Research Article

Statistics in Medicine

Received 10 July 2013, Accepted 1 October 2015 Published online 24 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6811

Multivariate test power approximations for balanced linear mixed models in studies with missing data

Brandy M. Ringham,[a*†] Sarah M. Kreidler,[b] Keith E. Muller[c] and Deborah H. Glueck[a]

Multilevel and longitudinal studies are frequently subject to missing data. For example, biomarker studies for oral cancer may involve multiple assays for each participant. Assays may fail, resulting in missing data values that can be assumed to be missing completely at random. Catellier and Muller proposed a data analytic technique to account for data missing at random in multilevel and longitudinal studies. They suggested modifying the degrees of freedom for both the Hotelling–Lawley trace *F* statistic and its null case reference distribution. We propose parallel adjustments to approximate power for this multivariate test in studies with missing data. The power approximations use a modified non-central *F* statistic, which is a function of (i) the expected number of complete cases, (ii) the expected number of non-missing pairs of responses, or (iii) the trimmed sample size, which is the planned sample size reduced by the anticipated proportion of missing data. The accuracy of the method is assessed by comparing the theoretical results to the Monte Carlo simulated power for the Catellier and Muller multivariate test. Over all experimental conditions, the closest approximation to the empirical power of the Catellier and Muller multivariate test is obtained by adjusting power calculations with the expected number of complete cases. The utility of the method is demonstrated with a multivariate power analysis for a hypothetical oral cancer biomarkers study. We describe how to implement the method using standard, commercially available software products and give example code. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: balanced linear mixed models; data missing completely at random; Hotelling–Lawley trace power approximation; multilevel and longitudinal studies

1. Introduction

1.1. Balanced linear mixed models for complete and missing data

Mixed models are used often in biomedical research to account for correlations between repeated measures of outcomes, even with missing data. Throughout, we use the term 'repeated measures' in the most general sense as multiple measurements of any form taken on a single independent sampling unit. Repeated measurements can be taken across time (longitudinal) or space (spatial), within clusters (multilevel), using different scales (multivariate outcomes), or in designs with combinations of these factors. For convenience, we will call each measurement taken on an independent sampling unit a 'unit of observation'. Each independent sampling unit may have several such units of observation.

In this work, we consider the problem of calculating power for experiments, which may have outcome data missing completely at random (MCAR). We limit the discussion to balanced linear mixed models as described by Muller and co-authors [1, 2]. We describe the model for the complete data first, and then explain how the missing data process may affect power calculations.

Muller *et al.* [1] make the following assumptions about balanced linear mixed models: (i) each independent sampling unit has the same number and choice of response variables, (ii) each independent sampling

[a]*Department of Biostatistics and Informatics, University of Colorado Denver, Aurora, CO, U.S.A.*
[b]*Neptune and Company, Lakewood, CO, U.S.A.*
[c]*Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, U.S.A.*
**Correspondence to: Brandy M. Ringham, University of Colorado Anschutz Medical Campus, 13001 E. 17th Place, B119, Bldg 500, Room W3125 Aurora, CO 80045, U.S.A.*
[†]*E-mail: brandy.ringham@ucdenver.edu*

unit has the same covariance structure, and (iii) covariates have the same value for the entire independent sampling unit, no matter which unit of observation is considered. Note that we shall relax Assumption 1 shortly to allow for missing data. Assumption 3 corresponds to using a single value of each predictor for each independent sampling unit. This means that the models we consider cannot have time-varying or space-varying covariates in longitudinal or spatial studies, different treatment assignments within clusters for multilevel studies, or different predictors for various outcome variables in multivariate studies. To clarify the definition of balanced linear mixed models used in this work, we present examples in the succeeding text that include descriptions of study design, study goals, independent sampling units, units of observation, and predictors.

An example of a longitudinal study with complete data for which a balanced linear mixed model is appropriate is a study of the shape of the weight gain curve of women during pregnancy. Each woman is measured nine times on the first day of each month of gestation. Each woman is an independent sampling unit, and the unit of observation is the weight of the woman at each month. We assume that the covariance between the repeated measurements is the same for all women. The model coefficients represent the average weight at each time point, if only an intercept is used as a predictor. Assumption 3 means that the model could include pre-pregnant body mass index, which has the same value for each of the nine weight measurements for one participant. If, however, the investigators wished to add a time-varying covariate such as total dietary fat intake measured for each month of gestation for each pregnant woman, a balanced linear mixed model would not be appropriate. The investigators would need a more complex modeling strategy.

An example of a single level, group-randomized trial with complete data for which a balanced linear mixed model is appropriate is a trial comparing the effectiveness of two workplace alcohol reduction programs. Here, we assume that each workplace is randomized to one of the two programs, and that the workplace is the independent sampling unit. Each workplace has many workers. The unit of observation is the self-reported average amount of alcohol consumed by each worker in each workplace in the week following the program. The group-randomized trial can be modeled using a balanced linear mixed model if every group (or cluster) includes the same number of participants, if the covariance structure is the same in each workplace, and if the predictors in the model are the same for every worker in a single workplace. In the balanced linear mixed model describing such an experiment, the coefficients are estimates of the average amount of alcohol consumed for each alcohol reduction program. Because workplaces rather than workers are randomized to a program, the main predictor, treatment group, can be used in the balanced linear mixed model. If, however, the investigators wanted to adjust for previous participation in an alcohol reduction program, they would need to use a different modeling strategy because the participation variable has a potentially different value for each worker.

