Open Access Medical Statistics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

Comparison of empirical study power in sample size calculation approaches for cluster randomized trials with varying cluster sizes – a continuous outcome endpoint

This article was published in the following Dove Press journal:
Open Access Medical Statistics
25 February 2016
Number of times this article has been viewed

Mavuto Mukaka[1,2]
Lawrence H Moulton[1]

[1]Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA; [2]Clinical Trials Support Group, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand

Background: Cluster randomized trials (CRTs) are a popular trial design. In most CRTs, researchers assume equal cluster sizes when calculating sample sizes. When clusters vary, assuming equal sized clusters may result in low study power. There are two common approaches to sample size calculations for varying cluster sizes. One approach uses a harmonic mean ($\bar{m}_H$) of cluster sizes, while the other incorporates the squared coefficient of variation (cv^2) of cluster sizes. We performed simulations to compare empirical power between the two methods as well as the arithmetic mean method for a continuous endpoint.

Study design: We considered cluster sizes that follow uniform distributions and performed 20,000 simulations under each scenario. Endpoints were analyzed using: 1) an individual-level linear regression model with Gaussian random intercepts for clusters; 2) an individual-level *t*-statistic with cluster-robust standard errors; 3) a generalized estimating equations (GEE) model with exchangeable correlation structure; and 4) a GEE model with independent correlation structure and robust standard errors.

Results: When the Gaussian random effects or the GEE model with exchangeable correlation structure was considered, the $\bar{m}_H$ method had 80% power. The cv^2 method had power of 85%–88%. However, when the data were analyzed using a *t*-statistic or the GEE model with independent correlation structure, the power of cv^2 method was 80%. The $\bar{m}_H$ method produced power of 71%–76%.

Conclusion: The performance of the sample size methods depends on the data analysis approaches. The degree of disparity in power depends also on the intracluster correlation coefficient. These findings emphasize the maxim that researchers should consider methods of analysis when designing CRTs to allow for appropriate sample size calculations.

Keywords: cluster randomized trial, varying cluster sizes, empirical power, harmonic mean, coefficient of variation, continuous endpoint

Introduction

Cluster randomized trial (CRT) designs are commonly used to evaluate the impact of health interventions between two or more treatments.[1,2] In CRTs, groups of individuals are randomized to different treatments. The groups that are typically randomized are households, health centers, villages, and intensive care units to mention a few. The broad reasons for choosing a CRT design include the practical challenges of conducting the intervention at the individual level and the need to obtain cluster level information on the effect of an intervention.[3]

Correspondence: Mavuto Mukaka
Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, 60th Anniversary Chalermprakiat Building, 3rd Floor, 420/6 Ratchawithi Rd, Ratchathewi District, Bangkok, 10400, Thailand
Tel +66 2 99 640 3239
Email mmukaka@gmail.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/OAMS.S96508

Researchers often assume an equal number of subjects within each cluster when calculating sample sizes for CRTs.[4–7] In such cases, the required sample size for an individually randomized study design is simply multiplied by an inflation factor popularly known as design effect (DEff) to account for clustering in the sample size for a CRT.[4] The elements of the inflation factor are the intracluster correlation coefficient (ICC), ρ, and the cluster size, m, DEff = $(1+ (m-1)\rho)$. The main advantage of this assumption is that it simplifies the calculations. In some cases, there are only slight differences in cluster sizes and the assumption of equal sized clusters may not be an issue. An arithmetic mean, $\bar{m}$, of the cluster size is commonly used in place of the cluster size m in such scenarios. The DEff then becomes: $(1+(\bar{m}-1)\rho)$, where $\bar{m}$ is the arithmetic mean of the cluster size. The sample size for assessing a difference in means between two treatment/intervention groups becomes:

$$c = \frac{(Z_{\alpha/2} + Z_{\beta})^2 [2\ (\sigma_w^2 + \sigma_b^2)\ (1+\{\bar{m}-1\}\rho)]}{\bar{m}(\mu_0 - \mu_1)^2} \qquad (1)$$

where c is the total number of clusters, $Z_{\alpha/2}$ and Z_{β} are the standard normal values corresponding to the upper tail probabilities of $\alpha/2$ and β, respectively; α is the two-sided significance level, and $1-\beta$ is the study power, with β the probability of making type II error; μ_0 and μ_1 are the means in the control and intervention arms respectively; σ_w and σ_b are within cluster and between cluster standard deviations of the outcome, respectively.

