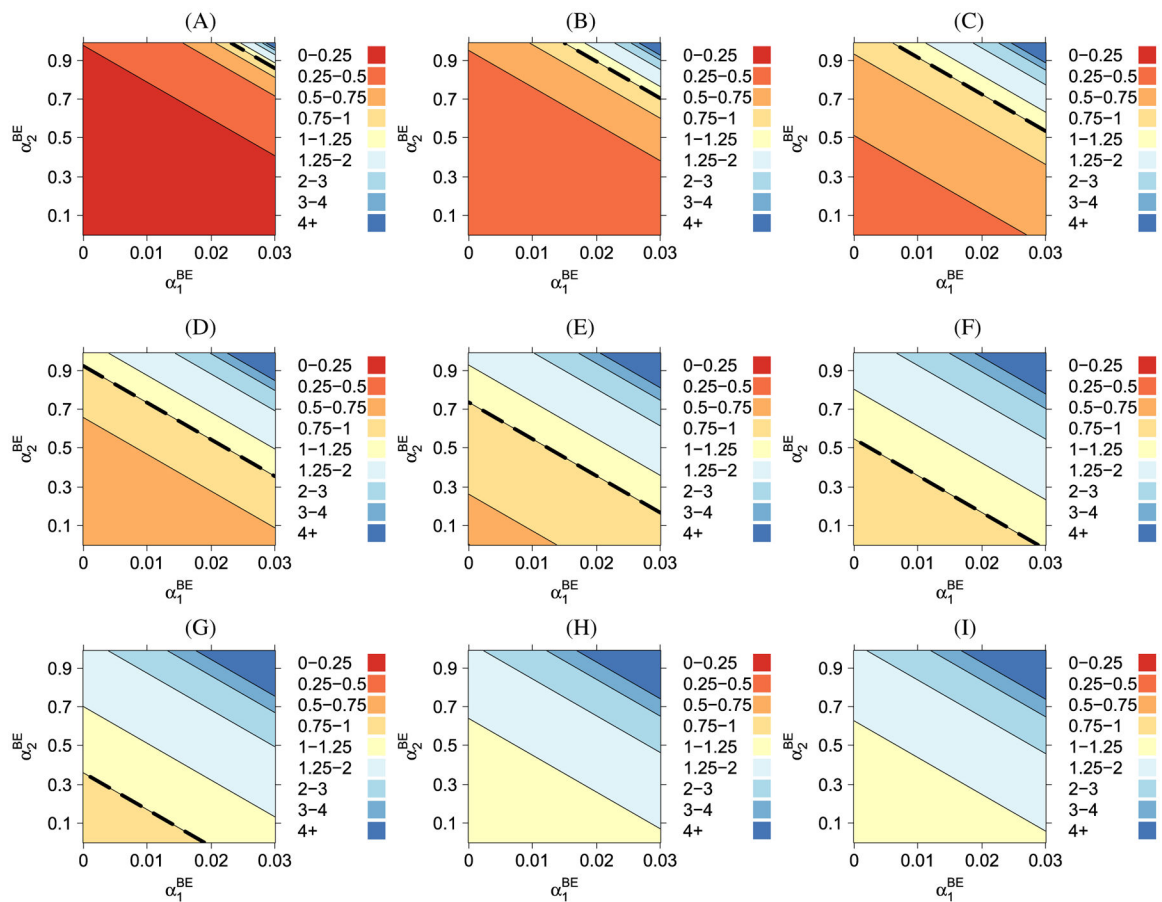
FIGURE 3.

Contour plots for the relative variance obtained under the proportional decay correlation model and the block exchangeable correlation model, for varying values of the proportional decay model decay parameter d and the block exchangeable model correlation parameters α_1^{BE}, α_2^{BE}. In all panels, the within-period correlation $\tau = 0.03$, the number of periods $T = 4$, and the cohort size $N = 20$. The dashed thick line indicates the equality of variances. A, $\tau = 0.03$ d = 0.1; B, $\tau = 0.03$ d = 0.2; C, $\tau = 0.03$ d = 0.3; D, $\tau = 0.03$ d = 0.4; E, $\tau = 0.03$ d = 0.5; F, $\tau = 0.03$ d = 0.6; G, $\tau = 0.03$ d = 0.7; H, $\tau = 0.03$ d = 0.8; I, $\tau = 0.03$ d = 0.9

TABLE 1

Example correlation structures in the literature on designing stepped wedge cluster randomized trials

