Published in final edited form as:

Stat Methods Med Res. 2015 December ; 24(6): 1009–1029. doi:10.1177/0962280212437452.

A Comparison of Power Analysis Methods for Evaluating Effects of a Predictor on Slopes in Longitudinal Designs with Missing Data

Cuiling Wang, Charles B. Hall, and Mimi Kim

Department of Epidemiology and Population Health, Albert Einstein College of Medicine of Yeshiva University, 1300 Morris Park Ave, NY 10461, U.S.A.

Abstract

In many longitudinal studies, evaluating the effect of a binary or continuous predictor variable on the rate of change of the outcome, i.e., slope, is often of primary interest. Sample size determination of these studies, however, is complicated by the expectation that missing data will occur due to missed visits, early drop out and staggered entry. Despite the availability of methods for assessing power in longitudinal studies with missing data, the impact on power of the magnitude and distribution of missing data in the study population remain poorly understood. As a result, simple but erroneous alterations of the sample size formulae for complete/balanced data are commonly applied. These "naive" approaches include the average sum of squares (ASQ) and average number of subjects (ANS) methods. The goal of this paper is to explore in greater detail the effect of missing data on study power and compare the performance of naive sample size methods to a correct maximum likelihood based method using both mathematical and simulation based approaches. Two different longitudinal aging studies are used to illustrate the methods.

1. Introduction

The primary goal in many longitudinal studies is to evaluate the effect of a binary or continuous predictor variable on the rate of change in the outcome. For example, in a proposed randomized clinical trial of the effect of cognitive remediation (CR) on mobility in sedentary seniors, the objective was to estimate the effect of the intervention on the change of normal gait velocity. Another example is from the Einstein Aging Study, in which interest centered on the effect of cerebral vascular function measured at baseline, a continuous variable, on the rate of cognitive decline.

When there are no missing data, simple and widely used methods are available (e.g. [1]) for computing the power and sample size of such studies. In reality, however, drop out, missed visits and staggered entry result in missing data. In the above CR trial, 20% drop out is expected at the post-intervention visit. Missing data is also a concern in the Einstein Aging Study since subjects must undergo five annual cognitive assessments.

Address for correspondence: Cuiling Wang, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461. cuiling.wang@einstein.yu.edu..

Conflict of interest: None declared.

Approaches for assessing the power of longitudinal and cluster designs have been extensively investigated for both balanced designs (e.g., [2-17]) and designs which are unbalanced due to missing data (e.g., [18-31]). Specifically, methods for unbalanced clustered designs have been proposed for between group comparisons of overall means [18-19], means at a specific time point during follow-up [20-21], slopes [22-26], and other aspects of longitudinal data which may include testing of slopes as a special case [27-31]. Both maximum likelihood estimate (MLE) and generalized estimating equation (GEE, [32]) methods have been considered in these approaches.

In spite of the availability of these methods, it remains unclear how the extent and distribution of missing data actually impact the power in longitudinal studies. Consequently, simple and naive alterations of conventional approaches for the complete data case are often used in practice. We consider two such methods: the average sum of squares (ASQ) and average number of subjects (ANS) methods, and evaluate in detail how they perform relative to a correct MLE based method based on the linear mixed effects models (LME) [33].

The remainder of this paper is organized as follows. In Section 2, we describe a correct MLE based method, evaluate the impact of missing data on the power estimated with this approach, and consider how the ASQ and ANS methods deviate from the MLE based method. Simulation studies to compare the performance of the different approaches are described in Section 3. In Section 4 the methods are applied to two different studies in the elderly. Finally, we conclude the paper with a discussion in Section 5.

2. Method

2.1 Background and Notation

Consider a longitudinal study of m independent subjects, indexed by i , with n planned visits, indexed by j . Here n is considered fixed so that the asymptotic property of the estimators are applicable for large m . For subject i , $i = 1, \dots, m$, let Z_i denote the time independent predictor, Y_{ij} denote the j^{th} outcome measure of subject i , and t_{ij} denote the time of the j th measurement, $j = 1, \dots, n$. For complete data, $\mathbf{Y}_i^c = (Y_{i1}, \dots, Y_{in})^T$, where superscript T denotes matrix transposition. Let $\mathbf{X}_i^c$ be the design matrix for the fixed effects with dimension $n \times p$, where p is the dimension of the parameter β in the linear mixed effects model

$$\mathbf{Y}_i^c = \mathbf{X}_i^c \beta + \epsilon_i,$$

where ϵ_i is the $n \times 1$ residual term, possibly consisting of random effects and measurement error, assumed to be from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{V}$ of dimension $n \times n$. At the design stage, t_{ij} is assumed fixed with value denoted by t_j , $t = (t_1, \dots, t_n)$, and V is a known $n \times n$ variance-covariance matrix. A common structure used for $\mathbf{V}$ is compound symmetry with common variance σ^2 and a correlation matrix $\mathbf{C}$ with a common correlation coefficient ρ , i.e., $\mathbf{V} = \sigma^2 \mathbf{C}$, and $\mathbf{C} = (1 - \rho) \mathbf{I}_n + \rho \mathbf{1}_n \mathbf{1}_n^T$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{1}_n$ is the $n \times 1$ vector with unit elements. This is also

the variance-covariance structure of $\mathbf{Y}$ conditioning on $\mathbf{X}$ when $\epsilon_{ij} = u_i + e_{ij}$ in model (1), where u_i is the random intercept with variance σ_u^2 , e_{ij} is the independent measurement error

with variance σ_e^2 , in which case $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$.

In many longitudinal studies, researchers are often interested in the effect of the predictor Z on the slope or rate of change in the outcome, which can be evaluated in the following model

$$Y_{ij} = \beta_0 + \beta_1 Z_i + \beta_2 t_{ij} + \beta_3 Z_i t_{ij} + \epsilon_{ij}. \quad (1)$$

In model (1), the j^{th} row of the design matrix $\mathbf{X}_i^c$ is $\mathbf{X}_{ij} = (1, Z_i, t_{ij}, Z_i t_{ij})$. The parameters β_0 and β_2 are the intercept and slope for $Z_i = 0$; β_1 measures the association between Z and the outcome Y at $t_j = 0$ or baseline; and β_3 measures the effect of Z on the slope. If Z is binary, e.g., $Z_i = 0$ for the control and 1 for the intervention group, then β_3 is the expected difference in the slope between the intervention and the control groups; if Z is continuous, β_3 is the expected difference in the slope of change in the outcome when Z increases by 1 unit.

In reality, longitudinal data are often unbalanced due to missing data. Denote $\mathbf{R}_i = (R_{i1}, \dots, R_{in})^T$ as the observation indicator for subject i , where $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. For simplicity, we assume that $P(R_{i1} = 1) = 1$ since the data are usually fully observed at baseline, and $P(R_{in} = 1) > 0$ because otherwise, we are reduced to the case with a smaller planned or maximum number of measurements. The missing data pattern consists of all the possible values of $\mathbf{R}_i$. For example, a common missing data pattern in longitudinal studies is monotone missing, where $R_{ik} \geq R_{il}$ for $k < l$, $k, l = 1, \dots, n$. In this case, the missing data pattern can be fully described by the number of observed measurements

denoted by $K_i = \sum_{j=1}^n R_{ij}$. If the probability of missing Y_{ij+1} given that Y_{ij} is observed, $j = 1, \dots, n-1$, is $pd_j = P(R_{ij+1} = 0 | R_{ij} = 1, Z_i)$, then $P(K = 1 | Z) = pd_1$,

$P(K = k) = \prod_{j=1}^{k-1} (1 - pd_j) pd_k$, $k = 2, \dots, n-1$, and $P(K = n | Z) = \prod_{j=1}^n (1 - pd_j)$.

Denote q as the number of possible values that $\mathbf{R}_i$ can take. In the case of a monotone missing data pattern, $q = n$; in the extreme scenario that each component of $\mathbf{R}_i$ except the first can take either 0 or 1, $q = 2^{n-1}$. Denote $\mathbf{Y}_i$ as the observed outcome, $\mathbf{X}_i$ as its corresponding design matrix, and $\mathbf{V}_i$ as the variance of $\mathbf{Y}_i$ which is the corresponding sub-matrix of $\mathbf{V}$. The missing data process is assumed not informative, i.e., the missing data mechanism is missing at random (MAR) and the parameter that governs the missing data is disjoint from the parameter β . Under this assumption, the missing data process can be ignored and the inference for β can be based on the observed data [34].

