HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2023 May 19.

Published in final edited form as:

Stat Med. 2022 April 15; 41(8): 1376–1396. doi:10.1002/sim.9283.

Accounting for unequal cluster sizes in designing cluster randomized trials to detect treatment effect heterogeneity

Guangyu Tong[1,2], **Denise Esserman**[1,2], **Fan Li**[1,2,3]

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

[2]Yale Center for Analytical Sciences, Yale School of Public Health, New Haven, Connecticut,

[3]Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, Connecticut,

Abstract

Unequal cluster sizes are common in cluster randomized trials (CRTs). While there are a number of previous investigations studying the impact of unequal cluster sizes on the power for testing the average treatment effect in CRTs, little is known about the impact of unequal cluster sizes on the power for testing the heterogeneous treatment effect (HTE) in CRTs. In this work, we expand the sample size procedures for studying HTE in CRTs to accommodate cluster size variation under the linear mixed model framework. Through analytical derivation and graphical exploration, we show that the sample size for the HTE with an individual-level effect modifier is less affected by unequal cluster sizes than with a cluster-level effect modifier. The impact of cluster size variability jointly depends on the mean and coefficient of variation of cluster sizes, covariate intraclass correlation coefficient (ICC) and the conditional outcome ICC. In addition, we demonstrate that the HTE-motivated analysis of covariance framework can be used for analyzing the average treatment effect, and offer a more efficient sample size procedure for studying the average treatment effect adjusting for the effect modifier. We use simulations to confirm the accuracy of the proposed sample size procedures for both the average treatment effect and HTE in CRTs. Extensions to multivariate effect modifiers are provided and our procedure is illustrated in the context of the Strategies to Reduce Injuries and Develop Confidence in Elders trial.

Keywords

average treatment effect; coefficient of variation; heterogeneous treatment effect; linear mixed model; sample size calculation; variable cluster sizes

1 | INTRODUCTION

Cluster randomized trials (CRTs) are one type of study design that randomizes entire clusters or groups of individuals to treatment arms.[1] These trials are conducted either

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

because the intervention itself is designed to be implemented at the cluster level, or to prevent treatment contamination; logistical and administrative issues may also play a part in adopting a CRT.[2] Because individuals nested in clusters share the same physical environment or social connections, the outcomes measured from individuals in the same cluster tend to be more alike than those measured for individuals from different clusters. This creates a positive outcome intraclass correlation coefficient (outcome-ICC) that inflates the variance of the average treatment effect estimator. Specifically, a simple design effect, defined as 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 CRT, for estimating the average treatment effect is given by

$$\text{design effect} = 1 + (m-1)\rho_y, \tag{1}$$

where ρ_y is the outcome-ICC and m is the assumed common cluster size.[2] While the average treatment effect has been the cornerstone in comparative effectiveness research with CRTs, interest is growing in understanding whether the treatment effect varies among pre-specified patient subgroups, such as those defined by baseline demographic or clinical characteristics. In this context, the concept of heterogeneous treatment effect (abbreviated as HTE hereafter) refers to potentially different treatment effects across patient subgroups that can arise due to various reasons, such as diverse practices, varying responses to treatment, or differing vulnerability to certain diseases.[3,4] Pre-planned HTE analyses of CRTs enable a rigorous understanding of how care innovations may impact outcomes for vulnerable or other important subpopulations, and facilitate the development of targeted interventions to reduce known disparities in health outcomes. On this front, a recent systematic review by Stark et al[5] reported that 16 out of 64 CRTs examined HTE among demographic patient subgroups. They further noticed a lack of guidance on designing CRTs with pre-planned HTE analyses, for which new statistical methods are needed.

To assist in the planning of CRTs to detect HTE(s), Yang et al[6] recently developed a new sample size and power formula and clarified the associated design effect. In particular, they found that the outcome-ICC and the covariate-ICC (equivalently ICC of the effect modifier) jointly determine the power of the HTE analyses in CRTs. As a counterpart of outcome-ICC, the covariate-ICC captures the fraction of between-cluster covariate variation relative to the total of between- and within-cluster covariate variation,[7,8] and plays an important role in sample size determination for HTE analyses. Furthermore, while the design effect (1) for estimating the average treatment effect is monotonically increasing with a larger outcome-ICC, the design effect for estimating a HTE with a individual-level effect modifier is bounded from above, showing that relative to an individually randomized trial, the inflation of sample size for studying the average treatment effect often needs to be larger than that for studying an individual-level HTE in a CRT. This leads to a greater chance of sufficiently powering both the average treatment effect and the HTE(s) in CRTs. However, a key assumption of their sample size procedure is that the cluster sizes are equal, whereas, in practice, the cluster sizes are often variable. A systematic review conducted by Eldridge et al[9] indicated two-thirds of CRTs have unequal sized clusters. The impact of unequal cluster sizes for the average treatment effect analysis has been investigated in, for example, Kerry et al,[10] van Breukelen et al,[11] Candel and van Breukelen,[12] and Li and Tong;[13,14] a

good review can also be found in Eldridge et al.[15] These authors found that while cluster size variation can lead to a notable loss of efficiency or power for cluster-level analyses,[10] they lead to smaller loss of efficiency in CRTs with linear mixed model analyses.[11] In most cases, the loss of power is compensated for by the addition of 10% to 14% more clusters, depending on the magnitude of variation in cluster sizes.[11,12] Whereas sample size methodology has been relatively well studied in CRTs with average treatment effect analysis, it is currently unclear how to account for unequal cluster sizes in designing CRTs to detect treatment effect heterogeneity.

In this article, we generalize the results in Yang et al[6] to develop modified variance expressions of the HTE estimator in CRTs with unequal cluster sizes. The new variance expressions clarify the implication of varying cluster size on the power of the HTE test, and provide a closed-form solution to adjust for it in the design stage. When the effect modifier is measured at the individual level, we show that the variance expression for the HTE parameter is generally insensitive to unequal cluster sizes, and the sample size methods in Yang et al[6] provide a reasonable approximation. On the other hand, unequal cluster sizes have a larger impact on the power of the HTE test with a cluster-level effect modifier, and a proper correction factor is needed for sample size determination. In addition, we point out that the linear mixed model with treatment-by-covariate interactions provides an unbiased estimator of the average treatment effect and derive the variance expression of this covariate-adjusted average treatment effect estimator in CRTs. Interestingly, with a continuous outcome, we show that the design effect expression for the covariate-adjusted average treatment effect analysis has an identical form (up to a correction factor due to cluster size variability) as that for the unadjusted average treatment effect analysis (Equation 1), except that the marginal outcome-ICC, ρ_y, will be replaced by the covariate-adjusted, conditional outcome-ICC. This finding not only helps unify the sample size considerations for the covariate-adjusted average treatment effect and the HTE analyses of CRTs, but also provides a new perspective on the possible efficiency improvement from covariate adjustment in the design and analysis of CRTs for the average treatment effect.

The rest of this article is organized as follows. Section 2 introduces the analytical model, and develops the sample size formula for HTE analysis of CRTs with a single effect modifier adjusting for unequal cluster sizes. The new sample size formula is expressed as a function of outcome-ICC and covariate-ICC as well as the coefficient of variation (CV) of cluster sizes. The methodology is extended to accommodate multiple effect modifiers in Section 3. We conduct a series of simulation studies to examine the small-sample performance of the proposed sample size formulas, both for the HTE analysis and for covariate-adjusted average treatment effect analysis. Under certain conditions, we demonstrate numerically that the adjusted average treatment effect analysis based on the HTE model can substantially reduce the required number of clusters in CRTs, which further illustrates the merits of the HTE model. We illustrate our sample size procedure in Section 4 in the context of the Strategies to Reduce Injuries and Develop Confidence in Elders (STRIDE) study,[16,17] which was a two-arm parallel CRT with pre-specified HTE analysis. Section 5 concludes with a discussion.

2 | STATISTICAL METHODS WITH A UNIVARIATE EFFECT MODIFIER

2.1 | Sample size requirement for the test of the HTE assuming equal cluster sizes

We first review the linear mixed model that allows for a test of the HTE in CRTs as well as the associated sample size procedure in Yang et al[6] assuming equal cluster sizes. Consider a two-arm CRT with a total of n clusters, each with cluster size $m_i, i = 1, \ldots, n$. We define Y_{ij} as the quantitative outcome of the jth individual ($j = 1, \ldots, m_i$) in the ith cluster. Denote the treatment variable as W_i, which is randomized at the cluster level. We write $W_i = 1$ if cluster i is randomized to the treatment arm, and $W_i = 0$ if cluster i receives usual care. Typical analytical models for CRTs require the adjustment for between-cluster variability, and the linear mixed model with a cluster-level random-effect is commonly used in these circumstances to estimate the treatment effect. To allow for the test of the HTE, we let X_{ij} denote an individual-level effect modifier of interest (we will use effect modifier and covariate interchangeably in what follows), such as gender or racial group indicator. The linear mixed model with main effects of cluster-level treatment, effect modifier, as well as their interaction can be expressed as

$$Y_{ij} = \beta_1 + \beta_2 W_i + \beta_3 X_{ij} + \beta_4 W_i X_{ij} + \lambda_i + \epsilon_{ij}, \quad \lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2). \tag{2}$$

In this model, β_1, β_2, β_3, β_4 stand for the intercept (grand mean), main effect of treatment, main effect of the potential effect modifier, and the treatment-by-covariate interaction effect. Further, λ_i is the random intercept that measures the cluster-specific departure from the overall mean and is assumed to follow $\mathcal{N}(0, \sigma_\lambda^2)$; ϵ_{ij} is the normal residual error, and is assumed to be independent from λ_i. By definition, the outcome-ICC conditional on the effect modifier is written as $\rho_{y|x} = \sigma_\lambda^2/(\sigma_\lambda^2 + \sigma_\epsilon^2)$. We start with the individual-level effect modifier and then study the cluster-level effect modifier as a special case where $X_{ij} = X_i$ for all j.

Based on model (2), the test for the HTE can be formulated by testing $H_0: \beta_4 = 0$. Putting H_0 in the context of a binary effect modifier (eg, $X_{ij} = 1$ if female; $X_{ij} = 0$ otherwise), the parameter β_4 encodes the difference in treatment effect across subgroups (eg, female and non-female), and a two-sided test of the HTE can be conducted using the Wald statistic, $\hat{\beta}_4/se(\hat{\beta}_4)$, based on a reference normal distribution under the null. To obtain the sample size required for a test of the HTE, Yang et al[6] derived the large-sample variance of the generalized least squares estimator, $\hat{\beta}_4$, assuming all clusters have the same size, $m_i = m$. In brief, we define $\bar{W} = E(W_i)$ as the proportion of clusters treated, $\sigma_w^2 = \bar{W}(1 - \bar{W})$ as the variance of the treatment indicator, which equals 1/4 under a balanced design, σ_x^2 as the marginal variance of the effect modifier, $\sigma_{y|x}^2 = \sigma_\lambda^2 + \sigma_\epsilon^2$ as the total variance of the outcome conditional on the effect modifier, and ρ_x as the covariate-ICC, which can be defined as $\rho_x = \mathrm{Cov}(X_{ij}, X_{ij'})/\sigma_x^2$. Here, $\mathrm{Cov}(X_{ij}, X_{ij'})$ represents the common covariance between effect modifiers observed for any two individuals j and j' in each given cluster i.[7] As a specific example, if $X_{ij} = \mu_1 + b_i + c_{ij}$, with μ_1 as the global mean, $b_i \sim \mathcal{N}(0, \sigma_b^2)$ and $c_{ij} \sim \mathcal{N}(0, \sigma_c^2)$, then the covariate-ICC is given by $\rho_x = \sigma_b^2/(\sigma_b^2 + \sigma_c^2)$. Given the outcome-ICC and covariate-ICC, the large-sample variance of $\sqrt{n}\hat{\beta}_4$ is shown to be

$$\sigma_4^2 = \frac{\sigma_{y|x}^2(1-\rho_{y|x})\{1+(m-1)\rho_{y|x}\}}{m\sigma_w^2\sigma_x^2\{1+(m-2)\rho_{y|x}-(m-1)\rho_x\rho_{y|x}\}}. \quad (3)$$

Under the equal cluster size assumption, given a pre-specified HTE effect size δ, the required number of clusters that ensures $100(1-\zeta)\%$ power with an α-level test is

$$n = \frac{(z_{1-\alpha/2}+z_{1-\zeta})^2\sigma_{y|x}^2(1-\rho_{y|x})\{1+(m-1)\rho_{y|x}\}}{m\delta^2\sigma_w^2\sigma_x^2\{1+(m-2)\rho_{y|x}-(m-1)\rho_x\rho_{y|x}\}}, \quad (4)$$

where z_q is the q-quantile of the standard normal distribution.