However, it is very common to work with unequal cluster sizes in practice, such that the cluster sizes may vary considerably. In situations where all members of a cluster are studied, cluster sizes will more likely vary.[1] A good example is randomizing health centers, where the interest is in the patients who are on antiretroviral therapy. Clearly, the numbers of antiretroviral therapy patients would vary from one health center to another. When cluster sizes vary, sample size calculations that assume equal cluster size and those that utilize the arithmetic mean cluster size may not yield sizes that have enough power to detect a desired effect.[1,2,6–8] In general, this simplicity in sample size calculations is done at the expense of reduced study power when clusters vary.

Furthermore, a method of analysis of the endpoint may potentially have an impact on the power. For example, the data analyzed at individual level may not necessarily have the same power as cluster level summary analysis. Similarly, population-averaged estimates may not necessarily yield the same power as subject-specific estimates. In general, less attention is paid to the implication of the method of analysis of the primary endpoint on power and inference. In a 2004 review of CRTs, Varnell et al[9] found that about 20.3% of the reviewed articles reported inappropriate analyses according to the study designs. A review of methods of analyses are present in a 2004 review by Murray et al.[5] In general, researchers tend to focus only on cluster sizes, ICC, and the nature of outcome whether continuous or binary when making sample size calculations for CRTs. The statistical analysis section in study protocols often tends to state the methods of analysis that will be used based on what other publications have routinely outlined without further reflection on the study power. Ignoring the method of analysis at the design stage may result in study power implications during analysis.

For varying cluster sizes, there are two commonly used sample size calculation methods. One approach uses a harmonic mean of cluster sizes in the DEff instead of the arithmetic mean to calculate sample sizes.[1] Let m_i be the number of individuals in the ith cluster for i=1, 2, …, c, where c is the total number of clusters available for randomization. The harmonic mean of cluster sizes is:

$$\bar{m}_{\mathrm{H}} = \frac{1}{\sum_{i=1}^{c} (1/m_i)/c}. \qquad (2)$$

That is, to calculate a harmonic mean, one first obtains the arithmetic mean of the reciprocal of each of the cluster sizes. Then, one takes the reciprocal of the resulting arithmetic mean.

Alternative studies of sample size calculations for CRTs with varying cluster sizes suggest using a modified DEff that includes the squared coefficient of variation of cluster sizes.[6–8] The following DEff is used:

$$1+\{(cv^2+1)\bar{m}-1\}\rho \qquad (3)$$

where cv is the coefficient of variation of cluster sizes and $\bar{m}$ is the arithmetic mean of the cluster sizes. The cv of cluster sizes is the ratio of the standard deviation of the cluster sizes to the arithmetic mean ($\bar{m}$) of the cluster sizes.[6,7] Eldridge et al[8] suggest ways of estimating the cv for this approach depending on the distribution of the cluster sizes.

We note that researchers have the liberty to choose one of the two methods of sample size calculations as long as they cite the appropriate reference. We also note that depending on the distribution of cluster sizes, these two sample size calculation methods may lead to different sample size estimates for the same scientific question and parameters. This

suggests that one of the two methods may be underpowered or may provide a very conservative high power. In addition, the method of analysis is rarely considered when making sample size calculations. The rationale for this study was, therefore, to compare empirical power from these two sample size calculation approaches as well as the arithmetic mean method taking the methods of analysis into account, in order to provide informed guidance on their use in practice.