The parameter $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ in the linear mixed effects model (1) is estimated by solving the following estimating equation

$$\mathbf{U}(\beta) = \sum_{i=1}^m \mathbf{X}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i) = 0.$$

The maximum likelihood estimate (MLE) of β can be expressed as

$$\hat{\beta} = \left\{ \sum_{i=1}^m \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right\}^{-1} \sum_{i=1}^m \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{Y}_i,$$

where $\hat{\mathbf{V}}$ is the estimate of $\mathbf{V}$. And the asymptotic variance of $\hat{\beta}$ is

$$\Sigma_{\beta} = \frac{1}{m} \left\{ E \left(\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right) \right\}^{-1}. \quad (2)$$

Expression (2) provides the basis for the inference about β . For example, Tu [28] proposed a power analysis method for testing any linear combination of the parameter, $\mathbf{A}\beta$, where $\mathbf{A}$ is a known full rank $s \times 4$ matrix, based on the variance of $\mathbf{A}\hat{\beta}$ which can be expressed as $\mathbf{A}\Sigma_{\beta}\mathbf{A}^T$. Here we focus on the power analysis for testing β_3 which is the parameter of interest in most longitudinal studies. The variance of $\hat{\beta}_3$ is the $(4, 4)^{th}$ element of Σ_{β} , denoted as V_b .

From the variance expression (2), in which the expectation is taken over the joint distribution of $\mathbf{R}_i$ and Z_i , it can be seen that Σ_{β} depends on the distribution of $\mathbf{R}_i$ given Z_i . We assume the marginal distribution of $\mathbf{R}_i|Z_i$ is known, with $P(\mathbf{R}_i = \mathbf{r}_l|Z_i = z) = p_l(z)$, $l = 1, \dots, q$. Note that this does not imply that $\mathbf{R}_i$ is independent from the observed outcome $\mathbf{Y}_i$, i.e., the data is missing completely at random (MCAR). The missing data process is assumed MAR and thus can depend on the observed outcome $\mathbf{Y}_i$, but for power analyses it is not necessary to specify the full underlying missing data process, $P(\mathbf{R}_i|Z_i, \mathbf{Y}_i)$; it is only the marginal distribution of $\mathbf{R}_i|Z_i$ that matters [35].

Given planned observation times $t = (t_1, \dots, t_n)$, the residual variance-covariance matrix V , the distribution of Z , and the missing data distribution $P(\mathbf{R}|Z)$, the variance of $\hat{\beta}_3$, V_b , can be calculated using expression (2). Here we focus on the case of monotone missing data pattern in which the distribution of missing data is expressed by the random variable K . For a given $K = k$, X_i is a $4 \times k$ matrix whose j^{th} column is $(1, Z_i, t_j, Z_i t_j)^T$, $j = 1, \dots, k$; denote $\phi_{lj}^{(k)}$ as the $(l, j)^{th}$ element of $\mathbf{V}_i^{-1}$, $l, j = 1, \dots, k$, $\mathbf{t}^{(k)} = (1, \dots, t_k)^T$, and

$$a_k = \mathbf{1}_k^T \mathbf{V}_i^{-1} \mathbf{1}_k = \sum_{l=1}^k \sum_{j=1}^k \phi_{lj}^{(k)}, \quad b_k = \mathbf{1}_k^T \mathbf{V}_i^{-1} \mathbf{t}^{(k)} = \sum_{l=1}^k \sum_{j=1}^k \left(\phi_{lj}^{(k)} t_l \right), \text{ and}$$

$$c_k = \mathbf{t}^{(k)T} \mathbf{V}_i^{-1} \mathbf{t}^{(k)} = \sum_{l=1}^k \sum_{j=1}^k \left(\phi_{lj}^{(k)} t_l t_j \right). \text{ Then } \Sigma_{\beta}(2) \text{ can be expressed as}$$

$$\frac{1}{m} \begin{pmatrix} E(a_k) & E(Z_i a_k) & E(b_k) & E(Z_i b_k) \\ E(Z_i a_k) & E(Z_i^2 a_k) & E(Z_i b_k) & E(Z_i^2 b_k) \\ E(b_k) & E(Z_i b_k) & E(c_k) & E(Z_i c_k) \\ E(Z_i b_k) & E(Z_i^2 b_k) & E(Z_i c_k) & E(Z_i^2 c_k) \end{pmatrix}^{-1}.$$

If K and Z are independent, i.e., the distribution of missing data does not depend on Z , then Σ_{β} can be further simplified as

$$\frac{1}{m} \begin{pmatrix} E(a_k)B & E(b_k)B \\ E(b_k)B & E(c_k)B \end{pmatrix}^{-1},$$

where

$$B = \begin{pmatrix} 1 & E(Z) \\ E(Z) & E(Z^2) \end{pmatrix},$$

thus

$$V_b = \frac{E(a_k)}{m\sigma_z^2 [E(a_k)E(c_k) - \{E(b_k)\}^2]}. \quad (3)$$

Similar derivations have been shown in the GEE setting using an independent working correlation matrix [23]. The formula (3) can be applied to any variance structure. For example, if a random intercept and random slope model is used, and the variances of the random intercept and slope are σ_0^2 and σ_1^2 , respectively, with a correlation of ρ between the two random effects, and the variance of the random error is σ_e^2 , then the j^{th} diagonal element of $\mathbf{V}$ is $\sigma_0^2 + t_j^2\sigma_1^2 + 2\rho\sigma_0\sigma_1 t_j + \sigma_e^2$, $j=1, \dots, n$; and the $(j, l)^{th}$ non-diagonal element of $\mathbf{V}$ is $\sigma_0^2 + t_j t_l \sigma_1^2 + \rho\sigma_0\sigma_1 (t_j + t_l)$ for $j \neq l, j, l = 1, \dots, n$. Application of (3) is straightforward since the inverse of a matrix is easily calculated in most statistical software programs for any given $\mathbf{V}$ and the distribution of K .

To calculate the sample size m needed or the power $1 - \gamma$ for testing the null hypothesis $H_0: \beta_3 = 0$ versus $H_1: \beta_3 = d$ using a two-sided α level test, we solve the equation $d = V_b(Z_{1-\alpha/2} + Z_{1-\gamma})^2$, where Z_p is the p^{th} percentile of the standard normal distribution.

We consider the compound symmetry variance structure because it is common and closed form formulae for V_b can be easily obtained. We first examine power based on the asymptotic properties of the MLE for the parameter of interest, β_3 , and evaluate how power depends on the distribution of the missing data in section 2.2. We then examine two alternative power calculation methods commonly used in practice in section 2.3. Although we only consider the unadjusted model (1), the methods can be easily extended using the approach of [36] for controlling for confounders.

2.2 MLE based method and impact of missing data

2.2.1 MLE based method for power analysis with missing data—The power for testing β_3 in model (1) in the presence of missing data has been studied (e.g., [22-31]), and explicit formulas for sample size and power calculations are available in [22-26], but not as a function of the distribution of K for linear mixed effects models. We express the power and sample size formulae in terms of the distribution of K so that it can be more easily compared to the other conventional methods considered here. We assume that the missing

data pattern is monotone where the distribution of $\mathbf{R}$ is equivalently expressed by the number of measurements K . The distribution $\mathbf{p} = (p_1, \dots, p_n)$ of K , with $p_k = P(K = k|Z = z)$, $k = 1, \dots, n$, is assumed known. The variance matrix V is assumed compound symmetry with common variance σ^2 and correlation coefficient ρ .

Since the inverse of a $k \times k$ compound symmetry matrix V_i is

$\frac{1}{\sigma^2(1-\rho)} \left(\mathbf{I}_k - \frac{\rho}{1-\rho+K\rho} \mathbf{1}_k \mathbf{1}_k^T \right)$, the asymptotic variance V_b of $\hat{\beta}_3$ can be expressed as

$$V_b = \frac{\sigma^2}{m \Delta \sigma_z^2}, \quad (4)$$

$$\Delta \equiv \Delta(n, \mathbf{p}, t, \rho)$$

$$= \mathbf{E} \left(S_K^2 \right) / (1 - \rho) + \delta, \quad S_K^2 = \sum_{j=1}^K t_j^2 - \left(\sum_{j=1}^K t_j \right)^2$$

$$/K, \quad \delta = \mathbf{E} \left\{ \frac{\left(\sum_{j=1}^K t_j \right)^2}{K(1-\rho+K\rho)} \right\}$$

$$- \left\{ \mathbf{E} \left(\frac{\sum_{j=1}^K t_j}{1-\rho+K\rho} \right) \right\}^2$$

$$\text{where } \mathbf{E} \left(\frac{K}{1-\rho+K\rho} \right)$$

. As shown in Appendix A, δ is non-negative, and equals zero only when $K \equiv n$, i.e., there are no missing data in a balanced design.

