The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 9

Sample Size Estimation for Repeated Measures Analysis in Randomized Clinical Trials with Missing Data

Kaifeng Lu, *Merck & Co.*
Xiaohui Luo, *Merck & Co.*
Pei-Yun Chen, *Merck & Co.*

Recommended Citation:
Lu, Kaifeng; Luo, Xiaohui; and Chen, Pei-Yun (2008) "Sample Size Estimation for Repeated Measures Analysis in Randomized Clinical Trials with Missing Data," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 9.
DOI: 10.2202/1557-4679.1098

Sample Size Estimation for Repeated Measures Analysis in Randomized Clinical Trials with Missing Data

Kaifeng Lu, Xiaohui Luo, and Pei-Yun Chen

Abstract

In designing longitudinal studies, researchers must determine the number of subjects to randomize based on the power to detect a clinically meaningful treatment difference and a proposed analysis plan. In this paper, we present formulas for sample size estimation and an assessment of statistical power for a two-treatment repeated measures design allowing for subject attrition. These formulas can be used for comparing two treatment groups across time in terms of linear contrasts. Subjects are assumed to drop out of the study at random so that the missing data do not alter the parameters of interest.

KEYWORDS: completers analysis, missing at random, subject attrition

Author Notes: The authors would like to thank an anonymous reviewer and the editor for their careful review and helpful comments, which greatly improved the presentation of the article.

1. INTRODUCTION

In clinical trials, subjects may drop out early before study completion for a variety of reasons. One approach to the resulting missing data problem is to carry forward the last observed value to the time point of interest. This is often known as the last observation-carried-forward (LOCF) approach. However, it is well-known that this approach is not guaranteed to produce "conservative" results (e.g., Verbeke and Molenberghs 2000; Mallinckrodt et al. 2001a, 2001b). Another approach is to only use data from subjects who completed the study (completers or complete cases). This is often referred to as the completers analysis. It is well-known that the completers may not be representative of the entire population if the missing data are not missing completely at random (MCAR), and even when the MCAR assumption is met, the completers analysis is often less efficient as a result of ignoring information from the observed data prior to drop-out (e.g., Liu and Gould 2002; Little and Rubin 2002; Schafer and Graham 2002).

Under the assumption of "missing-at-random" (MAR) missing data mechanism, repeated measures analysis can often provide valid statistical inference (e.g., Little and Rubin 2002; Gadbury et al. 2003; Mallinckrodt et al. 2003). It has been adopted as the primary analysis in place of the LOCF approach in some clinical programs (see, for example, Kelsey et al. 2004, Mallinckrodt et al. 2004).

Although there is rich literature regarding the sample size and power calculations for repeated measures designs (e.g. Rochon 1991; Overall and Doyle 1994; Rochon 1998; Patel and Rowe 1999; Liu et al. 2002; Jung and Ahn 2003), the sample size and power issues for repeated measures with missing data have not been studied extensively. Hedeker et al. (1999) presented formulas for estimating sample sizes for a comparison of two groups in terms of single degree-of-freedom contrasts of population means across study time points for longitudinal designs with attrition, but their formulas are only applicable to estimates based on complete cases at each time point instead of the maximum likelihood estimates from the repeated measures model.

In this paper, we derive formulas for sample size estimation based on the maximum likelihood estimates from repeated measures analysis. We focus on a comparison between two treatment groups in terms of mean difference at the last time point, but the approach can be easily extended to treatment comparisons across time. In Section 2, the notation and assumptions are described. In Section 3, we derive the sample size formulas and compare them to those based on the completers analysis. Results from simulation studies are presented in Section 4 to assess the sensitivity of the sample size formulas to the missing data assumption. In Section 5, we apply the sample size formulas to an example in calcium

supplementation study, followed by a discussion and concluding remarks in Section 6.

2. NOTATION AND ASSUMPTIONS

The sample size formulas proposed in this article are based on a family of repeated measures analyses often referred to as "mixed model for repeated measures" (MMRM) (e.g., Mallinckrodt et al. 2001a; Mallinckrodt et al. 2003; Lane 2008). The MMRM method is likelihood-based and includes an unstructured modeling of time and the within-subject error correlation structure. Specifically, the model includes factors of treatment, time point as a categorical variable, and interaction of treatment by time, and an unstructured covariance matrix is used to model the correlation among the repeated measurements.

Suppose there are $n_{a1} \geq n_{a2} \geq \ldots \geq n_{aJ}$ subjects in treatment group a (=1, 2) at time points $t_1 < t_2 < \ldots < t_J$. Without loss of generality, assume the first n_{11} subjects are from treatment group 1 and the remaining n_{21} subjects are from treatment group 2. Denote $n_{1,J+1} = n_{2,J+1} = 0$, the observed data can be represented as $\{(a_i, y_i), i = 1, \ldots, n_{11} + n_{21}\}$, where a_i is the treatment indicator with $a_i = 1$ for the first n_{11} subjects and $a_i = 2$ for the remaining n_{21} subjects, respectively; and y_i denotes the repeated measurements observed on subject i so that $y_i = (y_{i1}, \ldots, y_{ij})'$ for a subject who dropped out right after the jth time point. For simplicity we have assumed a monotone missingness and considered the situation where treatment is the only variable in the model (besides the categorical time variable). It is straightforward to generalize the sample size formulas to repeated measures models that include other baseline variables (e.g., baseline measurement for change from baseline analysis) as we will show in the simulation and discussion sections.

