Published in final edited form as:
Stat Med. 2009 March 15; 28(6): 1017–1027. doi:10.1002/sim.3527.

Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials

Moonseong Heo[1] and **Andrew C. Leon**[2,3]

[1]Division of Biostatistics Department of Epidemiology and Population Health Albert Einstein College of Medicine Bronx, NY, USA

[2]Department of Psychiatry Weill Medical College of Cornell University New York, NY, USA

[3]Department of Public Health Weill Medical College of Cornell University New York, NY, USA

Abstract

In designing a longitudinal cluster randomized clinical trial (cluster-RCT), the interventions are randomly assigned to clusters such as clinics. Subjects within the same clinic will receive the identical intervention. Each will be assessed repeatedly over the course of the study. A mixed-effects linear regression model can be applied in a cluster-RCT with three level data to test the hypothesis that the intervention groups differ in the course of outcome over time. Using a test statistic based on maximum likelihood estimates, we derived closed form formulae for statistical power to detect the intervention by time interaction and the sample size requirements for each level. Importantly, the sample size does not depend on correlations among second level data units and the statistical power function depends on the number of second and third level data units through their product. A simulation study confirmed that theoretical power estimates based on the derived formulae are nearly identical to empirical estimates.

Keywords

longitudinal cluster RCT; *three level data*; *power*; *sample size*; *intervention by time interaction*; *effect size*

1. Introduction

A longitudinal cluster randomized trial (cluster-RCT) assumes a three level data structure in that the time-specific outcome assessments are nested within subjects who in turn, are nested within the randomized clusters. For instance, consider a study designed to test the effect of an experimental intervention of physician training on the reduction of severity of patients' symptoms of depression over time. In this design, primary care clinics are randomly assigned to either experimental or control intervention and each physician within an experimental clinic is trained to detect and treat depression. Each physician will treat multiple subjects, who, in turn, repeatedly measured on severity of depression symptoms over time.

The primary hypothesis in such a study would focus on the difference in declines of symptom severity over time between subjects who were treated by physicians with and

Address correspondence to: Moonseong Heo, Ph.D. Division of Biostatistics Department of Epidemiology and Population Health Albert Einstein College of Medicine 1300 Morris Park Avenue, Belfer 1303E Bronx, NY 10461 Phone (718) 430 8838 Fax (718) 430 8780 e-mail: mheo@aecom.yu.edu.

without the experimental intervention. The three level data in a longitudinal cluster-RCT could test the significance of the intervention by time interaction using a mixed-effects linear regression model [1-3].

Sample size determination and power calculations are essential in designing a cluster-RCT. The number of clusters that is required for a target statistical power must be estimated at the experimental design stage. To this end, we build on sample size formulae for two level data structures [4-6] to derive explicitly closed form power function and sample size formulae for detecting a hypothesized interaction effect. The derivations are based on a distribution of a test statistic that used the maximum likelihood estimate of the interaction effect. A simulation study followed to verify the statistical power achieved with the estimated sample sizes.

2. Statistical Model

A three level mixed-effects linear model for outcome Y can be written as follows:

$$Y_{ijk}=\beta_0+\xi X_{ijk}+\tau T_{ijk}+\delta X_{ijk}T_{ijk}+u_i+u_{j(i)}+e_{ijk}, \tag{1}$$

where i =1,2,…,2N_3 is the index for the level three unit (e.g., clinic); j = 1,…, N_2, is the index for the level two unit (e.g., subject) nested within each i; and k = 1, 2, …, N_1, is the index for the level one unit (e.g., repeated outcome observations) within each j. The intervention assignment indicator variable X_{ijk} = 0 if the i-th level three unit is assigned to a control intervention and X_{ijk} = 1 if assigned to an experimental intervention; therefore X_{ijk} = X_i for all j and k. Furthermore, here a balanced design is assumed in that $\Sigma_i X_i = N_3$. The time variable is denoted by T_{ijk}. In this study, it is assumed that $T_{ijk} = T_k$ for all i and j, and that the time increase from 0 (the baseline) to $T_{\text{end}} = N_1$ - 1 (the last time point) by 1 with equal time intervals. Therefore, the parameter ξ represent the intervention effect at the baseline, and the parameter τ represents the slope of time effect, that is, decline in symptom severities over time. Finally, the intervention by time effect δ is of primary interest representing the slope difference in outcome Y between the intervention groups, or additional decline in the experimental group. The overall fixed intercept is denoted by β_0.