Meanwhile it can also be shown that δ is far less than the first component of ,

$$\mathbf{E} \left(S_K^2 \right) / (1 - \rho) \text{ (see Appendix B for proof).}$$

When there are no missing data, $\delta = 0$, $\mathbf{E} \left(S_K^2 \right) = S_n^2 = \sum_{j=1}^n t_j^2 - \left(\sum_{j=1}^n t_j \right)^2 / n$, so

$$\Delta = S_n^2 / (1 - \rho) \text{ and (4) reduces to}$$

$$V_b^{full} = \frac{(1 - \rho) \sigma^2}{m S_n^2 \sigma_z^2}. \quad (5)$$

The sample size m needed to achieve power $1 - \gamma$ to detect $H_1: \beta_3 = d$, using a two-sided α level test, is calculated by

$$m = \frac{\sigma^2 \left(Z_{1-\alpha/2} + Z_{1-\gamma} \right)^2}{\Delta d^2 \sigma_z^2}. \quad (6)$$

Equivalently, if the sample size m is given, then the power $1 - \gamma$ can be calculated as

$$1 - \gamma = \Phi \left(\frac{\sigma_z}{\sigma} \sqrt{\frac{d^2}{m\Delta}} - Z_{1-\alpha/2} \right) \equiv \Phi \left(c - Z_{1-\alpha/2} \right), \quad (7)$$

where $c = \frac{\sigma_z}{\sigma} \sqrt{\frac{d^2}{m\Delta}}$. When there are no missing data, i.e., $P(K = n) = 1$, (6) and (7) reduce to the familiar complete data formulae

$$m = \frac{(1 - \rho) \sigma^2 (Z_{1-\alpha/2} + Z_{1-\gamma})^2}{S_n^2 d^2 \sigma_z^2}, \quad (8)$$

$$1 - \gamma = \Phi \left(\frac{\sigma_z}{\sigma} \sqrt{\frac{d^2 (1 - \rho)}{m S_n^2}} - Z_{1-\alpha/2} \right). \quad (9)$$

If the predictor Z is binary with mean μ_z , e.g., as in a longitudinal study comparing two treatment groups, $\sigma_z^2 = \mu_z (1 - \mu_z)$.

To illustrate how the crucial term in (4), (6) and (7) is calculated, we give an example for $n = 2$. Suppose $t = (0, 1)$ (e.g., $t=0$ and 1 indicate baseline and follow-up, respectively), the intraclass correlation coefficient is $\rho = 0.5$, and the drop out rate at follow-up is 30% so $\mathbf{p} =$

$$\begin{aligned} S_2^2 &= 0.5, \quad E(S_K^2) \\ &= 0(0.30) + 0.5(0.70) = 0.35, \quad E \left\{ \frac{(\sum_{j=1}^K t_j)^2}{K(1 - \rho + K\rho)} \right\} \\ &= 0 \times 0.3 + \frac{1}{2(1+0.5)} \\ &\quad \times 0.7 = 0.233, \quad E \left(\frac{\sum_{j=1}^K t_j}{1 - \rho + K\rho} \right) \end{aligned}$$

$(0.30, 0.70)$. Since $S_1^2 = 0$ and $= 0 \times 0.3 + \frac{1}{1+0.5} \times 0.7 = 0.467$, and

$E \left(\frac{K}{1 - \rho + K\rho} \right) = 1 \times 0.3 + \frac{2}{1+0.5} \times 0.7 = 1.233$, hence $\delta = 0.233 - 0.467^2 / 1.233 = 0.056$, $= 0.35 / 0.5 + 0.056 = 0.756$. If $\sigma^2 = 1$ and $\mu_z = 0.5$ so $\sigma_z^2 = 0.25$, then the sample size needed to detect $d = 0.5$ under the alternative hypothesis with 90% power using two-sided α level 0.05 test is, from (6),

$$m = \sigma^2 (Z_{1-\alpha/2} + Z_{1-\gamma})^2 / (\Delta d^2 \sigma_z^2) = (1.96 + 1.28)^2 / (0.756 \times 0.5^2 \times 0.25) = 223.$$

2.2.2 Impact of missing data—Expressions (4), (6), and (7) suggest that the impact of missing data on the power for testing β_3 is through the function δ , which depends on the maximum number of measurements n , correlation coefficient ρ , the value of time of measurements t , and the distribution $\mathbf{p}$ of the number of measurements K . For a given distribution of K , the efficiency and power increase as n and ρ increase, and as the time

intervals between repeated measurements increase. Less missing data results in smaller and thus more efficient and powerful estimate of β_3 . Studies with a higher proportion of drop out at each time interval, or the same proportions of drop out but occurring earlier (e.g., 40% drop out at the last time interval versus 40% drop out at the first time interval), clearly have more missing data and thus smaller and power.

However, without calculating , the impact of the severity of missing data when $n > 2$ and with a variable drop out rate across visits is not obvious. In practice, the proportion of subjects with complete data, p_n , or the mean number of repeated measures, $E(K)$, are commonly used as a summary of missing data. Although a large p_n or $E(K)$ intuitively seems better for power, in fact the entire distribution of K needs to be known in order to calculate and hence the power or sample size. To illustrate this using studies with $n = 5$, $t = (0, 1, 2, 3, 4)$ and $\rho=0.5$, we list values of in Table 1 for different values of $\mathbf{p}$, as well as the asymptotic power (Asym) estimate using the MLE based method and the empirical power (Emp) from 2000 simulations for detecting $\beta_3 = -0.02$, with $m = 500$, $\sigma^2 = \sigma_z^2 = 1$, and values of 0.25, 0.50 and 0.75 for the correlation coefficient, ρ . The first and last rows are the complete data cases with $n = 5$ and $n = 4$, respectively, as references. Rows 2-3 correspond to the cases of the least and most missing data given $p_5 = 0.6$, i.e., all drop outs before the last visit occur at the 4th and 1st visit, respectively. Similar results are shown for $p_5 = 0.2$. Rows 6-9 are examples of distributions of K with in decreasing order and $E(K) = 4$. The empirical and asymptotic power estimates are similar.

2.2.3 Adjusting for confounders—To adjust for a set of confounders, W , and the interaction between W and time when examining the effect of the main predictor variable on the slope in model (1), we can follow the method in [36] and simply replace σ_z^2 with $\sigma_z^2 (1 - R_{z|w}^2)$, the conditional variance of Z given W , in all the power calculation formulas.

2.3 Alternative methods

In practice, alternative approaches that simply modify the complete data formula (8) or (9) are often used.

2.3.1 Average sum of squares (ASQ) method—Since the power analysis formulae for complete data, (8) and (9), depend on n only through S_n^2 , it seems reasonable to replace S_n^2 in (8) or (9) by its expectation $E(S_K^2)$ in the presence of missing data; that is,

$$m^{ASQ} = \frac{(1 - \rho) \sigma^2 (Z_{1-\alpha/2} + Z_{1-\gamma})^2}{E(S_K^2) d^2 \sigma_z^2},$$

and the power estimate $1 - \gamma^{ASQ}$ is

$$1 - \gamma^{ASQ} = \Phi \left(\frac{\sigma_z}{\sigma} \sqrt{\frac{d^2 (1 - \rho)}{m E(S_K^2)}} - Z_{1-\alpha/2} \right).$$

This method is referred to as the average sum of squares (ASQ) approach. We have shown that $\delta \geq 0$. This implies that $E(S_K^2) \leq E(S_K^2) + (1 - \rho)\delta$. Since ASQ method ignores the nonnegative component, δ , in the denominator of V_b , it is conservative in that the sample size is always overestimated or power is always underestimated when there are missing data and $\rho < 1$. However, as δ is less than $E(S_K^2) / (1 - \rho)$, and almost ignorable when the drop out rate is not too high, the ratio of the sample size estimates from ASQ versus the MLE based methods, $m^{ASQ}/m = 1 + (1 - \rho)\delta/E(S_K^2) \equiv r^{ASQ}$, is greater than or equal to 1 but less than 2, and is close to 1 when the drop out rate is small or moderate. In term of the power estimates for given sample size m , the ASQ power estimate, $1 - \gamma^{ASQ} = \Phi(c/\sqrt{r^{ASQ}} - Z_{1-\alpha/2})$, is less than or equal to the MLE based power estimate, $1 - \gamma = \Phi(c - Z_{1-\alpha/2})$. This can be seen intuitively by noting that the subjects with only one measurement do not contribute to the power analysis since $S_1^2 = 0$.

