DOI: 10.1111/j.1541-0420.2007.00769.x

Sample Size Determination for Hierarchical Longitudinal Designs with Differential Attrition Rates

Anindya Roy,[1,2] Dulal K. Bhaumik,[1,3,4] Subhash Aryal,[1,3,4] and Robert D. Gibbons[1,3,4,*]

[1]Center for Health Statistics, University of Illinois at Chicago, 1601 W. Taylor St., Chicago, Illinois 60612, U.S.A.

[2]Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland 21250, U.S.A.

[3]Department of Psychiatry, University of Illinois at Chicago, 1601 W. Taylor St., Chicago, Illinois 60612, U.S.A.

[4]Division of Epidemiology and Biostatistics, University of Illinois at Chicago, 1601 W. Taylor St. Chicago, Illinois 60612, U.S.A.

[*]*email:* rdgib@uic.edu

Summary. We consider the problem of sample size determination for three-level mixed-effects linear regression models for the analysis of clustered longitudinal data. Three-level designs are used in many areas, but in particular, multicenter randomized longitudinal clinical trials in medical or health-related research. In this case, level 1 represents measurement occasion, level 2 represents subject, and level 3 represents center. The model we consider involves random effects of the time trends at both the subject level and the center level. In the most common case, we have two random effects (constant and a single trend), at both subject and center levels. The approach presented here is general with respect to sampling proportions, number of groups, and attrition rates over time. In addition, we also develop a cost model, as an aid in selecting the most parsimonious of several possible competing models (i.e., different combinations of centers, subjects within centers, and measurement occasions). We derive sample size requirements (i.e., power characteristics) for a test of treatment-by-time interaction(s) for designs based on either subject-level or cluster-level randomization. The general methodology is illustrated using two characteristic examples.

Key words: Cost analysis; Dropouts; Mixed effects; Power analysis; Profile analysis; Three-level nested design.

1. Introduction

In the medical and health sciences, multicenter randomized clinical trials have become the design of choice for large-scale studies of the effects of medical interventions on health outcomes. In such trials a primary focus of the analysis is testing the single degree of freedom treatment by linear time interaction. In this design, treatment and treatment-by-time interactions are treated as fixed effects in the statistical model, whereas the intercept and slope of the time trends are considered to be random at both the center level and the subject level.

When designing a longitudinal multicenter clinical trial, there are several fundamental questions that must be answered: (i) What are the minimum number of subjects (n), centers (c), and time points (T) that are required in order to attain a prespecified power level for testing significance of the treatment-by-time interaction? (ii) If the number of centers falls below the minimum number, can one still achieve the desired level of statistical power (e.g., 80%) by increasing the number of subjects enrolled in each center? (iii) How should the sample size be adjusted if dropouts are expected? (iv) If there is a budget constraint, what is the optimum allocation of c, n, and T which achieves the desired power level?

At present there are no definitive guidelines that can simultaneously answer all of these questions for three-level nested designs. Questions (i) and (iii) can be implicitly solved using a computer-intensive method which iteratively computes the power and solves for the sample size. One can use the form of the noncentrality parameter of the test for specific models; see Liu and Liang (1997), Verbeke and Molenberghs (2000), and Diggle et al. (2002). For question (ii), there is no clear answer and the findings of this article will show that the actual answer disproves the general perception in the practice of multicenter clinical trials. Such findings are in the same vein with those in the group randomization trial literature (Donner, 1992; Donner and Klar, 2000) where the trade-off between cluster size and number of clusters is investigated using results by Kish and Frankel (1974) and others in the context of multistage sampling. The cost analysis question is addressed by many investigators (e.g., Bloch [1986], Lui and Cumberland [1992], and Raudenbush and Liu [2000]) for some simpler designs and statistical models than the ones considered here.

Sample size determination is an important problem as insufficient sample size can lead to inadequate sensitivity delaying important discoveries, whereas an excessive sample size can be a waste of resources and may have ethical implications.

The objective of this article is to explicate the power characteristics of the two general types of three-level studies discussed above (i.e., subject randomization versus cluster randomization) while allowing for dropouts. We answer the above questions under the umbrella of a general three-way nested model which considers S treatments comprised of K factors ($S \geq K$), with potentially different sampling proportions. In terms of the time trends, we consider up to Q components, which may include up to a Qth degree polynomial ($Q < T$) or any possible set of contrasts, including those used in profile analyses (Fitzmaurice, Laird, and Ware, 2004). The literature of determining sample size for repeated measures is extensive for both univariate and multivariate linear models. Muller and Barton (1989) and Muller et al. (1992) discuss statistical power computation for longitudinal studies. Overall and Doyle (1994), Kirby, Galai, and Munoz (1994), Ahn, Overall, and Tonidandel (2001) among others provide formula for sample size estimation for repeated measures models involving two groups. Snijders and Bosker (1999) give approximate power calculations for testing fixed effects in mixed models involving repeated observations. Researchers have attempted to model the effect of dropouts on power computation (Diggle et al., 2002). Verbeke and Molenberghs (2000) used a dropout process model for deriving the noncentrality parameter of an F-test that is used for testing fixed effects in linear mixed models. Hedeker, Gibbons, and Waternaux (1999) derived sample size formula for two-level nested designs with a constant attrition rate. However, no results for sample size determination are available for three-level designs, specifically when there is attrition. Several researchers have investigated the problem of determining sample size for maximizing power under budget constraints. Bloch (1986) and Lui and Cumberland (1992) determine the sample size and the number of time points for a specific power level in a repeated measurement study in order to minimize suitable cost functions. Raudenbush and Liu (2000) determine the number of centers and number of subjects per center in a multicenter randomized clinical trial by incorporating costs at each level of the design.

The article is organized as follows. In Section 2, we introduce the model and discuss the relevant hypotheses and test statistics. We derive the sample size formula for subject-level randomization (Theorem 1) and center-level randomization (Theorem 2) using differential attrition rates. Some special cases like no attrition and constant attrition are also discussed in this section. We illustrate our results using two motivating examples in Section 3 and provide numerical results. Section 4 provides a cost analysis in order to minimize the total cost of the study under the constraint of a prespecified power level. We summarize our findings in Section 5. Proofs of the main results (Theorems 1 and 2) are relegated to a Web Appendix which can be found in the supplementary materials posted on the *Biometrics* website.