Suppose responses from different subjects are independent and responses from the same subject are distributed as multivariate normal with means, $E(y_{ij} \mid a_i = a) = \mu_{aj}$ (a = 1, 2), and variances and covariances, $\text{cov}(y_{ij}, y_{ik} \mid a_i = a) = \sigma_{ajk}$, which are assumed to be known or can be estimated consistently. In matrix notation, the mean vector and the variance-covariance matrix for the repeated measures in treatment group a are $\mu_a = (\mu_{a1}, \ldots, \mu_{aJ})$ and $\Sigma_a = (\sigma_{ajk} : j, k = 1, \ldots, J)$, respectively. In general, it is often assumed that $\Sigma_1 = \Sigma_2 = \Sigma = (\sigma_{jk} : j, k = 1, \ldots, J)$.

To facilitate the presentation of the sample size formulas, we also define the following quantities. Let $r_{aj} = n_{aj}/n_{a1}$ denote the retention rate at the jth time point for treatment group a, hence the corresponding attrition rate $a_{aj} = 1 - r_{aj}$, also let $\lambda = n_{11}/n_{21}$ denote the allocation ratio of sample sizes at the first time point. Furthermore, let $\sigma_j = \sigma_{jj}^{1/2}$ denote the standard deviation of y_{ij}, and $\rho_{jk} = \sigma_{jk}/(\sigma_j\sigma_k)$

denote the correlation between y_{ij} and y_{ik}. In matrix notation, for treatment group a, we will use $r_a = (r_{aj} : j = 1; \ldots, J)$ and $R = (\rho_{jk} : j, k = 1, \ldots, J)$ to describe the retention and the correlation structure between repeated measures, respectively. Common choices include exponential or linear retention, compound symmetry (CS) or first-order autoregressive (AR(1)) correlation structure.

Suppose we are interested in estimating sample size based on the following contrast of the treatment means across the J time points:

$$\theta_c = \sum_{j=1}^{J} c_j (\mu_{1j} - \mu_{2j}),$$

The choice of the contrast depends on the analysis that is planned for the study. For example, if we are interested in the comparison of treatment means at the last time point, then $c_j = I(j = J)$ $(j = 1, \ldots, J)$. In some circumstances, we might be interested in testing the treatment difference in time-weighted average mean changes from baseline, in which case, $c_j \propto (t_j - t_{j-1})$, with $t_0 = 0$. Similarly, taking $c_j \propto (j - (J + 1)/2)$ provides a test of treatment by linear time interaction, or equivalently, the difference in the slopes of the mean responses between the two treatment groups.

In this article, we will focus on the treatment difference at the last time point, so that

$$\theta = \mu_{1J} - \mu_{2J}.$$

Let $\hat{\mu}_a$ be an estimate of μ_a $(a = 1, 2)$, and take $c = (0, \ldots, 0, 1)'$, then $\mathrm{var}(\hat{\mu}_{aJ}) = c' \mathrm{var}(\hat{\mu}_a) c$, and $\mathrm{var}(\hat{\theta}) = \mathrm{var}(\hat{\mu}_{1J}) + \mathrm{var}(\hat{\mu}_{2J})$.

To test $H_0 : \theta = 0$ versus $H_1 : \theta = \pm\delta$ with power $1 - \beta$ at significance level α, the sample size estimation based on a two-sample z-test is given by

$$| E_{H_1}(\hat{\theta}) | / \sqrt{\mathrm{var}_{H_1}(\hat{\theta})} = z_{1-\alpha/2} + z_{1-\beta}, \qquad (1)$$

where z_γ denotes the γ-percentile of the standard normal distribution. In general, $E_{\theta=\delta}(\hat{\theta}) = \delta$ provided that the estimator is consistent.

In addition, we assume that the process by which subjects dropped out of the study only depends on the data observed prior to drop-out, i.e., the missing data are missing at random, so that the MMRM model which ignores the missing data process can provide valid inference. For example, in diabetes studies, patients not meeting specific glycemic goals after prespecified time points will have rescue therapy initiated, resulting in missing data on glycemic endpoints for the primary analysis which excludes data after rescue. Patients with higher baseline hemoglobin A1c (HbA_{1c}) values are more likely to be rescued than patients with lower baseline HbA_{1c} values, and patients randomized to the placebo group are more likely to be rescued than patients randomized to the active treatment group. The missing data as a result of protocol-specified rescue procedures can be considered as being missing at random.