Methods

Simulations

A simulation study was performed in Stata 13 (StataCorp LP, College Station, TX, USA) to compare the empirical study power from the two sample size calculation approaches for CRTs with varying cluster sizes as well as the standard formula which factors the arithmetic mean into the DEff. In one set of simulations, we considered cluster sizes that follow a uniform distribution, U[10,100], giving a mean cluster size of 55, harmonic mean 38.3, variance of 675, and hence *cv* of cluster sizes is 0.47 (or cv^2=0.22). In another set of simulations, we examined cluster sizes that follow a uniform distribution, U[5,100], resulting in an arithmetic mean cluster size of 52, harmonic mean of 31, variance of 752, which gives a *cv* of cluster sizes of 0.53 (or cv^2=0.28). The uniform distribution was chosen so the results would be compared with previous researchers who had used the uniform distribution.[6,7] The endpoint of interest was set to be a continuous outcome. In all the scenarios considered, the sample sizes were calculated to detect a change in mean of 15 units between the control and the intervention groups. The within cluster variance of the outcome was set at 2,000. The between cluster variances were varied to achieve the different ICC (ρ) levels ranging from 0.1 to 0.7. The wide range of ICC (ρ) used in these simulations are consistent with the literature.[6,7] Nominal power was set at 80% and 5% type I error rate was allowed in sample size calculations. The following sample size formulas were used to calculate the number of clusters, *c*, per arm for the *cv* and harmonic mean methods, respectively:

$$c = \frac{(Z_{\alpha/2} + Z_{\beta})^2[2(\sigma_w^2 + \sigma_b^2)(1 + \{(1 + cv^2)\bar{m} - 1\}\rho)]}{\bar{m}(\mu_0 - \mu_1)^2} \quad (4)$$

and

$$c = \frac{(Z_{\alpha/2} + Z_{\beta})^2[2(\sigma_w^2 + \sigma_b^2)(1 + (\bar{m}_H - 1)\rho)]}{\bar{m}_H(\mu_0 - \mu_1)^2} \quad (5)$$

In these simulations, clusters were randomized either to the treatment or control group. For comparison, we also calculated the sample size corresponding to cv^2=0, that is, just using the arithmetic mean cluster size. We performed 20,000 simulations under each scenario being investigated.

Three commonly used methods of analysis in CRTs were used for outcome data analysis in these simulations. The three methods that were used to analyze the simulated data sets were: 1) an individual-level linear regression model with Gaussian random intercepts for clusters (estimated via maximum likelihood); 2) an individual-level *t*-statistic with cluster-robust standard errors to account for clustering, and 3) the generalized estimating equations (GEE) with exchangeable correlation structure. We also considered the GEE with independent correlation structure and robust standard errors. This GEE model specification is close to that producing the *t*-statistic with robust standard errors. The cluster-robust standard errors for the individual-level *t*-statistic method were estimated by decomposing the Huber–White matrix at cluster level (ie, using cluster level components) rather than using individual-level components.

Let y_{ij} denote the outcome of individual i in cluster j, i=1, 2, 3, …, n; j=1, 2, 3, …, k. The Gaussian random intercepts model is given as:

$$y_{ij} = \beta_0 + \beta_1 treat_j + e_{ij} \quad (6)$$

$$\beta_0 = u + \alpha_j \quad (7)$$

where u is the overall mean (a constant), α_j is the cluster j effect, random, ~N(0, τ^2), β_1 is the treatment effect, e_{ij}~N(0, σ^2), α_j and e_{ij} are independent.

In this model, the standard error for the estimation of β is estimated via maximum likelihood estimation.

On the other hand, the model for the Student's *t*-test with robust standard errors is specified as a linear regression model as:

$$y_{ij} = \beta_0 + \beta_1 treat_j + e_{ij} \quad (8)$$

where β_0 is the overall mean (a constant), β_1 is the treatment effect, e_{ij}~N(0, σ^2).

In this model, the standard errors for the estimation of β are estimated using the Huber–White–Royall expression. This robust variance estimator in a cluster design with K clusters C_1, C_2, C_3, ….. C_K is given as:

$$\hat{\Omega} = \hat{v}\left(\sum_{j=1}^{K} h_j^{(C)'} h_j^{(C)}\right)\hat{v} \quad (9)$$