2. Models and Results

Suppose there are c centers, and n subjects are nested within each center. The total nc subjects are divided into S treatment groups. Suppose the treatment index for the sth treatment group is a $1 \times (K + 1)$ vector whose first element is one, representing a baseline treatment factor and the other K elements are the levels of K factors (e.g., a treatment could be a cocktail drug comprised of different dose levels of K main drugs) contributing to the treatment. Each subject is assigned to a particular treatment. The treatments can be assigned in two different ways. In the first assignment scheme each subject in each center is randomized into one of S treatment groups and suppose the allocation proportions of the treatment groups are $\pi = (\pi_1, \ldots, \pi_S)$. Thus, $n\pi_s$ subjects are assigned to the sth treatment group. We will assume that the allocation proportion vectors remain the same across centers. In the literature, this assignment is known as subject-level randomization where the subjects are nested within the treatment groups which in turn are nested in the centers. The other option is to randomly assign the c centers to the S treatments with allocation proportion vector π. Thus, $c\pi_s$ number of centers will receive the sth treatment. In that case, all subjects in a given center will receive the treatment that has been assigned to the center. This is known as center-level randomization where the subjects are nested in centers which in turn are nested in the treatment groups. It is understood that the quantities $n\pi_s$ and the $c\pi_s$ in the above discussion are all integers.

Let us assume that a longitudinal study has been conducted at T different time points. In this article, we incorporate the possibility that subjects may dropout from the experiment at any time point. The dropout process can be either modeled as a stochastic process (Verbeke and Molenberghs, 2000) or empirically estimated. However, at the design stage it is unlikely that a good stochastic model for the dropouts is available. We let the dropouts be prespecified by the experimenter. For the development of our model we assume a monotone dropout rate. We permit different dropout rates at different time points and across different treatment groups. Specifically, let $\xi_{s,t}$ denote the fraction of subjects in the sth treatment group with responses at only the first t time points. We will call the vector of such fractions for a particular treatment group, $\xi_s = (\xi_{s,1}, \ldots, \xi_{s,T})'$ as the attrition vector of that treatment group. Our ultimate objective is to determine the sample size for the test of interaction between the treatment and time. Unless some subjects stay beyond the first time point, testing of treatment-by-time interaction is not a feasible proposition. Thus, we will assume that $\xi_{s,1} < 1$ for all s. To capture the evolution of the treatment responses over time, we consider a time trend matrix T_R of order $T \times (Q + 1)$. We will take the first column of the T_R matrix to be one, to capture a baseline effect. The remaining Q columns each represent a particular functional effect of time. For example, if we are considering only a linear trend, then Q is one and the second column of T_R will be $(1, \ldots, T)'$. We will assume the columns of T_R to be linearly independent. The basic time variable t is assumed to be equally spaced but because we are considering very general functions of t, the spacings can be adjusted by changing the trend functions. Because subject-level randomization and center-level randomization give rise to two distinct designs, we will consider these two cases separately. The development of the methodology relies heavily on matrix notation. The notation is explained in Table 1.

Table 1
Glossary of notation

Notation	Meaning
c, n, T	# of centers, subjects, measurement occasions
K, S, Q	# of treatments, treatment factors, time trends
Σ_γ, Σ_δ, Σ_ϵ	center, subject, error variance component
τ_*	alternative value for single degree of freedom test
κ	$(z_\alpha + z_\beta)^2/\tau_*^2$ where z_α are N(0,1) percentiles
T_R; T_{Rt}	time matrix; first t rows of T_R
f_{22}	(2,2)th element of $(T'_R \Sigma_\epsilon^{-1} T_R)^{-1}$
$\mathcal{T}$	$(T'_{R1}, T'_{R2}, \ldots, T'_{RT})'$
u_s	row vectors of treatment indices
1_r; I_r	$r \times 1$ vector of ones; $r \times r$ identity matrix
$\bigoplus_{i=1}^r B_i$	blockdiagonal matrix with blocks B_1, $B_2, \ldots, B_r$
$\{_c x_i\}_{i=1}^r$	stacking column vectors x_1, $x_2, \ldots, x_r$ vertically
$A \otimes B$	Kronecker product of the matrices A and B
$\begin{pmatrix} A_{11} & \cdot\cdot & A_{1r} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \\ A_{m1} & \cdot\cdot & A_{mr} \end{pmatrix} \overset{\bullet}{\otimes} \begin{pmatrix} B_{11} & \cdot\cdot & B_{1r} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \\ B_{m1} & \cdot\cdot & B_{mr} \end{pmatrix}$	$\begin{pmatrix} A_{11} \otimes B_{11} & \cdots & A_{1r} \otimes B_{1r} \\ \cdots & \cdots & \cdots \\ A_{m1} \otimes B_{m1} & \cdots & A_{mr} \otimes B_{mr} \end{pmatrix}$
U; V_s; $\mathcal{U}_s$	$(u'_1 : u'_2 : \cdots : u'_S)'$; $\{\{_c 1_{n\xi_{s,t}}\}_{t=1}^T\}_{s=1}^S$; $u'_s \otimes V_s$
ζ_s; ζ; Φ; X_c	$\{_c 1_{n\pi_s \xi_{s,t}}\}_{t=1}^T$; $\{_c \zeta_s\}_{s=1}^S$; $\zeta \overset{\bullet}{\otimes} (1_S \otimes \mathcal{T})$; $(U \overset{\bullet}{\otimes} \Phi)$
Λ	$[\bigoplus_{s=1}^S \bigoplus_{a=1}^T I_{n\pi_s \xi_{s,a}} \otimes T_{Ra}]$
$\Sigma_{\epsilon,tt}$	upper left $t \times t$ block of Σ_ϵ
G_t; R_s	$\Sigma_{\epsilon,tt} + T_{Rt}\Sigma_\delta T'_{Rt}$; $\sum_{t=1}^T \xi_{s,t} T'_{Rt} G_t^{-1} T_{Rt}$
$\tilde{\Omega}$	$(U \otimes I_{Q+1})'(\bigoplus_{s=1}^S \pi_s R_s)(U \otimes I_{Q+1})$
$\tilde{\Omega}_n$	$(U \otimes I_{Q+1})'(\bigoplus_{s=1}^S \pi_s [n^{-1} R_s^{-1} + \Sigma_\gamma]^{-1})(U \otimes I_{Q+1})$