3. SAMPLE SIZE FORMULAS

To derive the sample size formulas, we first establish a model to estimate the treatment means, from which we obtain the variances for estimates of the treatment means; then we find the variance for estimate of the treatment difference at the last time point and use (1) to solve for the required sample size.

3.1 Repeated measures analysis

3.1.1 Information decomposition. Let $\hat{\mu}_a$ be the maximum likelihood estimate (MLE) of μ_a based on the MMRM model, then the asymptotic variance of $\hat{\mu}_a$ is given by I_a^{-1}, where

$$I_a = \sum_{j=1}^{J} (n_{aj} - n_{a,j+1}) \begin{pmatrix} \Sigma_j^{-1} & 0_{j\times(J-j)} \\ 0_{(J-j)\times j} & 0_{(J-j)\times(J-j)} \end{pmatrix} \tag{2}$$

is the total information about μ_a, Σ_j is the leading $j \times j$ sub-matrix of Σ, and 0_{kl} denotes the $k \times l$ matrix of 0's (a null matrix when $l = 0$). The jth term in the summation represents the information contribution from subjects who had and only had the first j measurements, that is, subjects who dropped out after time point t_j , also recall $n_{a,J+1} = 0$. The proof of (2) is given in the Appendix.

3.1.2 Inflation factors. If we let R_j denote the leading $j \times j$ sub-matrix of the correlation matrix R, and denote

$$I_a^* = \sum_{j=1}^{J} (r_{aj} - r_{a,j+1}) \begin{pmatrix} R_j^{-1} & 0_{j\times(J-j)} \\ 0_{(J-j)\times j} & 0_{(J-j)\times(J-j)} \end{pmatrix}, \tag{3}$$

then by (2), we have $\text{var}(\hat{\mu}_{aJ}) = \varphi_a(\sigma_J^2 / n_{a1})$, where

$$\varphi_a = [I_a^{*-1}]_{JJ} \tag{4}$$

is the (J, J) element of I_a^{*-1} and is defined as the variance (or sample size) inflation factor with respect to the estimation of μ_{aJ} for treatment group a. The rationale for this definition is as follows. The variance of $\hat{\mu}_{aJ}$ in the presence of missing data is φ_a times that had there been no missing data. Alternatively, by (1), the sample size required to detect the same deviation of μ_{aJ} from zero in the presence of missing data is φ_a times that required had there been no missing data. To indicate the dependence of the inflation factors on the retention rates and the correlation structure, we write $\varphi_a = \varphi_a(J, r_a, R)$.

3.1.3 Effective sample sizes. The effective sample size is defined as the number of completers that gives the complete cases estimate of the treatment mean at the last time point the same precision as that of the MLE. It follows that the effective sample size at the last time point for treatment group *a* is

$$n_{aJ}^{*} = n_{a1} / \varphi_a . \tag{5}$$

Alternatively, the effective sample size can be interpreted as the number of patients needed in the absence of missing data to give the MLE of the treatment mean at the last time point the same precision as that of the MLE in the presence of missing data.

If we use $A \geq B$ to denote that $A - B$ is a symmetric and semi-positive definite matrix for any two symmetric and semi-positive definite matrices A and B, then by (3), $I_a^{*} \geq r_{aJ} R^{-1}$, hence $I_a^{*-1} \leq r_{aJ}^{-1} R$, so that by (4), $\varphi_a \leq r_{aJ}^{-1}$, and by (5), $n_{aJ}^{*} \geq n_{aJ}$. In other words, the repeated measures analysis provides more information at the last time point than the completers analysis, resulting in a smaller inflation factor and a larger effective sample size at the last time point for each treatment group.

3.1.4 Sample size formulas. The asymptotic variance of the estimate of treatment difference at the last time point based on the repeated measures analysis is

$$\mathrm{var}(\hat{\theta}) = (\varphi_1 + \lambda\varphi_2)(\sigma_J^2 / n_{11}) .$$

If we use a two-sample z-test to test $H_0 : \mu_{1J} = \mu_{2J}$ versus $H_1 : \mu_{1J} - \mu_{2J} = \pm\delta_J$, then by (1), to achieve power $1 - \beta$ at significance level α, we need the following sample size for the first treatment group,

$$n_{11} = (\varphi_1 + \lambda\varphi_2)(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_J^2 / \delta_J^2 . \tag{6}$$

Alternatively, we can compute n_{11} by making use of the effective allocation ratio λ^{*}, which is defined as the ratio of the effective sample sizes for the two treatment groups. It follows from (5) that

$$\lambda^{*} = n_{1J}^{*} / n_{2J}^{*} = \lambda\varphi_2 / \varphi_1 .$$

To compute n_{11}, first use λ^{*} as the effective allocation ratio at the last time point in

$$n_{1J}^{*} = (1 + \lambda^{*})(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_J^2 / \delta_J^2$$

to obtain n_{1J}^{*}, then multiply n_{1J}^{*} by the inflation factor φ_1.