We examined the ratio of the sample size estimates using the ASQ method versus the MLE based method, m^{ASQ}/m , in a longitudinal study with $n = 2$ and $n = 5$, various values of the correlation coefficient ρ , and a constant drop out rate. Results are shown in Figures 1 and 2 for $n = 2$ and $n = 5$, respectively. The figures show that the ratio m^{ASQ}/m is always greater than or equal to 1, increases as ρ decreases or as the drop out rate increases, and increases faster as the drop out increases when the correlation ρ is smaller. As the drop out rate approaches 1, the ratio reaches a limit less than 2 for both $n = 2$ and $n = 5$, i.e., ASQ overestimates sample size by no more than 100%. When ρ is large and the drop out rate is moderate or small, the ratio is close to 1 and thus the ASQ method is a good approximation to the MLE based method. For example, when $\rho = 0.8$ and the drop out rate is 30% in the case of $n = 2$, it only overestimates the sample size by about 3%.

2.3.2 Average number of subjects (ANS) method—Another commonly used approach for computing power with missing data is to equate it to a balanced longitudinal study without missing data that has the same total number of repeated measurements

$\sum_{i=1}^m K_i = mE(K)$ (e.g., [17]). We call it the average number of subjects (ANS) method. To calculate power, m is replaced in the complete data formula (9) with

$m^* = \sum_{i=1}^m K/n = mE(K)/n$, i.e., the ANS estimate of power

$$1 - \gamma^{ANS} = \Phi \left(\frac{\sigma_z}{\sigma} \sqrt{\frac{nd^2 (1 - \rho)}{m E(K) S_n^2}} - Z_{1-\alpha/2} \right).$$

To calculate sample size, m^* is first determined using the complete data formula (8) and then the sample size is calculated as $nm^*/E(K)$; that is, the result from (8) is multiplied by $n/E(K)$, which yields

$$m^{ANS} = \frac{n(1-\rho)\sigma^2(Z_{1-\alpha/2} + Z_{1-\gamma})^2}{E(K)S_n^2 d^2 \sigma_z^2}.$$

The function S_k^2 , the realization of S_K^2 for $K = k$, is an increasing cubic function of k . For example, if $n = 5$ and $t_j = j - 1, j = 1, \dots, 5$, $S_k^2 = (k^3 - k)/12, k = 1, \dots, 5$, i.e., $(S_1^2, S_2^2, S_3^2, S_4^2, S_5^2) = (0, 0.5, 2, 5, 10)$. Hence the contribution to $E(S_K^2)$, the dominant component in the denominator of V_b , from a subject with n measurements, is very much larger than that from a subject with number of measurements $k < n$. Replacing the power estimate by the complete data formula of a balanced study with the same number of total measurements assumes that the contribution to V_b of subjects with $k < n$ number of measurements is the same as that of subjects with n measurements. This results in overestimated power or underestimated sample size, and its difference from the MLE based estimate is larger when there are more missing data, especially for larger n as the differences between S_n^2 and the other $S_k^2, k < n$, increases as n increases.

We examined the ratio of the sample size calculated using the ANS method versus the

correct method, $\frac{m^{ANS}}{m} = \frac{n(1-\rho)\Delta}{E(K)S_n^2} \equiv r^{ANS}$, for $n = 2$ and $n = 5$, various values of ρ , and a constant drop out rate. The results, shown in Figure 3 and 4 for $n = 2$ and $n = 5$, respectively, indicate that the ratio m^{ANS}/m is always less than or equal to 1, decreases as ρ increases and as the drop out rate increases, and decreases faster as the drop out rate increases when ρ is larger and when n is larger. When the drop out rate is approaching 1, the ratio is approaching 0. This implies that the ANS method can severely underestimate sample size, and thus should be avoided.

When the sample size m is fixed, the ANS power estimate,

$1 - \gamma^{ANS} = \Phi\left(c/\sqrt{r^{ANS}} - Z_{1-\alpha/2}\right)$, is greater than or equal to the MLE based power estimate, $1 - \gamma = \Phi(c - Z_{1-\alpha/2})$.

2.3.3 Other methods—Other commonly used methods in practice include using only the subset of subjects with all measurements $k = n$ and then applying the complete data formula. Here only a proportion p_n of the subjects are used, so it clearly is too conservative, especially when p_n is small and n is moderate or large. In the case of $n = 2$, it is equivalent to the ASQ method. As this method has already been studied in the literature (e.g. [24]), we do not consider it further in this paper.

3. Simulation studies

3.1 Simulation study 1

In the first simulation study, the missing data were generated completely at random, and our goal was to examine the performance of the MLE based methods and the conventional ASQ and ANS methods. We considered longitudinal studies with 2 and 5 visits, representing short and moderate lengths of follow-up, respectively. For $n = 2$, the planned follow-up times $t = (0, 1)$, representing baseline and follow-up visits; for $n = 5$ we have $t = (0, 1, 2, 3, 4)$, representing baseline and 4 follow-up visits with equal time intervals (e.g, annually). We considered values of 0.25, 0.50 and 0.75 for the correlation coefficient ρ , representing low, medium and high levels of correlation, respectively. The predictor variable, Z , was generated from a Bernoulli distribution with $P(Z = 1) = 0.5$ for the case of $n = 2$, and standard normal for the case of $n = 5$. The repeatedly measured outcome Y was generated from model (1), with $\beta_0 = \beta_1 = \beta_2 = 0$, $\sigma^2 = 1$. The first (or baseline) measure Y_1 was observed for every subject. For the missing data distribution, the observation indicator R_2 for the follow-up measure Y_2 in the case of $n = 2$ was generated from a Bernoulli distribution with $P(R_2 = 1|Z, Y_1) = 1 - p_d$. In the case of $n = 5$, a constant drop out rate p_d was considered, with $P(R_{ij} = 1|R_{ij-1} = 1, Z_i, Y_{i1}, \dots, Y_{ij-1}) = 1 - p_d, j = 2, \dots, 5$. Values of 0.1, 0.2 and 0.3, commonly observed in longitudinal studies, were considered for the drop out rate p_d . We evaluated the power for given sample size m of 100, 200, 300, 400 and 500, with $\beta_3 = -0.3$ and -0.02 for $n = 2$ and $n = 5$, respectively. Empirical power estimates from 2000 simulations (Emp) were obtained, and compared to the power estimates using the MLE based method (Asym), average sum of squares method (ASQ) and average number of subjects method (ANS). Sample size estimates from different methods for given power of 80% and 90% were also examined, with $\beta_3 = 0.3$ and 0.04 for $n = 2$ and $n = 5$, respectively, along with the empirical power estimates from 2000 simulations (Emp) using the estimated sample sizes. Extreme drop out rates of 0.5, 0.7 and 0.9 were also considered. For assessing power estimates, sample sizes of $m = 500$ for the case of $n = 2$ and $m = 10000$ for the case of $n = 5$ were used. For evaluating sample size estimates for a given power, the target power was set at 80% with $\beta_3 = 0.3$ and 0.05 for $n = 2$ and $n = 5$, respectively. The situation of no missing data was also considered as a reference, in which case all three power analysis methods were reduced to the complete data formulae.

Results of the performance of different power estimates are shown in Table 2a. Overall, the MLE based estimate of power is close to the empirical power estimate. The power estimate using the ASQ method is similar to the empirical estimate when $p_d = 10\%$. The trend of ASQ underestimating power becomes slightly more apparent as the missing data proportion becomes larger, especially for small ρ . For example, for $n = 2$, when $p_d = 20\%$ or 30% , ASQ underestimates power by more than 10% compared to the empirical estimates when $m = 100$ and $\rho = 0.25$. For $n = 5$, when $p_d = 30\%$, the degree of underestimation of power by the ASQ method reaches a maximum of 22% when $\rho = 0.25$. Under the smallest missing data proportion considered, $p_d = 10\%$, the ANS power estimate is similar to the empirical power estimate when $n = 2$, and is slightly greater when $n = 5$. The degree of overestimation of the ANS power estimate is more severe when the drop out rate increases, especially for large ρ and n . For example, for $n = 2$, the power is overestimated by 13% when $\rho = 0.75$ and $m =$

100 or 200; and for $n = 5$ and $\rho = 0.75$, the magnitude of overestimation is between 15% and 29% when $p_d = 20\%$, and strikingly over 45% when $p_d = 30\%$ and $m \geq 200$.

Sample size estimates for a given power were also compared across the different approaches and results are shown in Table 2b. Overall, the empirical power using the sample size estimate from the MLE based method (Asym) is close to the target power value. The sample size estimate from the ASQ method is close to that from the MLE based method when ρ is large and p_d is small. Overestimation of the ASQ method in the sample size reaches maximums of 13% and 23% for $n = 2$ and $n = 5$, respectively, when $p_d = 30\%$.