2.1 *Subject-Level Randomization*

Let y denote the vector of observations and let θ, γ, and δ denote the vector of fixed effects, the vector of center-level random effects and the vector of subject-level random effects, respectively. The coefficient matrices are $X = 1_c \otimes X_c$ for the fixed effects, $Z_\gamma = I_c \otimes \Phi$ for center-level random effects, and $Z_\delta = I_c \otimes \Lambda$ for the subject-level random effects, where X_c, Φ, Λ, I_c and T_{Ra} are defined in Table 1. Then a mixed-model for the responses can be written as

$$y = X\theta + Z_\gamma \gamma + Z_\delta \delta + \epsilon. \tag{1}$$

where ϵ is the vector of model errors. The problem of interest is to test a linear hypothesis about the treatment-by-time interaction which reduces to a general linear hypothesis about the fixed-effect parameters:

$$H_0 : L\theta = 0 \quad \text{vs.} \quad H_1 : L\theta \neq 0, \tag{2}$$

where rows of L describe d linearly independent linear hypotheses about the fixed-effect parameters θ. Note that some frequently used hypotheses like treatment-by-time interaction, profile analysis based on the end time point are all included in (2). When $d = 1$ the concept of one-sided tests is applicable. To determine a critical region for the hypothesis (2) we need to make distributional assumptions about the random components of the model. Let

$$\begin{aligned} (\gamma_{0,i}, \ldots, \gamma_{Q,i})' &\sim N_{Q+1}(0, \Sigma_\gamma), \\ (\delta_{isj0}, \ldots, \delta_{isjQ})' &\sim N_{Q+1}(0, \Sigma_\delta). \end{aligned} \tag{3}$$

For any $a = 1, 2, \ldots, T$, let the errors $(\epsilon_{isaj1}, \ldots, \epsilon_{isaja})' \sim N_a(0, \Sigma_{\epsilon,aa})$, where $\Sigma_{\epsilon,aa}$ is defined in Table 1. We will assume that the random effects are independent of each other and the variance components Σ_γ, Σ_δ, and Σ_ϵ are specified. Under these assumptions, the distribution of the responses is $y \sim N(X\theta, \Sigma)$, where Σ can be explicitly defined.

Let us now define the appropriate critical region for testing (2), based on the assumption that Σ is completely known. The appropriate test statistic is the pivot $F = (L\hat{\theta})'[L\Omega L']^{-1}(L\hat{\theta})$, where $\hat{\theta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ is the generalized least squares (GLS) estimator of θ, $\Omega = (X'\Sigma^{-1}X)^{-1}$ is the covariance matrix of $\hat{\theta}$. Let $\chi^2_{d,\alpha}$ be the upper α percentile of the χ^2-distribution with d degrees of freedom. The most powerful critical region at level α for testing (2) is

$$\mathfrak{C} = \{y : F > \chi^2_{d,\alpha}\}. \tag{4}$$

In practice, an appropriate test statistic for the hypothesis (2) will be a Wald-type pivotal statistic based on a feasible version of the GLS estimator (FGLS) of θ where the variance components are replaced by their estimators. Kenward and Roger (1997) suggested a scaled F distribution as a finite sample approximation to the null distribution of an FGLS test. The scale and the degrees of freedom of the approximate F involve estimated parameters and also involve n, c, and T in a highly nonlinear way. The distribution of the FGLS test under the alternative hypothesis is not known. Helms (1992), and Verbeke and Molenberghs (2000) provide approximate degrees of freedom and the expression for noncentrality parameters for the approximate F test (including dropouts rates that change over time but not over treatment groups) under the alternative hypothesis. This methodology cannot properly determine the minimum number of centers needed in the study and hence is not likely to yield the optimum cost solution of (c, n, T). This is the basis of choosing the test (4) for this study. Even if we consider the feasible version of the test, our results continue to hold approximately if we include only the noncentrality parameter into the power computation and not the estimated degrees of freedom. We acknowledge that in practice only feasible versions of the GLS test can be used. In the supplementary materials section for this article which is posted on the *Biometrics* website, we report the results of a simulation study investigating the performance of the power function for a sample whose size has been determined using the test (4), via a limited simulation study for the setup of the first example given in the examples section. When the number of independent units (e.g., centers in center-level randomization) is small relative to the total sample size, the χ^2 test may have problems maintaining the nominal level due to the dependence among the observations. The problem of interest is to test significance of the treatment-by-time interactions. Thus, each row of the hypothesis matrix L is of the form $l_U \otimes l_T$ where l_U is generally a treatment contrast and l_T is a time contrast. In most situations we are not interested in the baseline time effect and the fixed parameters. Hence the fixed parameters $\theta_1, \ldots, \theta_{Q+1}$, do not enter into the hypothesis. This amounts to a treatment contrast l_U whose first element is zero. This assumption about the rows of L amounts to the constraint $Le_1 = 0$, where e_1 is the vector with one at the first place and zero everywhere else. This additional constraint will lead to a certain simplification of our results without compromising much generality in practice. Let $G(d, \alpha, \beta)$ be the noncentrality parameter of a noncentral χ^2 distribution with d degrees of freedom, whose lower β percentile is the upper α percentile of a central χ^2 distribution with d degrees of freedom.