If the workplaces are clustered within neighborhoods, the design becomes a multi-level trial. As long as the same number of workplaces appear in each neighborhood, the correlation between and within workplaces is the same across neighborhoods, and the randomization is at the neighborhood level, the resulting trial can be modeled with a balanced linear mixed model.

The final example, the one for which we present a power analysis in Section 6 of this paper, is a study with multiple outcomes. The goal of the study is to assess the diagnostic value of three oral cancer screening biomarkers, measured on each participant. Here, the study participant is the independent sampling unit, and the unit of observation is the measurement of each biomarker. Each biomarker is sampled from individuals, some with diagnosed cancer of the oral cavity or pharynx (cases) and some with no previous oral cancer diagnoses (controls). The predictors in the model are indicator variables for cases and controls. The study investigator plans to compare biomarker levels between cases and controls and needs to account for the correlation between biomarkers and potentially missing data points. A balanced linear mixed model is appropriate for this study because the same number and type of measurements are taken on each participant, the disease is diagnosed on the level of the participant, and one can assume that the covariance structure of the biomarkers is the same for each participant. The coefficients of the model are the average levels of each biomarker for those with cancer, and those without cancer. When using a balanced linear mixed model, the investigator can include participant level information as predictors or covariates, such as the disease state. However, the investigator would need a different and more complex model if they would like to adjust for a variable measured on the biomarker level.

We have described balanced linear mixed models for experiments with complete data. However, in many longitudinal, spatial, multivariate, and multilevel experiments, missing outcome data is common. In this manuscript, we describe power analysis for experiments that when planned, fulfill all the assumptions

for a balanced linear mixed model, but when observed, may have missing outcome data. This relaxes Assumption 1: the observed experiment may have a different number of units of observation for each independent sampling unit. For the example with women during pregnancy, this means that each woman may have a different number of weight measurements, a subset of the full nine months of weight measurements. For the workplace alcohol treatment experiment, it means that each workplace may have a different number of workers. For the biomarkers example, it means that each study participant may have one, two, or three of the planned biomarker values.

We assume that the data is MCAR [3] with some constant probability, π. The MCAR assumption means that the chance that each unit of observation is missing is unrelated to observed or unobserved characteristics of the independent sampling unit, or the temporal or spatial details of the unit of observation. The MCAR assumption is appropriate when data are missing due to a random event, such as participants relocating or a failure of the instruments used to conduct the experiment. The assumption is not appropriate if data are missing due to some process that is correlated with the data values themselves. An example of violating the MCAR assumption occurs if an instrument only records values above a certain level or if participants drop out of a therapeutic drug study because the drug worsens symptoms more than placebo. We provide a more formal definition of the missing data process in the Methods section.

1.2. Multivariate hypothesis testing for balanced linear mixed models

A standard data analysis approach for studies with multiple correlated outcomes is to use a mixed model Wald test with Kenward–Roger degrees of freedom [4]. The test may be used for studies with either complete or missing data. However, mixed model data analysis using the Wald test has two problems. Depending on the experiment, the models may have a low rate of convergence or an inflated Type I error rate. Convergence problems may occur because of the difficulty of estimating the multiple parameters needed for an unstructured covariance structure (Table I) [2, 5–7]. The observed Type I error rate may exceed the target Type I error rate if the analyst misspecifies the true covariance (Table II) [7]. Inflation occurs in simulation studies even with sample sizes of 100, as in the experiment considered by Gurka *et al*. [7] (shown in Table II). More importantly, Gurka *et al*. [7] showed mathematically that the inflation can occur with infinitely large sample sizes.

For balanced linear mixed models with complete outcome data, Muller *et al.* recommend that researchers choose a multivariate hypothesis test instead of the Wald test. They argue that a multivariate test 'always controls test size and has a good power approximation, in sharp contrast to mixed model tests' [1]. In studies with complete outcome data and a balanced design, the mixed model can be recast as a general linear multivariate model, and the mixed model Wald test with Kenward–Roger degrees of freedom becomes equivalent to the multivariate Hotelling–Lawley trace test.

Calculation of the Hotelling–Lawley trace statistic requires complete data, limiting the utility of this multivariate approach. To address this gap, Catellier and Muller [2] provided a modification to the

Table I. Convergence rates for mixed models.

Study	% Converged
Catellier and Muller (2000)	10
Serrano (2008)	37
Gurka, Edwards, Muller (2011)	61
Fouladi and Shieh (2004)	84

Table II. Type I error for mixed model with correct and incorrect covariance models, $\alpha = 0.05$, $N_t = 100$, $p = 5$, 20% missing data.