Decay	Design feature	Correlation structure	Example references
No	Cross-sectional	Nested exchangeable	Hemming et al,[7] Hooper et al,[8] and Li et al[5]
	Closed-cohort	Block exchangeable	Hooper et al[8] and Li et al[5,14]
Yes	Cross-sectional	Exponential decay	Kasza et al,[10] Kasza and Forbes,[11] and Grantham et al[12]
	Closed-cohort	Proportional decay	This article

TABLE 2

Examples of the proportional decay structure for cohort designs and the exponential decay structure for cross-sectional designs. The illustration is based on a stepped wedge trial with $T = 3$ periods and $N_i = 2$ measurements per cluster period. Define $y_i = (y_{i11}, y_{i12}, y_{i13}, y_{i21}, y_{i22}, y_{i23})'$

	Proportional decay structure	Exponential decay structure
corr(y_i)	$\begin{pmatrix} 1 & \rho & \rho^2 & \tau & \tau\rho & \tau\rho^2 \\ \rho & 1 & \rho & \tau\rho & \tau & \tau\rho \\ \rho^2 & \rho & 1 & \tau\rho^2 & \tau\rho & \tau \\ \tau & \tau\rho & \tau\rho^2 & 1 & \rho & \rho^2 \\ \tau\rho & \tau & \tau\rho & \rho & 1 & \rho \\ \tau\rho^2 & \tau\rho & \tau & \rho^2 & \rho & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & \tau\rho & \tau\rho^2 & \tau & \tau\rho & \tau\rho^2 \\ \tau\rho & 1 & \tau\rho & \tau\rho & \tau & \tau\rho \\ \tau\rho^2 & \tau\rho & 1 & \tau\rho^2 & \tau\rho & \tau \\ \tau & \tau\rho & \tau\rho^2 & 1 & \tau\rho & \tau\rho^2 \\ \tau\rho & \tau & \tau\rho & \tau\rho & 1 & \tau\rho \\ \tau\rho^2 & \tau\rho & \tau & \tau\rho^2 & \tau\rho & 1 \end{pmatrix}$

TABLE 3

Summary of bias-corrected sandwich variance estimators for $\hat{\theta}$

Label	Correction	C_i	B_i	References
BC0	none	I	I	Liang and Zeger[19]
BC1	less	I	$(I - H_i)^{-1/2}$	Kauermann and Carroll[28]
BC2	more	I	$(I - H_i)^{-1}$	Mancl and DeRouen[29]
BC3	less	$\mathrm{diag}\left\{\left(1 - \min\left\{\zeta, \left[D_i' V_i^{-1} D_i \Omega_1^{-1}\right]_{jj}\right\}\right)^{-1/2}\right\}$	I	Fay and Graubard[30]

TABLE 4

Percent relative bias of the correlation parameters based on uncorrected quasi-least squares (QLS) and matrix-adjusted QLS (MAQLS) for each simulation scenario when the treatment effect is zero

Correlations		Effect size	Design resource			Percent bias for τ		Percent bias for ρ	
τ	ρ	δ	I	N	T	QLS	MAQLS	QLS	MAQLS
0.03	0.2	0	18	10	7	−26.6	3.5	−0.6	0.2
0.03	0.2	0	18	24	4	−16.0	5.5	−0.3	0.0
0.03	0.2	0	20	14	5	−20.2	3.4	−0.5	0.0
0.03	0.2	0	21	8	4	−29.4	3.2	−0.4	0.2
0.03	0.2	0	15	8	4	−41.3	4.7	−0.6	0.3
0.03	0.8	0	16	12	5	−26.2	4.4	−0.6	0.0
0.03	0.8	0	24	7	5	−27.7	0.6	−0.6	−0.1
0.03	0.8	0	12	8	5	−49.4	8.2	−2.1	−0.4
0.03	0.8	0	12	5	4	−74.2	6.6	−2.3	−0.5
0.03	0.8	0	10	5	3	−91.6	−0.4	−1.5	−0.5
0.10	0.2	0	21	11	8	−9.2	3.9	−0.6	0.6
0.10	0.2	0	24	11	7	−8.1	3.4	−0.4	0.6
0.10	0.2	0	15	16	6	−11.8	7.1	−0.6	1.1
0.10	0.2	0	18	8	7	−12.7	3.9	−0.8	0.7
0.10	0.2	0	16	7	5	−16.7	3.7	−0.7	1.0
0.10	0.8	0	20	18	5	−8.0	5.0	−0.1	0.2
0.10	0.8	0	15	9	4	−15.2	3.1	−0.2	0.2
0.10	0.8	0	10	20	3	−16.9	6.5	−0.1	0.3
0.10	0.8	0	12	5	5	−27.0	2.6	−1.5	0.0
0.10	0.8	0	9	7	4	−29.7	3.8	−1.0	0.1