It is assumed that the error term e_{ijk} is normally distributed as $N\left(0,\sigma_e^2\right)$, the level two random intercept $u_{j(i)}\sim N\left(0,\sigma_2^2\right)$ and the level three random intercept $u_i\sim N\left(0,\sigma_3^2\right)$. Among those random components, it is further assumed that $u_i \perp u_{j(i)} \perp e_{ijk}$, i.e., these three random components are mutually independent. In addition, *conditional independence* is assumed for all $u_{j(i)}$ and for all e_{ijk}, whereas as u_i are *unconditionally* independent. That is, $u_{j(i)}$ are independent conditional on u_i, and e_{ijk} are independent conditional on both u_i and $u_{j(i)}$. After all, β_0, ξ, τ and δ are fixed effect parameters and the last three terms in model (1) are random effects.

As the parameter δ is of the primary interest, the null hypothesis to be tested is:

$$\mathrm{H}_0{:}\delta=0 \tag{2}$$

Under model (1), with its accompanying assumptions such as conditional independence among random components, it can be shown that the elements of the mean vector are

$$E\left(Y_{ijk}\right)=\beta_0+\xi X_i+\tau T_k+\delta X_i T_k \tag{3}$$

and that the elements of the covariance matrix are:

$$Cov\left(V_{ijk}, Y_{i'j'k'}\right)=1\left(i=i'\,\&\,j=j'\,\&k=k'\right)\sigma_e^2+1\left(i=i'\,\&\,j=j'\right)\sigma_2^2+1\left(i=i'\right)\sigma_3^2, \tag{4}$$

where 1(.) is an indicator function. This yields in particular,

$$\sigma^2 \equiv Var\left(Y_{ijk}\right)=Cov\left(Y_{ijk}, Y_{ijk}\right)=\sigma_e^2+\sigma_2^2+\sigma_3^2.$$

Therefore, the correlation among level two data can be written for $j \neq j'$ as follows.

$$\rho_2=Corr\left(Y_{ijk}, Y_{ij'k'}\right)=\frac{\sigma_3^2}{\sigma_e^2+\sigma_2^2+\sigma_3^2}=\frac{\sigma_3^2}{\sigma^2}. \tag{5}$$

And, the correlation among level one data can be written for $k \neq k'$,

$$\rho_2=Corr\left(Y_{ijk}, Y_{ijk'}\right)=\frac{\sigma_2^2+\sigma_3^2}{\sigma_e^2+\sigma_2^2+\sigma_3^2}=\frac{\sigma_2^2+\sigma_3^2}{\sigma^2}. \tag{6}$$

It can be easily seen that $\rho_1 \geq \rho_2$ with equality when $\sigma_2^2=0$.

3. Maximum Likelihood Estimate and its Variance

The maximum likelihood estimate (MLE)$\widehat{\delta}$ of the interaction effect is indeed the slope difference between the two groups: that is,

$$\widehat{\delta}=\widehat{\eta}_1-\widehat{\eta}_0, \tag{7}$$

where $\widehat{\eta}_g\ (g=0,1)$ is the MLE of the slope for the outcome Y in the g-th group, in which $X_i = g$. Specifically, for i in the g-th group,

$$\begin{aligned}\widehat{\eta}_g \quad &= \sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1}\left(T_k-\bar{T}\right)\left(Y_{ijk}-\bar{Y}_g\right)/\sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1}\left(T_k-\bar{T}\right)^2\\ &= \sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1}\left(T_k-\bar{T}\right)\left(Y_{ijk}-\bar{Y}_g\right)/N_3N_2N_1Var_p\left(T\right),\end{aligned} \tag{8}$$

where: 1)$\bar{Y}_g\ (g=0,1)$ is the overall group mean of the outcome Y for the g-th group; 2) $\bar{T}=\Sigma_{k=1}^{N_1}T_k/N_1$ is the "mean" time point; and 3) $Var_p\left(T\right)=\Sigma_{k=1}^{N_1}\left(T_k-\bar{T}\right)^2/N_1$ is the "population variance" of the time variable T. In fact, the slope estimate (8), but not the variance of the slope estimate, is the same as that of an ordinary linear regression with $u_i = u_{j(i)} = 0$ in model (1). The reason for this, on a heuristic level, is that weights assigned to data points Y_{ijk} in estimation of the slopes are identical and the slopes do not depend on random intercepts of any data level. Indeed, the ordinary least square estimate (8) is the mle under a perfectly balanced design [2] that we are considering in this paper.