Theorem 1. *Suppose there are n subjects divided into S treatment groups, in each of the c centers. Suppose the treatment allocation proportions in each center are given by the allocation vector π. Let the attrition vector in the sth treatment group in each center be ξ_s. Then to attain power of at least $(1 - \beta)$ for the test* (4) *at an alternative value of θ at θ_*, a lower bound for the required number of subjects per center is given by*

$$n \geq G(d, \alpha, \beta)/\left(c\theta_*' L'[L\tilde{\Omega}^{-1}L']^{-1}L\theta_*\right), \tag{5}$$

where $\tilde{\Omega}$ is defined in Table 1.

The sample size formula (5) does not involve parameters of the covariance matrix Σ_γ of the center-level random components. This is true when our inference is regarding the treatment-by-time interaction parameters and not the baseline treatment parameters alone. The formula (5) can be written as $c \geq G(d, \alpha, \beta)/(n\theta_*' L'[L\tilde{\Omega}^{-1}L']^{-1}L\theta_*)$. Thus, the formula can be also used to determine the number of centers to be included in the study when the number of subjects in each center, n, is specified. Hence, the roles of c and n are interchangeable when randomization is done at the subject level. More generally the particular allocation of n and c is not important, as long as their product attains the lower bound $G(d, \alpha, \beta)/(\theta_*' L'[L\tilde{\Omega}^{-1}L']^{-1}L\theta_*)$. Under certain simpler error covariance structures (e.g., $\Sigma_\epsilon = \sigma_\epsilon^2 I_T$), the elements of $\tilde{\Omega}$ can be written as a function of T and in those cases the formula (5) can also be used to determine the number of time points T required to attain a prespecified power level when n and c are fixed.

2.1.1 *Sample size determination for single degree of freedom test with no attrition.* A scenario that is often encountered in practice is when a treatment is compared with a placebo or a control and there is a single time trend. In this case the treatment matrix U is $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ and a typical element of the second column of the time matrix T_R is $g(t)$, the trend at time t. Let y_{isajt} denote the response at the tth time point of the jth subject from the ath attrition group nested in the sth treatment group of the ith center. Then a model for the response y_{isajt} is,

$$\begin{aligned} y_{isajt} &= \eta_0 + \eta_1 g(t) + \tau_0 x_{ijk} + \tau_1 x_{ijk} g(t) + \gamma_{i0} + \gamma_{i1} g(t) \\ &\quad + \delta_{isj0} + \delta_{isj1} g(t) + \epsilon_{isajt}, \end{aligned} \tag{6}$$

where γ's are the center-level random components and δ's are those at the subject-level. The treatment indicator $x_{ijk} = 1$ if the subject was assigned to test condition and equal to 0 otherwise. The problem of interest is to test whether there is any treatment-by-time interaction, i.e.,

$$H_0 : \tau_1 = 0 \quad \text{vs.} \quad H_1 : \tau_1 > 0. \tag{7}$$

Let π_1 and π_2 be the sampling proportions and let there be no attrition. Let the alternative value be $\tau_* = L\theta_*$ and let $\kappa = (z_\alpha + z_\beta)^2/\tau_*^2$ where z_α and z_β are the standard normal percentiles. Let f_{22} and $\sigma_{\delta,22}$ denote the (2,2)th element of the matrices $(T_R' \Sigma_\epsilon^{-1} T_R)^{-1}$ and Σ_δ respectively.

Corollary 1. *The sample size formula for the single degree of freedom test is*

$$n \;\geq\; \frac{(f_{22} + \sigma_{\delta,22})\kappa}{\pi_1 \pi_2 c}. \tag{8}$$

In a single degree of freedom test, for a one-sided alternative, the value $(z_\alpha + z_\beta)^2$ matches with $G(1, 2\alpha, \beta)$ of the general formula. The U matrix for the above scenario is nonsingular. The sample size formula in (5) can be simplified further whenever U is nonsingular, even in presence of differential attrition rates. The sample size formula in (8) depends on the subject-level covariance matrix only through the variance of the subject-level random slope.

2.2 *Center-Level Randomization*

Next we consider the case in which the centers are randomly assigned to the treatments and all subjects in a given center receive the treatment assigned to that center. The centers are randomly assigned to S treatments according to allocation proportions π, i.e., all the subjects in $\pi_s c$ centers will receive treatment s. Suppose the subjects in a center receiving the sth treatment drop out according to the attrition vector ξ_s. Because the observations across the centers are independent, it is enough to define the model center by center. The fixed-effect matrix for a center receiving treatment s is given by $X_s = \mathcal{U}_s \overset{\bullet}{\otimes} \mathcal{T}$, where $\mathcal{U}_s$ and $\mathcal{T}$ are defined in Table 1. Then the sample size formula associated with the hypothesis (2) for a specific alternative is given by the following theorem.

Theorem 2. *Suppose there are c centers divided into S treatment groups according to the allocation vector π. Let each* *center have n subjects and let the dropout rates in the centers receiving the sth treatment be ξ_s. Then to attain a power of at least $(1-\beta)$ for the test (4) at an alternative value θ_*, a lower bound for the required number of subjects per center is given by*

$$n \ \geq \ \min\{i : f(i) \geq G(d,\alpha,\beta)/c\}, \qquad (9)$$

where $f(n)$ is a strictly increasing function of n defined by $f(n) = \theta_*' L'[L\tilde{\Omega}_n^{-1}L']^{-1}L\theta_*$ and $\tilde{\Omega}_n$ is defined in Table 1. The formula (9) has a feasible solution for all values of c such that

$$c \geq G(d,\alpha,\beta) \Big/ \left(\theta_*' L' \left[L\tilde{\Omega}_\infty^{-1}L'\right]^{-1} L\theta_*\right), \qquad (10)$$

where $\tilde{\Omega}_\infty = (U'\Delta_\pi U) \otimes \Sigma_\gamma^{-1}$, and Δ_π is a diagonal matrix with diagonal elements π_s.

The condition (10) puts a restriction on the minimum number of centers. Formula (10) shows that if the number of centers fall below a critical level, you can never compensate the shortage in number of centers by increasing the number of subjects per center. Such trade-offs between the cluster size and number of clusters can be also found in the group randomization trial literature and are analyzed via the intracluster correlation; see Donner and Klar (2000) and the references therein. The results have roots in the pioneering work by Kish and Frankel (1974) in the context of multi-stage sampling. To determine the values of n and c using the approximate F distribution under the alternative an iterative method is needed. The current methodology provides a lower bound for values of c in the iterations. Note that at the lower bound for c, n can be inordinately large, and so somewhat larger values of c should be considered in practice.