We can make several key conclusions based on the Equation (4). First, the sample size formula for testing the HTE only depends on the effect size of the HTE but not on the main effect of treatment or the effect modifier. Second, different from previous sample size formulas in the education literature for designing school-based CRTs,[18,19] the above formula explicitly clarifies that the conditional outcome-ICC, $\rho_{y|x}$, and the marginal covariate-ICC, ρ_x, jointly determine the power of the HTE analysis. Furthermore, the relationship between n and $\rho_{y|x}$ is parabolic in that n first increases as $\rho_{y|x}$ increases from zero, and then decreases after a critical point. The sample size also decreases with a smaller ρ_x and larger covariate variance σ_x^2. Finally, while the usual design effect (1) for estimating the average treatment effect is monotonically increasing with a larger (unconditional) outcome-ICC, the design effect for estimating the HTE with an individual-level effect modifier is bounded from above and can even decrease with a larger outcome-ICC.[6] An important implication from this observation is that, relative to an individually randomized trial, studying the average treatment effect in a CRT often requires a larger sample size than that for studying the HTE, leading to a greater chance of having sufficient power for detecting both the average treatment effect and the HTE in CRTs.

2.2 | Sample size requirement for the test of the HTE allowing for unequal cluster sizes

We derive a more general sample size expression for testing the HTE with a single effect modifier assuming the cluster sizes arise from a common, non-degenerate distribution, namely, $m_i \sim f(m_i)$ with bounded first and second moments. We first assume that the effect modifier is measured at the individual level such that $\rho_x < 1$. To facilitate the derivation, we mean-center the cluster-level treatment in the model as follows,

$$Y_{ij} = b_1 + b_2(W_i - \bar{W}) + b_3X_{ij} + b_4(W_i - \bar{W})X_{ij} + \lambda_i + \epsilon_{ij},$$

where $b_1 = \beta_1 + \beta_2\bar{W}$, $b_2 = \beta_2$, $b_3 = \beta_3 + \beta_4\bar{W}$, $b_4 = \beta_4$. Define the collection of design points $\boldsymbol{Z}_{ij} = \left(1, (W_i - \bar{W}), X_{ij}, (W_i - \bar{W})X_{ij}\right)^T$, and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{im_i})^T$. Given the value of σ_λ^2 and σ_ϵ^2, the scaled maximum likelihood estimator of $\boldsymbol{b} = (b_1, b_2, b_3, b_4)^T$, $\sqrt{n}\left(\hat{\boldsymbol{b}} - \boldsymbol{b}\right)$, converges to a multivariate normal distribution with mean zero and variance $\boldsymbol{\Sigma}_n = n\sigma_{y|x}^2\boldsymbol{U}_n^{-1}$, where $\boldsymbol{U}_n = \sum_{i=1}^n \boldsymbol{Z}_i^T\boldsymbol{R}_i^{-1}\boldsymbol{Z}_i$, and $\boldsymbol{R}_i = (1-\rho_{y|x})\boldsymbol{I}_{m_i} + \rho_{y|x}\boldsymbol{J}_{m_i}$ is the exchangeable correlation matrix for

cluster i implied from model (2), where $\boldsymbol{I}_{m_i}$ and $\boldsymbol{J}_{m_i}$ are $m_i \times m_i$ identity matrices and matrices of ones, respectively. Then the large-sample variance of $\sqrt{n}\hat{b}_4$ is the lower-right element of the variance matrix $\lim_{n\to\infty}\boldsymbol{\Sigma}_n = \sigma^2_{y\mid x}\left(\lim_{n\to\infty} n^{-1}\boldsymbol{U}_n\right)^{-1}$.

To derive an explicit variance expression, we notice the inverse of the exchangeable correlation matrix is

$$\boldsymbol{R}_i^{-1} = \frac{1}{1-\rho_{y\mid x}}\boldsymbol{I}_{m_i} - \frac{\rho_{y\mid x}}{(1-\rho_{y\mid x})\{1+(m_i-1)\rho_{y\mid x}\}}\boldsymbol{J}_{m_i}.$$

Plugging this inverse expression into the generalized least squares variance, we show in Web Appendix A that an approximate sample size formula for testing $H_0: \beta_4 = 0$ under unequal cluster sizes is given by

$$nVar\left(\hat{\beta}_4\right) = \frac{\sigma^2_{y\mid x}(1-\rho_{y\mid x})}{\sigma^2_w\sigma^2_x\{\bar{m}+(1-\rho_x)\bar{p}+\rho_x\bar{q}\}}, \tag{5}$$

where

$$\bar{q} = E\left\{\frac{-m_i^2\rho_{y\mid x}}{1+(m_i-1)\rho_{y\mid x}}\right\} \quad \text{and} \quad \bar{p} = -\left\{\frac{\bar{m}\rho_{y\mid x}}{1+(\bar{m}-1)\rho_{y\mid x}}\right\}\left[1-\mathrm{CV}^2\frac{\bar{m}\rho_{y\mid x}(1-\rho_{y\mid x})}{\{1+(\bar{m}-1)\rho_{y\mid x}\}^2}\right].$$

Therefore, we obtain the required number of clusters for a two-sided α-level z-test to achieve $100(1-\zeta)\%$ power as

$$n = \frac{(z_{1-\alpha/2}+z_{1-\zeta})^2\sigma^2_{y\mid x}(1-\rho_{y\mid x})\{1+(\bar{m}-1)\rho_{y\mid x}\}}{\bar{m}\delta^2\sigma^2_w\sigma^2_x\{1+(\bar{m}-2)\rho_{y\mid x}-(\bar{m}-1)\rho_x\rho_{y\mid x}\}} \times \underbrace{\left[1-\mathrm{CV}^2\frac{\bar{m}\rho_{y\mid x}(1-\rho_{y\mid x})(\rho_x-\rho_{y\mid x})}{\{1+(\bar{m}-2)\rho_{y\mid x}-(\bar{m}-1)\rho_x\rho_{y\mid x}\}\{1+(\bar{m}-1)\rho_{y\mid x}\}^2}\right]^{-1}}_{\text{Correction Factor}\,\theta_1(\mathrm{CV})}, \tag{6}$$

which, compared to (4), includes an additional correction factor, $\theta_1(\mathrm{CV})$, due to cluster size variation. Depending on the magnitude of the covariate-ICC and the conditional outcome-ICC, the correction factor may be larger, equal, or smaller than *unity*. To be more specific, when $\rho_x > \rho_{y\mid x}, \theta_1(\mathrm{CV}) > 1$, and the sample size will be inflated due to unequal cluster sizes. In contrast, when $\rho_x < \rho_{y\mid x}, \theta_1(\mathrm{CV}) < 1$, and cluster size variation reduces the required sample size. Finally, when $\rho_x = \rho_{y\mid x}, \theta_1(\mathrm{CV}) = 1$ and the required sample size requirement becomes invariant to cluster size variation. Despite such explicit relationships, the correction factor should in general be close to one, especially when the mean cluster size is moderate to large (eg, 50 or above) because $\lim_{\bar{m}\to\infty}\theta_1(\mathrm{CV}) = 1$. Given the CV of cluster sizes rarely exceed one in CRTs, when the average cluster size is not too small (eg, < 20), Equation (6) suggests that unequal cluster sizes should have minimal impact on the sample size requirement for the test of the HTE with an individual-level effect modifier.

Unequal cluster sizes, however, can have a larger impact on the required sample size when testing for a HTE when there is a cluster-level effect modifier (as compared with the results with an individual-level effect modifier). To see this, recall that a cluster-level effect modifier attributes the same value for each individual in a cluster, and therefore becomes a special case of the above derivation when the covariate-ICC, $\rho_x = 1$. In this case, the sample size formula (6) reduces to

$$n = \frac{(z_{1-\alpha/2} + z_{1-\zeta})^2 \sigma^2_{y|x}\{1 + (\bar{m}-1)\rho_{y|x}\}}{\bar{m}\delta^2\sigma^2_w\sigma^2_x} \times \underbrace{\left[1 - \mathrm{CV}^2 \frac{\bar{m}\rho_{y|x}(1-\rho_{y|x})}{\{1 + (\bar{m}-1)\rho_{y|x}\}^2}\right]^{-1}}_{\text{Correction Factor } \theta_2(\mathrm{CV})}, \quad (7)$$

whose correction factor due to unequal cluster sizes, $\theta_2(\mathrm{CV})$, has a form identical to the equations derived in van Breukelen et al[11] and Candel and van Breukelen,[12] except that their unconditional outcome-ICC is replaced by the conditional outcome-ICC, $\rho_{y|x}$. Beyond this difference, the generic findings in van Breukelen et al[11] hold for the test of the HTE with a cluster-level effect modifier, namely, the efficiency loss in estimating β_4 due to unequal cluster sizes is within 15% with a moderate CV of cluster sizes (eg, 0.6). Only when the CV of cluster sizes is large (eg, 0.9) would we expect to see an efficiency loss over 20%. In addition, the correction factor, $\theta_2(\mathrm{CV})$, is a parabolic function of $\rho_{y|x}$, and peaks when $\rho_{y|x} = 1/(\bar{m}+1)$.[11,13]

To further illustrate these results, we present in Figure 1 the correction factor as a function of average cluster sizes, $\bar{m} \in \{20,100\}$ and CV of cluster sizes $\in \{0.3, 0.9\}$, representing a moderate and an extreme degree of cluster size variation. We vary the covariate−ICC $\in [0,1]$, with the lower limit representing a fully independent individual-level covariate and the upper limit representing a cluster-level covariate. The conditional outcome-ICC is varied between 0 and 0.2, as this value rarely exceeds 0.2 for commonly reported health outcomes.[9] In Figure 1A where $\bar{m} = 20$ and CV = 0.3, the correction function is very close to 1 even when the covariate-ICC approaches 1 (ie, the case with a cluster-level effect modifier). With a larger CV = 0.9 in Figure 1B, the correction factor remains close to 1 except when the covariate-ICC approaches 1 and the outcome-ICC is small. In this case, the correction factor can reach slightly over 1.24, indicating a 24% efficiency loss in power and increase in sample size for a highly correlated individual-level effect modifier or a cluster-level effect modifier. Increasing the average cluster size, as shown in Figure 1C,D, notably, can further reduce the sensitivity of $\mathrm{Var}(\hat{\beta}_4)$ to cluster size variation, because the peak contour region (where correction factor reaches its largest magnitude) moves to the upper left corner with the largest covariate-ICC and smallest conditional outcome-ICC.

2.3 | Sample size requirement for studying the covariate-adjusted average treatment effect allowing for unequal cluster sizes

While the linear mixed model (2) is primarily motivated by the analysis of the HTE, it also enables the covariate-adjusted analysis of the average treatment effect. Without loss of generality, if we assume that the effect modifier has global mean $\mu_1 = 0$ (otherwise one could mean center the covariates without altering the interpretation of

β_4), the main effect parameter β_2 can be interpreted as an average treatment effect parameter that characterizes the marginal treatment effect for the overall population. Model (2) implies that $E(Y_{ij} \mid W_i = 1) = \beta_1 + \beta_2$ whereas $E(Y_{ij} \mid W_i = 0) = \beta_1$, and therefore $\beta_2 = E(Y_{ij} \mid W_i = 1) - E(Y_{ij} \mid W_i = 0)$ defines a valid average treatment effect parameter due to the fact that treatment indicator Z_i is randomized at the cluster level. Similar observations were previously discussed in the context of individually randomized trials with analysis of covariance (ANCOVA) models,[20,21] for which model (2) is a generalization to allow for cluster-level random effects and ICCs. Based on the derivation in Section 2.1, we show in Web Appendix A that

$$n\mathrm{Var}\left(\hat{\beta}_2\right) = \frac{\sigma^2_{y\mid x}\left[1 + (\bar{m} - 1)\rho_{y\mid x}\right]}{\sigma^2_w \bar{m}} \times \left[1 - \mathrm{CV}^2 \frac{\bar{m}\rho_{y\mid x}(1 - \rho_{y\mid x})}{\left\{1 + (\bar{m} - 1)\rho_{y\mid x}\right\}^2}\right]^{-1}. \tag{8}$$