Correspondingly, the empirical power estimates using the ASQ sample size is larger than the target power level under these circumstances. The ANS method underestimates sample size compared to the MLE based method, especially when p_d and ρ are larger. The most severe underestimation occurs at the highest drop out rate, $p_d = 30\%$, and the lowest $\rho = 0.25$ considered, with 14% and 37% for $n = 2$ and $n = 5$, respectively. Correspondingly, the empirical power estimates using the ANS sample size are lower than the target power level, e.g., 59% when the target power level is 80% for $n = 5$, $p_d = 30\%$ and $\rho = 0.75$.

Results under extremely high drop out rates are shown in Table 3a for evaluating power estimates for given sample size, and Table 3b for evaluating sample size estimate for given power. As in the regular case of low to moderate drop out rates, for given sample sizes, the MLE based estimates of power are close to the empirical power estimates; and for given power, the empirical power using the sample size estimate from the MLE based method is close to the target power level. It is more evident that the ASQ underestimates power for a given sample size, and overestimates sample size for a given power but the magnitudes of under and over estimation are limited. For example, for given sample sizes (Table 3a), in all the scenarios with $\rho = 0.25$, ASQ underestimates power by 12% to 27%; and for given power, the sample size is overestimated by a maximum of 56% when $n = 5$, $p_d = 90\%$ and $\rho = 0.25$. On the other hand, the ANS method dramatically overestimates power for given sample size, and underestimates sample size for given power, especially in the case of $n = 5$. For example, when $p_d = 0.9$ and $\rho = 0.75$, the ANS power estimate is 100% with the assumed sample size while the empirical power is only 21% (Table 3a), and for the given power, the ANS sample size estimate is underestimated by 97%, and the empirical power using the ANS sample size is only 8.6% in contrast to the target 80% level of power (Table 3b).

3.2 Simulation study 2

As introduced in section 2.1, the missing data mechanism is assumed to be MAR in the linear mixed effects model, i.e., R can depend on the observed outcome Y , but the efficiency and power of the MLE depend only on the marginal distribution of the missing data, $p = P(R|Z)$. To examine the performance of the MLE based method under MAR, we performed the second simulation study with Z and Y generated as described in the previous simulation study, but generated the observation indicator R under MAR, and compared the empirical power with that generated from MCAR with the same marginal distribution. Specifically, for $n = 2$, we generated R_2 with $P(R_2 = 1|Y_1 = 0, Z) = \pi_1$, and $P(R_2 = 1|Y_1 < 0, Z) = \pi_2$, so that the marginal distribution $P(R_2 = 1|Z) = (\pi_1 + \pi_2)/2$ since Y_1 had mean 0 and was

unrelated to Z . When $\pi_1 = \pi_2$, the missing data mechanism was MAR. Different values of (π_1, π_2) were considered. The value (1,0.8) corresponds to an overall 10% drop out at follow up, values (1, 0.6) and (0.9, 0.7) correspond to an overall 20% drop out, and values of (1, 0.4) and (0.8, 0.6) correspond to an overall 30% drop out. For $n = 5$, we considered for simplicity the situation that drop out occurred only at the first follow up visit, with $P(R_{i2} = 1 | Y_1 = 0, Z) = \pi_1$, $P(R_{i2} = 1 | Y_1 = 1, Z) = \pi_2$, and $R_{i5} = R_{i4} = R_{i3} = R_{i2}$. The marginal distribution of R_2 was therefore $P(R_{i2} = 1 | Z) = (\pi_1 + \pi_2)/2$ since Y_1 had mean 0 and was unrelated to Z , which correspond to $p_d = 1 - (\pi_1 + \pi_2)/2$ drop out at the first follow up visit, so $p = (p_d, 0, 0, 0, 1 - p_d)$. Values of (0.5, 0.5), (0.3, 0.7), (0.2, 0.8), (0.1, 0.9) were considered for (π_1, π_2) , which all corresponded to $p_d = 0.5$ at the first follow up. The cases of $\pi_1 = \pi_2 = 0.5$ correspond to the MCAR missing data case. The results are summarized in Tables 4a and 4b for $n=2$ and 5, respectively.

From Tables 4a and 4b, the empirical power estimates from different MAR cases with the same marginal drop out rate are all similar to that from the MCAR case, and all empirical power estimates are similar to the MLE based power estimates which only used the marginal distribution of the missing data.

4. Examples

We apply the MLE based method and the ASQ and ANS methods to the power analysis of two longitudinal studies in the elderly. Empirical power estimates from 2000 simulations are also obtained.

4.1 Example 1: Between group comparison of change before and after treatment in a clinical trial with 2 visits

In designing a randomized clinical trial to test the efficacy of cognitive remediation (CR) in improving mobility in sedentary seniors[37], eligible participants will be randomized equally either to CR or health education control group. The primary mobility outcome is the change in gait velocity from baseline at the end of the intervention. The primary hypothesis is that subjects in the CR intervention group will show better improvement in gait velocity than those in the control group.

The linear mixed effects model (1) will be used to analyze the data. Here the predictor Z , an indicator of the intervention group, is binary with $P(Z_i = 1) = 0.5$. The goal was to estimate the sample size needed to detect, with 85% power, a difference of 4 cm/sec in the change of gait velocity between the intervention and control group, with an expected drop out rate of 20% after intervention.

Based on previous studies[38], the value of σ is assumed to be 24 cm/sec and the intraclass correlation coefficient ρ is 0.875. Using formula (6) from the MLE based method, the estimate of sample size needed is $m = 400$. Using the average sum of squares (ASQ) method, the sample size estimate is 406, close to the correct value because ρ is large and the expected drop out is not severe. With the average number of subjects (ANS) method, the estimated sample size is 360, which underestimates the correct sample size by 10%. Simulation studies with 2000 repetitions were performed to evaluate the empirical power

using these sample size estimates. The empirical power from sample sizes of 400, 406 and 360, estimated using the MLE based, ASQ, and ANS methods, respectively, are 85.2%, 85.1% and 81.2%. The MLE based and ASQ sample size estimates are similar and both yield empirical power estimates similar to the target level of 85%. As expected, the underestimated sample size using the ANS approach results in an empirical power estimate smaller than the target value.

4.2 Example 2: Effect of a continuous predictor on the slope in a longitudinal study with 5 visits

In designing one of the projects in the renewal application of the Einstein Aging Study (EAS)[39], a long term study of aging, it is assumed that 390 subjects with baseline measures of cerebral vascular reactivity would be available. The primary hypothesis is that this variable would be associated with the rate of cognitive decline as measured by Free and Cued Selective Reminding Test (FCSRT)[40]. The study cohort will consist of a mix of current study participants and new enrollees. The outcome will be measured at baseline and each annual follow up visit during a 5 year study period. Given the study design and the expected attrition rate, it is projected that 30%, 9%, 24%, 19% and 18% of the sample will have 1, 2, 3, 4 and 5 repeated measures in FCSRT, respectively; the intra class correlation ρ is 0.50, and the conditional standard deviation σ is 6.4.

The goal was to determine the power for detecting a difference of 0.35 points per year in the rate of cognitive decline in FCSRT corresponding to a 1 SD unit difference in the cerebral vascular reactivity measure using a two-sided α level 0.05 test and 390 total subjects.

The power estimates are 0.83, 0.79 and 0.95 using the MLE based, ASQ and ANS methods, respectively. The empirical power estimate from 2000 simulations is 0.813, close to the MLE based estimate. The ASQ method slightly underestimated power, while the ANS method overestimated power by 17%.

5. Discussion

Missing data are common in longitudinal studies and therefore should be taken into account when assessing power for testing the effect of a predictor variable on the rate of change in the outcome. Despite the availability of methods for evaluating power with missing data, it has remained unclear how missing data actually affect power and erroneous approaches are still often used in practice. We have shown that the average sum of squares (ASQ) method ignores a nonnegative component of the function and thus results in an overestimated sample size or underestimated power, especially when the correlation between repeated measures is small and the rate of missing data is high. For small to moderate degrees of missing data and large intra-class correlation, the ASQ is similar to the MLE based approach. The average number of subjects (ANS) method, on the other hand, assumes subjects with incomplete and complete data contribute to the power analysis equally, and thus can result in seriously overestimated power or underestimated sample size, especially when the degree of missing data is severe and the number of complete measures and the correlation among repeated measures are moderate or large. Extensive simulation studies have confirmed these conclusions. The power analysis method we considered is based on a

linear mixed effects model which assumes that the data are missing at random (MAR), i.e., the missing data can depend on the observed outcome given the predictor. However, only the marginal distribution of the missing data, i.e., the distribution of missing data given the predictor, is needed for determining power. Our findings suggest that the ANS method can be greatly misleading and should not be used. The conservative ASQ method can be a good approximation when the correlation is large and the extent of missing data is moderate. Based on our simulation studies, the asymptotically correct MLE based method performs well with limited sample sizes and is recommended especially with large amounts of missing data.