	Type I error	
Correlation model	Correct	Incorrect
Autoregressive (AR)	0.047	0.064
Linear in time	0.048	0.088
Linear in time & AR	0.044	0.117

Hotelling–Lawley trace test reference distribution. The modification permits the use of the multivariate approach even for studies with missing outcome data, so long as the planned study analysis fits the assumptions for a balanced linear mixed model. The Catellier and Muller [2] approach controls the Type I error rate in many experimental scenarios, even in the presence of missing outcome data.

1.3. Multivariate power approximations for balanced linear mixed models

A multivariate data analysis requires an aligned multivariate power analysis [8]. Current data and power analysis techniques for balanced linear mixed models and multivariate hypothesis tests assume complete data [1]. For analysts facing the possibility of missing outcome data, and choosing to use a Catellier and Muller multivariate test, a new power method is needed. In this work, we propose new power approximations for the Catellier and Muller multivariate test. We compare members of a class of power approximations and suggest a specific power approximation that yields power values with accuracy to the second decimal place for many common experimental designs with missing data.

In the current work, the new power approximations are described in eight sections. Section 2 contains general notation. Section 3 reviews known methods for data analysis of the balanced linear mixed model, both with complete data and missing data. Section 4 describes new power approximations for the Catellier and Muller test for the balanced linear mixed model with potentially missing data. In addition, code is provided for implementation of the method. Section 5 presents simulation results for the approximations. Section 6 demonstrates the utility of the method for planning an oral cancer biomarkers study. Section 7 discusses the implications of the work and provides recommendations for implementing the method, including guidance on executing the method using commercial software. Section 8 describes future directions of the research.

2. General notation

Throughout, notation is similar to that used in Muller and Stewart [9]. Let $\boldsymbol{A} = \{a_{ij}\}$ be a matrix with dimensions $(r \times c)$ and transpose $\boldsymbol{A}' = \{a_{ji}\}$. Indicate row i of $\boldsymbol{A}$ as $\boldsymbol{A}_i$ and column j of $\boldsymbol{A}$ as $\boldsymbol{a}_j$. Let $\text{vec}(\boldsymbol{A}) = \left[\boldsymbol{A}_1\ \boldsymbol{A}_2\ \ldots\ \boldsymbol{A}_n\right]'$. The direct product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is $\boldsymbol{A} \otimes \boldsymbol{B} = \{a_{ij}\boldsymbol{B}\}$. Let $\boldsymbol{I}_n$ denote the identity matrix with dimensions $(n \times n)$. Write $\mathbf{1}_{(r,c)}$ and $\mathbf{0}_{(r,c)}$ to denote $(r \times c)$ matrices with all elements equal to 1 and 0, respectively. Denote the trace of $\boldsymbol{A}$ as $\text{tr}(\boldsymbol{A})$. The rank of $\boldsymbol{A}$ is indicated by $\text{rank}(\boldsymbol{A})$. For $\boldsymbol{A}$ square and full rank, denote $\boldsymbol{A}^{-1}$ as the unique and full-rank inverse. Write the expected value of the random variable X as $E(X)$.

Let $X \sim F(\nu_1, \nu_2, \omega)$ indicate that the random variable X has a non-central F distribution with numerator degrees of freedom ν_1, denominator degrees of freedom ν_2, and non-centrality parameter ω. Write $F_F(x; \nu_1, \nu_2, \omega)$ to indicate the probability that $X \sim F(\nu_1, \nu_2, \omega)$ falls in the interval $[0, x)$. Similarly, write $f = F_F^{-1}(p; \nu_1, \nu_2, \omega)$ to indicate that $F_F(f; \nu_1, \nu_2, \omega) = p$ for probability $p \in [0, 1]$ [9]. Similar notation is used for the central F distribution, the only difference being the absence of the non-centrality parameter.

3. Known methods for multivariate hypothesis testing for balanced linear mixed models

3.1. Multivariate hypothesis testing with complete data

Consider a complete balanced linear mixed model [1] with N_t independent sampling units [9, p. 101] and p repeated measures on each independent sampling unit. Let $\boldsymbol{Y} = \{y_{ij}\}$ be an $(N_t \times p)$ response matrix with $p \ll N_t$, $\boldsymbol{X}$ an $(N_t \times q)$ matrix of fixed effects of rank $r \leqslant q$, $\boldsymbol{B}$ a $(q \times p)$ matrix of fixed effect coefficients, $\boldsymbol{\Sigma}$ a $(p \times p)$ full rank, finite, positive definite, symmetric matrix, and $\boldsymbol{E}$ an $(N_t \times p)$ error matrix. Let $\boldsymbol{C}$ be an $(a \times q)$ contrast matrix for comparisons made between independent sampling units, and $\boldsymbol{U}$ a $(p \times b)$ contrast matrix for comparisons made between the repeated measures within an independent sampling unit [8].