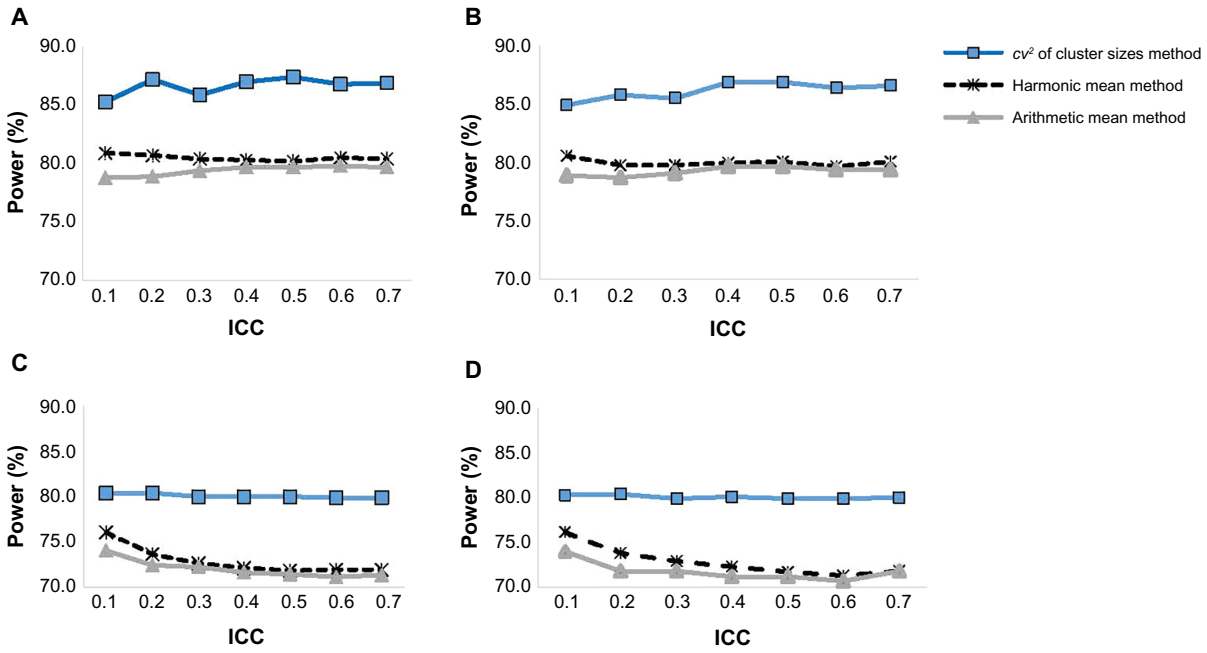

Figure 1 Empirical power for the three sample size calculation methods and four different data analysis approaches over a range of ICCs, cluster sizes ~U[10,100].
Notes: (**A**) Gaussian random effects maximum likelihood linear regression model was used to analyze data. (**B**) GEE with exchangeable correlation structure was used to analyze data. (**C**) An individual-level *t*-statistic with robust standard errors to adjust for clustering was used. (**D**) GEE with independent correlation structure and robust standard errors was used.
Abbreviations: GEE, generalized estimating equations; ICC, intracluster correlation coefficient; cv^2, the squared coefficient of variation of cluster sizes.

where $\hat{v}$ is the conventional estimate of the variance and $h_j^{(c)}$ is the contribution of the *j*th cluster to variance estimation.

The *t*-statistic was obtained as the ratio of the estimate of treatment effect to the clustered-robust standard errors. It should be noted that the *t*-test may also be obtained via cluster level summaries, however, such an approach discourages inclusion of covariates in the analysis.

Results

For the cluster sizes that follow a U[10,100], at all levels of the ICCs, the *cv* method resulted in the highest required sample sizes followed by the harmonic mean method and then finally the arithmetic mean method. When data were analyzed using the individual-level linear regression model with Gaussian random intercepts for clusters (estimated via maximum likelihood), the study power consistently remained around 80% for all levels of ICC (Figure 1). The same pattern was observed for those analyses performed using the GEE with exchangeable correlation structure (Figure 1). This was in agreement with the power input parameter in the sample size calculations, which was set at 80%. The arithmetic mean resulted in slight underpower yielding power of about 78%–79%, while the *cv* method had a conservatively high power of 85%–87%, substantially higher than what was optimally set in sample size calculations.