Based on equations (3) and (8), it can easily be shown that the MLE $\widehat{\delta}$ is unbiased, i.e., $E\left(\widehat{\delta}\right)=E\left(\widehat{\eta}_1-\widehat{\eta}_0\right)=(\tau+\delta)-\tau=\delta$. The variance of a slope MLE $\widehat{\eta}_\delta$ can be obtained based on equation (4) as follows (see Appendix for a proof):

$$Var\left(\widehat{\eta}_g\right)=\frac{\sigma_e^2}{N_3N_2N_1Var_p(T)}=\frac{(1-\rho_1)\,\sigma^2}{N_3N_2N_1Var_p(T)}. \tag{9}$$

Therefore, the variance of $\widehat{\delta}$ is

$$Var\left(\widehat{\delta}\right)=Var\left(\widehat{\eta}_1-\widehat{\eta}_0\right)=Var\left(\widehat{\eta}_1\right)+Var\left(\widehat{\eta}_0\right)=\frac{2(1-\rho_1)\,\sigma^2}{N_3N_2N_1Var_p(T)}. \tag{10}$$

Observe that $\widehat{\eta}_1$ and $\widehat{\eta}_0$ are independent each other. It is notable, however, that the variance of $\widehat{\delta}$ depends only on the residual variance σ_e^2, and none of σ_3^2, σ_2^2, or ρ_2. Therefore, *for a given total variance* σ^2, it decreases with decreasing σ_e^2 or increasing ρ_1, the correlation among the first level data.

4. Power and sample size

The following test statistic *D*, based on (7) and (10), can be used to test the null hypothesis (2):

$$D=\frac{\widehat{\delta}}{se\left(\widehat{\delta}\right)}=\frac{\widehat{\delta}}{\sqrt{Var\left(\widehat{\eta}_1\right)+Var\left(\widehat{\eta}_0\right)}}=\frac{\sqrt{N_3N_2N_1Var_p(T)}\left(\widehat{\eta}_1-\widehat{\eta}_0\right)}{\sigma\sqrt{2(1-\rho_1)}}. \tag{11}$$

If the three variance components—σ_2^2, σ_3^2 and σ_e^2— are known, then the test statistic *D* is normally distributed with mean$\delta/se\left(\widehat{\delta}\right)$ and variance 1. When those three variance components are unknown and replaced by their MLE's, the test statistic *D* becomes a Wald test statistic and its *asymptotic* distribution is normal based on a large sample theory [7]. Thus, under the null hypothesis (2), *D* ~ *N*(0, 1) and under an alternative hypothesis of $\delta\neq0, D{\sim}N\left(\delta/se\left(\widehat{\delta}\right),1\right)$.

The power of the test statistic *D*, denoted by φ, can therefore be written as follows:

$$\varphi=1-\beta=\Phi\left[\frac{\delta}{\sigma}\sqrt{\frac{N_3N_2N_1Var_p(T)}{2(1-\rho_1)}}-\Phi^{-1}(1-\alpha/2)\right], \tag{12}$$

where α is a two-sided significance level; β represents the probability of type II error; Φ is the cumulative distribution function (CDF) of a standard normal distribution and Φ^{-1} is its inverse. From now on, it is understood that: 1) δ = |δ| > 0; and 2) the probability below a critical value, $\Phi^{-1}(\alpha/2)$, in the other side under the alternative hypothesis is negligible and thus assumed to be 0. When the slope difference is expressed in pooled within-group standard deviation (SD) units, i.e., when expressed in terms of a standardized effect size

$$\Delta_\delta = \delta / \sigma,$$

the power function can be expressed as follows:

$$\varphi = \Phi\left[\Delta_\delta \sqrt{N_3 N_2 N_1 Var_p(T)/2(1-\rho_1)} - \Phi^{-1}(1-\alpha/2).\right] \quad (13)$$

It follows that when the hypothesis testing is based on D with a two-sided significance level of α, the third level unit sample size N_3 per group for a desired statistical power φ = 1 - β can be calculated from equation (12) as:

$$N_3 = \frac{2\left(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right)^2 (1-\rho_1)\sigma^2}{N_2 N_1 Var_p(T)\delta^2}, \quad (14)$$

or equivalently in terms of the standardized effect size Δ_δ from equation (13)

$$N_3 = \frac{2\left(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right)^2 (1-\rho_1)}{N_2 N_1 Var_p(T)\Delta_\delta^2}. \quad (15)$$

More precisely, N_3 is the smallest integer greater than the right hand side of equation (14) or (15). It can be observed that the level 3 sample size is a deceasing function of increasing ρ_1 and $Var_p(T)$ in particular. Stated differently, more follow-up with more consistent (as opposed to erratic) observations within subjects over time will increase the power (15) and at the same time will reduce sample size required of N_3 or N_2 for the same anticipated power.