Under the assumption that $\text{vec}(\boldsymbol{E}') \sim \mathcal{N}_{N_t p}(\mathbf{0}, \boldsymbol{I}_{N_t} \otimes \boldsymbol{\Sigma})$, the complete balanced mixed model can be written as

$$\text{vec}(\boldsymbol{Y}') = (\boldsymbol{X} \otimes \boldsymbol{I}_p)\,\text{vec}(\boldsymbol{B}') + \text{vec}(\boldsymbol{E}'). \quad (1)$$

The complete balanced mixed model in Equation (1) can be written as an equivalent general linear multivariate model [9, p. 245] as shown in the following:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}. \tag{2}$$

For either model form, the Hotelling–Lawley trace test for the general linear hypothesis $H_0 : \boldsymbol{CBU} = \boldsymbol{\Theta} = \boldsymbol{\Theta}_0$ can be tested using $K\left(\nu_e\right) = \mathrm{tr}\left[\boldsymbol{S}_h\boldsymbol{S}_e^{-1}\left(\nu_e\right)\right]$. Here, $\nu_e = N_t - r$, $\boldsymbol{S}_e\left(\nu_e\right) = \nu_e\boldsymbol{U}'\hat{\boldsymbol{\Sigma}}\boldsymbol{U}$, $\hat{\boldsymbol{\Sigma}}\left(\nu_e\right) = \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)'\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)/\nu_e$, and $\boldsymbol{S}_h = \left(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0\right)'\left[\boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{C}'\right]^{-1}\left(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0\right)$ [1]. Under the null hypothesis, the test statistic

$$F\left(\nu_e\right) = \frac{\left(\nu_e - b - 1\right)K\left(\nu_e\right)/\nu_1}{\left[\nu_2\left(\nu_e\right) - 2\right]/\nu_2\left(\nu_e\right)} \tag{3}$$

has an approximate central $F\left[\nu_1, \nu_2\left(\nu_e\right)\right]$ distribution [10], with $\nu_1 = ab$ and

$$\nu_2\left(\nu_e\right) = 4 + \frac{\left(ab+2\right)\left[\nu_e^2 - \nu_e\left(2b+3\right) + b\left(b+3\right)\right]}{\nu_e\left(a+b+1\right) - \left(a+2b+b^2-1\right)}. \tag{4}$$

3.2. Multivariate hypothesis testing with missing data

Catellier and Muller [2] suggested a data analytic technique for the general linear multivariate model with missing data only in $\boldsymbol{Y}$. The approach maintains accurate Type I error rate in small samples.

To implement the approach, define missing data summary statistics N_{mk}, for $k \in \{1, 2, 9\}$ as follows. Here, the subscripts parallel Catellier and Muller [2, Table I]. For $\boldsymbol{D} = \left\{d_{ij}\right\}$, let $d_{ij} = 1$ if y_{ij} is non-missing and 0 otherwise. Let $\mathcal{S} = \left\{\left(j, j'\right) : j \in \left\{1, 2, \ldots, j'-1\right\} \cap j' \in \left\{1, 2, \ldots, p\right\}\right\}$. The number of non-missing pairs of observations in columns j and j' of $\boldsymbol{Y}$ is $N_{jj'} = \boldsymbol{d}_j'\boldsymbol{d}_{j'}$. Let $N_{m1} = \sum_{i=1}^{N_t}\prod_{j=1}^{p} d_{ij}$, $N_{m2} = \min_{\mathcal{S}}\left\{N_{jj'}\right\}$, and $N_{m9} = \bar{N}_{jj'}$. Notice that with complete data, N_{m1}, N_{m2}, and N_{m9} all collapse to N_t.

With $\nu_{mk} = \left(N_{mk} - r\right)$, three possible test statistics for missing data are

$$F\left(\nu_{mk}\right) = \frac{\left(\nu_{mk} - b - 1\right)K\left(\nu_{mk}\right)/\nu_1}{\left[\nu_2\left(\nu_{mk}\right) - 2\right]/\nu_2\left(\nu_{mk}\right)}, \tag{5}$$

which have approximate central $F\left[\nu_1, \nu_2\left(\nu_{mk}\right)\right]$ distributions [2], with $\nu_1 = ab$ and

$$\nu_2\left(\nu_{mk}\right) = 4 + \frac{\left(ab+2\right)\left[\nu_{mk}^2 - \nu_{mk}\left(2b+3\right) + b\left(b+3\right)\right]}{\nu_{mk}\left(a+b+1\right) - \left(a+2b+b^2-1\right)}. \tag{6}$$

Of the set of 11 statistics considered by Catellier and Muller, we focus only on N_{m1}, N_{m2}, and N_{m9}. The missing data summary statistic N_{m1} provides a lower bound for the effective sample size. Catellier and Muller [2] recommended N_{m2} to control test size in data analyses using the Hotelling–Lawley trace. Tu *et al.* [11] suggested using $E\left(N_{m9}\right)$, the trimmed sample size, to calculate an upper bound for power.