This variance expression (8) leads to several key findings. First, if all cluster sizes are equal so that CV = 0, the variance expression $n\mathrm{Var}\left(\hat{\beta}_2\right)$ has the identical form of the unadjusted linear mixed model analysis without covariates,[1, 11] except that the marginal outcome-ICC ρ_y will be replaced by the conditional outcome-ICC $\rho_{y\mid x}$. In the presence of cluster size variation, the corresponding sample size formula for an average treatment effect of $\beta_2 = \Delta$ becomes

$$\begin{aligned} n = {} & \frac{\left\{t_{1-\alpha/2}(n-2) + t_{1-\zeta}(n-2)\right\}^2 \sigma^2_{y\mid x}\left[1 + (\bar{m} - 1)\rho_{y\mid x}\right]}{\bar{m}\Delta^2\sigma^2_w} \\ & \times \underbrace{\left[1 - \mathrm{CV}^2 \frac{\bar{m}\rho_{y\mid x}(1 - \rho_{y\mid x})}{\left\{1 + (\bar{m} - 1)\rho_{y\mid x}\right\}^2}\right]^{-1}}_{\text{Correction Factor } \theta_2(\mathrm{CV})}, \end{aligned} \tag{9}$$

where $t_q(n-2)$ denotes the q-quantile of the t-distribution with $n-2$ degrees of freedom. Here, we choose the t-test and the between-within degrees of freedom (# of clusters—# of cluster-level covariates) as it frequently carries the nominal type I error rate even with a limited number of clusters.[22] Note Equation (9) includes the same correction factor as suggested in van Breukelen et al[11] for linear mixed model analysis without covariates, except that the marginal outcome-ICC ρ_y is replaced with the conditional outcome-ICC $\rho_{y\mid x}$. In practice, it is often the case that adjusting for a prognostic covariate leads to a smaller conditional ICC such that $\rho_{y\mid x} < \rho_y$, and this will inevitably reduce the variance for estimating the average treatment effect with model (2) compared to its counterpart without X_{ij}. The reduction in variance can directly translate into a smaller required sample size to achieve the same level of power, providing a more cost-effective approach for studying the treatment effect in CRTs by leveraging baseline covariates. Finally, the correction factor, $\theta_2(\mathrm{CV})$, in Equation (9) is identical to the correction factor in (7). This is somewhat expected because the treatment indicator in a CRT can be regarded as a cluster-level covariate and subject to the same efficiency impact due to cluster size variation.

3 | GENERALIZATION TO MULTIPLE OR MULTIVARIATE EFFECT MODIFIERS

3.1 | Generic variance expressions for the HTE and average treatment effect estimators

While our main focus is to elucidate the impact of unequal cluster sizes for tests of the HTE and average treatment effect with a univariate effect modifier, it is possible to extend the sample size methodology in the presence of multivariate effect modifiers. Suppose $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ is a p-dimensional vector of effect modifiers, the linear mixed model (2) can be extended as

$$Y_{ij} = \beta_1 + \beta_2 W_i + \boldsymbol{\beta}_3^T \boldsymbol{X}_{ij} + \boldsymbol{\beta}_4^T \boldsymbol{X}_{ij} W_i + \lambda_i + \epsilon_{ij}, \quad \lambda_i \sim \mathcal{N}\left(0, \sigma_\lambda^2\right), \quad \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right). \tag{10}$$

The specification of $\boldsymbol{X}_{ij}$ includes the following cases: (i) $\boldsymbol{X}_{ij}$ includes multiple univariate effect modifiers as linear terms; (ii) $\boldsymbol{X}_{ij}$ involves a univariate, individual-level effect modifier but is further decomposed into cluster-level and individual-level components to address different effects for the aggregated and lower-level variations, or the so-called contextual effect; this is further addressed in Section 3.2; (iii) $\boldsymbol{X}_{ij}$ involves a univariate, individual-level continuous effect modifier but with both linear and nonlinear terms; an example of this is given in Section 5 and Web Appendix D; (iv) combinations of the above. Below, we proceed with a general specification of $\boldsymbol{X}_{ij}$.

Suppose we are interested in jointly testing the null hypothesis of no HTE, $H_0: \boldsymbol{\beta}_4 = 0_{p \times 1}$, and we denote $\hat{\boldsymbol{\beta}}_4$ as the maximum likelihood estimator of the interaction parameters. Defining $\boldsymbol{\Omega}_4 = n\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_4\right)$ and $\hat{\boldsymbol{\Omega}}_4$ as its consistent estimator, we can proceed with the Wald statistic, $Q = n\hat{\boldsymbol{\beta}}_4^T \hat{\boldsymbol{\Omega}}_4^{-1} \hat{\boldsymbol{\beta}}_4$. Under the null, the test statistic Q follows a central χ^2-distribution with p degrees of freedom. Therefore, the sample size requirement for testing a HTE of size $\boldsymbol{\beta}_4 = \boldsymbol{\delta}$ can be expressed by

$$1 - \theta \leq \int_{\chi^2_{1-\alpha(p)}}^{\infty} \varphi\left(u; p, n\boldsymbol{\delta}^T \boldsymbol{\Omega}_4^{-1} \boldsymbol{\delta}\right) du, \tag{11}$$

where $\varphi(u; p, \zeta)$ is the density of the non-central χ^2-distribution with p degrees of freedom and non-centrality parameter ζ. The characterization of the sample size formula, therefore, depends on an explicit expression of the variance matrix $\boldsymbol{\Omega}_4$, or equivalently, the concentration matrix $\boldsymbol{\Omega}_4^{-1}$, as it plays a key role in determining the non-centrality parameter.

Similar to Section 2.2, we assume the cluster sizes m_i are sampled from a non-degenerate distribution with bounded first and second moments. This allows us to derive in Web Appendix B

$$\boldsymbol{\Omega}_4 = n\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_4\right) = \frac{\sigma^2_{y|x}(1-\rho_{y|x})\{1+(\bar{m}-1)\rho_{y|x}\}}{\bar{m}\sigma_w^2} \boldsymbol{\Lambda}_x^{-1/2} \boldsymbol{\Theta}\left(\mathrm{CV}\right)\left\{\boldsymbol{\Gamma}_x^1 + (\bar{m}-2)\rho_{y|x}\boldsymbol{\Gamma}_x^1 - (\bar{m}-1)\rho_{y|x}\boldsymbol{\Gamma}_x^0\right\}^{-1} \boldsymbol{\Lambda}_x^{-1/2}, \tag{12}$$

where the $p \times p$ diagonal matrix $\mathbf{\Lambda}_x = \text{diag}\left(\sigma_{x_1}^2, \ldots, \sigma_{x_p}^2\right)$ contains the marginal variance of each effect modifier, $\mathbf{\Gamma}_x^1 = \mathbf{\Lambda}_x^{-1/2}\left\{E\left(\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^T\right) - E\left(\boldsymbol{X}_{ij}\right)E\left(\boldsymbol{X}_{ij}\right)^T\right\}\mathbf{\Lambda}_x^{-1/2}$ is the marginal correlation matrix between p covariates (the diagonal elements are all 1 by definition of a marginal correlation matrix), and $\mathbf{\Gamma}_x^0 = \mathbf{\Lambda}_x^{-1/2}\left\{E\left(\boldsymbol{X}_{ij}\boldsymbol{X}_{ik}^T\right) - E\left(\boldsymbol{X}_{ij}\right)E\left(\boldsymbol{X}_{ik}\right)^T\right\}\mathbf{\Lambda}_x^{-1/2}$ is the multivariate counterpart of the covariate-ICC in the univariate case. The diagonal element of $\mathbf{\Gamma}_x^0$ is the covariate-ICC of each covariate, while the off-diagonal elements represent the intraclass cross-correlations between two different covariates. The middle multiplier in (12), $\mathbf{\Theta}(\text{CV})$, refers to a correction matrix, and can be considered as a multivariate extension of the correction factor, $\theta_1(\text{CV})$:

$$\begin{aligned}&\mathbf{\Theta}\left(\text{CV}\right)\\&= \left[\boldsymbol{I}_{p\times p} - \text{CV}^2\frac{\bar{m}\rho_{y\,|\,x}(1-\rho_{y\,|\,x})}{\{1+(\bar{m}-1)\rho_{y\,|\,x}\}^2}\left\{\mathbf{\Gamma}_x^1 + (\bar{m}-2)\rho_{y\,|\,x}\mathbf{\Gamma}_x^1 - (\bar{m}-1)\rho_{y\,|\,x}\mathbf{\Gamma}_x^0\right\}^{-1}\left(\mathbf{\Gamma}_x^0 - \rho_{y\,|\,x}\mathbf{\Gamma}_x^1\right)\right]^{-1},\end{aligned} \tag{13}$$

where $\boldsymbol{I}_{p\times p}$ is a $p \times p$ identity matrix. We notice that the variance expression (5) is a special case of (12) when $p = 1$, because in this case, $\mathbf{\Gamma}_x^1 = 1, \mathbf{\Gamma}_x^0 = \rho_x$, and $\mathbf{\Lambda}_x = \sigma_x^2$. With multivariate effect modifiers and $p \geq 2$, the impact of unequal cluster sizes on the non-centrality parameter and hence the statistical power is characterized by $\mathbf{\Theta}(\text{CV})$, whose limit is given by $\lim_{\bar{m}\to\infty}\mathbf{\Theta}(\text{CV}) = \boldsymbol{I}_{p\times p}$. Therefore, it is reasonable to expect that unequal cluster sizes generally have minimal impact on the sample size requirement for the joint test of the HTE with individual-level effect modifiers as long as the mean of the cluster sizes is not too small. Finally, in the special case when $\mathbf{\Gamma}_x^0 = \rho_{y\,|\,x}\mathbf{\Gamma}_x^1$ (resembling the condition $\rho_x = \rho_{y\,|\,x}$ in the univariate case), we have $\mathbf{\Theta}(\text{CV}) = \boldsymbol{I}_{p\times p}$ and the impact of unequal cluster sizes for testing the HTE is negligible regardless of other design parameters.

If the multivariate effect modifiers are all at the cluster level, then by definition $\mathbf{\Gamma}_x^1 = \mathbf{\Gamma}_x^0 = \mathbf{\Gamma}_x$, which all describe the marginal correlation matrix between the p cluster-level covariates. In this case, variance matrix (12) becomes

$$\begin{aligned}\mathbf{\Omega}_4 = n\text{Var}\left(\hat{\boldsymbol{\beta}}_4\right) &= \frac{\sigma_{y\,|\,x}^2\{1+(\bar{m}-1)\rho_{y\,|\,x}\}}{\bar{m}\sigma_w^2}\mathbf{\Lambda}_x^{-1/2}\mathbf{\Gamma}_x^{-1}\mathbf{\Lambda}_x^{-1/2}\\&\times\underbrace{\left[1 - \text{CV}^2\frac{\bar{m}\rho_{y\,|\,x}(1-\rho_{y\,|\,x})}{\{1+(\bar{m}-1)\rho_{y\,|\,x}\}^2}\right]^{-1}}_{\text{Correction Factor}\,\theta_2(\text{CV})},\end{aligned} \tag{14}$$

where we observe the correction matrix degenerates to a univariate correction factor $\theta_2(\text{CV})$, which takes the exact same form as the correction factor with a single cluster-level effect modifier.

While the main motivation of the linear mixed model (10) is to study a multivariate HTE, similar to Section 2.3, the model permits a covariate-adjusted estimator for the average treatment effect. Without loss of generality, we assume the multivariate effect modifiers are global mean centered such that $E\left(\boldsymbol{X}_{ij}\right) = 0_{p\times 1}$. In this case, $\hat{\beta}_2$ can be interpreted as a

covariate-adjusted average treatment effect estimator, similar to its counterpart in ANCOVA models for individually randomized trials.[21] We show in Web Appendix B that

$$n\mathrm{Var}\left(\hat{\beta}_2\right) = \frac{\sigma^2_{y\mid x}\left[1+(\bar{m}-1)\rho_{y\mid x}\right]}{\sigma^2_w \bar{m}} \times \theta_2\left(\mathrm{CV}\right), \tag{15}$$

which has the identical form as (8) except that $\rho_{y\mid x}$ is interpreted as the outcome-ICC conditional on the set of multivariate effect modifiers. Intuitively, adjustment for $\boldsymbol{X}_{ij}$ may reduce the conditional total variance of the outcome $\sigma^2_{y\mid x}$ and the outcome-ICC $\rho_{y\mid x}$ compared to their marginal counterparts (due to explained variation and explained clustering), hence leading to a more efficient average treatment effect estimator and a smaller required sample size to reach the same level of power. In this regard, our study of the sample size and variance expressions also provides an alternative perspective to justify the necessity for covariate adjustment in CRTs. We will further demonstrate this efficiency perspective for the average treatment effect analysis in our simulation study in Section 4.