The sample size N_2 has a reciprocal relationship with N_3 in a sense that the power depends through N_2N_3 because both are free each other and of the other parameters. Therefore, sample size N_2 for the level two data can immediately be determined from equation (15) as follows:

$$N_2 = \frac{2\left(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right)^2 (1-\rho_1)}{N_3 N_1 Var_p(T)\Delta_\delta^2}. \quad (16)$$

The sample size N_1 for the level one data should, however, be determined in an iterative manner because $Var_p(T)$ is a function of N_1. Specifically, an iterative solution for N_1 must satisfy the following equation:

$$N_1 = \frac{2\left(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right)^2 (1-\rho_1)}{N_3 N_2 Var_p(T)\Delta_\delta^2}. \quad (17)$$

5. Simulation study specification

We conducted simulation studies to verify the sample size N_3 (15) and the power function (13) using SAS PROC MIXED, which is suitable for fitting the three-level mixed-effects linear model (1). For a two-sided significance level $\alpha = 0.05$ and a desired power $\varphi = 0.8$, the following combinations of the simulation parameters were prespecified: $\Delta_\delta T_{\text{end}} = \Delta_\delta(N_1 - 1) = 0.3, 0.4, 0.5$; $N_2 = 5, 10, 20, 30$; $N_1 = 3, 6, 12$; $\rho_1 = 0.4, 0.5, 0.6$ while without loss of generality $\sigma = 1$, $\rho_2 = 0.05$, $\beta_0 = \xi = 0$, and $\tau = -1$ (in model (1)) remained fixed. This 3×4×3×3 factorial design scheme yielded a total of 108 combinations of those parameters. In particular, the effect size of the interaction, or the between-group slope difference Δ_δ, is specified in a way that it would yield a standardized between-group mean difference $\Delta_\delta T_{\text{end}}$ at the end of trial, i.e., when $T = T_{\text{end}} = N_1 - 1$.

To generate simulated data, we first estimated N_3 using equation (15) for a given combination (see step 2 below). Specifically, for each combination we followed the following steps for simulations:

1. Calculate the variance of time, $Var_p(T)$, for given N_1;
2. Calculate N_3 (15) with the computed $Var_p(T)$ and given α, φ, N_1, N_2, and Δ_δ;
3. Calculate variance components, σ_2^2, and σ_3^2 based on equations (5) and (6) for given ρ_1, ρ_2 and σ^2; Specifically, $\sigma_2^2 = (\rho_1 - \rho_2)\,\sigma^2$ and $\sigma_3^2 = \rho_2\sigma^2$;
4. Calculate $\sigma_e^2 = \sigma^2 - \left(\sigma_3^2 + \sigma_2^2\right)$;
5. Calculate $\delta = \sigma\Delta_\delta$ for the given σ^2 and Δ_δ;
6. Generate the random intervention assignment indicator $X_i = 0$ or 1 for each $i = 1,2,.., 2N_3$ in a balanced manner so that $\Sigma_i X_i = N_3$;
7. Generate u_i from $N\left(0, \sigma_2^2\right)$ independently for each $i = 1,2,\ldots,2N_3$ (Unconditional independence assumption);
8. For each u_i, generate $u_{j(i)}$ from $N\left(0, \sigma_2^2\right)$ independently for $j = 1,2,\ldots,N_2$ (Conditional independence assumption);
9. For each combination of u_i and $u_{j(i)}$, generate e_{ijk} from $N(0, \sigma_e{}^2)$ independently for $k = 1,2, \ldots,N_1$ (Conditional independence assumption);
10. Generate outcome data set for $Y_{ijk} = \beta_0 + \xi X_i + \tau T_k + \delta X_i T_k + u_i + u_{j(i)} + e_{ijk}$ (1);
11. Fit the data set with the three-level linear mixed-effects model (1);
12. Retain a p-value, denoted by $p_s(\delta)$ for the s-th simulated data set, obtained from testing the null hypothesis (2);
13. Repeat the steps 6-12 for 1000 times (i.e., $s = 1, 2, \ldots, 1000$) for each combination of the simulation parameters.

Let us denote the empirical power by $\tilde{\varphi}$ that is obtained from the 1000 simulations as follows:

$$\tilde{\varphi} = \sum_{s=1}^{1000} 1\left\{p_s(\delta) < \alpha\right\} / 1000. \qquad (18)$$

This empirical power is compared with the theoretical power φ that is computed based on N_3 obtained in step 2 above, but not with the prespecified power of 0.8. It should be noted that the theoretical power φ obtained in that way is never less than the prespecified power of 0.8 since N_3 is the smallest integer greater than the right hand side of equation (15).