4. New multivariate power approximations for balanced mixed models with complete or missing data

For complete data, Muller *et al.* [1, 8] proposed calculating power for the Hotelling–Lawley trace using a non-central F approximation. With ν_1 and $\nu_2\left(\nu_e\right)$ as described in Section 3.1, Muller *et al.* [1] defined $f_{\mathrm{crit}} \approx F_F^{-1}\left[1-\alpha;\ \nu_1,\ \nu_2\left(\nu_e\right)\right]$ and the non-centrality parameter as $\omega\left(\nu_e\right) = \nu_e K\left(\nu_e\right)$. Power was approximated as $P\left(\nu_e\right) = 1 - F_F\left[f_{\mathrm{crit}}\left(\nu_e\right); \nu_1, \nu_2\left(\nu_e\right), \omega\left(\nu_e\right)\right]$.

For power calculations, the missing data process must be explicitly defined. With $\pi \in [0, 1)$, the population proportion of missing data assumes that d_{ij} are independently and identically distributed Bernoulli (π) random variables. Further assume that d_{ij} are independent of the values in $\boldsymbol{Y}$. Note that the assumed missing data process gives rise to data that are MCAR [3].

If the process that creates missing data is random, and independent of the values of the data itself, one can imagine a series of possible realizations of $\boldsymbol{D}$. Each realization of $\boldsymbol{D}$ corresponds to a $\boldsymbol{Y}$ matrix with a varying number of missing values, at varying locations. In turn, each $\boldsymbol{Y}$ matrix yields a different power for the experiment. Calculating the expected power yields the average power over all possible realizations of $\boldsymbol{D}$.

Under the assumption that the power function is approximately linear in N_{mk} in a neighborhood of $E\left(N_{mk}\right)$, approximate $E\left[P\left(N_{mk}\right)\right] \approx P\left[E\left(N_{mk}\right)\right]$, where the expectation is calculated over the distribution of $\boldsymbol{D} = \left\{d_{ij}\right\}$.

It can be shown that

$$E\left(N_{m1}\right) = N_t\left(1-\pi\right)^p \tag{7}$$

and

$$E\left(N_{m9}\right) = N_t\left(1-\pi\right). \tag{8}$$

We provide the results of a regression model for $E\left(N_{m2}\right)$ (Table A.1, Appendix A) as the calculation proved intractable. Notice that for complete data, $E\left(N_{m1}\right)$, $E\left(N_{m2}\right)$, and $E\left(N_{m9}\right)$ reduce to N_t. A SAS/IML module to calculate $E\left(N_{m1}\right)$, $E\left(N_{m2}\right)$, and $E\left(N_{m9}\right)$ is included in Appendix C. In addition, a version of the free, open-source code appears at *www.SampleSizeShop.org*.

Each N_{mk} yields a separate power approximation. Calculate the approximations as follows:

Step 1: For the null hypothesis $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, define values for α, $\boldsymbol{X}$, $\boldsymbol{C}$, $\boldsymbol{U}$, $\boldsymbol{B}$, and $\boldsymbol{\Sigma}$.
Step 2: For $\nu_{*mk} = E\left(N_{mk}\right) - r$, calculate $f_{\text{crit}}\left(\nu_{*mk}\right) \approx F_F^{-1}\left\{1-\alpha;\ \nu_1,\ \nu_2\left(\nu_{*mk}\right)\right\}$.
Step 3: Calculate the non-centrality parameter, $\omega\left(\nu_{*mk}\right) = \nu_{*mk}K\left(\nu_{*mk}\right)$.
Step 4: Calculate power as a function of the non-central F distribution as

$$\text{Power}\left(\nu_{*mk}\right) \approx 1 - F_F\left[f_{\text{crit}}\left(\nu_{*mk}\right); \nu_1, \nu_2\left(\nu_{*mk}\right), \omega\left(\nu_{*mk}\right)\right]. \tag{9}$$

5. Numerical evaluations

5.1. Methods

To evaluate the accuracy of the power adjustment, theoretical power was compared with simulated empirical power for a range of experimental designs. As in Catellier and Muller [2], the designs defined $\alpha = 0.05$, $\pi \in \{0,\ 0.05,\ 0.10\}$, and hence expected percentage of missing data $(100 \cdot \pi) \in \{0\%,\ 5\%,\ 10\%\}$, $p \in \{3,\ 6\}$, $N_t \in \{12, 24, 48, 96, 192, 384\}$, $\boldsymbol{C} = \left[\boldsymbol{0}_{(3,\,1)}\ \boldsymbol{I}_3\right]$ and $\boldsymbol{U} = \boldsymbol{I}_p$. The case with $N_t = 12$, and $p = 6$ was omitted due to being implausibly small. For predictors, $\boldsymbol{X} = \boldsymbol{X}_e \otimes \boldsymbol{1}_{(N_t/4,\,1)}$ with $\boldsymbol{X}_e$ a (4×4) matrix containing an intercept, and the linear, quadratic, and cubic orthogonal trends over the repeated measures. The hypothesis of interest was the presence of a linear, quadratic, or cubic trend over the response variables for each group.