On the other hand, when data were analyzed using the individual-level *t*-statistic or the GEE with independent correlation structure and with cluster-robust standard errors to account for clustering in either case, the study power from the *cv* method consistently remained around 80% for all levels of ICC (Figure 1). This was in agreement with the power input parameter in the sample size calculations, which was set at 80%. Both the harmonic mean and arithmetic mean sample size approaches resulted in substantially lower than the optimal 80% yielding power of about 70%–76% (Figure 1).

For the cluster sizes that follow a U[5,100], which gives higher *cv*, lower harmonic and arithmetic means, respectively than those of the U[10,100] distribution, we observed the similar patterns of power that depended on the method of analysis (Figure 2). Figure 2 presents the power trends for the three methods of analysis over a range of ICC for the cluster sizes that follow a U[5,100].

The detailed summary of sample sizes and power for cluster sizes that follow a U[10,100] is presented in Table 1, according to the method of data analysis. At each ICC level, the cv^2 method resulted in the highest required sample sizes, while the arithmetic mean method had the least. The harmonic mean method gives slightly higher sample sizes than the arithmetic mean method (Table 1).

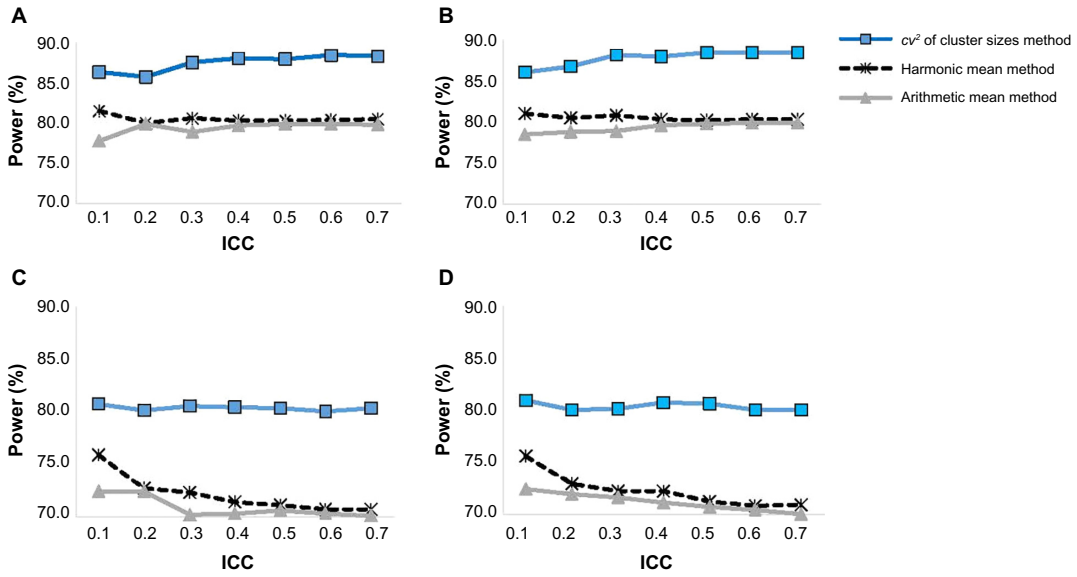

Figure 2 Empirical power for the three sample size calculation methods and four different data analysis approaches over a range of ICCs, cluster sizes ~U[5,100].
Notes: (**A**) Gaussian random effects maximum likelihood linear regression model was used to analyze data. (**B**) GEE with exchangeable correlation structure was used to analyze data. (**C**) An individual-level *t*-statistic with robust standard errors to adjust for clustering was used. (**D**) GEE with independent correlation structure and robust standard errors was used.
Abbreviations: GEE, generalized estimating equations; ICC, intracluster correlation coefficient; cv^2, the squared coefficient of variation of cluster sizes.