Acknowledgement

This work was supported by National Institutes of Health [P01-AG03949 and R01-AG02511903].

Appendix A: Proof that δ is non-negative

Denote $a = \frac{\sum_{j=1}^K t_j}{\sqrt{K(1-\rho+K\rho)}}$, $b = \frac{K}{1-\rho+K\rho}$. By the Cauchy-Schwartz Inequality, $\{E(ab)\}^2 \leq E(b^2)E(a^2)$, hence

$$\delta = E(a^2) - \{E(ab)\}^2 / E(b)^2 \geq 0,$$

where the equality holds when K is a constant, $K \equiv n$, i.e., there are no missing data in a balanced design.

Appendix B: Proof that δ is far less than the $E(S_K^2)/(1-\rho)$

Because δ increases while $E(S_K^2)/(1-\rho)$ decreases with ρ , it suffices to show that

$\delta < E(S_K^2)/(1-\rho)$ when $\rho = 0$. When $\rho = 0$,

$$\begin{aligned} \delta &= E \left\{ \frac{\left(\sum_{j=1}^K t_j \right)^2}{K} \right\} - \frac{\left\{ E \left(\sum_{j=1}^K t_j \right) \right\}^2}{E(K)} \\ &\leq E \left\{ \left(\sum_{j=1}^K t_j / \sqrt{K} \right)^2 \right\} - \left\{ E \left(\sum_{j=1}^K t_j \right) \right\}^2 E(1/K) \\ &\leq E \left\{ \left(\sum_{j=1}^K t_j / \sqrt{K} \right)^2 \right\} - \left\{ E \left(\sum_{j=1}^K t_j / \sqrt{K} \right)^2 \right\} \\ &= E \left\{ K \text{Var} \left(\bar{t}_k | K=k \right) \right\}, \end{aligned}$$

where the first inequality follows as $E(1/K) = 1/E(K)$ from Jensen's inequality, the second

inequality follows because $\left\{ E \left(\sum_{j=1}^K t_j \right) \right\}^2 E(1/K) \geq \left\{ E \left(\sum_{j=1}^K t_j / \sqrt{K} \right) \right\}^2$ based on

Cauchy-Schwartz Inequality. The variable being taken expectation of in δ is the variation of $t_j / \sqrt{k}$ for a given $K = k$. While the first component of δ is

$$E(S_K^2) = E\left\{\left(\sum_{j=1}^K t_j^2\right) - \left(\sum_{j=1}^K t_j\right)^2 / K\right\} = E\{k \text{Var}(t|K=k)\}. \text{ Since the variation in } t \text{ is larger than the mean } \bar{t}_k \text{ given } K=k, \delta < E(S_K^2).$$

References

1. Diggle, PD.; Heagerty, P.; Liang, KY.; Zeger, SL. Analysis of longitudinal data. 2nd. Oxford University Press; New York: 2002.
2. Donner, A.; Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold; London, England: 2000.
3. Bloch DA. Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Statistics in Medicine*. 1986; 5:663–667. [PubMed: 3823673]
4. Rochon J. Sample size calculations for two-group repeated- measures experiments. *Biometrics*. 1991; 47:1383–1398.
5. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine*. 1998; 17:1643–1658. [PubMed: 9699236]
6. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement design. *Controlled Clinical Trials*. 1994; 15:100–123. [PubMed: 8205802]
7. Overall JE, Shobaki G, Anderson CB. Comparative evaluation of two models for estimating sample sizes for tests on trends across repeated measurements. *Controlled Clinical Trials*. 1998; 19:188–197. [PubMed: 9551283]
8. Lipsitz SR, Fitzmaurice GM. Sample size for repeated measures studies with binary responses. *Statistics in Medicine*. 1994; 13:1233–1239. [PubMed: 7973204]
9. Muller KE, LaVange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*. 1992; 87:1209–1226. [PubMed: 24790282]
10. Schouten HJA. Planning group sizes in clinical trials with a continuous outcome and repeated measures. *Statistics in Medicine*. 1999; 18:255–264. [PubMed: 10070672]
11. Kim H, Williamson JM, Lyles CM. Sample size calculations for studies with correlated ordinal outcomes. *Statistics in Medicine*. 2005; 24:2977–2987. [PubMed: 16149125]
12. Tu XM, Kowalski J, Zhang J, Lynch KG, Crits-Christoph P. Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine*. 2004; 23:2799–2815. [PubMed: 15344187]
13. Brooks RJ, Cottenden AM, Fader MJ. Sample sizes for designed studies with correlated binary data. *The Statistician*. 2003; 52:539–551.
14. Hendricks SA, Wassell JT, Collins JW, Sedlak SL. Power determination for geographically clustered data using generalized estimating equations. *Statistics in Medicine*. 1996; 15:1951–1960. [PubMed: 8888487]
15. Patel H, Rowe E. Sample size for comparing linear growth curves. *Journal of Biopharmaceutical Statistics*. 1999; 9:339–350. . [PubMed: 10379697]
16. Liu G, Liang KY. Sample size calculations for studies with correlated observations. *Biometrics*. 1997; 53:937–947. [PubMed: 9290224]
17. Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. *Statistics in Medicine*. 2002; 21:1787–1801. [PubMed: 12111912]
18. Manatunga AK, Hudgens MG. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*. 2001; 43:75–86.
19. Taljaard M, Donner A, Klar N. Accounting for expected attrition in the planning of community intervention trials. *Statistics in Medicine*. 2007; 26:2615–2628. [PubMed: 17068842]

20. 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.
21. Lu K, Luo X, Chen P. Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *The International Journal of Biostatistics*. 2008 Article 9.
22. Wu MC. Sample size for comparison of changes in the presence of right censoring caused by death, withdrawal, and staggered entry. *Controlled Clinical Trials*. 1988; 9:32–46. [PubMed: 3356151]
23. Jung S, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine*. 2003; 22:1305–1315. [PubMed: 12687656]
24. Ahn C, Jung SH. Effect of dropouts on sample size estimates for test on trends across repeated measurements. *Journal of Biopharmaceutical Statistics*. 2005; 15:33–41. [PubMed: 15702603]
25. Lefante JJ. The power to detect differences in average rates of change in longitudinal studies. *Statistics in Medicine*. 1990; 9:437–446. [PubMed: 2362980]
26. Dawson JD. Sample size calculations based on slopes and other summary statistics. *Biometrics*. 1998; 54:323–330. [PubMed: 9544525]
27. Dawson JD, Lagakos SW. Size and power of two-sample tests of repeated measures data. *Biometrics*. 1993; 49:1022–1032. [PubMed: 7906957]
28. Tu XM, Zhang J, Kowalski J, Shults J, Feng C, Sun W, Tang W. Power analyses for longitudinal study designs with missing data. *Statistics in Medicine*. 2007; 26:2958–2981. [PubMed: 17154250]
29. Jung SH, Ahn CW. Sample size for a two-group comparison of repeated binary measurements using GEE. *Statistics in Medicine*. 2005; 24:2583–2596. [PubMed: 16118812]
30. Lu K, Mehrotra DV, Liu GH. Sample size determination for constrained longitudinal data analysis. *Statistics in Medicine*. 2009; 28:679–699. [PubMed: 19051207]
31. 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]
32. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
33. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics*. 1992; 38:963–974. [PubMed: 7168798]
34. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
35. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag; New York: 2000.
36. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*. 1998; 17:1623–1634. [PubMed: 9699234]
37. Verghese J, Mahoney J, Ambrose AF, Wang C, Holtzer R. Effect of cognitive remediation on gait in sedentary seniors. *Journal of Gerontology Series A: Biological Sciences and Medical Sciences*. 2010; 65:1338–1343.
38. Rolita L, Holtzer R, Wang C, Derby C, Lipton RB, Verghese J. Homocysteine and mobility decline in older adults. *Journal of American Geriatrics Society*. 2010; 58:545–550.
39. Lipton RB, Katz MJ, Kuslansky G, Sliwinski M, Stewart W, Verghese J, Crystal H, Buschke H. Screening for dementia by telephone using the memory impairment screen. *Journal of the American Geriatrics Society*. 2003; 51:1382–1390. [PubMed: 14511157]
40. Buschke H. Cued recall in amnesia. *Journal of Clinical and Experimental Neuropsychology*. 1984; 6:433–440.

Figure 1.