6. Simulation study results

Table 1 summarizes the specified (N_2 and N_1) and estimated (N_3) sample sizes, the empirical power $\tilde{\varphi}$ (18) and the theoretical power φ (13) based on the estimated N_3. Although the empirical power is negligibly underestimated as reflected on the mean differences in the last row in Table 1, it is virtually identical to the theoretical power. For instance, among the 108 combinations (Table 1), the maximum absolute difference $|\varphi - \tilde{\varphi}|$ was 0.027, which is tolerable given that the width of the 95% confidence interval for simulation estimates is $\pm 1.96\sqrt{0.8 \times 0.2/1000} = \pm 0.025$. Thus, the derived formulae for sample size and the power are very accurate under the conditions that were examined. In each case, the theoretical power is no less than 0.8, since the power calculations were based on "integer" values of N_3.

As expected, the sample size N_3 for the identical power decreases with increasing correlation ρ_1 when the other design parameters are held the same. For example, when N_2 = 5, N_1 = 6, and $\Delta_\delta T_{end}$ = 0.3, (or Δ_δ = 0.3/5 = 0.06) the respective sample sizes requirements for 80% power, for the level three data (N_3), were 30, 25, and 20 for ρ_1 = 0.4, 0.5, and 0.6. Furthermore, the theoretical power is identical for various combinations of N_2 and N_3 that yield an equivalent product, assuming other design parameters are held constant. For instance, as shown in Table 1, each the following pairs of N_2 and N_3 with a product of 210 yielded identical power of 0.801 when N_1 = 3, ρ_1 = 0.4, $\Delta_\delta T_{end}$ = 0.3 (or Δ_δ= 0.3/2 = 0.15): N_2 = 5 and N_3 = 42; N_2 = 10 and N_3 = 21; N_2 = 30 and N_3 = 7.

7. Application

The results in Table 1 can be applied to designing a longitudinal cluster-RCT. Consider, for instance, a longitudinal cluster-RCT that compares an innovative primary care level intervention with a usual primary care practice on depression outcome of subjects as conducted in the PROSPECT [8,9] and the RESPECT [10] trials. To test whether the course of depressive symptoms over time depends on the care that the subjects receive, it is anticipated that primary clinics can accommodate 20 subjects (N_2) for the research purpose and each patient would be followed up for 6 times (N_1) for assessments. The results presented in Table 1 can be applied to estimating number of primary clinics, i.e., level 3 units (N_3), for 80% power. If ρ_1 = 0.5, then four clinics (N_3) for each of the two intervention groups, or a total of 160 subjects, would be needed to detect an effect size $\Delta_\delta T_{end}$ = 5Δ_δ = 0.4 (or Δ_δ = 0.4/5 = 0.08) with at least 80% statistical power (Table 1). Sample size requirements for other design parameters can be obtained from Table 1. For other combinations of design specification that were not presented in Table 1, the sample size formula (18) can be applied.

8. Discussion

The derived power function (13) and level 3 unit sample size formula (15) requirements to detect an intervention by time interaction are shown to be accurate compared to empirical estimates based on a simulation study. Therefore, sample size formulae (16, 27) for number of level 2 and level 1 data units are also accurate because they are different expressions of equation (15). Importantly, the sample size did not depend on correlations among second level data units and the statistical power function depends on the number of second and third level data units through their product. Furthermore, when either N_3 or N_2 is equal to one, it reduces the level 3 data structure to that of level 2 data with the number of second level data as N_2 or N_3 correspondingly. In either case, the variance σ_3^2 of the level three random intercept can be considered to be 0 and thus ρ_2 can be assumed to be 0. This reduces the sample size formula (14) to equation (2.4.1) in Diggle et al [6] on its page 29, as it should. In Diggle et al's formula too, it can be found that the power function is increasing in ρ_1.

Collectively, therefore, as far as testing the intervention by time interaction is concerned, the design can be very flexible for the same statistical power depending on feasibility. For example, when N_3N_2 = 200 subjects per group is needed for 80% power, then sample sizes for N_3 and N_2 can be determined depending on availability of recruitment of level two and level three units regardless of an anticipated ρ_2. To this end, if recruitment of 10 subjects (N_2) per clinic was feasible, then the investigators could try to enlist 20 clinics (N_3) per intervention group. On the other hand, if only 5 clinics (N_3) were available per intervention group, then recruitment of 40 subjects (N_2) per clinic would be required. In an extreme case where only one clinic (N_3=1) is available, one could recruit 200 subjects (N_2) from the single clinic.

Although the empirical power was based on unknown variance components of random effects, it was virtually identical to the theoretical power derived with known variance components in the test statistic D (11). Therefore, derivation of power function with unknown variances may not be necessary even for small N_3, although it might be possible through application of CDFs of central and non-central t distributions [11] replacing the standard normal CDF Φ and its inverse Φ^{-1} in equation (14) or (15).