3.1.5 Optimal allocation ratio. The optimal allocation ratio yields the smallest total number of subjects required to achieve the same power. Given the retention rates and the correlation structure for the repeated measures in each treatment group, the optimal allocation ratio, λ^{OPT}, can be obtained from the sample size formulas. In fact, by (6) and the Cauchy-Schwartz inequality, the total number of subjects, $n_{11} + n_{21} = (1 + \lambda^{-1})(\varphi_1 + \lambda\varphi_2)(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_J^2 / \delta_J^2$, is minimized at λ^{OPT}

$= (\varphi_1/\varphi_2)^{½}$. In particular, the optimal allocation ratio equals 1 if the retention rates and the correlation structure are the same for the two treatment groups. Other allocation strategies include: "proportional to inflation factor" allocation ($\lambda = \varphi_1/\varphi_2$), equal allocation ($\lambda = 1$), and "more patients on drug" allocation (say, $\lambda = 2$).

3.1.6 Unknown covariance parameters. When the variance-covariance parameters are unknown, sample size estimation can be based on two-sample *t*-tests. Let $T(\nu;\xi)$ denote a random variable from a noncentral *t*-distribution with ν degrees of freedom and non-centrality parameter ξ, and $t_{\nu,1-\alpha/2}$ denote the $(1-\alpha/2)$-percentile of the central *t*-distribution with ν degrees of freedom, then the sample size equation is

$$P\{T(\nu;|\, E_{H_1}(\hat{\theta})\,|\,/\sqrt{\mathrm{var}_{H_1}(\hat{\theta})}) \geq t_{\nu,1-\alpha/2}\} = 1-\beta\,. \qquad (7)$$

There are numerous methods for computing the degrees of freedom ν in (7). For example, the SAS Mixed procedure offers containment method, between-within method, residual degrees-of-freedom method, Satterthwaite approximation, and Kenward & Roger approximation. In particular, if the unstructured correlation structure is used, then by default, the Mixed procedure uses the degrees-of-freedom $\nu_1 = n_{11} + n_{21} - 2 = n_{11}(1 + 1/\lambda) - 2$. Alternatively, we can use the effective sample sizes (or the effective allocation ratio) for sample size estimation, in which case, the associated degrees of freedom for the *t*-test is $\nu_2 = n^*_{1J} + n^*_{2J} - 2 = n^*_{1J}(1+1/\lambda^*) - 2$. We refer to the above *t*-tests as the 1-step *t*-test and the 2-step *t*-test based on the effective allocation ratio λ^*, respectively. Table 1 compares the sample size estimation based on the *z*-test and the two *t*-tests. It can be seen from the table that the *t*-tests are more conservative than the *z*-test; the 2-step *t*-test is more conservative than the 1-step *t*-test. We recommend using the 2-step *t*-test in the presence of missing data because it makes use of the effective sample sizes instead of the more optimistic sample sizes prior to patient attrition.

From Table 1, we also notice the effect of allocation ratio on the sample size estimation. In particular, the optimal allocation ratio, $\lambda^{\mathrm{OPT}} = (\varphi_1/\varphi_2)^{½}$, has the potential to reduce the required total sample size compared with other allocation ratio strategies.

DOI: 10.2202/1557-4679.1098

Table 1: Estimated total sample size, $n_{11} + n_{21}$, using a z-test and two t-tests for treatment difference at the last time point with significance level $\alpha = 0.05$, power $1 - \beta = 0.90$ and effect size $\delta_J/\sigma_J = 0.5$.

				t-test	
φ_1	φ_2	λ	z-test	1-step	2-step
1	1	1	168	170	170
		2	189	192	192
	2	$(1/2)^{1/2}$	245	247	248
		1/2	252	255	255
		1	252	254	256
		2	315	318	318
2	2	1	336	338	340
		2	378	381	384

3.1.7 AR(1) correlation structure. The inflation factor defined by (3) and (4) requires a series of matrix inversion. It can be shown that a simple expression exists for the autoregressive correlation structure. Specifically, assume that

$$\rho_{jk} = \rho^{|t_j - t_k|},$$

then the inflation factor for treatment group a is given by

$$\varphi_a = \frac{1}{r_{aJ}} - \sum_{j=1}^{J} \rho^{2|t_J - t_j|} \left(\frac{1}{r_{a,j+1}} - \frac{1}{r_{aj}} \right). \tag{8}$$