3.2 | Application to a univariate effect modifier allowing for between- and within-cluster effects

The derivations in Section 3.1 also have implications for analyses with a single individual-level effect modifier, if the analysis proceeds by distinguishing the between-cluster and within-cluster effects from the effect modifier. We assume the effect modifier is global mean centered without loss of generality ($\mu_1 = 0$). In this case, the linear mixed model (10) becomes

$$Y_{ij} = \beta_1 + \beta_2 W_i + \beta_{31}\bar{X}_i + \beta_{32}\left(X_{ij}-\bar{X}_i\right) + \beta_{41}\bar{X}_i W_i + \beta_{42}\left(X_{ij}-\bar{X}_i\right)W_i + \lambda_i + \epsilon_{ij}, \quad \lambda_i \sim \mathcal{N}\left(0,\sigma^2_\lambda\right), \quad \epsilon_{ij} \sim \mathcal{N}\left(0,\sigma^2_\epsilon\right),$$

in which case the null hypothesis of no HTE is given by $H_0: \beta_{41} = \beta_{42} = 0$. When it is assumed that $\beta_{31} = \beta_{32}$ and $\beta_{41} = \beta_{42}$, this model reduces to model (2) assuming homogeneous between- and within-cluster effects. The idea of distinguishing between-cluster and within-cluster effects was discussed, for example, in Neuhaus and Kalbfleisch[23] and Kreft et al[24] The association parameters of the cluster-specific mean $\bar{X}_i$ has also been used to describe the contextual effect in a CRT.[7,25] We extend those ideas to include full treatment-by-covariate interaction terms to describe the HTE arising from the between-cluster and within-cluster variations of the individual-level effect modifier.

To facilitate the derivation of a more explicit variance matrix for sample size determination, we reparameterize the above model as follows:

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 W_i + \beta^*_{31}\bar{X}_i + \beta_{32}X_{ij} + \beta^*_{41}\bar{X}_i W_i + \beta_{42}X_{ij}W_i + \lambda_i + \epsilon_{ij}, \quad \lambda_i \sim \mathcal{N}\left(0,\sigma^2_\lambda\right), \\ &\epsilon_{ij} \sim \mathcal{N}\left(0,\sigma^2_\epsilon\right), \end{aligned} \tag{16}$$

where $\beta^*_{31} = \beta_{31} - \beta_{32}$ and $\beta^*_{41} = \beta_{41} - \beta_{42}$. It has been argued that β^*_{31} and β^*_{41} may be more interpretable than β_{31} and β_{41} as contextual effect parameters.[26,27] Based on the reparameterization, the null hypothesis of no HTE is given by $H_0: \beta^*_{41} = \beta_{42} = 0$. For sample size determination, the generic variance expression (12) derived in Section 3.1 can be

applied with some simplifications. For brevity, we derive in Web Appendix C the form of the concentration matrix, $\mathbf{\Omega}_4^{-1}$, which is

$$\mathbf{\Omega}_4^{-1} = \frac{\bar{m}\sigma_w^2\sigma_x^2}{\sigma_{y|x}^2(1-\rho_{y|x})\{1+(\bar{m}-1)\rho_{y|x}\}} \begin{bmatrix} r(\mathrm{CV})(1-\rho_{y|x})\theta_2^{-1}(\mathrm{CV}) & r(\mathrm{CV})(1-\rho_{y|x})\theta_2^{-1}(\mathrm{CV}) \\ r(\mathrm{CV})(1-\rho_{y|x})\theta_2^{-1}(\mathrm{CV}) & \{1+(\bar{m}-2)\rho_{y|x}-(\bar{m}-1)\rho_x\rho_{y|x}\}\theta_1^{-1}(\mathrm{CV}) \end{bmatrix}. \tag{17}$$

This explicit expression demonstrates that the impact of unequal cluster sizes for testing the HTE based on model (16) depends on the CV of cluster sizes only through $r(\mathrm{CV}), \theta_1(\mathrm{CV})$, and $\theta_2(\mathrm{CV})$. To understand the impact of unequal cluster sizes in the HTE analysis based on model (16), we plot the relative change in the determinant of the covariance matrix (or equivalently, concentration matrix), $\det(\mathbf{\Omega}_4^{-1}) = 1/\det(\mathbf{\Omega}_4)$, under unequal vs equal cluster sizes when average cluster sizes, $\bar{m} \in \{20,100\}$ and CV of cluster sizes $\in \{0.3,0.9\}$ in Figure 2. The choice of design parameters in Figure 2 follows exactly those in Figure 1. Mathematically, the relative change in determinant due to unequal cluster sizes is defined as

$$\begin{aligned} \frac{\det(\mathbf{\Omega}_4)}{\det(\mathbf{\Omega}_4)|_{\mathrm{CV}=0}} &= \frac{\{1+(\bar{m}-2)\rho_{y|x}-(\bar{m}-1)\rho_x\rho_{y|x}\}-r(0)(1-\rho_{y|x})}{\{1+(\bar{m}-2)\rho_{y|x}-(\bar{m}-1)\rho_x\rho_{y|x}\}\theta_1^{-1}(\mathrm{CV})\theta_2^{-1}(\mathrm{CV})-r(\mathrm{CV})(1-\rho_{y|x})\theta_2^{-2}(\mathrm{CV})} \\ &= \theta_2\Bigg(\mathrm{CV}\Bigg)\Bigg[1-\mathrm{CV}^2\frac{(1-\rho_{y|x})^2}{(\bar{m}-1)\{1+(\bar{m}-1)\rho_{y|x}\}^2}\left\{1-\mathrm{CV}^2\frac{\bar{m}\rho_{y|x}}{1+(\bar{m}-1)\rho_{y|x}}\right\}\Bigg]^{-1}, \end{aligned} \tag{18}$$

which surprisingly turns out to be free of the covariate-ICC due to the inclusion of a cluster-level mean covariate in model (16). The detailed derivation of this expression is given in Web Appendix C. From Figure 2, we observe that the relative change in the determinant of $\mathbf{\Omega}_4$ due to cluster size variation is minimum when the CV of cluster sizes is not large (eg, < 0.3), or when the mean cluster size is large and the conditional outcome-ICC is not small (eg, > 0.1), which is more similar to what we observe when testing the HTE with a cluster-level effect modifier. Because the determinant is a generic measure of the total information in a covariance matrix (also sometimes referred to as the generalized variance in the multivariate statistics literature[28]), it is reasonable to expect that cluster size variation has a larger impact on power for the test of the HTE based on model (16) when the CV is large, average cluster size is small and outcome-ICC is relatively small.

4 | SIMULATION STUDY

We conduct a simulation study to investigate the accuracy of our proposed sample size and power procedure for studying both the HTE and the average treatment effect in CRTs with unequal cluster sizes. The purpose of the study is several fold. First, we wish to confirm numerically that the predicted power by our formula agrees well with the empirical power of the Wald test for the HTE, provided that the test has a valid type I error rate. This exercise also enables us to empirically check the sensitivity of empirical power of the test of the

HTE to cluster size variation. Second, as the HTE model implies an unbiased estimator of the average treatment effect, we also investigate the accuracy of the proposed sample size procedure based on the adjusted average treatment effect estimator to confirm our analytical derivations. Finally, we wish to demonstrate, from a sample size saving perspective, that the covariate-adjusted average treatment effect estimator can be more efficient than the conventional unadjusted average treatment effect estimator. By quantifying the exact amount of sample size saving to achieve the same power, our study provides useful guidance when limited resources are available to design a CRT. To concentrate on the main idea, we focus on the scenario with an individual-level (continuous or binary) effect modifier, similar to Yang et al.[6]

4.1 | Simulation design

We consider two-arm CRTs with equal allocation of treatment groups, such that σ_w^2 is fixed at 1/4. Throughout, we fix the nominal type-I error rate at $\alpha = 5\,\%$ and the desired power at $1 - \zeta = 80\,\%$. We assume one individual-level effect modifier that is either continuous or binary. In either case, we consider the following parameters to compute the required number of clusters to achieve the desired level of power. The covariate ICC, $\rho_x \in \{0.10,0.25,0.50\}$ and the conditional outcome-ICC (of the outcome), $\rho_{y|x} \in \{0.01,0.05,0.10\}$. The covariate ICCs are chosen to reflect mildly to moderately correlated effect modifiers, and the outcome-ICCs are chosen to reflect the common values reported in CRTs, which rarely exceed 0.2.[8,29] The mean cluster size is chosen to be $\bar{m} \in \{20,50,100\}$, and the degree of cluster size variation is chosen to be $\mathrm{CV} \in \{0,0.3,0.6,0.9\}$. These values are in accordance with previous simulations of CRTs with unequal cluster sizes.[13,14] When the effect modifier is continuous, the marginal covariate variance, σ_x^2, is set to 1, and the true HTE parameter, $\beta_4 = \delta$, is among $\{0.10, 0.15, 0.25\}$. When the effect modifier is binary, we set the marginal prevalence to be $\mu_1 = 0.3$ and therefore $\sigma_x^2 = \mu_1(1 - \mu_1) = 0.21$; the true HTE parameter $\beta_4 = \delta \in \{0.25,0.35,0.45\}$. These various combinations of parameters for the binary and continuous effect modifiers lead to roughly similar sample size requirements holding all other parameters equal. Overall, there are 324 simulation scenarios in our study.

To confirm whether the predicted power by our formula agrees well with the empirical power of the test of the HTE given each combination of design parameters, the simulation proceeds as follows: (i) We calculate the required number of clusters n by solving Equation (6) and taking the smallest even number above to ensure an equal allocation of treatment groups. (ii) We obtain the predicted power based on our variance formula in Equation (5) (the result could be slightly over 80% due to rounding). (iii) When CV of cluster size is 0, we set the cluster size to $\bar{m}$. When CV is above zero, we simulate cluster sizes from a Gamma distribution with the shape parameter set to CV^{-2} and rate parameter set to $\bar{m}-1\mathrm{CV}^{-2}$. (iv) Given ρ_x, we generate covariates as follows: For a binary effect modifier, we first simulate the cluster-specific prevalence π_i from a beta distribution with shape parameters $a = \mu_1(\rho_x^{-1} - 1)$ and $b = (1 - \mu_1)(\rho_x^{-1} - 1)$, and then generate the individual-level covariate from Bernoulli (π_i). This ensures the marginal prevalence and the covariate-ICC to be μ_1 and ρ_x. For a continuous covariate, the cluster-specific mean μ_i is generated from $\mathcal{N}(0, \rho_x\sigma_x^2)$, and then the individual-level covariate is simulated from $\mathcal{N}(0.5 + \mu_i, (1 - \rho_x)\sigma_x^2)$.

(v) The treatment W_i is assigned at the cluster level with equal probability to treatment and control. (vi) We simulate the individual-level outcomes following model (2), with the intercept, $\beta_1 = 0$; the main effect, $\beta_2 = 0.25$; the main covariate effect, $\beta_3 = 0.1$, and the specified HTE parameter $\beta_4 = \delta$. The conditional total variance is set to be $\sigma_{y|x} = 1$. (vii) For each simulated trial, the linear mixed model (2) is fitted by the restricted maximum likelihood approach (REML), and the *P*-value for testing no HTE is obtained from the Wald test. Steps (iii)-(vii) are repeated 5000 times for each simulation scenario, and the empirical type I error rate (false rejection rate under $H_0: \beta_4 = 0$) and the empirical power (correct rejection rate under the alternative with $\beta_4 = \delta \neq 0$) are recorded. Based on the Bernoulli model with a fixed target probability (0.05 or 0.8), the Monte Carlo SE with 5000 simulations is 0.006 for empirical power and 0.003 for empirical type I error. Therefore, an empirical type I error rate within [0.044, 0.056] is considered nominal, and the difference between empirical power and predicted power within [−0.011, 0.011] is considered in close agreement. All analyses are done in R (version 4.0.1) using nlme package.[30] Example simulation code can be found at https://github.com/ttyale/HTE.