It should be noted that the sample size formula is to detect a slope difference *per se* but not an expected between-group difference at T_{end}, the end of a study. In other words, the sample formula (15) derived herein is not appropriate to detect an intervention effect at a prespecified time point such as the end of a trial. It is because the variance of this effect is not equal to $T_{end}^2 Var(\widehat{\eta}_1 - \widehat{\eta}_0)$, even if the estimated quantities are the same. Thus, this intervention effect, $\Delta_\delta T_{\text{end}}$, served as the basis for estimating a hypothesized slope difference Δ_δ.

Other sample size formulae are available. For instance, Liu et al [12] derived sample size formulas for the slope difference using generalized estimating equations. Murray et al [13] presented detectable effect sizes based on expected mean square errors using random coefficients analysis for the nested cohort design. Roy et al [14] derived general-form sample size determinations using a mixed-effects linear model, taking into account for potential attrition rates and more general correlation structures. Heo and Leon [15] derived an algorithm for sample size requirements to detect a main effect of group using a linear mixed effects model for three level data. Although comparisons of sample sizes assuming different modeling approaches would provide better insight in designing a cluster-RCT, the sample size equations presented above (15,16,17) are more readily implemented.

The sample size determinations derived here have limitations. First, the formulae were derived assuming fixed numbers of units for all levels although number of subjects per clinic will likely vary, i.e., j = 1, 2, …, n_i, depending the i-th clinic. Furthermore, the number of assessments per subjects will also vary (i.e., k = 1, 2, …, n_{ij}, depending on both clinics and subjects) because attrition of subjects during a trial in reality is the norm rather than exception [16,17]. Nevertheless, our derivation based on non-varying cluster sizes provides a useful approximation and, further, can serve as a basis for deriving a sample size algorithm for varying cluster sizes. For instance, if the variation in the cluster sizes is completely at random in the missing data analysis framework [18], a replacement of the varying cluster sizes with an average cluster size has been shown to be effective for sample size and statistical power with varying cluster sizes under two level binary outcome data [19]. Second, for pragmatic reasons the covariance structure (4) considered here was based on the conditional independence assumption. Therefore, robustness of the derived formulae under alternative covariance structure, such as autocorrelation or unstructured covariance matrix, is unknown.

In conclusion, the derived formulae for sample sizes (15,16,17) and power functions (12,13) can be useful in designing community based longitudinal cluster-randomized clinical trials that compare slopes of outcomes over time between two intervention groups in a three level data structure.

Acknowledgments

We are grateful to Donald Hedeker Ph.D., two anonymous referees and an Associate Editor for their valuable suggestions. This study was supported in part by NIMH grants, P30MH068638 and R01MH060447.

Appendix

Proof of equation (9), *Var* $Var\left(\widehat{\eta}_g\right)=\frac{\sigma_e^2}{N_3N_2N_1Var_p\left(T\right)}=\frac{\left(1-\rho_1\right)\sigma^2}{N_3N_2N_1Var_p\left(T\right)}$. Let $W_k=\left(T_k-\bar{T}\right)$, then We have: $\Sigma_{k=1}^{N_1}W_k^2=N_1Var_p\left(T\right)$; $\Sigma_{k=1}^{N_1}W_k=0$; $\Sigma_{k'\neq k}^{N_1}W_{k'}=-W_k$ and $\widehat{\eta}_g=\Sigma_{i=1}^{N_3}\Sigma_{j=1}^{N_2}\Sigma_{k=1}^{N_1}W_k\left(Y_{ijk}-\bar{Y}_g\right)/N_3N_2N_1Var_p\left(T\right)=\Sigma_{i=1}^{N_3}\Sigma_{j=1}^{N_2}\Sigma_{k=1}^{N_1}W_kY_{ijk}/N_3N_2N_1Var_p\left(T\right)$. Observing that Y is independent over i, we decompose the variance of the numerator of $\widehat{\eta}_g$ as follows:

$$Var\left(\sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1}W_kY_{ijk}\right)=\underbrace{\sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1}W_k^2Cov\left(Y_{ijk},Y_{ijk}\right)}_{A}+\underbrace{\sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1}\sum_{k'\neq k}^{N_1}W_kW_{k'}Cov\left(Y_{ijk},Y_{ijk'}\right)}_{B}$$
$$+\underbrace{\sum_{i=1}^{N_3}\sum_{j=1}^{N_2}\sum_{j'\neq j}^{N_2}\sum_{k=1}^{N_1}\sum_{k'=1}^{N_1}W_kW_{k'}Cov\left(Y_{ijk},Y_{ij'k'}\right)}_{C}.$$

Now, recall equation (4), that is,

$$Cov\left(Y_{ijk},Y_{i'j'k'}\right)=1\left(i=i'\&j=j'\&k=k'\right)\sigma_e^2+1\left(i=i'\&j=j'\right)\sigma_2^2+1\left(i=i'\right)\sigma_3^2.$$