Table 2 displays the inflation factors for various combinations of number of time points, attrition rate, and the AR(1) correlation parameter, where we have assumed an exponential attrition. To facilitate comparison between different number of time points, we have used the attrition rate at the last time point, $a_J = 1 - r_J$, so that the attrition rate between adjacent time points is $1 - (1 - a_J)^{1/(J-1)}$. It can be seen from the table that as the attrition rate increases, the inflation factor also increases, indicating that a larger sample size is needed. For comparison between different numbers of time points, we have also used the correlation between the first and last time points, ρ_{1J}, so that the correlation between adjacent time points is $\rho_{1J}^{1/(J-1)}$. It can be seen from the table that the inflation factor decreases as the correlation increases, reflecting the increased information about the measurement at the last time point provided by the measurements from the previous time points. In addition, by adding more intermittent time points, we can also reduce the inflation factor. However, the reduction becomes negligible once the number of time points exceeds a certain level. In fact, given the overall attrition rate r_J and the correlation ρ_{1J}, the inflation factor tends to

$(1/r_{1J})\{1-(1-r_J\rho_{1J}^2)\log r_J/\log(r_J\rho_{1J}^2)\}$ as J goes to infinity. This limit lies between 1 and $1/r_J$.

Table 2: Inflation factor, $\varphi(J, r, R)$, for the repeated measures analysis for testing treatment difference at the last time point with J time points, attrition rate a_J, and the correlation ρ_{1J} for an AR(1) correlation structure.

		ρ_{1J}					
J	a_J	0	0.1	0.3	0.5	0.7	0.9
2	0.1	1.111	1.110	1.101	1.083	1.057	1.021
	0.2	1.250	1.247	1.227	1.188	1.128	1.047
	0.3	1.429	1.424	1.390	1.321	1.219	1.081
	0.4	1.667	1.660	1.607	1.500	1.340	1.127
4	0.1	1.111	1.101	1.083	1.063	1.040	1.014
	0.2	1.250	1.226	1.186	1.141	1.090	1.032
	0.3	1.429	1.386	1.317	1.240	1.152	1.053
	0.4	1.667	1.598	1.489	1.369	1.233	1.082

3.1.8 Compound symmetry correlation structure. Simple expressions also exist for the compound symmetry correlation structure. Specifically, assume that

$$\rho_{jk} = \rho,$$

then the inflation factor for treatment group a is given by

$$\varphi_a = (1-\rho)\left[\sum_{j=1}^{J-1}\frac{1}{r_{aj}}\frac{\rho^2}{(1+(j-2)\rho)(1+(j-1)\rho)}+\frac{1}{r_{aJ}}\frac{(1+(J-1)\rho)}{(1+(J-2)\rho)}\right]. \qquad (9)$$

The inflation factors for other correlation structures can be computed from (3) and (4).

3.2 Completers Analysis

For the completers analysis, the inflation factor for treatment group a is $1/r_{aJ}$. Therefore, assuming the same treatment effect and the standard deviation at the last time point, the percent reduction in required sample size by using the repeated measures analysis instead of the completers analysis for testing the treatment difference at the last time point is

$$p(J, r, R) = 100\ \{1 - r_J\,\varphi(J, r, R)\}.$$

Using the same specifications as in Table 2, Table 3 shows the percent reduction in sample size of the repeated measures analysis with respect to the completers analysis. It can be seen that the repeated measures analysis reduces to the completers analysis when there is no correlation among the repeated measures, in which case, no reduction in sample size can be achieved. However,

as the attrition rate increases, or as the number of time points increases, or as the correlation among repeated measures increases, the percent reduction in sample size increases. This demonstrates the ability of the repeated measures analysis to make efficient use of the information contained in the previous time points toward the estimation of treatment effect at the last time point.

Table 3: Percent reduction in estimated sample size, $p(J, r, R)$, for the repeated measures analysis with respect to the completers analysis for testing treatment difference at the last time point with J time points, attrition rate a_J, and the correlation ρ_{1J} for an AR(1) correlation structure.

		ρ_{1J}					
J	a_J	0	0.1	0.3	0.5	0.7	0.9
2	0.1	0	0.1	0.9	2.5	4.9	8.1
	0.2	0	0.2	1.8	5.0	9.8	16.2
	0.3	0	0.3	2.7	7.5	14.7	24.3
	0.4	0	0.4	3.6	10.0	19.6	32.4
4	0.1	0	0.9	2.5	4.3	6.4	8.7
	0.2	0	1.9	5.1	8.7	12.8	17.5
	0.3	0	3.0	7.8	13.2	19.4	26.3
	0.4	0	4.1	10.7	17.9	26.0	35.1

4. SIMULATION STUDIES

The results in Section 3 were derived by assuming a known variance-covariance matrix, which is usually unknown and needs to be estimated in practice; besides, the number of subjects at each time point is random but assumed to be fixed and known. To evaluate the performance of the formulas in the presence of violation of these assumptions, the following simulation studies were conducted.

Assume a clinical trial with two arms (treatment and placebo) and four post-baseline time points where the last time point was of primary interest; the endpoint had an AR(1) correlation structure and the standard deviations at the four time points were 0.7, 0.8, 0.9 and 1, respectively. The mean vector of the placebo group was assumed to be (0, 0, 0, 0). Given the mean vector of the treatment group, attrition rate between any two consecutive time points (assuming MCAR missingness and same attrition rates in both arms), and the targeted power, the sample size was determined using (7) and (8) at significance level 0.05; 1000 data sets were simulated with the determined sample size in each arm, and the power from the repeated measures ANOVA model (MMRM with treatment, time and treatment by time interaction) was estimated.