Choices for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ were modeled after Catellier and Muller [2] and Barton and Cramer [12]. Other factors included the diagonal elements of $\boldsymbol{\Sigma}$ (either equal, or unequal), and the correlation between the dependent variables (low or high). Values for $\boldsymbol{\Sigma}$ appear in Appendix B. For concentrated non-centrality [13], $\boldsymbol{B} = \boldsymbol{B}_c\left(p\right) = \Delta\left[\boldsymbol{0}_{(p,\,p)}\ \boldsymbol{1}_{(p,\,1)}\right]'$. For diffuse non-centrality [13], $\boldsymbol{B} = \boldsymbol{B}_d\left(p\right) = \Delta\left[\boldsymbol{0}_{(p,\,1)}\ \boldsymbol{\Sigma}^{1/2}\right]'$. In both cases, Δ was chosen to give power values $\in \{0.20,\ 0.50,\ 0.80,\ 0.90\}$ for complete data.

For each experimental design, theoretical power was calculated as in Equation (9) using three separate adjustments: $E\left(N_{m1}\right)$, $E\left(N_{m2}\right)$, and $E\left(N_{m9}\right)$. Empirical power was simulated as in Catellier and Muller [2]. For each design, a dataset was generated with random missing values. The missing data summary statistics, N_{m1}, N_{m2}, and N_{m9}, were tallied. The expectation-maximization (EM) algorithm [14] was used to impute missing data values and find estimates for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$. An observed F statistic was calculated for the Hotelling–Lawley trace with the modifications described in Section 3.2. The statistic was then compared with the modified null case reference distribution, also described in Section 3.2. For each combination of experimental factors, empirical power was calculated as the proportion of times the null hypothesis was rejected over 10,000 realizations of the data.

Experimental designs with less than 10% convergence (1000/10,000 trials) were excluded from the results. Convergence failure of the EM algorithm sometimes yields considerably fewer trials than

planned. Designs that lead to convergence failure may also cause convergence of the EM algorithm on local maxima, resulting in incorrect estimates. In addition, most data analysts prefer analytic methods with high convergence rates.

The accuracy of the power approximations was measured using either the deviation, the median deviation, or the maximum absolute deviation. The deviation was calculated as the theoretical power minus the empirical power. The median (or maximum) deviation was calculated as the median (or maximum) across experimental conditions with p, π, and complete data power held constant. Deviations, median deviations, or absolute deviations closer to zero indicated greater accuracy.

Raw deviations, rather than absolute deviations, were used in order to highlight the sign of the deviation. Similarly, the median, rather than the mean, was used because the median preserves information about signs.

5.2. *Results*

There were 26 experimental designs for which fewer than 10% of the 10,000 trials converged, approximately 3% percent of the 792 designs considered.

The effects of variance structure, sample size, and power on the deviations of the three power approximations are summarized in Table III and Figures 1 and 2, respectively. In general, the power approximation using N_{m1} had the best accuracy, with accuracy improving as sample size increased. Accuracy was about the same, no matter the complete data power, or the variance structure. For any of the three power approximations, the accuracy improved as the number of repeated measures and the percentage of missing data decreased.

The accuracy of all three power approximations was largely unaffected by changes in variance structure and type of non-centrality (Table III). However, the percentage of missing data and the number of repeated measures did affect accuracy for all three approximations. The power approximation using N_{m1} retained accuracy for designs involving $p = 3$, no matter the percentage of missing data.

Figure 1 shows the effect of sample size on the accuracy of the three power approximations. The designs considered were all chosen so that the complete data power was 90%. For designs with $p = 6$, the EM algorithm converged for fewer than 10% of trials for most designs with $N_t \leqslant 24$. Deviations were smaller for designs with fewer repeated measures and a smaller percentage of missing data. For power approximations using N_{m1}, deviations became smaller as sample size increased, with good performance by $N_t = 48$. The maximum absolute deviation for approximations using N_{m1} and designs with $p = 3$ was 0.02, even with 10% missing data. By contrast, the power approximations using N_{m2} and N_{m9} overestimated the empirical power across all sample sizes. None of the power approximations performed well for $p = 6$ and 10% missing data.

In Figure 2, the complete data power was demonstrated to have little effect on the accuracy of the power approximation using N_{m1}. Here, the focus was on designs with $N_t = 48$. For most experimental

Table III. Deviations for complete data power of 90% and $N_t = 48$.