It follows that $A=\sigma^2N_3N_2N_1Var_p(T)$ since $Var\left(Y_{ijk}\right)=\sigma^2=\sigma_e^2+\sigma_2^2+\sigma_3^2$. Further, $\Sigma_{k=1}^{N_1}\Sigma_{k'\neq k}^{N_1}W_kW_{k'}Cov\left(Y_{ijk},Y_{ijk'}\right)=-\left(\sigma_2^2+\sigma_3^2\right)\Sigma_{k=1}^{N_1}W_k^2$ since $\Sigma_{k'\neq k}^{N_1}W_{k'}=-W_k$. Therefore, $B=-\left(\sigma_2^2+\sigma_3^2\right)N_3N_2N_1Var_p\left(T\right)$. It is easy to see that $C=0$ since $\Sigma_{k=1}^{N_1}W_k=0$. Hence, we have $Var\left(\Sigma_{i=1}^{N_3}\Sigma_{j=1}^{N_2}\Sigma_{k=1}^{N_1}W_kY_{ijk}\right)=A+B=\sigma_e^2N_3N_2N_1Var_p\left(T\right)$. It follows that equation (9) above holds.

Reference

1. Goldstein, H. Multilevel Statistical Models. 2nd ed.. Wiley; New York: 1996.

2. Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd ed.. SAGE; Thousand Oaks: 2002.

3. Hedeker, D.; Gibbons, RD. Longitudinal Data Analysis. Wiley; Hoboken, NJ: 2006.

4. Donner A, Birkett N, Buck C. Randomization by clusters; Sample size requirements and analysis. American Journal of Epidemiology. 1981; 114:906–914. [PubMed: 7315838]

5. Donner A, Klar N. Statistical Consideration in the design and analysis of community intervention trials. Journal of Clinical Epidemiology. 1996; 49:435–439. [PubMed: 8621994]

6. Diggle, PJ.; Heagerty, P.; Liang, K-Y.; Zeger, SL. Analysis of Longitudinal Data. 2nd ed.. Oxford University Press; New York: 2002.

7. Serfling, RJ. Approximation Theorems of Mathematical Statistics. Wiley; New York: 1980.

8. Alexopoulos GS, Katz IR, Bruce ML, Heo M, Ten Have T, Raue PJ, Bogner HR, Schulberg HC, Mulsant BH, Reynolds CF III, the PROSPECT Group. Remission in depressed geriatric primary care patients: a report from the PROSPECT study. American Journal of Psychiatry. 2005; 62:718–724. [PubMed: 15800144]

9. Bruce ML, Ten Have TR, Reynolds CF III, Katz I, Schulberg HC, Mulsant BH, Brown GK, McAvay GJ, Pearson JL, Alexopoulos GS. Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. JAMA. 2004; 291:1081–1091. [PubMed: 14996777]

10. Dietrich AJ, Oxman TE, Williams JW Jr. Schulberg HC, Bruce ML, Lee PW, Barry S, Raue PJ, Lefever JJ, Heo M, Rost K, Kroenke K, Gerrity M, Nutting PA. Re-Engineering Systems for the Primary Care Treatment of Depression: A Randomized Controlled Trial. British Medical Journal. 2004; 329:602–605. [PubMed: 15345600]

11. Johnson, NL.; Kotz, S. Distributions in Statistics: Continuous Univariate Distributions-2. Houghton Mifflin; New York: 1970.

12. Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. Statistics in Medicine. 2002; 21:1787–1801. [PubMed: 12111912]

13. Murray DM, Blitstein JL, Hannan PJ, Baker WL, Lytle LA. Sizing a trial to alter the trajectory of health behaviors: Methods, parameter estimates, and their application. Statistics in Medicine. 2007; 26:2297–2316. [PubMed: 17044139]

14. Roy A, Bhaumik DK, Aryal S, Gibbons RD. Sample size determination for hierarchical longitudinal designs with differential attrition rates. Biometrics. 2007; 63:699–707. [PubMed: 17825003]

15. Heo M, Leon AC. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. Biometrics. in press.

16. Leon AC, Mallinckrodt CH, Chuan-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology. Biological Psychiatry. 2006; 59:1001–1005. [PubMed: 16503329]

17. Heo M, Leon AC, Meyers BS, Alexopoulos GS. Problems in statistical analysis of attrition in randomized controlled clinical trials of antidepressants for geriatric depression. Current Psychiatry Reviews. 2007; 3:178–185.

18. Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.

19. Heo M, Leon AC. Performance of a mixed effects logistic regression model with unequal cluster size. Journal of Biopharmaceutical Statistics. 2005; 15:513–526. [PubMed: 15920895]

Table 1

Sample size N_3 theoretical power φ and empirical power $\tilde{\varphi}$ for testing intervention group by time interaction effect in a three level mixed-effects linear regression analysis, based on 1000 simulations.