Second, we evaluate the performance of our sample size procedure for the covariate-adjusted average treatment effect analysis in Equation (9). We retain the aforementioned simulation procedure except for Steps (i) and (vii). Here, in Step (i), we compute the required number of clusters for testing the average treatment effect using Equation (9), and round the results to the smallest nearest even integer above. In step (vii), we still fit the linear mixed model (2) but ensures that the effect modifier X_{ij} is globally mean centered to have zero mean. This step does not affect the inference for the HTE parameter, but ensures that $\hat{\beta}_2$ can be interpreted as a (covariate-adjusted) average treatment effect estimator. The *P*-value for testing for no average treatment effect is obtained from the Wald *t*-test by fitting the linear mixed model. Finally, to illustrate the potential sample size saving for studying the average treatment effect under the HTE model, we conduct a parallel set of simulations where the outcomes are simulated from the linear mixed model (2) but the design and analysis proceed without X_{ij}. In this setting, we still resume the aforementioned simulation procedure, again, except for steps (i) and (vii). Here, in step (i), we compute the required number of clusters for testing the average treatment effect using Equation (9), but replace $\rho_{y|x}$ with the marginal outcome-ICC ρ_y, and $\sigma^2_{y|x}$ with the unadjusted variance σ^2_y. We use the formula derived in Yang et al[6] to arrive at $\sigma^2_y \approx \sigma^2_{y|x} + (\beta_3^2 + \beta_4^2/2 + \beta_3\beta_4)\sigma^2_x$, and to further induce the marginal outcome-ICC from the conditional outcome-ICC and covariate-ICC as $\rho_y \approx \omega\rho_{y|x} + (1-\omega)\rho_x$, where the weight is given by $\omega = \sigma^2_{y|x}/\sigma^2_y$. The required sample size is then estimated from the formula given ρ_y and compared with that estimated from formula (9) given $\rho_{y|x}$ to illustrate the potential sample size saving due to covariate adjustment with a single effect modifier. In Step (vii), we also fit the linear mixed model by omitting all terms with X_{ij} to verify the accuracy of the unadjusted sample size procedure.

4.2 | Simulation results

Web Table 1 summarizes the estimated number of clusters (n) using the proposed formula (9), the empirical type I error rate under the null of no HTE, the empirical power and the predicted power of the test of the HTE under $\delta = 0.15$ with a continuous individual-level

effect modifier. To highlight the main findings, the results on type I error and power are also graphically summarized in Figures 3 and 4. In general, the test of the HTE maintains a valid type I error rate, and carries empirical power that agrees well with the analytical predictions across all scenarios. In extreme cases with CV = 0.9, the empirical power may be slightly lower than the predicted power when the number of clusters is small (Figure 3D), which is expected given our sample size approximation relies on asymptotic theory where n becomes sufficiently large. As shown by comparisons across different levels of CV, the estimated numbers of clusters are generally not dramatically affected by the degree of cluster size variation (by comparing different panels of Figure 3), except when the average cluster size is small. For example, when $\bar{m} = 20$, a large CV of cluster size can mildly inflate the required sample size to achieve the same power, especially when the difference between ρ_x and $\rho_{y|x}$ is also large. However, when $\bar{m} = 100$, the impact of CV on required sample size becomes minimal, regardless of ICC parameters; this confirms our analytical findings in Section 2.2. Additional simulation results of the test of the HTE with different effect sizes, $\delta = 0.10$ and $\delta = 0.25$, and the same continuous effect modifier can be found in Web Tables 2 and 3, with qualitatively similar findings. Comparing across the results with different effect sizes, it appears that the required sample size for HTE analysis may be sensitive to effect size and becomes more sensitive to cluster size variation when the HTE is small. The parallel results for a binary effect modifier with $\delta = 0.25, \delta = 0.35$, and $\delta = 0.45$ are presented in Web Tables 4 to 6. The results and patterns for the a binary effect modifier are similar to that for a continuous effect modifier, and confirms the accuracy of the proposed sample size formula.

Web Table 7 presents the required number of clusters (n) for the test of the average treatment effect based on the HTE model with a treatment-by-covariate interaction term, the empirical type I error rate under the null of no average treatment effect, the empirical power and the predicted power for the test of the average treatment effect with a continuous effect modifier. Under the same parameter configuration as in Web Table 1, the induced average treatment effect from the HTE simulation model is fixed at 0.325 under $\delta = 0.15$. Across all scenarios, the empirical type I error rates for the covariate-adjusted test of the average treatment effect are close to 0.05, and the empirical power is consistent with the predicted power except for a few extreme cases when the required number of clusters is small and the CV of cluster size is largest (CV = 0.9). Compared to Web Table 1, Web Table 7 also shows that the sample size for average treatment effect analysis can be more sensitive to the conditional outcome-ICC, $\rho_{y|x}$, than that for the HTE analysis. The covariate-ICC, however, does not change the required sample size for adjusted average treatment effect analysis. Finally, we find that the sample size for the covariate-adjusted average treatment effect analysis can generally be more sensitive to the degree of cluster size variation than that for the HTE analysis, especially when the CV of cluster size becomes large. Web Tables 8 and 9 present the corresponding results for $\delta = 0.10$ with an induced average treatment effect of 0.30, and $\delta = 0.25$ with an induced average treatment effect of 0.375. Findings under these alternative effect sizes of the average treatment effect are similar to those in Web Table 7.

Web Table 10 presents the required number of clusters (n) for the test of the unadjusted average treatment effect ignoring the continuous effect modifier, the empirical type I error rate of the test, the empirical power and the predicted power for the test of the unadjusted

average treatment effect with an induced average treatment effect of 0.325($\delta = 0.15$). To help contrast with the results in Web Table 7, we plot the estimated sample size adjusting for the effect modifier vs that ignoring the effect modifier in Figure 5. Even with one continuous effect modifier, the estimated number of clusters ignoring the covariate is 2 to 6 larger than that needed when adjusting for the effect modifier. As we elaborated in Section 2.3, this is because the marginal outcome-ICC ρ_y can be larger than the conditional outcome-ICC given the effect modifier $\rho_{y|x}$. Likewise, the marginal total variance of the outcome σ_y^2 can be smaller than the conditional total variance $\sigma_{y|x}^2$, due to explained variation. These two sources of changes lead to reduction in needed sample size when a covariate is considered in the design and analysis of a CRT. Web Tables 11 and 12 summarize the corresponding results with alternative values of the δ, and therefore an induced average treatment effect of 0.300 and 0.375. The results are consistent with those in Web Table 10.

5 | ILLUSTRATIVE SAMPLE SIZE CALCULATION WITH THE STRIDE STUDY

To illustrate our proposed sample size procedure, we compute the required sample size for detecting the HTE and average treatment effect in the context of the design of the STRIDE trial. The STRIDE trial was a two-arm, pragmatic CRT with 5451 community-dwelling individuals aged 70 or above and at high risk for a serious fall injury. Individuals were clustered in 86 primary care practices. Intervention and enhanced usual care were randomized at the practice level with a 1:1 ratio. The intervention included comprehensive assessment, recommendations, and motivational interview on fall risk factors as well as developing and implementing an individualized fall care plan, whereas the enhanced usual care included a falls-information pamphlet and were encouraged to discuss fall risk with their primary care provider. The average cluster size of the study was $\bar{m} = 63$, with an estimated CV of cluster size around 0.5. Here, we focus on a continuous secondary outcome, concern score about falling, and two potential effect modifiers measured at the individual level, age and self-rated health (SRH). Additional details of the study can be found at the clinicaltrials.gov (NCT02475850) and elsewhere.[16,17,31]

The concern about falling outcome ranged from 10 to 40 and was assessed using a modified Fall Efficacy Scale at 24 months post intervention.[31] Throughout we consider two-sided tests with nominal 5% type I error rate and 20% type II error rate (80% power). Suppose we are interested in studying the potential effect modification with respect to the continuous individual-level covariate, age, which has marginal SD of 6.9 (assuming age is mean centered). We estimate the covariate-ICC of age using the STRIDE baseline data to be $\rho_x = 0.025$, and the conditional outcome-ICC to be $\rho_{y|x} = 0.01$. Using formula (6), the required number of clusters to detect the age-related HTE with a standardized effect size, $\delta\sigma_x/\sigma_{y|x} = 0.1$ (interpreted as the effect on SD unit increase in covariate on SD unit of the outcome) is estimated to be $n = 52$. In Table 1, we further vary the design parameters with conditional outcome-ICC, $\rho_{y|x} \in \{0.01, 0.05\}$, covariate-ICC, $\rho_x \in \{0.01, 0.05, 0.10, 0.20\}$, and three other values of CV of cluster sizes as a sensitivity analysis for the same target effect size. It is apparent that the required number of clusters is relatively insensitive to the cluster size variation (leading to requiring at most 2 more clusters under CV = 0.75), which is not

surprising as we have shown that the correction factor for testing an individual-level HTE due to unequal cluster sizes is generally close to 1. However, the sample size for the HTE will be jointly affected by the outcome-ICC and covariate-ICC in a non-monotone fashion, consistent with the findings in Yang et al[6] assuming equal cluster sizes. Furthermore, suppose we are interested in studying the potential effect modification with respect to the binary individual-level covariate, SRH, which measures whether one has good/excellent self-rated health. The marginal prevalence and the SD of SRH is 0.2 and 0.4, respectively; the covariate-ICC for this binary covariate is estimated by the modified moment estimator[32] to be $\rho_x = 0.05$. Assuming $\rho_{y|x} = 0.01, n = 80$ clusters are required to detect a HTE effect size of $\delta/\sigma_{y|x} = 0.2$ (interpreted as the effect from change in SRH on the SD unit of the outcome). In Table 1, our sensitivity analysis by varying design parameters indicate that the sample size is insensitive to the cluster size variation, but will jointly depend on the values of the covariate-ICC and outcome-ICC.

We then illustrate the calculation of the sample size for the average treatment effect on the concern about falling outcome, but consider using the conditional outcome-ICC according to Section 2.3. Our analysis of the STRIDE data suggests that the conditional outcome-ICC given either the age or SRH covariate is around 0.01. Following the original protocol, the required number of clusters is estimated to be $n = 12$ to detect a standardized effect size of $\Delta/\sigma_{y|x} = 0.3$ when the CV of cluster sizes is 0.5. A larger CV can require 2 more clusters as shown in Table 1. If the conditional outcome-ICC increases to 0.05, the required number of clusters is estimated to be $n = 76$, indicating the sensitivity of the average treatment effect sample size to outcome-ICC. However, in general, the estimated sample size based on the conditional outcome-ICC can be smaller than the conventional approach based on the marginal outcome-ICC, as we demonstrated in the simulation study in Section 4.

Finally, a potential limitation of our sample size calculation for the HTE analysis with age in Table 1 is that we have assumed a linear HTE model without higher-order terms for age. In the case where nonlinear effects of age are expected, the approach proposed in Section 3.1 can be applied to determine the required sample size for detecting the HTE. As a sensitivity analysis for the sample size to detect effect moderation by age, we assume a quadratic HTE model by additionally including quadratic age terms in both the main effect as well as the interaction effect (model details are provided in Web Appendix D), and combine Equations (11) and (12) to solve for the required number of clusters. We assume that the collection of age variables in each cluster is multivariate normal with mean zero, marginal variance σ_x^2 and common ICC ρ_x, under which condition we show in Web Appendix D that the variance of age squared is $2\sigma_x^4$. Therefore, $\mathbf{\Lambda}_x = \mathrm{diag}\{\sigma_x^2, 2\sigma_x^4\}$. Under this multivariate normal model, we further derive in Web Appendix D that

$$\mathbf{\Gamma}_x^1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{\Gamma}_x^0 = \begin{bmatrix} \rho_x & 0 \\ 0 & \rho_x^2 \end{bmatrix}.$$

Fixing the standardized effect size for linear age-related HTE as 0.1 as in Table 1, we vary the standardized effect size for quadratic age-related HTE from {0, 0.1, 0.2, 0.3} and present the sample size estimates in Table 2. If the true quadratic HTE is zero (or equivalently there

only exists linear HTE), an increased number of clusters is needed to attain 80% power based on the quadratic HTE model compared to the results under linear HTE model in Table 1. However, a smaller number of clusters is needed to achieve 80% power when there indeed exists a positive quadratic HTE based on age squared, and the sample size estimates can be further reduced with a stronger quadratic HTE effect size. In Web Table 13, we further obtain the required sample sizes based on the quadratic HTE model but assuming that the linear HTE of age is in fact 0. The resulting sample sizes are no smaller than those in Table 2, due to a reduced overall HTE signal, but may be either larger or smaller than those in Table 1. Taken together, these results suggest that the sample size estimates from a linear HTE model are generally conservative when there exists both linear and quadratic HTEs. However, the sample size estimates from a linear HTE model can be either smaller or larger when there only exists a quadratic HTE. Similar to Table 1, the CV of cluster sizes has negligible effect on the sample size for studying the HTE with an individual-level effect modifier.