				N_{m1}		N_{m2}		N_{m9}	
Non-centrality	π	ρ	σ^2	$p = 3$	$p = 6$	$p = 3$	$p = 6$	$p = 3$	$p = 6$
Concentrated	5	Low	=	0.022	−0.032	0.039	0.073	0.073	0.137
Concentrated	5	Low	≠	0.014	−0.034	0.032	0.071	0.065	0.135
Concentrated	5	High	≠	0.010	−0.020	0.027	0.085	0.061	0.149
Concentrated	10	Low	=	0.017	−0.114	0.060	0.152	0.139	0.286
Concentrated	10	Low	≠	0.018	−0.125	0.061	0.140	0.141	0.274
Concentrated	10	High	≠	0.019	−0.118	0.062	0.148	0.141	0.282
Diffuse	5	Low	=	−0.002	−0.042	0.015	0.063	0.049	0.127
Diffuse	5	Low	≠	0.007	−0.045	0.024	0.060	0.058	0.124
Diffuse	5	High	≠	0.006	−0.050	0.023	0.055	0.057	0.119
Diffuse	10	Low	=	0.001	−0.132	0.044	0.133	0.123	0.267
Diffuse	10	Low	≠	−0.001	−0.131	0.042	0.134	0.121	0.268
Diffuse	10	High	≠	−0.001	−0.150	0.042	0.115	0.121	0.249

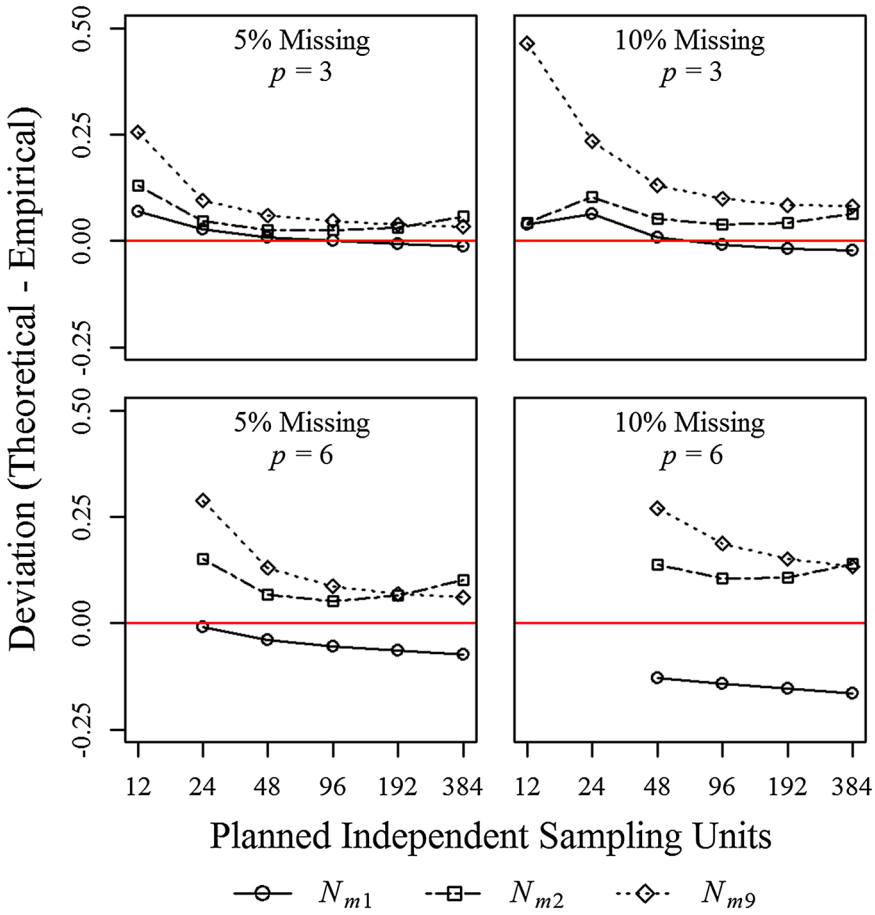

Figure 1. Median deviation for 90% complete data power.

scenarios, the power approximation using N_{m1} was within approximately 0.05 of the empirical power irrespective of the complete data power. Accuracy was best for designs with fewer repeated measures and less missing data. With $p = 6$, and 10% missing data, the performance of the approximation was poor.

6. Demonstration

To demonstrate the utility of the method, we conduct a power analysis for a hypothetical oral cancer biomarkers study. Cancer of the head and neck has a 50% five-year survival rate and up to two times the mortality rate in black males [15]. The high mortality rate is believed to be due to late stage diagnosis and treatment [15–17]. To facilitate earlier identification of disease, Elashoff *et al*. [18] evaluated 10 salivary biomarkers for the detection of oral squamous cell carcinoma. The biomarkers showed increased expression in cases over normal controls.

Suppose a researcher would like to validate three of the salivary biomarkers studied by Elashoff *et al*. [18], IL-1B, IL-8, and SAT, in the Veterans Affairs population. The researcher plans an unmatched case/control study. There will be 75 participants with diagnosed oral squamous cell carcinoma, and 75 participants without carcinoma. A saliva sample will be taken from each participant and analyzed to determine the levels of each biomarker. The researcher wishes to test the null hypothesis that there is no difference between cases and controls in the mRNA expression levels for any biomarker.