Ratio of sample size calculated using the average sum of squares (ASQ) method versus the MLE based method for $n=2$

Figure 2.
Ratio of sample size calculated using the average sum of squares (ASQ) method versus the MLE based method for $n=5$

Figure 3.
Ratio of sample size calculated using the average number of subjects (ANS) method versus the MLE based method for $n=2$

Figure 4.
Ratio of sample size calculated using the average number of subjects (ANS) method versus the MLE based method for $n=5$

Table 1

and power for some distributions of K from studies with $n = 5$

p	$E(K)$	$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$		
		Asym	Emp		Asym	Emp		Asym	Emp	
(0.0,0.0,1)	5	13.33	0.368	20	0.516	0.534	40	0.807	0.799	
(0.0,0.0,4,0.6)	4.6	10.81	0.330	16.10	0.434	0.451	32.07	0.717	0.707	
(0.4,0.0,0.0,6)	3.4	9.26	0.271	13.14	0.367	0.368	25.04	0.609	0.607	
(0.0,0.0,8,0.2)	4.2	8.10	0.250	12.07	0.342	0.345	24.05	0.592	0.596	
(0.8,0.0,0.0,2)	1.8	3.90	0.130	4.94	0.167	0.193	8.76	0.262	0.252	
(0.25,0.0,0.0,0.75)	4	10.88	0.314	15.83	0.428	0.425	30.79	0.699	0.699	
(0.15,0.0,0.05,0.3,0.5)	4	9.42	0.279	13.75	0.381	0.382	26.91	0.641	0.649	
(0.0,0.2,0.6,0.2)	4	7.42	0.229	10.96	0.316	0.332	21.72	0.550	0.554	
(0.0,0.1,0.8,0.1)	4	7.05	0.220	10.48	0.304	0.288	20.86	0.533	0.537	
(0.0,0.1,1.0)	4	6.67	0.210	10	0.293	0.277	20	0.516	0.509	

Empirical power from 2000 simulations and power estimates using the MLE based (Asym), the average sum of squares (ASQ) and average number of subjects (ANS) methods under different drop out rate p_d , correlation ρ and sample size m

Table 2a

ρ	m	$p_d = 0$			$p_d = 10\%$			$p_d = 20\%$			$p_d = 30\%$				
		Emp	Asym	Emp	Asym	ASQ	ANS	Emp	Asym	ASQ	ANS	Emp	Asym	ASQ	ANS
$n = 2, Z$ is Bernoulli with mean 0.5, $\beta_3 = -0.3$															
0.25	100	0.244	0.231	0.226	0.219	0.212	0.222	0.222	0.222	0.206	0.194	0.212	0.198	0.192	0.203
	200	0.420	0.410	0.384	0.388	0.376	0.393	0.373	0.364	0.341	0.376	0.325	0.337	0.305	0.358
	300	0.567	0.564	0.535	0.536	0.521	0.543	0.486	0.505	0.475	0.521	0.467	0.470	0.427	0.498
	400	0.675	0.688	0.675	0.659	0.642	0.666	0.639	0.625	0.591	0.642	0.572	0.585	0.536	0.617
	500	0.787	0.782	0.749	0.754	0.738	0.761	0.734	0.721	0.688	0.738	0.688	0.682	0.630	0.714
0.50	100	0.322	0.323	0.298	0.302	0.296	0.309	0.293	0.280	0.268	0.296	0.258	0.256	0.240	0.282
	200	0.558	0.564	0.547	0.531	0.521	0.543	0.488	0.495	0.475	0.521	0.452	0.454	0.427	0.498
	300	0.729	0.738	0.708	0.704	0.693	0.716	0.659	0.664	0.642	0.693	0.625	0.618	0.585	0.668
	400	0.855	0.851	0.832	0.822	0.812	0.833	0.790	0.786	0.765	0.812	0.748	0.742	0.709	0.790
	500	0.924	0.918	0.901	0.897	0.889	0.905	0.871	0.868	0.851	0.889	0.828	0.831	0.801	0.871
0.75	100	0.559	0.564	0.520	0.526	0.521	0.543	0.483	0.485	0.475	0.521	0.442	0.440	0.427	0.498
	200	0.865	0.851	0.818	0.817	0.812	0.833	0.768	0.776	0.765	0.812	0.733	0.725	0.709	0.790
	300	0.952	0.957	0.929	0.939	0.937	0.948	0.906	0.914	0.908	0.937	0.881	0.880	0.867	0.923
	400	0.990	0.989	0.980	0.982	0.981	0.985	0.970	0.970	0.967	0.981	0.948	0.951	0.944	0.975
	500	0.994	0.997	0.995	0.995	0.995	0.996	0.989	0.990	0.989	0.995	0.985	0.982	0.978	0.992
$n = 5, Z$ is standard normal, $\beta_3 = -0.02$															
0.25	100	0.125	0.109	0.106	0.093	0.090	0.097	0.086	0.080	0.074	0.087	0.071	0.068	0.062	0.078
	200	0.180	0.177	0.156	0.146	0.138	0.153	0.132	0.119	0.109	0.133	0.101	0.097	0.086	0.117
	300	0.248	0.244	0.209	0.197	0.186	0.208	0.163	0.158	0.142	0.178	0.129	0.125	0.109	0.154
	400	0.301	0.309	0.253	0.248	0.234	0.262	0.195	0.196	0.175	0.223	0.160	0.152	0.132	0.192
	500	0.375	0.372	0.293	0.298	0.281	0.315	0.224	0.234	0.209	0.267	0.188	0.179	0.154	0.229
0.5	100	0.147	0.143	0.116	0.117	0.114	0.125	0.092	0.096	0.092	0.110	0.085	0.079	0.074	0.098
	200	0.241	0.244	0.187	0.193	0.186	0.208	0.159	0.151	0.142	0.178	0.130	0.117	0.109	0.154

ρ	m	$\rho_d = 0$				$\rho_d = 10\%$				$\rho_d = 20\%$				$\rho_d = 30\%$			
		Emp		Asym		Emp		Asym		Emp		Asym		Emp		Asym	
		ANS	ASQ	ANS	ASQ	ANS	ASQ	ANS	ASQ	ANS	ASQ	ANS	ASQ	ANS	ASQ	ANS	ASQ
0.75	300	0.337	0.341	0.274	0.267	0.288	0.257	0.267	0.257	0.288	0.204	0.205	0.192	0.245	0.177	0.155	0.143
	400	0.434	0.432	0.343	0.338	0.367	0.327	0.338	0.327	0.367	0.257	0.258	0.241	0.311	0.187	0.192	0.176
	500	0.519	0.516	0.417	0.407	0.440	0.393	0.407	0.393	0.440	0.314	0.310	0.290	0.375	0.240	0.230	0.209
	100	0.249	0.244	0.188	0.189	0.208	0.186	0.189	0.186	0.208	0.147	0.146	0.142	0.178	0.128	0.113	0.109
	200	0.437	0.432	0.321	0.332	0.367	0.327	0.332	0.327	0.367	0.269	0.248	0.241	0.311	0.183	0.183	0.176
	300	0.587	0.591	0.447	0.463	0.509	0.456	0.463	0.456	0.509	0.344	0.348	0.337	0.435	0.254	0.253	0.242
	400	0.709	0.716	0.585	0.578	0.629	0.570	0.578	0.570	0.629	0.425	0.441	0.428	0.546	0.322	0.321	0.307
	500	0.808	0.807	0.692	0.674	0.726	0.666	0.674	0.666	0.726	0.522	0.526	0.512	0.640	0.374	0.386	0.370
																	0.558

Table 2b

Sample size estimates (empirical power from 2000 simulations) using the MLE based (Asym), the average sum of squares (ASQ) and average number of subjects (ANS) methods under different drop out rate p_d , correlation ρ and power