6 | DISCUSSION

In this article, we propose a set of sample size procedures to account for unequal cluster sizes in CRTs detecting treatment effect heterogeneity. With a univariate, individual-level effect modifier, we show that the correction factor due to cluster size variability is close to unity under certain parameter regions, such that the procedure in Yang et al[6] usually provides a reasonable approximation despite their assumption of equal cluster sizes. However, when the effect modifier is measured at the cluster level, the correction factor can often be larger than unity, and therefore the required sample size for detecting the HTE can be more sensitive to the degree of cluster size variability. By focusing on a univariate effect modifier, our numerical studies demonstrate that the proposed sample size procedure is accurate in finite samples under common values of the CV of cluster sizes. In any case, the correction factor we proposed can be used in the design stage to formally quantify the amount of additional clusters required to achieve the desired power for studying the HTE in CRTs with unequal cluster sizes. Finally, we have also generalized our procedure to accommodate multiple effect modifiers by developing a correction matrix (which reduces to a scalar correction factor with multiple cluster-level effect modifiers), discuss its potential application for designing CRTs with an anticipated contextual effect in Section 3.2, and illustrate its use in powering a nonlinear HTE in Section 5.

There has been an extensive effort in studying the impact of unequal cluster sizes in CRTs on the estimation and inference of the average treatment effect; see, for example, Kerry et al[10] and van Breukelen et al[11] with a continuous outcome; Candel et al,[12] Li and Tong[13] for a binary outcome; Li and Tong[14] for a count outcome. The consensus from these prior investigations is that the impact of cluster size variability depends on the choice of analysis. Whereas the impact of cluster size variability is much larger when the analysis proceeds by ignoring the outcome ICC (eg, cluster-level analysis or independence generalized estimating equations), the impact of cluster size variability can be milder when the estimation of treatment effect accounts for the outcome ICC (eg, a linear mixed model or generalized estimating equation with an exchangeable working correlation structure). In the latter case, the loss of power due to unequal cluster sizes needs to be compensated

for by including an additional 10% to 20% more clusters depending on the mean and CV of cluster sizes as well as the outcome ICC. However, these prior investigations have not considered covariates in the analytical models, and to the best of our knowledge, our work is the first to address the issue of unequal cluster sizes in CRTs when the interest lies in detecting the treatment-by-covariate interaction effect. When the covariate is measured at the cluster level, we show that the correction factor due to unequal cluster sizes has the same form of the expression as what was previously derived for studying the average treatment effect in CRTs by van Breukelen et al.[11] This is intuitive because the interaction term can be regarded as a cluster-level covariate, just like the treatment variable. On the other hand, if the covariate is measured at the individual level, the impact of unequal cluster sizes on power is often much smaller, perhaps due to the additional information gain by leveraging within-cluster contrasts and informing the individual-level interaction effect. As we show in Equation (6), the amount of potential sample size inflation jointly depends on the mean and CV of cluster sizes, covariate-ICC and conditional outcome-ICC. Through graphical exploration, we have demonstrated that the largest variance inflation due to unequal cluster sizes in studying an individual-level HTE occurs when the covariate-ICC approaches one (in which case the individual-level covariate is highly correlated in each cluster and behaves like a cluster-level covariate) and the mean cluster size is small. In the special case when the covariate-ICC equals the conditional outcome-ICC, the correlation factor equals to one and we expect minimal impact of unequal cluster sizes for studying an individual-level HTE.

In addition to developing new design formulas for studying HTEs, we have further provided insights into studying the average treatment effect in CRTs. In particular, models (2) and (10) can be regarded as multilevel analysis of covariance (ANCOVA) models, incorporating a cluster-level random effect into the classical ANCOVA model for independent and identically distributed observations. Connecting with the literature on efficient ANCOVA analysis of individually randomized trials,[6,21] we show that the HTE-motivated multilevel ANCOVA framework can also be used for potentially more efficient average treatment effect analysis. We provide a sample size formula for the covariate-adjusted average treatment effect analysis, which turns out to bear the same form as the sample size formula for the unadjusted average treatment effect even under unequal cluster sizes. A subtle difference, however, is that our sample size for the average treatment effect requires the conditional outcome variance, $\sigma_{y|x}$, and the conditional outcome-ICC, $\rho_{y|x}$, rather than the marginal counterpart in the traditional formula. Frequently, the adjustment of baseline covariates can explain residual variability as well as the degree of clustering in CRTs, as shown, for example, in the empirical study by Murray and Blistein.[8] The reduction of residual variability as well as the degree of clustering can then translate into sample size savings without compromising the study power. Therefore, from a sample size point of view, our results offer an alternative justification for leveraging baseline covariates in CRTs. In Section 4, our simulations demonstrate that one could save 2 to 6 clusters when powering on the average treatment effect by including a single continuous effect modifier with a small covariate effect and small to moderate covariate-ICC. We expect additional efficiency gain with more prognostic covariates, but this topic is subject to additional research. Relatedly, covariate adjustment has been recently written into a US FDA guidance document[33] for individually randomized trials, whereas no such guidance is currently available for

pragmatic CRTs; our findings may therefore also promote the application of the multilevel ANCOVA model in CRTs as a unifying framework for studying the HTE and the average treatment effect.

The investigation into the HTE is a trending topic for pragmatic CRTs, with continuing development of new statistical methodology. In this work, we focus on confirmatory HTE analysis with pre-specified effect modifiers and aim to quantify the design resources needed to achieve sufficient power for confirmatory HTE analysis. This sets us apart from the exploratory HTE analysis that is mostly data-driven and post-hoc, for which power analysis remains less relevant due to lack of pre-specification. To apply our sample size methodology, an additional assumption on the value of covariate-ICC is required beyond conventional assumptions in designing CRTs. However, reporting covariate-ICC has not yet become standard practice in CRTs and warrants additional research. Our view is that the lack of current covariate-ICC estimates does not invalidate our procedure, and we strongly encourage more empirical studies to report covariate-ICCs along with outcome-ICCs, perhaps by exploiting existing databases in order to facilitate the planning of future CRTs with a HTE objective. Korevaar et al,[34] have provided a good example of such a study by providing a comprehensive set of ICC estimates from the CLustered OUtcome Dataset (CLOUD) bank to assist in the planning of longitudinal and stepped wedge CRTs subject to complex correlation structures. On the other hand, we have also exemplified how sensitivity analysis can be carried out by varying values of the covariate-ICC in Section 5, and researchers may choose the most conservative sample size estimate without an accurate covariate-ICC estimate in the design stage. Alternatively, one can consider an internal pilot where the sample size for the larger study is re-estimated conditional on the updated covariate-ICC and outcome-ICC values from the pilot, extending the work of Lake et al[35] and van Schie and Moerbeek.[36]

One potential limitation of the current study is that we have focused on a continuous outcome. In practice, binary outcomes are also common in CRTs. While our proposed methods can provide a rough approximation when the HTE is measured based on the risk difference, they may not be directly applied when the HTE is measured based on the relative risk or odds ratio. To this end, we plan to conduct additional research to develop new sample size formulas for HTE analysis of CRTs based on generalized linear mixed models with non-identity link functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Research in this article was supported by a Patient-Centered Outcomes Research Institute Award® (PCORI® Award ME-2020C3-21072), and by CTSA Grant Number UL1 TR001863 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH). The statements presented in this article are solely the responsibility of the authors and do not necessarily represent the views of PCORI®, its Board of Governors or Methodology Committee, or the National Institutes of Health. The authors thank Drs. Erich Greene and Peter Peduzzi for their help in providing the summary data from the STRIDE trial for our illustrative application in Section 5.We thank the Associate Editor and two anonymous reviewers for their helpful comments.

Funding information

National Center for Advancing Translational Science (NCATS), Grant/Award Number: UL1 TR001863; Patient-Centered Outcomes Research Institute, Grant/Award Number: ME-2020C3-21072

DATA AVAILABILITY STATEMENT

The article is primarily focusing on study designs and therefore does not involve analysis of individual-level datasets.

REFERENCES

1. Murray DM. Design and Analysis of Group-Randomized Trials. 29th ed. Oxford, UK: Oxford University Press; 1998.
2. 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]
3. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q. 2004;82(4):661–687. [PubMed: 15595946]
4. Kent DM, Paulus JK, Van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med. 2020;172(1):35–45. [PubMed: 31711134]
5. Starks MA, Sanders GD, Coeytaux RR, et al. Assessing heterogeneity of treatment effect analyses in health-related cluster randomized trials: a systematic review. PLoS One. 2019;14(8):e0219894. [PubMed: 31404063]
6. Yang S, Li F, Starks MA, Hernandez AF, Mentz RJ, Choudhury KR. Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. Stat Med. 2020.
7. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. Psychol Methods. 1997;2(2):173.
8. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Eval Rev. 2003;27(1): 79–103. [PubMed: 12568061]
9. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin Trials. 2004;1(1):80–90. [PubMed: 16281464]
10. Kerry SM, Martin BJ. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med. 2001;20(3):377–390. [PubMed: 11180308]
11. van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicenter trials. Stat Med. 2007;26(13):2589–2603. [PubMed: 17094074]
12. Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. Stat Med. 2010;29(14):1488–1501. [PubMed: 20101669]
13. Li F, Tong G. Sample size estimation for modified Poisson analysis of cluster randomized trials with a binary outcome. Stat Methods Med Res. 2021;30(5):1288–1305. [PubMed: 33826454]
14. Li F, Tong G. Sample size and power considerations for cluster randomized trials with count outcomes subject to right truncation. Biom J. 2021;63(5):1052–1071. [PubMed: 33751620]
15. 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]
16. Bhasin S, Gill TM, Reuben DB, et al. A randomized trial of a multifactorial strategy to prevent serious fall injuries. N Engl J Med. 2020;383(2):129–140. [PubMed: 32640131]
17. Bhasin S, Gill TM, Reuben DB, et al. Strategies to reduce injuries and develop confidence in elders (STRIDE): a cluster-randomized pragmatic trial of a multifactorial fall injury prevention strategy: design and methods. J Gerontol Se A. 2018;73(8): 1053–1061.

18. Spybrook J, Kelcey B, Dong N. Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. J Educ Behav Stat. 2016;41(6):605–627.
19. Dong N, Kelcey B, Spybrook J. Power analyses for moderator effects in three-level cluster randomized trials. J Exp Educ. 2018;86(3): 489–514.
20. Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. Am Stat. 2001;55(4):314–321.
21. Lin W Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. Ann Appl Stat. 2013;7(1): 295–318.
22. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the generalized linearmixed model in analyzing binary outcome in small sample cluster-randomized trials. BMC Med Res Methodol. 2015;15(1):1–12. [PubMed: 25555466]
23. Neuhaus JM, Kalbfleisch JD. Between-and within-cluster covariate effects in the analysis of clustered data. Biometrics. 1998;54(2):638–645. [PubMed: 9629647]
24. Kreft IG, De Leeuw J, Aiken LS. The effect of different forms of centering in hierarchical linear models. Multivar Behav Res. 1995;30(1): 1–21.
25. Korendijk EJ, Hox JJ, Moerbeek M, Maas CJ. Robustness of parameter and standard error estimates against ignoring a contextual effect of a subject-level covariate in cluster-randomized trials. Behav Res Methods. 2011;43(4):1003–1013. [PubMed: 21512874]
26. Begg MD, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. Stat Med. 2003;22(16):2591–2602. [PubMed: 12898546]
27. Seaman S, Pavlou M, Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. Stat Med. 2014;33(30):5371–5387. [PubMed: 25087978]
28. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. Upper Saddle River, NJ: Prentice Hall; 2002.
29. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004;94(3):423–432. [PubMed: 14998806]
30. Pinheiro J, Bates D. Mixed-effects models in S and S-PLUS. Berlin, Germany: Springer Science & Business Media; 2006.
31. Gill TM, Bhasin S, Reuben DB, et al. Effect of amultifactorial fall injury prevention intervention on patient well-being: the STRIDE study. J Am Geriatr Soc. 2021;69(1):173–179. [PubMed: 33037632]
32. Kleinman JC. Proportions with extraneous variance: single and independent samples. J Am Stat Assoc. 1973;68(341):46–54.
33. Center for Drug Evaluation and Research, U.S. Food and Drug Administration. Adjusting for covariates in randomized clinical trials for drugs and biologics with continuous outcomes guidance for industry; 2002.
34. Korevaar E, Kasza J, Taljaard M, et al. Intra-cluster correlations from the clustered outcome dataset bank to inform the design of longitudinal cluster trials. Clin Trials. 2021;17407745211020852.
35. Lake S, Kammann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. Stat Med. 2002;21(10): 1337–1350. [PubMed: 12185888]
36. van Schie S, Moerbeek M. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. Stat Med. 2014;33(19):3253–3268. [PubMed: 24719285]