A similar simulation was done for the repeated measures ANCOVA model with baseline value as the covariate (MMRM with changes from baseline as the responses, and treatment, baseline value, time and treatment by time interaction as the model terms). The same variance-covariance structure mentioned above was adopted for the conditional distribution of changes from baseline given the baseline value. The baseline value followed $N(4, 5^2)$ distribution, and a regression coefficient of 0.1 was assumed for the baseline value in the repeated measures model. The results are summarized in Table 4 where the first part and the second part correspond to the repeated measures ANOVA model and the repeated measures ANCOVA model, respectively.

Table 4: Estimated power at significance level 0.05 under the MCAR missingness. Results based on 1000 replications; AR(1) correlation structure and exponential attrition assumed; four time points, with standard deviations (0.7, 0.8, 0.9, 1), placebo means (0, 0, 0, 0). The standard deviations were conditional on the baseline value for the repeated measures ANCOVA.

ρ of AR(1)	Attrition rate	Treatment means	Inflation factor	Sample size	Power (%) achieved/targeted
		Repeated Measures ANOVA			
0.7	0.10	(0.2, 0.15, 0.3, 0.5)	1.26	81	79.4 / 80
0.5	0.15	(0.6, 0.8, 0.7, 1.0)	1.56	36	91.4 / 90
0.3	0.10	(0.6, 0.8, 0.7, 1.2)	1.36	22	92.0 / 90
		Repeated Measures ANCOVA			
0.9	0.15	(0.5, 0.7, 0.8, 1.0)	1.20	21	82.7 / 80
0.8	0.05	(0.5, 0.7, 0.6, 0.8)	1.10	38	90.8 / 90
0.6	0.20	(0.2, 0.7, 1.0, 1.2)	1.77	29	91.9 / 90

In Table 4, ρ denotes the correlation between any two adjacent time points. For the repeated measures ANCOVA model, the means in the table were unconditional means while the standard deviations were conditional standard deviations. In the first case of the repeated measures ANOVA model, we assumed $\rho = 0.7$ in the AR(1) structure, treatment means were 0.2, 0.15, 0.3 and 0.5 at the four time points, respectively, and the attrition rate was 0.10 between any two adjacent time points. If there were no missing data, 64 subjects per arm would be required to detect a between-group difference of 0.5 at the last time point with 80% power at significance level 0.05. From (8) the inflation factor was 1.26; hence the required sample size for the repeated measures analysis was $81 = 64 \times 1.26$, which yielded a power of 79.4% in the simulation. Similarly, Table 4

confirms that (8) provided the desired power for the other scenarios considered above.

A simulation was also conducted for the MAR missingness. Again, we assumed two treatment arms and four time points in the clinical trials (for simplicity, no baseline value was involved) and AR(1) correlation structure with ρ = 0.6 and standard deviations of 0.7, 0.8, 0.9 and 1 at the four time points, respectively. Assume the mean vectors were (0, 0, 0, 0) and (0.3, 0.5, 0.8, 0.9) for the placebo and treatment groups, respectively. Suppose a patient would withdraw from the study whenever his/her measurement fell below −0.5 according to the protocol (MAR since the current measurement was still available). The retention rates were 76%, 63% and 52%, respectively, at the last three time points in the placebo group; while the retention rates were 87%, 81% and 78% for the treatment group. From (8), the inflation factors were 1.75 and 1.25 for the placebo and treatment groups, respectively. By assuming different sample size ratios of the two arms, four trials were designed to detect a between-group difference of 0.9 at the last time point with power 90% at significance level 0.05. The simulation results from 1000 data sets are summarized in Table 5. The first column is the ratio of the sample size of the treatment group to that of the placebo group; the next two columns are the sample size determined by (8); and the last column gives the achieved power. The first case was designed as the most efficient trial (requiring minimum total number of subjects) by setting the sample size proportional to the square root of inflation factor; 81 subjects were required. The second case was straightforward: 27 patients per arm were required when there was no attrition and the number of patients required in the presence of attrition was obtained by multiplying 27 by the inflation factor. The third case was designed as equal allocation of subjects to the two arms; 84 subjects in total provided 90.3% power. In the last case, there were as twice as many patients assigned to the treatment group as to the placebo group in order to obtain the safety profile for the drug; the cost was that we had to recruit 22% more patients compared with the most efficient case.

The standard errors for the achieved power ranged from 0.9% for ~90% achieved power to 1.3% for ~80% achieved power.