		power = 0.80			power = 0.90		
ρ	p_d	Asym	ASQ	ANS	Asym	ASQ	ANS
$n = 2$, Z is Bernoulli with mean 0.5, $\beta_3 = 0.3$							
0.25	0	524 (0.808)	524 (0.793)	524 (0.803)	701 (0.905)	701 (0.903)	701 (0.907)
	0.1	560 (0.789)	582 (0.815)	551 (0.787)	750 (0.897)	779 (0.915)	738 (0.892)
	0.2	606 (0.805)	655 (0.806)	582 (0.765)	810 (0.905)	876 (0.918)	779 (0.888)
	0.3	664 (0.812)	748 (0.856)	616 (0.774)	889 (0.897)	1001 (0.928)	825 (0.883)
0.5	0	349 (0.798)	349 (0.785)	349 (0.800)	467 (0.898)	467 (0.897)	467 (0.905)
	0.1	378 (0.806)	388 (0.804)	368 (0.797)	506 (0.898)	519 (0.905)	492 (0.890)
	0.2	415 (0.788)	437 (0.818)	388 (0.773)	555 (0.898)	584 (0.911)	519 (0.894)
	0.3	461 (0.805)	499 (0.823)	411 (0.755)	618 (0.896)	668 (0.919)	550 (0.861)
0.75	0	175 (0.803)	175 (0.795)	175 (0.798)	234 (0.897)	234 (0.903)	234 (0.894)
	0.1	192 (0.807)	194 (0.810)	184 (0.774)	257 (0.909)	260 (0.909)	246 (0.888)
	0.2	213 (0.791)	219 (0.814)	194 (0.755)	285 (0.896)	292 (0.907)	260 (0.873)
	0.3	240 (0.803)	250 (0.822)	206 (0.733)	322 (0.891)	334 (0.911)	275 (0.861)
$n = 5$, Z is standard normal, $\beta_3 = 0.04$							
0.25	0	368 (0.789)	368 (0.780)	368 (0.783)	493 (0.896)	493 (0.891)	493 (0.902)
	0.1	480 (0.777)	516 (0.818)	450 (0.766)	642 (0.896)	691 (0.915)	602 (0.870)
	0.2	646 (0.804)	745 (0.864)	548 (0.730)	864 (0.905)	997 (0.927)	733 (0.859)
	0.3	904 (0.807)	1111 (0.874)	664 (0.664)	1211 (0.901)	1487 (0.950)	889 (0.783)
0.5	0	246 (0.783)	246 (0.791)	246 (0.805)	329 (0.891)	329 (0.898)	329 (0.903)
	0.1	330 (0.791)	344 (0.806)	300 (0.775)	442 (0.889)	461 (0.915)	401 (0.877)
	0.2	458 (0.790)	497 (0.852)	365 (0.717)	613 (0.902)	665 (0.920)	489 (0.845)
	0.3	660 (0.801)	741 (0.840)	443 (0.625)	883 (0.891)	991 (0.924)	593 (0.734)
0.75	0	123 (0.797)	123 (0.788)	123 (0.792)	165 (0.902)	165 (0.888)	165 (0.897)
	0.1	169 (0.775)	172 (0.810)	150 (0.741)	226 (0.881)	231 (0.900)	201 (0.866)
	0.2	240 (0.787)	249 (0.810)	183 (0.678)	321 (0.890)	333 (0.897)	245 (0.805)
	0.3	352 (0.788)	371 (0.818)	222 (0.586)	471 (0.899)	496 (0.907)	297 (0.722)

Empirical power from 2000 simulations using the MLE based (Asym), the average sum of squares (ASQ) and average number of subjects (ANS) methods under extremely large constant drop out rate p_d

Table 3a

$p_d = 50\%$			$p_d = 70\%$			$p_d = 90\%$		
ρ	Emp	Asym	ASQ	ANS	Emp	Asym	ASQ	ANS
$n = 2$ ($m = 500$), Z is Bernoulli with mean 0.5, $\beta_3 = -0.3$								
0.25	0.562	0.575	0.491	0.660	0.400	0.416	0.323	0.598
0.50	0.714	0.718	0.660	0.828	0.530	0.525	0.451	0.772
0.75	0.935	0.934	0.918	0.984	0.774	0.776	0.738	0.969
$n = 5$ ($m = 10000$), Z is standard normal, $\beta_3 = -0.02$								
0.25	0.864	0.871	0.754	0.995	0.427	0.432	0.313	0.974
0.50	0.951	0.942	0.900	0.999	0.510	0.522	0.438	0.998
0.75	0.998	0.997	0.996	1.000	0.763	0.764	0.723	1.000

Table 3b

Sample size estimates (empirical power from 2000 simulations) given 80% power using the MLE based (Asym), the average sum of squares (ASQ) and average number of subjects (ANS) methods under extremely large constant drop out rate p_d

	$p_d = 50\%$			$p_d = 70\%$			$p_d = 90\%$		
ρ	Asym	ASQ	ANS	Asym	ASQ	ANS	Asym	ASQ	ANS
$n = 2$, Z is Bernoulli with mean 0.5, $\beta_3 = 0.3$									
0.25	851 (0.798)	1047 (0.879)	698 (0.724)	1287 (0.796)	1745 (0.899)	806 (0.609)	3467 (0.788)	5233 (0.935)	952 (0.309)
0.5	611 (0.799)	698 (0.839)	466 (0.678)	960 (0.789)	1163 (0.867)	537 (0.546)	2704 (0.802)	3489 (0.890)	635 (0.269)
0.75	328 (0.785)	349 (0.812)	233 (0.661)	531 (0.789)	582 (0.822)	269 (0.518)	1548 (0.793)	1745 (0.837)	318 (0.260)
$n = 5$, Z is standard normal, $\beta_3 = 0.05$									
0.25	1315 (0.780)	1795 (0.907)	608 (0.484)	3924 (0.786)	5793 (0.922)	827 (0.255)	21997 (0.792)	34375 (0.938)	1060 (0.109)
0.5	1006 (0.786)	1197 (0.856)	406 (0.411)	3093 (0.791)	3862 (0.874)	551 (0.212)	17557 (0.796)	22917 (0.907)	707 (0.088)
0.75	555 (0.798)	599 (0.817)	203 (0.385)	1750 (0.793)	1931 (0.811)	276 (0.191)	10146 (0.796)	11459 (0.836)	354 (0.086)

Table 4a

Empirical power from 2000 simulations under different (π_1, π_2) when missing data is MAR (n=2)

marginal		$p_d = 10\%$				$p_d = 20\%$				$p_d = 30\%$			
ρ	m	Asym	(0.9,0.9)	(1,0.8)	Asym	(0.8,0.8)	(1,0.6)	(0.9,0.7)	Asym	(0.7,0.7)	(1,0.4)	(0.8,0.6)	
0.25	100	0.219	0.226	0.222	0.206	0.222	0.201	0.211	0.192	0.198	0.199	0.193	
	200	0.388	0.384	0.397	0.364	0.373	0.377	0.363	0.337	0.325	0.323	0.336	
	300	0.536	0.535	0.533	0.505	0.486	0.499	0.501	0.470	0.467	0.478	0.458	
	400	0.659	0.675	0.671	0.625	0.639	0.611	0.422	0.585	0.572	0.597	0.582	
	500	0.754	0.749	0.760	0.721	0.734	0.727	0.715	0.682	0.688	0.680	0.672	
0.50	100	0.302	0.298	0.290	0.280	0.293	0.269	0.273	0.256	0.258	0.254	0.239	
	200	0.531	0.547	0.544	0.495	0.488	0.475	0.486	0.454	0.452	0.453	0.442	
	300	0.704	0.708	0.712	0.664	0.659	0.657	0.682	0.618	0.625	0.613	0.623	
	400	0.822	0.832	0.806	0.786	0.790	0.790	0.796	0.742	0.748	0.741	0.742	
	500	0.897	0.901	0.895	0.868	0.871	0.877	0.870	0.831	0.828	0.822	0.837	
0.75	100	0.526	0.520	0.538	0.485	0.483	0.490	0.487	0.440	0.442	0.436	0.451	
	200	0.817	0.818	0.823	0.776	0.768	0.779	0.788	0.725	0.733	0.718	0.709	
	300	0.939	0.929	0.947	0.914	0.906	0.910	0.919	0.880	0.881	0.869	0.877	
	400	0.982	0.980	0.983	0.970	0.970	0.973	0.970	0.951	0.948	0.954	0.948	
	500	0.995	0.995	0.994	0.990	0.989	0.987	0.993	0.982	0.985	0.979	0.980	

Table 4b

Empirical power from 2000 simulations under different (π_1, π_2) when missing data is MAR ($n=5$) (all correspond to marginal distribution $p = (0.5, 0, 0, 0, 0.5)$)

ρ	m	Asym	(0.5,0.5)	(0.3,0.7)	(0.2, 0.8)	(0.1, 0.9)
0.25	100	0.082	0.083	0.090	0.083	0.094
	200	0.124	0.105	0.140	0.132	0.123
	300	0.165	0.170	0.155	0.163	0.171
	400	0.206	0.194	0.193	0.214	0.222
	500	0.246	0.234	0.247	0.257	0.256
0.50	100	0.099	0.100	0.116	0.109	0.099
	200	0.156	0.171	0.138	0.159	0.159
	300	0.212	0.212	0.210	0.206	0.216
	400	0.268	0.272	0.260	0.268	0.262
	500	0.323	0.317	0.343	0.335	0.326
0.75	100	0.149	0.150	0.136	0.153	0.144
	200	0.255	0.253	0.279	0.274	0.260
	300	0.356	0.342	0.367	0.349	0.361
	400	0.451	0.455	0.458	0.430	0.448
	500	0.538	0.548	0.538	0.514	0.539