2.2.1 *Sample size determination for single degree of freedom test with no attrition.* Consider the scenario of Subsection 2.1.1 for center-level randomization.

Corollary 2. *Let $\tau_* = L\theta_*$ be the alternative value. Then the formula (9) reduces to*

$$n \geq \frac{(f_{22} + \sigma_{\delta,22})\kappa}{\pi_1\pi_2 c - \kappa\sigma_{\gamma,22}}, \qquad (11)$$

where $\sigma_{\gamma,2,2}$ is the (2, 2)th element of Σ_γ. A feasible solution to (11) exists provided the number of centers satisfies $c \geq \pi_1^{-1}\pi_2^{-1}\kappa\sigma_{\gamma,22}$.

The expression in (11) depends on the elements of Σ_γ only through $\sigma_{\gamma,22}$, the variance of the center-level random slope. Once a center-level random slope is present, to have a consistent test of the treatment-by-time interaction, one needs to accurately estimate the variance of the center-level slope. This entails treating the centers as a sample from a larger population of centers and thus to draw valid inference on a center-level component one needs a sufficiently large sample of centers. Accurate estimation of the center-level slope variance is not possible if the number of centers is inadequate. This disproves the myth that if you have few centers you can still achieve adequate power by increasing the number of subjects per center. For a model with only a random intercept at the center level, the denominator of the right-hand side of (11) does not depend on either elements of Σ_γ or n. In that case, given a prespecified number of centers and an alternative value τ_* one can always find the required number of subjects per center needed to attain any prespecified power level.

To determine the number of centers for a prespecified value of n, the formula (11) can be written as $c \geq \pi_1^{-1}\pi_2^{-1}[\kappa\sigma_{\gamma,qq} + (f_{22} + \sigma_{\delta,22})\kappa/n]$.

3. Examples

To help fix ideas about the three-level nested model using real world studies, consider the following two examples.

Example 1. Testing for treatment trends in mental health schizophrenia data.

The following data were collected as part of the National Institute of Mental Health schizophrenia collaborative study on treatment-related changes in overall severity using the Inpatient Multidimensional Psychiatric Scale (IMPS) (Lorr and Klett, 1966). Nine centers participated in this study, and within each center, subjects were randomly assigned to a placebo condition or one of three drug conditions (chlorpromazine, fluphenazine, or thioridazine). Hedeker and Gibbons (2006) analyzed data only from subjects assigned to either the placebo or the test condition. In this study there were seven time points, hence $T = 7$; however, for the purpose of illustration we use $T = 5$ time points. An appropriate model for the response is (6) and the hypothesis of interest is (7), significance of interaction of test condition and time. To linearize the time versus response function, Hedeker and Gibbons (2006) considered $g(t) = \sqrt{t-1}$. The estimates of the parameters obtained by Hedeker and Gibbons (2006) will be used as the parameter values. Hedeker and Gibbons (2006) assumed the center-level slope variance $\sigma_{\gamma,11}$ to be zero and also the error covariance matrix as $\Sigma_\epsilon = \sigma_\epsilon^2 I_T$. The estimates are $\sigma_{\delta,00} = 0.285, \sigma_{\delta,11} = 0.225, \sigma_{\delta,01} = 0.025, \sigma_{\gamma,00} = 0.039$, and $\sigma_\epsilon^2 = 0.570$. We will address the question of sample size determination when $\sigma_{\gamma,11} = 0.1368$, and $\sigma_{\gamma,01} = 0$. This is a crude method of moments estimator of the center-level slope variance obtained from the data. In the actual trial, subjects were randomized to treatments within centers; however, we can use these data, and estimates of the variance components to derive sample sizes for both subject-level and center-level randomizations. Treating the above estimates as the parameter values for model (6) we determine the sample size required to attain a power of at least 80% for the test (4) with $\alpha = 0.05$.

Case I. Subject-level randomization with no attrition.

The appropriate formula for the sample size calculation in this case is (8). Because we are assuming only a single trend (i.e., $Q = 1$) and $\Sigma_\epsilon = 0.57I$, we have $f_{22} = \sigma_\epsilon^2\sigma_{gT}^{-2}$ where $\sigma_{gT}^2 = \sum_{t=1}^T (g(t) - \bar{g}_T)^2$ is the variance of the trend column and $\bar{g}_T = T^{-1}\sum_{t=1}^T g(t)$ is the mean of the trend column. In this example, $\pi_1 = \pi_2 = 0.5$ and the number of centers is $c = 9$. The value of κ for $\alpha = 0.05$ and $\beta = 0.2$ is $\kappa = 6.1826/\tau_*^2$. The variance of an observation at time point t is $\sigma_{y,t}^2 = 0.5322 + 0.05\sqrt{t-1} + 0.3618\,t$. We can rewrite κ in terms of an effect size at time t. Then the effect size at time point t is $ES_t = \tau_*\sqrt{t-1}/\sigma_{y,t}$. In that case $\kappa = 6.1826/(\sigma_{y,t}^2 ES_t^2)$. Note that at $t = 5$, $\tau_* = 0.2343$, $\kappa_5 = 112.94$, $\sigma_{g5}^2 = 2.4452$,