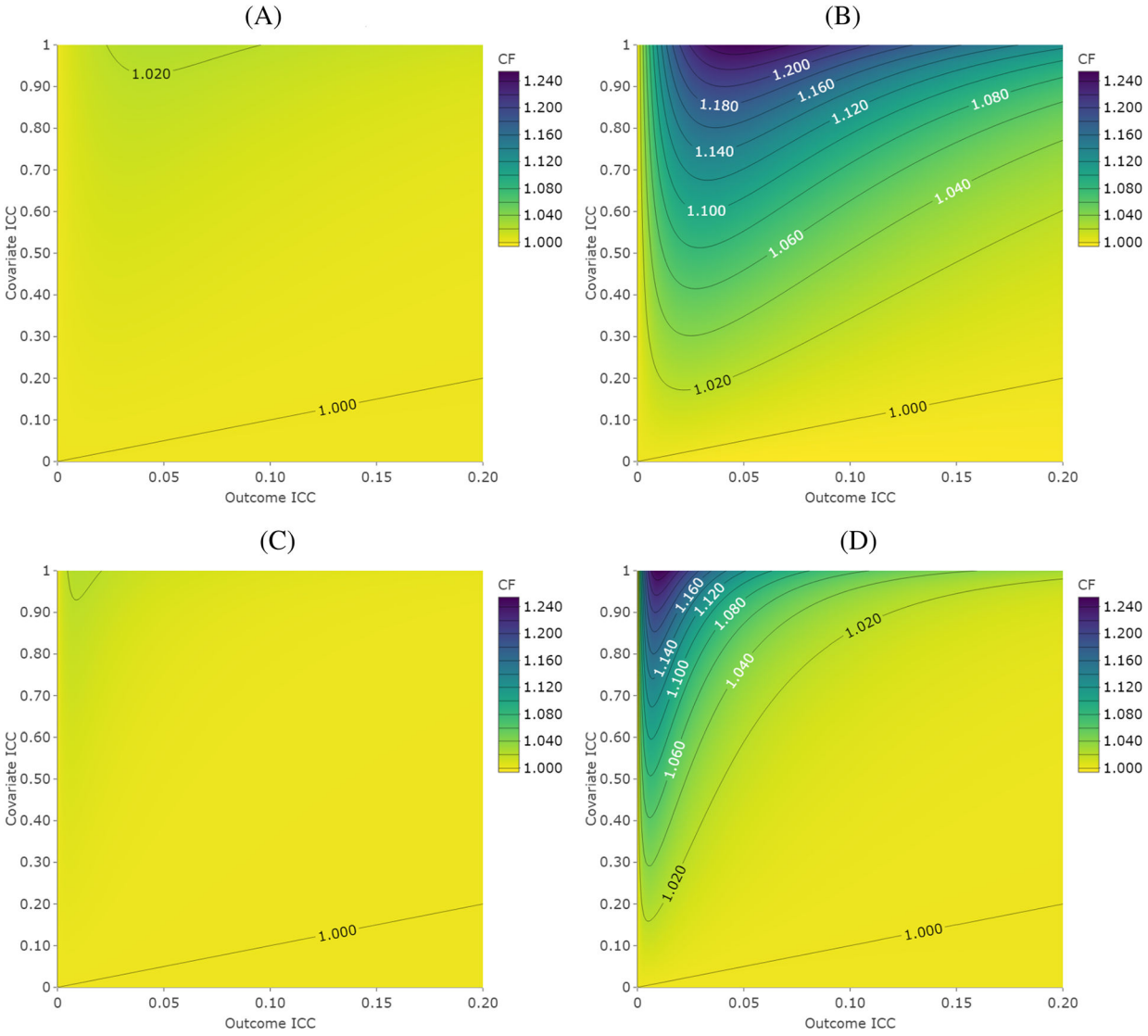

FIGURE 1.

The correction factor (CF = θ_1(CV)) as a function of average cluster sizes ($\bar{m} \in \{20,100\}$), coefficient of variation (CV) of cluster sizes $\in \{0.3, 0.9\}$, covariate ICC $\in [0, 1]$ and outcome ICC $\in [0, 0.2]$. When the covariate ICC $\rho_x = 1$, the effect modifier is a cluster-level covariate. (A) $\bar{m} = 20$, CV = 0.3; (B) $\bar{m} = 20$, CV = 0.9; (C) $\bar{m} = 100$, CV = 0.3; (D) $\bar{m} = 10$, CV = 0.9

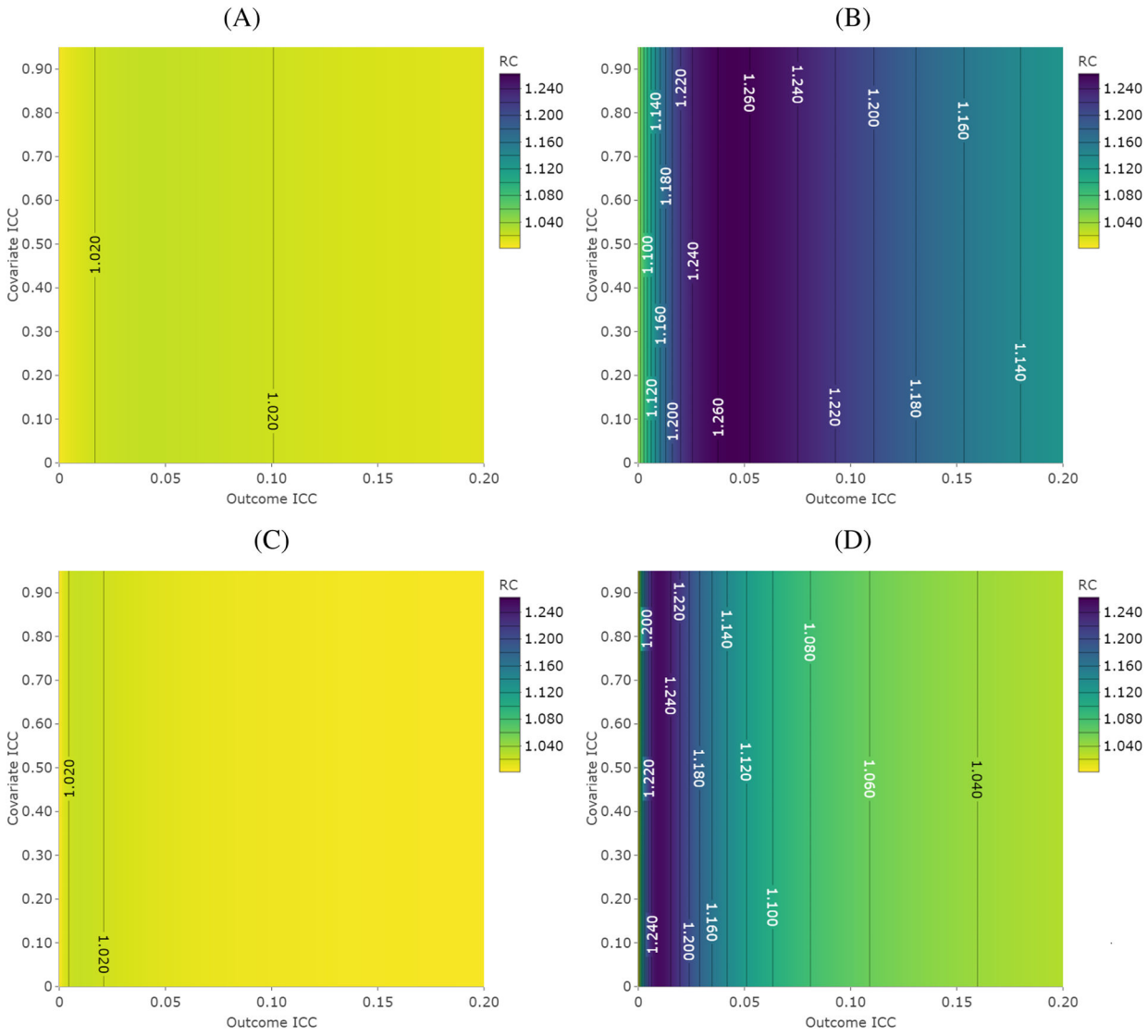

FIGURE 2.

The relative change (RC) in the determinant of the covariance matrix, $\det(\boldsymbol{\Omega}_4)/\{\det(\boldsymbol{\Omega}_4)|_{CV=0}\}$, as a function of average cluster sizes $\bar{m} \in \{20, 100\}$, coefficient of variation (CV) of cluster sizes $\in \{0.3, 0.9\}$, covariate ICC $\in [0, 0.95]$ and outcome ICC $\in [0, 0.2]$. We consider the largest covariate-ICC to be 0.95 because model (16) is not estimable if X_{ij} is a cluster-level covariate with $\rho_x = 1$. (A) $\bar{m} = 20$, CV = 0.3; (B) $\bar{m} = 20$, CV = 0.9;(C) $\bar{m} = 100$, CV = 0.3; (D) $\bar{m} = 100$, CV = 0.9

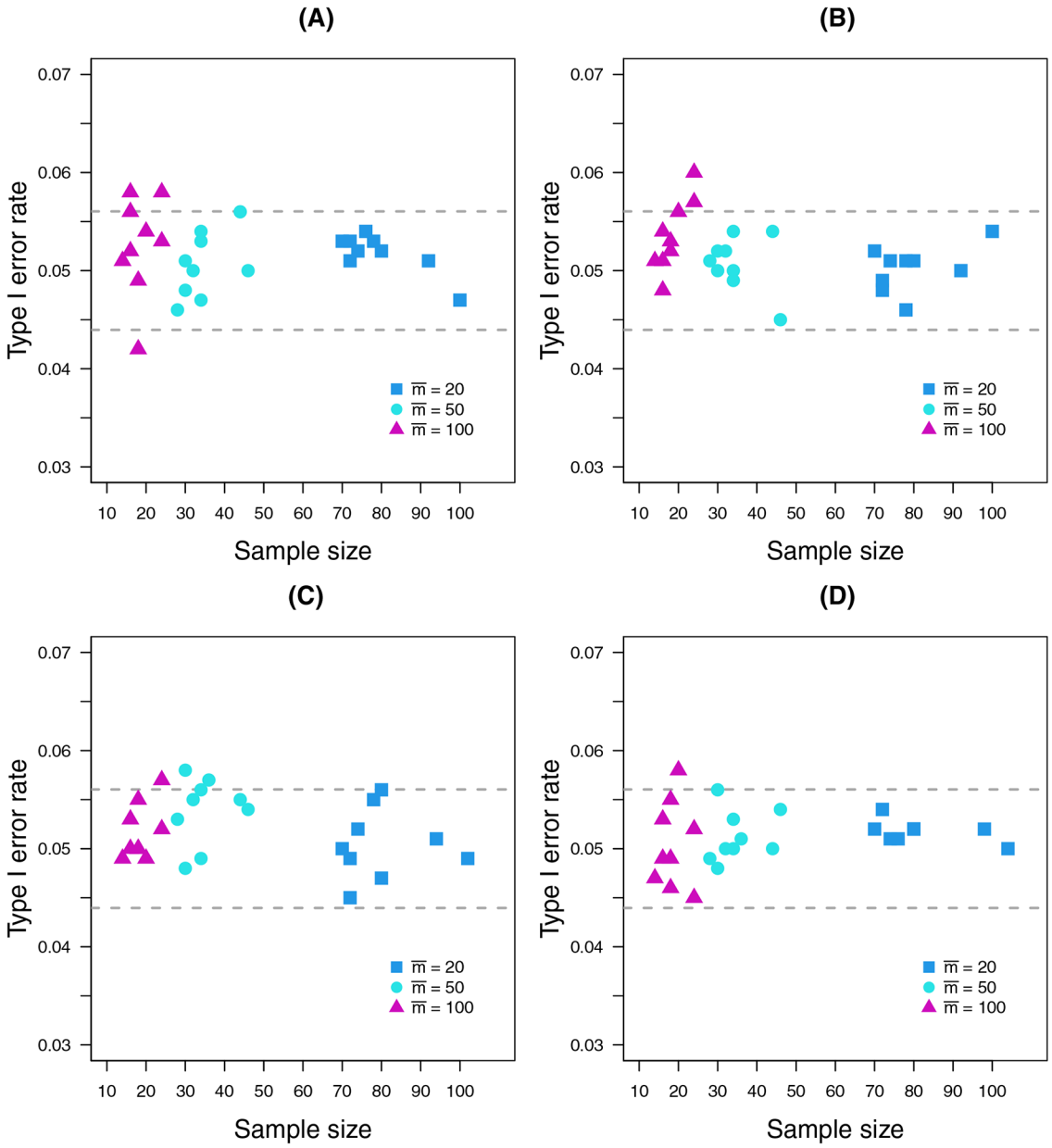

FIGURE 3.