TABLE 5

Simulation scenarios and nominal size, along with the empirical type I error rates corresponding to the MAQLS z-test and t-test (DoF $= I - 2$) with different variance estimators. Empirical type I error rates between 4.5% and 5.5% are highlighted in boldface and considered acceptable based on the margin of error from a binomial model with 10 000 Monte Carlo replications. Pred: nominal type I error rate; MB: model-based variance; BC0: uncorrected sandwich variance of Liang and Zeger (1986); BC1: bias-corrected sandwich variance of Kauermann and Carroll (2001); BC2: bias-corrected sandwich variance of Mancl and DeRouen (2001); BC3: bias-corrected sandwich variance of Fay and Graubard (2001)

								z-test						t-test			
τ	ρ	δ	I	N	T	Pred	MB	BC0	BC1	BC2	BC3	Pred	MB	BC1	BC2	BC2	BC3
0.03	0.2	0	18	10	7	5.0	**4.8**	8.4	6.8	**5.1**	7.1	5.0	3.4	6.3	**4.7**	3.8	**4.9**
0.03	0.2	0	18	24	4	5.0	**4.9**	8.1	6.5	**5.0**	6.5	5.0	3.4	6.0	**4.7**	3.6	**4.7**
0.03	0.2	0	20	14	5	5.0	**5.1**	8.3	6.9	5.8	7.0	5.0	3.8	6.5	**5.4**	4.4	**5.5**
0.03	0.2	0	21	8	4	5.0	**5.2**	7.9	6.6	**5.3**	6.6	5.0	3.8	6.2	**5.0**	4.0	**4.9**
0.03	0.2	0	15	8	4	5.0	5.7	9.7	7.7	5.8	7.7	5.0	3.7	7.0	**5.3**	4.1	**5.3**
0.03	0.8	0	16	12	5	5.0	5.6	9.0	7.3	5.6	7.3	5.0	3.8	6.7	**5.1**	3.8	**5.1**
0.03	0.8	0	24	7	5	5.0	**5.3**	7.7	6.6	**5.5**	6.6	5.0	4.3	6.2	**5.2**	4.3	**5.3**
0.03	0.8	0	12	8	5	5.0	6.0	10.6	7.9	5.8	7.9	5.0	3.5	7.0	**4.9**	3.2	**5.0**
0.03	0.8	0	12	5	4	5.0	5.7	10.3	7.8	5.7	7.5	5.0	3.4	6.9	**5.0**	3.5	**4.9**
0.03	0.8	0	10	5	3	5.0	6.3	11.2	8.2	5.6	7.1	5.0	3.0	7.0	**4.6**	3.1	3.8
0.10	0.2	0	21	11	8	5.0	**5.0**	8.1	6.9	5.7	7.0	5.0	3.8	6.6	**5.3**	4.4	5.6
0.10	0.2	0	24	11	7	5.0	**5.1**	7.7	6.4	**5.3**	6.5	5.0	3.8	6.1	**5.1**	4.3	**5.2**
0.10	0.2	0	15	16	6	5.0	**4.8**	9.7	7.5	5.7	7.8	5.0	3.1	6.8	**5.2**	4.0	**5.3**
0.10	0.2	0	18	8	7	5.0	**4.8**	8.8	7.1	5.6	7.3	5.0	3.4	6.7	**5.2**	4.0	**5.4**
0.10	0.2	0	16	7	5	5.0	**5.2**	9.1	7.2	**5.5**	7.3	5.0	3.4	6.6	**5.1**	3.8	**5.2**
0.10	0.8	0	20	18	5	5.0	**5.3**	7.9	6.4	**5.2**	6.5	5.0	3.8	6.1	**4.8**	3.9	**4.8**
0.10	0.8	0	15	9	4	5.0	**5.5**	9.1	7.2	**5.3**	7.0	5.0	3.7	6.6	**4.8**	3.7	**4.6**
0.10	0.8	0	10	20	3	5.0	5.6	11.7	8.5	6.0	7.4	5.0	2.8	7.3	**4.9**	3.0	4.0
0.10	0.8	0	12	5	5	5.0	5.8	10.5	7.9	5.7	7.9	5.0	3.3	6.9	**4.9**	3.3	**4.9**
0.10	0.8	0	9	7	4	5.0	6.0	12.0	8.6	5.7	8.3	5.0	2.8	7.2	**4.7**	2.9	**4.5**