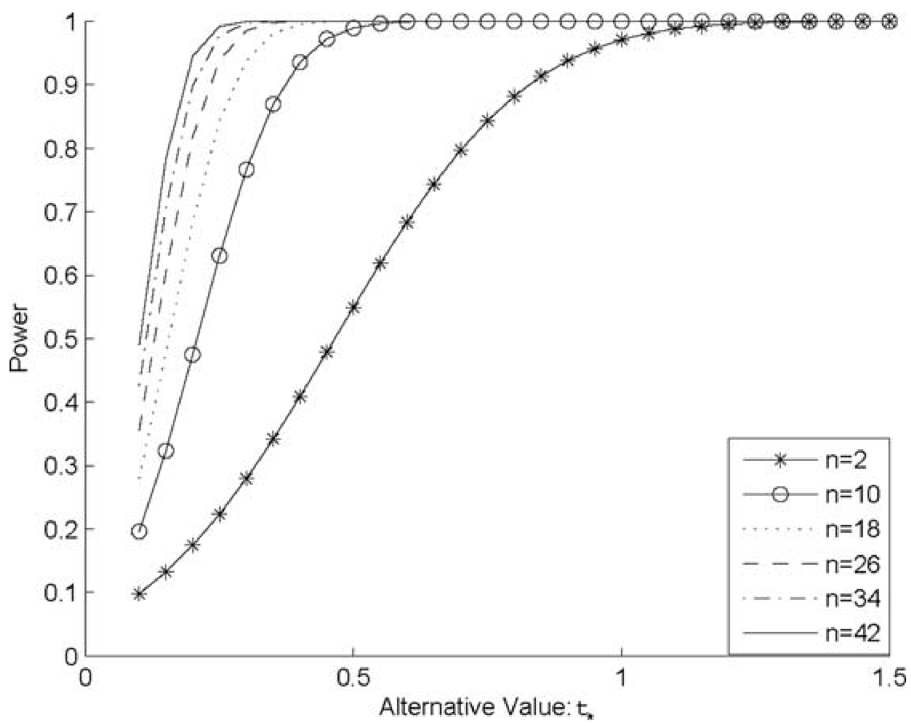

Figure 1. Power curve as functions of τ_* for various values of n in Example 1 with subject-level randomization.

$\sigma^2_{y,5} = 2.4412$. In this formulation, the value of the effect size is $ES_t = 0.30$. Hence the required number of subjects per center is $n \geq 4(6.1826)(0.233 + 0.225)/(9\sigma^2_{y,t}ES^2_t) = 23$, or 23 subjects in each of 9 centers.

In Figure 1 we see that power curves are monotonically increasing function of the alternative value τ_* and they change drastically when number of subjects increases from 2 to 42 in an increment of 8.

Case II. Subject-level randomization with constant attrition.

Suppose $\xi_A = \xi_B = (.1\,.1\,.1\,.1\,.6)'$ and $c = 9$. The formula (8) becomes $n \geq r_{22}\kappa/(\pi_A\pi_B c)$, where r_{22} is the (2,2) element of R_1^{-1} and R_1 is defined in Table 1. Assuming as in Case I that $t = 5$, $\tau_* = 0.2343$, the required number of subjects per center for this case is a total of 29 subjects.

Case III. Center-level randomization with no attrition.

The minimum number of centers obtained from the formula (2) is $c \geq \pi_A^{-1}\pi_B^{-1}\kappa\sigma_{\gamma,11} = 4(6.1826)(0.1368)/\tau_*^2 = 3.3831/\tau_*^2$, and the number of subjects with a value of c more than $3.3831/\tau_*^2$ is $n \geq 4(f_{22} + 0.225)(6.1826)/(c\tau_*^2 - 4(6.1826)(0.1368)) = 25.5168/(c\tau_*^2 - 3.3831)$. For $t = 5$ timepoints, there should be a minimum of 62 centers. For 62 centers, the required number of subjects per center is 552. Note that this is a huge number of subjects. As the number of centers increases beyond the required minimum number, the total number of subjects required decreases dramatically. For example, with 70 centers, only 26 subjects per center are required. With 100 centers, only 6 subjects per center are required. Compared to subject-level randomization (Case I), in which we needed a total of $9 \times 23 = 207$ subjects, center-level randomization requires a much larger number of subjects, for example, a high of 34,224 for 62 centers and 600 for 100 centers. Based on these results, it would seem logical to always use subject-level randomization; however, there are many cases in which this is not possible. For example, most school-based interventions require that randomization is at the level of the school, because only one condition (i.e., experimental or control) can be implemented at a given school.

Case IV. Center-level randomization with constant attrition.

Suppose $\xi_A = \xi_B = (.1\,.1\,.1\,.1\,.6)'$. Assuming 100 centers, 7 subjects per center are required (as compared to 6 subjects per center based on no attrition).

Example 2. Profile analysis for treatment of lead-exposed children (TLC) trial.

The TLC Trial Group was a placebo controlled double blind randomized trial. The purpose of this clinical trial was to compare the effect of lead chelation with succimer to placebo therapy. Data were collected at sites in Cincinnati, Ohio; Philadelphia, Pennsylvania; Baltimore, Maryland; and Newark, New Jersey. Complete data description is available at `http://dir.niehs.nih.gov/direb/studies/tlc/home.htm`.

In the TLC trial, one is interested in comparing the two treatment groups (succimer or placebo) in terms of their patterns of change from baseline in the mean blood lead levels. This question is equivalent to a profile analysis (Fitzmaurice et al., 2004). Model (6) can also be used for the TLC trial data with the four centers, Baltimore, Cincinnati, Newark, and Philadelphia. However, the null hypothesis for the profile analysis would be different from that of the single degree of freedom hypothesis of treatment-by-time interaction. In profile analysis, we first construct the contrast between the treatment means for each time point and then test the equality of these contrasts. Mathematically, this concept can be explained as follows: The treatment matrix is $U = \left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)$ and the time matrix is $T_R = (1_T : C_T)$ where C_T is a $T \times (T - 1)$ matrix whose columns are time contrasts and are orthogonal to the first column of T_R. In particular, the form of C_T appropriate for comparing every time point with the baseline time period is $C_T^{'} = (1_{T-1} : -I_{T-1})$. The objective of the profile analysis is to test the interaction between the treatment and the time contrast L_T. Suppose L_U is the vector $(1\,0)'$ and L_T is a $(T - 1) \times T$ time contrast matrix whose linearly independent rows are orthogonal to 1_T. The df involved in this test is $Q = (T - 1)$ and let $\Sigma_\epsilon = \sigma_\epsilon^2 I_T$. The required sample size at an alternative value θ at $\theta_* = (\theta_{1*}\,\theta_{2*})'$ is

Subject level: $n \geq G(T - 1, \alpha, \beta)/[\theta_{2*}^{'} L_T^{'}(L_T R^{-1} L_T^{'})^{-1} L_T \theta_{2*}]$.