Empirical type I error rate for studying the HTE as a function of the estimated sample size (number of clusters) with a continuous individual-level effect modifier, by four different coefficients of variation (CV) of cluster sizes. The dashed lines indicate the Monte Carlo error bounds based on 5000 simulations with a 5% nominal type I error rate. Details on the design parameters for each scenario are provided in Web Table 1. (A) CV = 0; (B) CV = 0.3; (C) CV = 0.6; (D) CV = 0.9

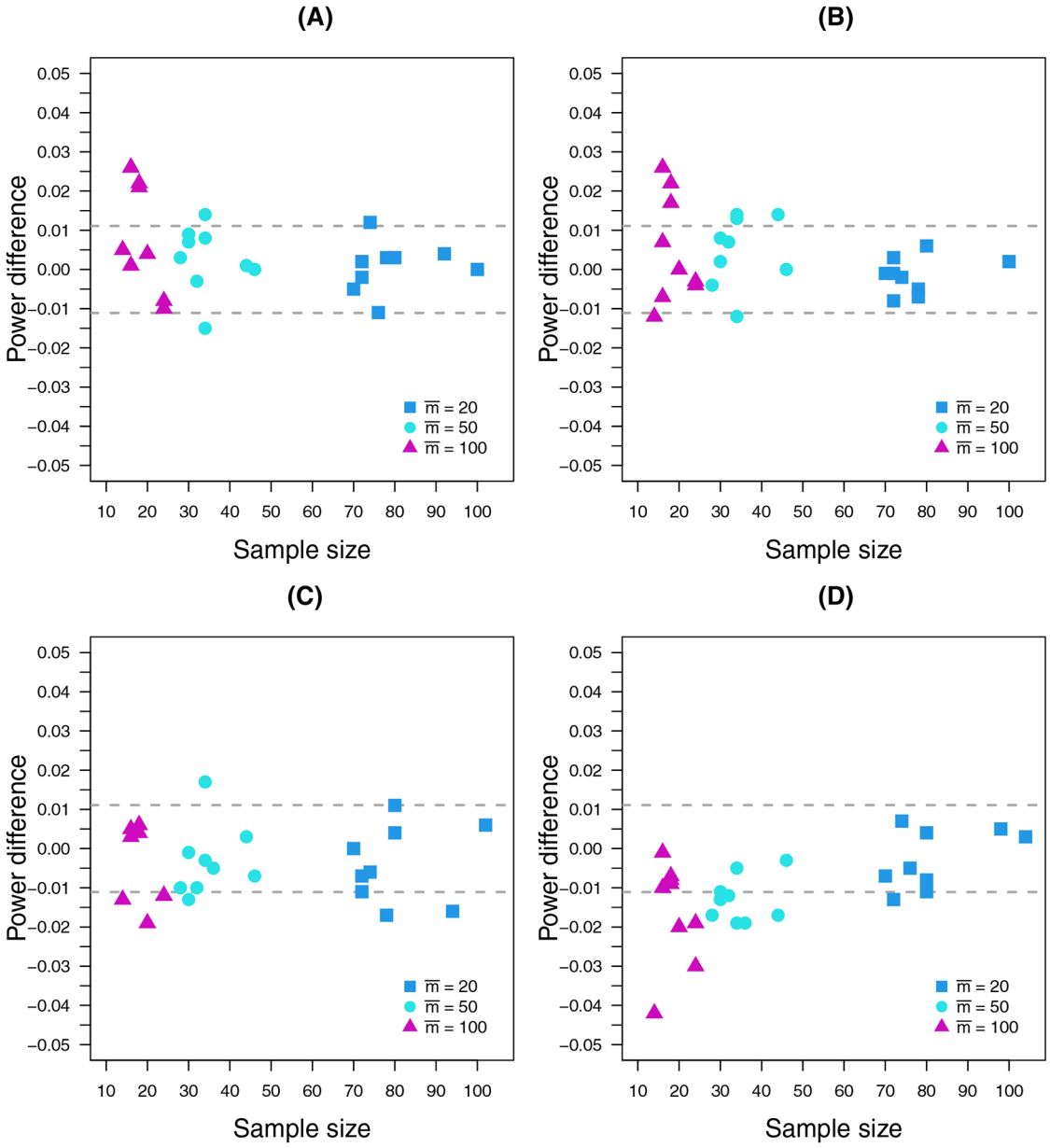

FIGURE 4.

Difference between the empirical and predicted power for studying the HTE as a function of the estimated sample sizes (number of clusters) with a continuous individual-level effect modifier, by four different coefficients of variation (CV) of cluster sizes. The effect size for power is set to be $\beta_4 = \delta = 0.15$. The dashed lines indicate the Monte Carlo error bounds based on 5000 simulations with 80% target power. Details on the design parameters for each scenario are provided in Web Table 1. (A) CV = 0; (B) CV = 0.3; (C) CV = 0.6; (D) CV = 0.9

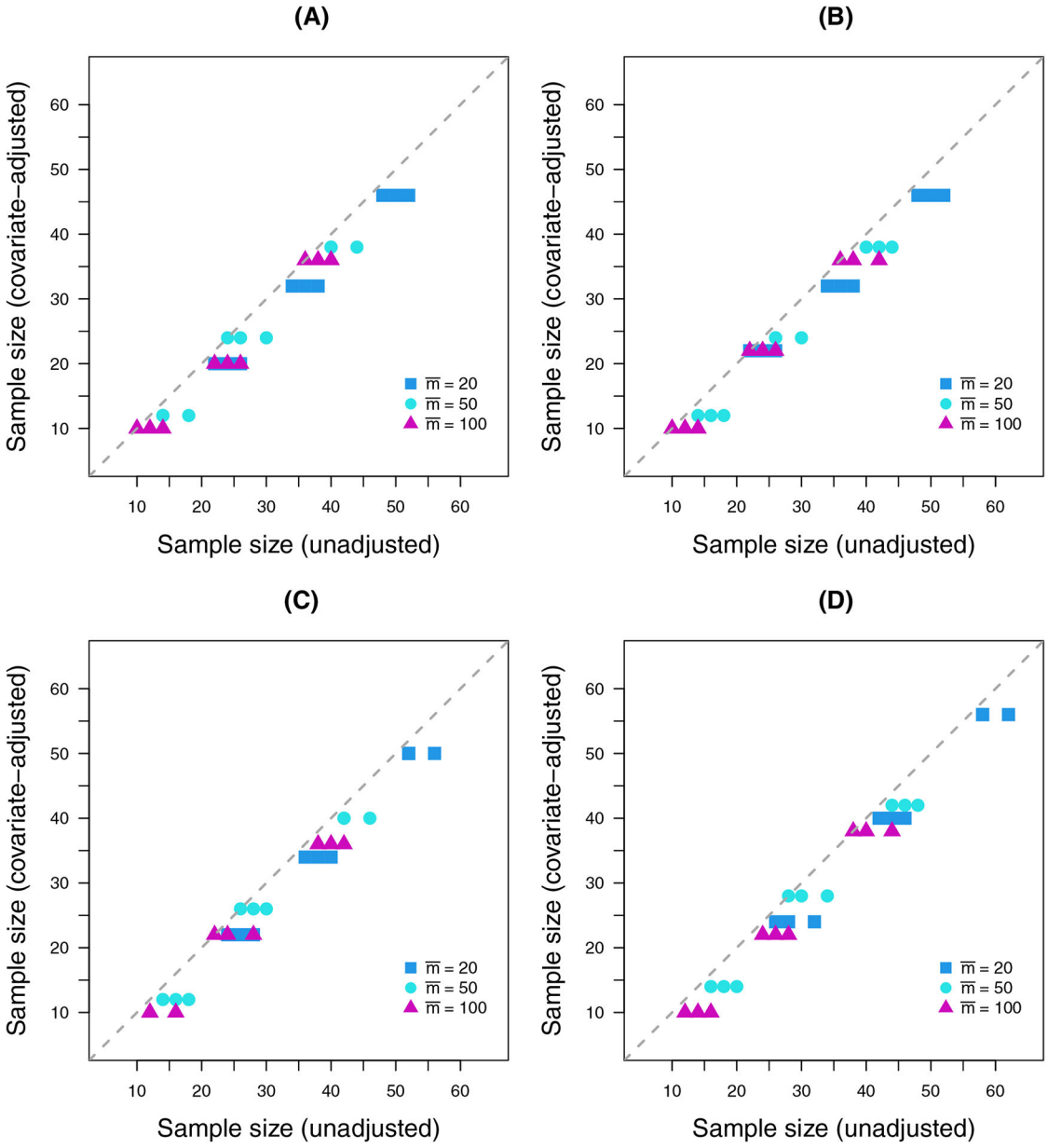

FIGURE 5.

Estimated sample size (number of clusters) for testing the average treatment effect based on the HTE model vs that based on the unadjusted linear mixed model when the outcomes are simulated with a continuous individual-level effect modifier. The effect size of the average treatment effect is induced from the HTE model and obtained as 0.325. The dashed lines indicate equal sample sizes. Details on the design parameters are provided in Web Tables 7 and 10. (A) CV = 0; (B) CV = 0.3; (C) CV = 0.6; (D) CV = 0.9

TABLE 1

Estimated required number of clusters (n) for detecting the HTE with respect to a continuous potential effect modifier (age) and a binary potential effect modifier (self-rated health), as well as the covariate-adjusted average treatment effect (ATE), based on the continuous outcome, concern about failing in the STRIDE study

		HTE (age)				**HTE (self-rated health)**				**ATE**			
		CV of cluster sizes				**CV of cluster sizes**				**CV of cluster sizes**			
$\rho_{y\|x}$	ρ_x	**0**	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**
0.01	0.01	58	58	58	58	78	78	78	78	12	12	**12**	14
	0.025	52	52	**52**	52	80	80	80	80				
	0.05	52	52	52	52	80	80	**80**	80				
	0.10	52	52	52	54	82	82	82	82				
	0.20	54	54	56	56	86	86	86	86				
0.05	0.01	50	50	50	50	50	50	50	50	76	76	76	76
	0.025	50	50	50	50	78	78	78	78				
	0.05	50	50	50	50	78	78	78	78				
	0.10	52	52	52	52	82	82	82	82				
	0.20	58	58	58	58	90	90	90	90				

Note: The nominal type I error rate is 5% and the nominal power is 80%. Bold values indicate estimation based on the specified design assumptions, and the CV of cluster sizes, covariate-ICC, and conditional outcome-ICC are varied as a sensitivity analysis for sample size.

TABLE 2

Estimated required number of clusters (n) based on the quadratic HTE model with a continuous potential effect modifier (age) and the continuous outcome, concern about failing in the STRIDE study

		HTE for Age^2 = 0.0				**HTE for Age^2 = 0.1**				**HTE for Age^2 = 0.2**				**HTE for Age^2 = 0.3**			
		CV of cluster sizes				**CV of cluster sizes**				**CV of cluster sizes**				**CV of cluster sizes**			
$\rho_{y\|x}$	ρ_x	**0**	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**	**0**	**0.25**	**0.5**	**0.75**
0.01	0.01	76	76	76	76	40	40	40	40	18	18	18	18	10	10	10	10
	0.025	76	76	**76**	76	40	40	**40**	40	18	18	**18**	18	10	10	**10**	10
	0.05	78	78	78	78	40	40	40	40	18	18	18	18	10	10	10	10
	0.10	78	78	78	80	40	40	40	40	18	18	18	18	10	10	10	10
	0.20	82	82	82	84	42	42	42	42	18	18	18	18	10	10	10	10
0.05	0.01	74	74	74	74	38	38	38	38	18	18	18	18	10	10	10	10
	0.025	74	74	74	74	38	38	38	38	18	18	18	18	10	10	10	10
	0.05	76	76	76	76	38	38	38	38	18	18	18	18	10	10	10	10
	0.10	78	80	80	80	40	40	40	40	18	18	18	18	10	10	10	10
	0.20	86	86	86	86	42	42	42	42	18	18	18	18	10	10	10	10

Note: The nominal type I error rate is 5% and the nominal power is 80%. Bold values indicate estimation based on the specified design assumptions, and the CV of cluster sizes, covariate-ICC, and conditional outcome-ICC are varied as a sensitivity analysis for sample size.