5. AN EXAMPLE

A trial studying the effect of calcium supplementation on bone acquisition in adolescent white girls was reported in Table 5.4.1 of Vonesh and Chinchilli (1997). One hundred twelve healthy adolescent girls were randomized to a daily calcium supplement (500 mg calcium citrate malate) or placebo with 1:1 ratio at Visit 1 (baseline), and then followed up every six months within the next 2 years.

The endpoint was change from baseline at Visit 5 in total body bone mineral density (TBBMD, g/cm^2). A monotone missingness was observed and the number of patients (as well as the retention percentage with respect to Visit 2) on each treatment arm across time is summarized in Table 6. The correlation matrix estimated from the observed data is provided in Table 7.

Table 5: Patient withdrawal due to measurement below threshold (MAR); 90% power targeted at significance level 0.05. Results based on 1000 replications; AR(1) correlation structure with $\rho = 0.6$; four time points; retention rates (100%, 76%, 63%, 52%) and (100%, 87%, 81%, 78%) in the placebo and treatment groups, respectively; standard deviations (0.7, 0.8, 0.9, 1); placebo means (0, 0, 0, 0), treatment means (0.3, 0.5, 0.8, 0.9).

Allocation ratio	Sample size determined		Achieved
(Treatment/Placebo)	Treatment	Placebo	power (%)
$(1.25/1.75)^{1/2}$	37	44	90.4
1.25/1.75	34	48	91.0
1	42	42	90.3
2	66	33	90.4

Table 6: Number of patients (retention rate (%) with respect to Visit 2) on each treatment arm across time for the calcium supplementation data from Vonesh and Chinchilli (1997). Visit 1 was the randomization visit and Visit 2 was the first postrandomization visit.

	Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
Calcium	55	52 (100)	48 (92)	46 (88)	44 (85)
Placebo	57	53 (100)	51 (96)	48 (91)	47 (89)

Table 7: Correlation coefficients of TBBMD change from baseline for the calcium supplementation data from Vonesh and Chinchilli (1997).

	Visit 2	Visit 3	Visit 4
Visit 3	0.75		
Visit 4	0.69	0.87	
Visit 5	0.65	0.77	0.86

From (4), the Visit 5 inflation factors with respect to Visit 2 were 1.08 and 1.06 for the calcium group and the placebo group, respectively. Taking into account of the subject attrition between Visit 1 and Visit 2, the inflation factors

with respect to Visit 1 were 1.15 and 1.14 for the calcium group and the placebo group, respectively that is about 15% more subjects were required to compensate the attrition. The small inflation factors were due to the high retention rates and strong correlations as seen in Table 6 and Table 7. Had the completers analysis been employed, 23% more subjects would be needed to offset the drop-out.

6. DISCUSSION

Based on the simulation results, we can draw the following conclusions. The 2-step *t*-test based on the effective sample sizes (or the effective allocation ratio) are adequate to account for the uncertainty associated with the estimation of variance parameters. In randomized clinical trials, covariates other than treatment are generally balanced across treatment groups. If the conditional variances given the treatment and the covariates in the model are used, ignoring covariates other than treatment in sample size determination is still appropriate as shown in the simulation studies.

We have focused on the comparison between treatment groups at the last time point. For other types of treatment comparisons, such as the time-weighted average mean response over time or the treatment by linear time interaction effect, the derivation of sample size formulas is straightforward as the variance of $\hat{\theta}_c = c'(\hat{\mu}_1 - \hat{\mu}_2)$ is given by c' $(I_1^{-1} + I_2^{-1})$ c for the appropriate vector of coefficients c.

The sample size formulas derived in Section 3.1 assume a known correlation structure. In practice, there is often only limited information at the design stage of clinical trials. For many clinical endpoints collected over time, e.g., body weight for obesity studies, HbA_{1c} for diabetes studies, blood pressure (systolic or diastolic) for hypertension studies, the correlations between repeated measurements are positive and decreasing over increasing time gaps. Under these assumptions, the inflation factors based on the compound symmetry correlation structure with correlation estimated from that between baseline and the last time point are easy to calculate (see equation (9)) and provide a less conservative sample size requirement than the complete-case analysis.

Although the sample size formulas do not impose an upper bound on the amount of missing data, the assumptions for the missing data mechanism and the analysis model become more questionable when more data points are missing. Therefore, it is critical to minimize patient attrition in clinical trials.

A SAS macro which implements the power/sample size calculations for the MMRM model is available upon request.

APPENDIX A

A.1 Proof of Equation (4)

For the repeated measures analysis (MMRM), the maximum likelihood estimate (MLE) is the same as the generalized least squares (GLS) estimate. To express the repeated measures analysis in the linear model notation, let I_p denote the $p \times p$ identity matrix, $0_{p\times q}$ denote the $p \times q$ matrix of 0's (a null matrix when $q = 0$), 0_p denote the p vector of 0's, and similarly define $1_{p\times q}$ and 1_p. For simplicity, also let col(·) denote the vertical concatenation operation, diag(·) denote the operation that creates a diagonal matrix, and ⊗ denote the direct (Kronecker) product operator.