Center level: $n = \min\{i\colon \theta_{2*}^{'} L_T^{'}[L_T(R^{-1}/i + \Sigma_\gamma)L_T^{'}]^{-1} L_T \theta_{2*} \geq G(T - 1, \alpha, \beta)/c\}$,

where the number of centers satisfies $c \geq \theta_{2*}^{'} \Sigma_\gamma^{-1} \theta_{2*}$.

4. Cost Analysis

An experimenter may increase the power of a three-level study in several ways. The variables are T, n, and c. We have treated $(\pi_1, \ldots, \pi_S)$, as known constants in our results. In practice, optimal determination of the proportions may be the prime goal of the study. Thus, the variables for power/cost analysis are c, n, $\pi_1, \ldots, \pi_{S-1}$, and T. However, there is a cost associated with each of the variables. Such trade-off analysis can be also found in multi-stage sampling literature. The fundamental goal of the cost analysis is to minimize the total cost of the study under the constraint on the variables obtained from the sample size formula. In this section, we first formulate the cost analysis problem in terms of a general cost function under the assumption that the sampling proportions are given. Special attention is paid to a particular cost function that is similar to those considered in Bloch (1986) and Lui and Cumberland (1992). The cost analysis problem for this special cost function is solved in light of Example 1.

Let the cost function be denoted by $Q(c, n, T)$. Then the cost analysis problem can be written as an optimization problem $\min_{\delta(\theta_*, c, n, T) \geq G(d, \alpha, \beta)} Q(c, n, T)$, where $(c, n, T) \in \mathbb{Z}_+^3$ are numbers on the positive integer lattice, and $\delta(\cdot, \cdot, \cdot, \cdot)$ is the noncentrality parameter of the distribution of the test statistic under the alternative $\theta = \theta_*$. After doing some linear algebra, and if the error covariance matrix Σ_ϵ is of a simpler form (e.g., independent, intra-class correlation, autoregressive or more generally Toeplitz form) it can be shown that for center-level randomization (the case of subject-level randomization has an easier form) the optimization problem reduces to

$$\min_{c(y'(n^{-1}I + \Delta)^{-1} y) \geq \phi(T)} Q(c, n, T), \qquad (12)$$

where $y \in \mathbb{R}^d, \Delta$ is a diagonal matrix and $\phi(T)$ is a function of T. All quantities depend on the parameters of the problem, the alternative value and the function G. The form (12) can be solved using integer programming routines. Given that most cost functions Q will be increasing in all arguments, the solution will lie on the boundary of the constraint space (integer boundary). As we are solving only in a three-dimensional space, the problem (12) can be satisfactorily solved through repeated function evaluation. In simpler

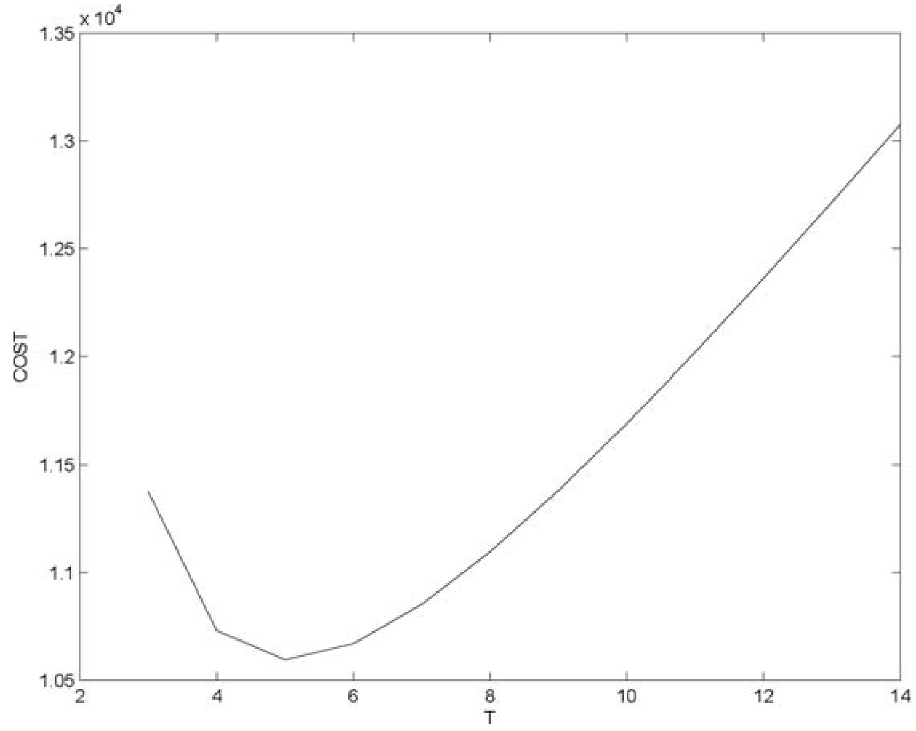

Figure 2. Cost as a function of T when n and c are given at their optimal values.

problems, an approximate optimal solution can be found by solving the corresponding continuous problem.