Table 1 Empirical power of sample size approaches according to method of data analysis – cluster sizes ~U[10,100] based on 20,000 simulations

Method of data analysis	ICC	cv^2 of cluster sizes method		Harmonic mean method		Arithmetic mean method	
		C*	Power (%)	C*	Power (%)	C*	Power (%)
Gaussian random intercepts model	0.1	44	85.3	38	80.9	36	78.8
	0.2	90	87.2	76	80.7	74	78.9
	0.3	158	85.9	132	80.4	130	79.4
	0.4	236	87.0	196	80.3	194	79.7
	0.5	346	87.4	286	80.2	284	79.7
	0.6	516	86.8	426	80.5	424	79.8
	0.7	840	86.9	690	80.4	688	79.7
GEE model with	0.1	44	85.3	38	80.8	36	79.1
(exchangeable correlation structure)	0.2	90	86.2	76	80.0	74	78.9
	0.3	158	85.9	132	80.0	130	79.3
	0.4	236	87.3	196	80.2	194	79.9
	0.5	346	87.3	286	80.3	284	79.9
	0.6	516	86.8	426	79.9	424	79.6
	0.7	840	87.0	690	79.6	688	79.6
GEE model with	0.1	44	80.3	38	76.2	36	74.0
(independent correlation)	0.2	90	80.4	76	73.8	74	71.8
and robust standard errors	0.3	158	79.9	132	72.9	130	71.8
	0.4	236	80.1	196	72.3	194	72.1
	0.5	346	79.9	286	71.7	284	71.2
	0.6	516	79.9	426	71.3	424	70.7
	0.7	840	80.0	690	71.8	688	71.8
t-Statistic with robust standard errors to	0.1	44	80.4	38	76.1	36	74.0
adjusted for clustering	0.2	90	80.4	76	73.6	74	72.4
	0.3	158	80.0	132	72.6	130	72.2
	0.4	236	80.0	196	72.1	194	71.6
	0.5	346	80.0	286	71.8	284	71.4
	0.6	516	79.9	426	71.9	424	71.1
	0.7	840	80.0	690	71.9	688	71.3

Note: C* is the total number of clusters in both arms.
Abbreviations: GEE, generalized estimating equations; ICC, intracluster correlation coefficient; cv^2, the squared coefficient of variation of cluster sizes.

Table 2 Empirical power of sample size approaches according to method of data analysis – cluster sizes ~U[5,100] based on 20,000 simulations

Method of analysis	ICC	cv^2 of cluster sizes method		Harmonic mean method		Arithmetic mean method	
		C*	Power (%)	C*	Power (%)	C*	Power (%)
Gaussian random intercepts	0.1	46	86.3	40	81.5	38	77.8
	0.2	94	85.7	78	80.0	76	79.9
	0.3	166	87.5	134	80.6	130	78.9
	0.4	246	88.0	198	80.3	194	79.7
	0.5	362	87.9	288	80.3	284	79.9
	0.6	540	88.4	428	80.4	424	79.9
	0.7	878	88.3	692	80.5	688	79.8
GEE model with	0.1	46	86.0	40	81.5	38	78.5
(exchangeable correlation structure)	0.2	94	86.7	78	80.5	76	78.8
	0.3	166	88.1	134	80.8	130	78.9
	0.4	246	87.9	198	80.3	194	79.6
	0.5	362	88.4	288	80.2	284	79.8
	0.6	540	88.4	428	80.3	424	79.9
	0.7	878	88.4	692	80.3	688	79.9
GEE model with	0.1	46	80.9	40	75.6	38	72.4
(independent correlation)	0.2	94	80.0	78	72.9	76	71.9
and robust standard errors	0.3	166	80.1	134	72.2	130	71.6
	0.4	246	80.7	198	72.2	194	71.1
	0.5	362	80.6	288	71.2	284	70.7
	0.6	540	80.0	428	70.8	424	70.4
	0.7	878	80.3	692	70.9	688	70.0
t-Statistic with robust standard errors to	0.1	46	80.7	40	75.9	38	72.4
adjusted for clustering	0.2	94	80.1	78	72.7	76	72.4
	0.3	166	80.5	134	72.3	130	70.2
	0.4	246	80.4	198	71.4	194	70.3
	0.5	362	80.3	288	71.1	284	70.6
	0.6	540	80.0	428	70.7	424	70.3
	0.7	878	80.3	692	70.7	688	70.1

Note: C* is the total number of clusters in both arms.
Abbreviations: GEE, generalized estimating equations; ICC, intracluster correlation coefficient; cv^2, the squared coefficient of variation of cluster sizes.