			$\Delta_\delta T_{end} = 0.3$			$\Delta_\delta T_{end} = 0.4$			$\Delta_\delta T_{end} = 0.5$		
N_2	N_1	ρ_1	N_3	φ	$\tilde{\varphi}$	N_3	φ	$\tilde{\varphi}$	N_3	φ	$\tilde{\varphi}$
5	3	0.4	42	0.801	0.798	24	0.807	0.787	16	0.823	0.838
		0.5	35	0.801	0.804	20	0.807	0.810	13	0.813	0.806
		0.6	28	0.801	0.818	16	0.807	0.810	11	0.834	0.839
	6	0.4	30	0.801	0.803	17	0.804	0.807	11	0.808	0.836
		0.5	25	0.801	0.775	15	0.826	0.835	9	0.801	0.798
		0.6	20	0.801	0.810	12	0.826	0.802	8	0.841	0.847
	12	0.4	18	0.806	0.831	10	0.801	0.800	7	0.835	0.832
		0.5	15	0.806	0.821	9	0.831	0.829	6	0.845	0.837
		0.6	12	0.806	0.794	7	0.820	0.817	5	0.860	0.855
10	3	0.4	21	0.801	0.801	12	0.807	0.788	8	0.823	0.829
		0.5	18	0.812	0.811	10	0.807	0.804	7	0.841	0.852
		0.6	14	0.801	0.791	8	0.807	0.802	6	0.865	0.865
	6	0.4	15	0.801	0.809	9	0.826	0.845	6	0.841	0.843
		0.5	13	0.816	0.809	8	0.849	0.855	5	0.841	0.846
		0.6	10	0.801	0.795	6	0.826	0.822	4	0.841	0.829
	12	0.4	9	0.806	0.788	5	0.801	0.814	4	0.881	0.875
		0.5	8	0.831	0.822	5	0.868	0.869	3	0.845	0.834
		0.6	6	0.806	0.784	4	0.868	0.878	3	0.914	0.912
20	3	0.4	11	0.819	0.825	6	0.807	0.793	4	0.823	0.826
		0.5	9	0.812	0.809	5	0.807	0.793	4	0.885	0.878
		0.6	7	0.801	0.784	4	0.807	0.805	3	0.865	0.863
	6	0.4	8	0.826	0.816	5	0.863	0.862	3	0.841	0.853
		0.5	7	0.844	0.822	4	0.849	0.826	3	0.900	0.903
		0.6	5	0.801	0.800	3	0.826	0.838	2	0.841	0.839
	12	0.4	5	0.845	0.850	3	0.868	0.857	2	0.881	0.866
		0.5	4	0.831	0.842	3	0.920	0.919	2	0.930	0.927

			$\Delta_\delta T_{end} = 0.3$			$\Delta_\delta T_{end} = 0.4$			$\Delta_\delta T_{end} = 0.5$		
N_2	N_1	ρ_1	N_3	φ	$\tilde{\varphi}$	N_3	φ	$\tilde{\varphi}$	N_3	φ	$\tilde{\varphi}$
		0.6	3	0.806	0.800	2	0.868	0.871	2	0.970	0.964
30	3	0.4	7	0.801	0.823	4	0.807	0.800	3	0.865	0.863
		0.5	6	0.812	0.806	4	0.873	0.867	3	0.918	0.901
		0.6	5	0.828	0.824	3	0.851	0.846	2	0.865	0.851
	6	0.4	5	0.801	0.801	3	0.826	0.829	2	0.841	0.839
		0.5	5	0.867	0.867	3	0.888	0.887	2	0.900	0.900
		0.6	4	0.867	0.868	2	0.826	0.828	2	0.952	0.956
	12	0.4	3	0.806	0.819	2	0.868	0.857	2	0.970	0.964
		0.5	3	0.872	0.871	2	0.920	0.929	1	0.845	0.847
		0.6	2	0.806	0.801	2	0.965	0.964	1	0.914	0.916
Mean				0.815	0.814		0.840	0.837		0.866	0.865

N_1 = the number of level one units (repeated measures) per subjects; N_2 = the number of level two units (subjects) per clinic; N_3 = the number of level three units (clinics) per group, i.e., the sample size obtained from equation (15); T_{end}d = N_1 - 1; ρ_1 = correlation among level one data (5); φ = theoretical power based on the formula (13); $\tilde{\varphi}$ = empirical power based on equation (18); Δ_δ = standardized effect size of the slope difference that yields an intervention efect Δ_δ T_{end} at the end of a study.