For example, let us consider the single degree of freedom test situation discussed in Example 1 and let the cost structure be of the form

$$Q_1(c, n, T) = a_1 + a_2T + a_3c + a_4Tc + a_5nc + a_6Tnc,$$

where a_1 is the initial cost, a_2 is the cost involved with each time point, a_3 is the initial cost for enrolling each center for the study, a_4 is the incremental center cost at each time point, a_5 is the per subject enrollment cost, and a_6 is the incremental cost for each subject at each time point. When there is attrition, the total number of measurements is smaller than Tnc and the coefficient of a_6 is potentially smaller than the factor Tnc considered in the cost function. For more details about cost functions we refer to Bloch (1986) and Lui and Cumberland (1992). We will write f_{22} in Theorem 1 as $f_{22,g}(T)$ to indicate its dependence on the trend function $g(t)$ and T. In Example 1, $g(t) = \sqrt{t-1}$. Then the optimization problem can be written as $\min_{n(c-c_{\min})\geq\phi(T)} Q_1(c, n, T)$, where $\phi(T) = 4(f_{22,g}(T) + \sigma_{\delta,22})\kappa$ and $c_{\min} = 0$ for subject-level randomization and $c_{\min} = 4\kappa\sigma_{\gamma,22}$ for center-level randomization. Let us consider the case of subject level randomization. If we optimize over $\{c \geq 1; n \geq 1; T \geq 2\}$, then the solution is $n_{\text{opt}} = \phi(T_{\text{opt}})/c_{\text{opt}}$ where c_{opt} and T_{opt} are obtained by minimizing $[a_1 + a_2T + a_3c + a_4cT + a_5\phi(T) + a_6T\phi(T)]$ for $T > 1$ and $1 \leq c \leq \phi(T)$. For a given value of T, the above cost function is linear in c and hence the solution occurs at $c = 1$. Thus, the optimum number of time points, T_{opt} is a solution to

$$\min_{T>1}[(a_1 + a_3) + (a_2 + a_4)T + (a_5 + a_6T)\phi(T)]. \qquad (13)$$

We chose hypothetical values of the coefficient vector $(a_1\, a_2\, a_3\, a_4\, a_5\, a_6)$ as (1000 200 1000 100 10 1) and optimized the cost in (13). The optimal solution lies at $(c = 1, n = 207, T = 5)$. Thus, the total number of subjects and the number of time points matches with those in Example 1, case I; however, due to the cost structure, the optimum cost solution puts all subjects in one center instead of having nine centers with 23 subjects in each center. Figure 2 verifies that the cost is minimized at $T_{\text{opt}} = 5$ when n and c are held at their optimal values. The minimum cost at $T = 5$ is about $10,600 with the current relative cost structure.

5. Conclusions

In this article, we have developed a general model for sample size determination for multicenter longitudinal studies. This methodology includes multicenter randomized longitudinal clinical trials with randomization at either the subject level or the center level. We consider two-group and multiple-group comparisons. With respect to time, we considered any possible set of contrasts, but focused in particular on polynomial contrasts, with a linear trend model as the simplest case, and simple contrasts to baseline as a form of profile analysis. Our model is also general with respect to treatment group allocation proportions, as well as dropout rates, which are allowed to vary both between groups and over time.

A very important finding of this study is that for multicenter longitudinal studies with subject-level randomization, the center-level variance components play no role whatsoever in computing power. While initially somewhat counterintuitive, further reflection reveals that because interest is in the treatment-by-time interaction, and both treatment and time are nested within centers, variability in intercepts and trend coefficients across centers plays no role in the power characteristic of the statistical test of the treatment-by-time interaction. This is not the case, however, for center-level randomization. Here, center-level trend coefficients play a significant role in sample size determination, such that different combinations between centers and number of subjects within centers, can lead to quite different power characteristics, even if the total number of subjects is identical. Furthermore, we note that for center-level randomization, there is a minimum number of centers, below which certain levels of statistical power (e.g., 0.8) are unattainable, irrespective of the number of subjects per center.

Taken as a whole, the findings of this study provide a more rigorous statistical foundation for the design of multicenter longitudinal randomized clinical trials. The results presented here only apply to the case of linear mixed-effects regression models. Future work in this area should examine statistical power characteristics of nonlinear two- and three-level mixed-effects models for binary, ordinal, nominal, and count response data. Also, one should investigate similar results for other tests used in longitudinal clinical trials and incorporate the issue of controlling Type I error in the cost optimization exercise.

6. Supplementary Materials

Web Appendices for proofs of Theorems 1 and 2 and Web Tables for some simulation results on sensitivity analysis of the proposed formula referenced in Sections 1 and 2, respectively, are available under the Paper Information link at the *Biometrics* website `http://www.tibs.org/biometrics`.

Acknowledgements

The authors thank an associate editor and two referees for their insightful comments that significantly improved the quality of this article. This work was supported in part by a grant from the National Institute of Mental Health (R01 MH069353-01).

References

Ahn, C., Overall, J. E., and Tonidandel, S. (2001). Sample size and power in repeated measurement analysis. *Computer Methods and Programs in Biomedicine* **64,** 121–124.

Bloch, D. A. (1986). Sample size requirements and the cost of a randomized clinical trial with repeated measurement. *Statistics in Medicine* **5,** 663–667.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* New York: Oxford University Press.

Donner, A. (1992). Sample size requirements for cluster randomization designs. *Statistics in Medicine* **11,** 743–750.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomised Trials in Health Research.* London: Arnold.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis.* New York: Wiley.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis.* New York: Wiley.

Hedeker, D., Gibbons, R. D., and Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics* **24,** 70–93.

Helms, R. W. (1992). Intentionally incomplete longitudinal designs: Methodology and comparison of some full span designs. *Statistics in Medicine* **11,** 1889–1913.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed-effects from restricted maximum likelihood. *Biometrics* **53,** 983–997.

Kirby, A. J., Galai, N., and Munoz, A. (1994). Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials* **15,** 165–172.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B* **36,** 1–37.

Liu, G. and Liang, K. Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53,** 937–947.

Lorr, M. and Klett, C. J. (1966). *Inpatient Multidimensional Psychiatric Scale: Manual.* Palo Alto, California: Consulting Psychologists Press.

Lui, K. J. and Cumberland, W. G. (1992). Sample size requirement for repeated measurements in continuous data. *Statistics in Medicine* **11,** 633–641.

Muller, K. E. and Barton, C. N. (1989). Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association* **84,** 549–555.

Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association* **87,** 1209–1226.

Overall, J. E. and Doyle, S. R. (1994). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials* **15,** 100–123.

Raudenbush, S. W. and Liu, X. F. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods* **5,** 199–213.

Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel Analysis.* London: Sage Publications.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Received May 2006. *Revised September* 2006.
Accepted November 2006.