Table 2 provides a comprehensive summary of the power findings for cluster sizes that follow a U[5,100], according to the method of data analysis and method of sample size calculation. The increased cv^2 in the U[5,100] cluster size distribution compared with a U[10,100] resulted in a huge increase in the resulting sample sizes by this method than the respective increases for the harmonic and the arithmetic mean sample size calculation methods (Table 2).

Discussion

In this simulation work, it has been observed that the performance of the sample size calculation methods for CRTs with varying cluster sizes depends on the method of analysis. This is consistent with a recent discussion by Rutterford et al.[10] When the random effects or the GEE model with exchangeable correlation structure was considered, sample sizes of $\bar{m}_H$ method yielded desired power to detect the difference in means between two groups. The cv^2 method had very high power. This is consistent with the observation of Eldridge et al.[8]

However, when the data were analyzed using an individual-level *t*-statistic or the GEE model with independent correlation structure with clustered-robust standard errors in both cases, the empirical power of cv^2 method was 80% as expected. The $\bar{m}_H$ method produced empirical power, which is lower than expected under this setting. The power findings were of similar patterns for the two *cv*s that were considered. As expected, the sample size substantially increased with the increased *cv*.

The literature review for the sample size approaches considered in this study reveals that the sample size calculation methods for CRTs with varying cluster sizes work best depending on the data analysis method that is employed. The degree of disparity in power between the two sample size methods seems to depend not only on the method of analysis, but also on ICC (ρ). Sample sizes that use the arithmetic mean method were underpowered for the *t*-test,

GEE, and random intercepts models. The arithmetic mean method is highly underpowered when a *t*-test or the GEE with independent correlation structure is used. The unacceptably observed low power of the arithmetic mean method in case of a *t*-test agrees with the observation of Manatunga et al.[6] In this case, the harmonic mean method makes an important correction that retains the study power at 80%. Rutterford et al[10] have recently discussed sample sizes for a wide range of CRT designs and have summarized the methods of data analysis under which the different sample size methods may be appropriate.

In summary, the $\bar{m}_H$ method is ideal for the random intercepts and the GEE models with exchangeable correlation structure. When the individual-level *t*-test or the GEE model with independent correlation structure with robust standard errors is the method of analysis of choice for a CRT, the use of the cv^2 method should be encouraged. Despite its simplicity, the use of the arithmetic mean cluster size should be discouraged when cluster sizes are expected to vary. Moreover, a harmonic mean of cluster sizes can easily be estimated from the available clusters; and Eldridge et al[8] provide suggestions for estimating the *cv* of cluster sizes.

Conclusion

The performance of the sample size methods depends on the method of data analysis. The degree of disparity in power depends also on the ICC. These simulation findings emphasize the fundamental principle that researchers should consider methods of analysis when designing CRTs to allow for appropriate sample size calculations. Moreover, it can be important to account for variability of cluster size. Alternatively, one can obtain a conservative estimate by employing a minimum cluster size in standard calculations.

Acknowledgment

This work was supported by the Johns Hopkins University Center for AIDS Research (Grant Number 1P30AI094189) from the National Institute of Allergy And Infectious Diseases.

Author contributions

MM: Substantial contributions to conception and design, analysis, and interpretation of data, drafting the article, and the final approval of the version to be published. LHM: Substantial contributions to conception and design, interpretation of data, drafting the article, revising it critically for important intellectual content, and the final approval of the version to be published.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. Boca Raton: Chapman and Hall/CRC Press; 2009.
2. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Chichester, UK: John Wiley and Sons, Ltd; 2012.
3. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials*. 2004;1(3):297–305.
4. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol*. 1981;114(6):906–914.
5. 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.
6. Manatunga A, Hudgens M, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biom J*. 2001;43:75–86.
7. Kang S, Ahn C, Jung S. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. *Drug Inf J*. 2003;37:109–114.
8. 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.
9. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health*. 2004;94(3):393–399.
10. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol*. 2015;44(3):1051–1067.