TABLE 6

Simulation scenarios and predicted power, along with the difference between empirical and predicted power corresponding to the MAQLS z-test and t-test (DoF = $I-2$) with different variance estimators. Differences from the prediction within 0.8% are highlighted in boldface and considered acceptable. Pred: predicted power; MB: model-based variance; BC0: uncorrected sandwich variance of Liang and Zeger (1986); BC1: bias-corrected sandwich variance of Kauermann and Carroll (2001); BC2: bias-corrected sandwich variance of Mancl and DeRouen (2001); BC3: bias-corrected sandwich variance of Fay and Graubard (2001)

								z-test						t-test			
τ	ρ	δ	I	N	T	Pred	MB	BC0	BC1	BC2	BC3	Pred	MB	BC0	BC1	BC2	BC3
0.03	0.2	0.3	18	10	7	89.9	**−0.5**	1.1	−1.0	−3.4	−0.9	86.0	**0.2**	2.4	**−0.1**	−3.4	**0.2**
0.03	0.2	0.3	18	24	4	88.6	**−0.4**	1.5	**−0.4**	−2.9	**−0.5**	84.4	**0.5**	3.2	**0.5**	−2.5	**0.4**
0.03	0.2	0.3	20	14	5	89.7	**0.0**	1.5	**−0.5**	−2.6	**−0.3**	86.2	**0.3**	2.6	**0.2**	−2.5	**0.4**
0.03	0.2	0.4	21	8	4	87.5	**−0.5**	1.3	**−0.8**	−3.4	−0.9	83.9	**−0.1**	2.2	**−0.5**	−3.2	**−0.6**
0.03	0.2	0.5	15	8	4	90.7	−1.1	1.0	−1.2	−3.9	−1.1	85.9	**−0.3**	3.0	**−0.2**	−4.3	**−0.3**
0.03	0.8	0.2	16	12	5	88.6	**−0.5**	1.2	−1.4	−4.4	−1.3	83.8	**0.1**	2.5	**−0.6**	−4.4	**−0.6**
0.03	0.8	0.2	24	7	5	88.2	**−0.5**	1.1	**−0.6**	−2.4	**−0.7**	85.2	**0.2**	1.8	**0.0**	−2.3	**0.0**
0.03	0.8	0.3	12	8	5	94.1	−1.1	**0.4**	−1.8	−4.6	−1.9	88.7	**0.3**	2.7	**−0.4**	−4.8	**−0.3**
0.03	0.8	0.4	12	5	4	95.2	**−0.8**	**0.7**	−1.2	−3.8	−1.5	90.3	**0.5**	2.8	**−0.3**	−4.1	**−0.6**
0.03	0.8	0.5	10	5	3	94.6	**−0.7**	1.1	−1.1	−4.8	−2.3	87.8	**0.8**	4.3	**−0.3**	−6.0	−2.4
0.10	0.2	0.3	21	11	8	87.8	**−0.6**	1.5	**−0.4**	−2.8	**−0.1**	84.3	**−0.1**	2.6	**0.1**	−2.8	**0.5**
0.10	0.2	0.3	24	11	7	87.8	**−0.6**	1.2	**−0.3**	−2.5	**−0.1**	84.8	**0.0**	2.1	**0.2**	−2.2	**0.4**
0.10	0.2	0.4	15	16	6	90.6	−1.3	1.2	−1.0	−4.3	**−0.8**	85.8	**−0.8**	2.9	**−0.3**	−4.1	**0.1**
0.10	0.2	0.4	18	8	7	91.6	**−0.1**	1.5	**−0.4**	−2.7	**−0.1**	87.9	**0.4**	2.7	**0.3**	−2.5	**0.6**
0.10	0.2	0.5	16	7	5	88.6	**−0.8**	1.9	**−0.6**	−3.5	**−0.4**	83.8	**−0.1**	3.4	**0.3**	−3.5	**0.6**
0.10	0.8	0.2	20	18	5	86.1	**−0.5**	1.8	**−0.5**	−3.0	**−0.6**	82.1	**−0.3**	2.7	**0.0**	−3.2	**0.0**
0.10	0.8	0.3	15	9	4	89.5	−1.0	1.0	−1.5	−4.6	−1.8	84.5	**−0.1**	2.6	**−0.8**	−4.8	−1.1
0.10	0.8	0.4	10	20	3	94.4	**−0.6**	1.6	**−0.8**	−4.3	−2.1	87.5	**0.4**	4.6	**0.6**	−5.3	−1.8
0.10	0.8	0.4	12	5	5	93.2	**−0.5**	0.9	−1.3	−4.4	−1.4	87.5	**0.7**	3.2	**−0.3**	−4.8	**−0.3**
0.10	0.8	0.5	9	7	4	97.3	**−0.6**	**0.4**	−1.1	−4.2	−1.4	91.4	1.1	3.5	**−0.2**	−5.7	**−0.7**