Define

$$X_{1j} = 1_{n_{1j}-n_{1,j+1}} \otimes (I_j, 0_{j\times(J-j)}, 0_{j\times J}),$$
$$X_{2j} = 1_{n_{2j}-n_{2,j+1}} \otimes (0_{j\times J}, I_j, 0_{j\times(J-j)}),$$
$$y_{1j} = \mathrm{col}\{y_{n_{1,j+1}+1}, \ldots, y_{n_{1j}}\},$$
$$y_{2j} = \mathrm{col}\{y_{n_{11}+n_{21,j+1}+1}, \ldots, y_{n_{11}+n_{21j}}\},$$

and

$$X = \mathrm{col}\{X_{1J}, \ldots, X_{11}, X_{2J}, \ldots, X_{21}\},$$
$$y = \mathrm{col}\{y_{1J}, \ldots, y_{11}, y_{2J}, \ldots, y_{21}\},$$
$$\beta = \mathrm{col}\{\mu_1, \mu_2\},$$

then the repeated measures model can be written as $y = X\beta + \varepsilon$, where ε is normally distributed with mean $0_{n_{11}+n_{21}}$ and variance-covariance matrix given by

$$V = \mathrm{diag}\{\Sigma_{1J}, \ldots, \Sigma_{11}, \Sigma_{2J}, \ldots, \Sigma_{21}\},$$

where

$$\Sigma_{aj} = I_{n_{aj}-n_{a,j+1}} \otimes \Sigma_j, \quad a = 1, 2,$$

and Σ_j is the leading $j \times j$ sub-matrix of $\Sigma = (\sigma_{st}, s, t = 1, \ldots, J)$. Therefore, the GLS estimate of β can be written as

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

with variance given by

$$\mathrm{var}(\hat{\beta}) = (X'V^{-1}X)^{-1} = \mathrm{diag}\{I_1^{-1}, I_2^{-1}\},$$

where I_a is given in (2). Hence $\hat{\beta} = (\hat{\mu}_1', \hat{\mu}_2')'$, $\hat{\mu}_a \sim N(\mu_a, I_a^{-1})$, and $\hat{\mu}_1$ is independent of $\hat{\mu}_2$, which leads to (3) and (4).

Note that the retention rates, r_{aj}, have been fixed and the covariance matrix, Σ, has been assumed known in the above derivation. When the process by which subjects drop out of the study only depends on the data observed prior to drop-out, i.e., the missing data are missing at random, the covariance parameters can be

consistently estimated via the restricted maximum likelihood and r_{aj} can be viewed as the average retention rates. This completes the proof.

REFERENCES

Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. *Obesity Reviews* 2003, 4: 175-184.

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

Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine* 2003, 22: 1305-1315.

Kelsey DK, Sumner CR, Casat CD, Coury DL, Quintana H, Saylor KE, Sutton VK, Gonzales J, Malcolm SK, Schuh KJ, Allen AJ. Once-daily atomoxetine treatment for children with attention-deficit/hyperactivity disorder, including an assessment of evening and morning behavior: a double-blind, placebo-controlled trial. *Pediatrics* 2004, 114 (1): e1-e8.

Lane P. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharmaceutical Statistics* 2008, 7(2): 93-106.

Little RJA, Rubin DB. *Statistical Analysis With Missing Data* 2002, Wiley, New York, second edition.

Liu A, Boyett JM, Xiong X. Sample size calculation for planning group sequential trials. *Statistics in Medicine* 2000, 19: 205-220.

Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics* 2002, 12: 207-226.

Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics* 2001, 11: 9-21.

—. Type I error rates from mixed effects model repeated measures compared with fixed effects model ANOVA with missing values imputed via LOCF. *Drug Information Journal* 2001, 35: 1215-1225.

Mallinckrodt CH, Raskin J, Wohlreich MM, Watkin JG, Detke MJ. The efficacy of duloxetine: a comprehensive summary of results from MMRM and LOCF_ANCOVA in eight clinical trials. *BMC Psychiatry* 2004, 4-26.

Mallinckrodt CH, Sanger TM, Dubé S, DeBrota DJ, Molenberghs G, Carroll RJ, Potter WZ, Tollefson GD. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry* 2003, 53: 754-760.

Overall J, Doyle S. Estimating sample size for repeated measurement designs. *Controlled Clinical Trials* 1994, 15: 100-123.

Patel HI, Rowe E. Sample size for comparing linear growth curves. *Journal of Biopharmaceutical Statistics* 1999, 9: 339-350.

Rochon J. Sample size calculations for two-group repeated-measures experiments. *Biometrics* 1991, 47: 1383-1398.

—. Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine* 1998, 17: 1643-1658.

Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002, 7(2): 147-177.

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

Vonesh EF, Chinchilli VM. *Linear and Nonlinear Models for the Analysis of Repeated Measurements* 1997, Marcel Dekker Inc, New York.