Per the recommendations of Muller *et al*. [1], the researcher plans to use a general linear multivariate model (Equation (2)) and the Hotelling–Lawley trace. The planned $\boldsymbol{Y}$ is a (150×3) matrix containing the mRNA expression levels for the three biomarkers and $\boldsymbol{X} = \boldsymbol{I}_2 \otimes \mathbf{1}_{(150/2,\,1)}$. For power analysis, the researcher plans to use $\alpha = 0.05$, $\boldsymbol{C} = \begin{bmatrix} 1 & -1 \end{bmatrix}$, $\boldsymbol{U} = \boldsymbol{I}_3$, and a compound symmetric $\boldsymbol{\Sigma} = \sigma^2 \left[\mathbf{1}_{(3,\,1)} \mathbf{1}'_{(3,\,1)} \rho + \boldsymbol{I}_3 (1 - \rho) \right]$, with $\sigma^2 = 2.9$ and $\rho = 0.4$. In addition, the researcher considers

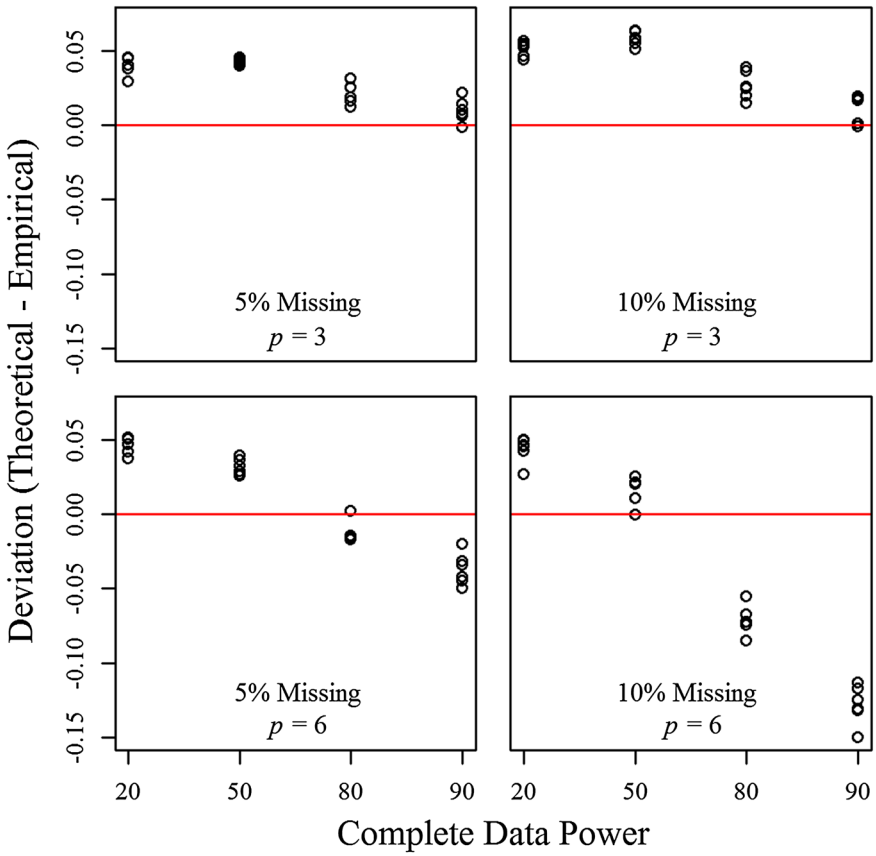

Figure 2. Deviations for power approximations using N_{m1} and experimental designs with $N_t = 48$.

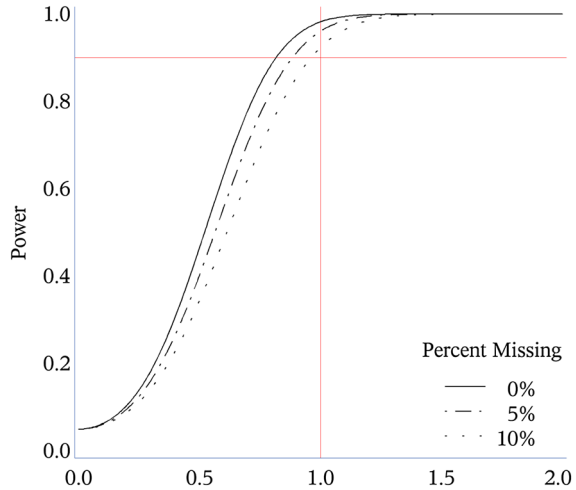

Figure 3. Power curves for a hypothetical oral cancer biomarkers study.

$\boldsymbol{B} = [\boldsymbol{\mu}_c \ \boldsymbol{\mu}_n]'$, with control means $\boldsymbol{\mu}_n' = [20.1 \ 19.8 \ 21.3]$, $\boldsymbol{\delta}' = [-1.3 \ -2.1 \ -1.4]$ and case means $\boldsymbol{\mu}_c' = K\boldsymbol{\delta}' + \boldsymbol{\mu}_n'$. The values for σ^2, $\boldsymbol{\mu}_c$, $\boldsymbol{\mu}_n$, and $\boldsymbol{\delta}$ are extrapolated from Table III in Elashoff *et al*. [18]. The parameter ρ is chosen arbitrarily. In a real power analysis, ρ would be based either on the literature or on the previous experience of the researcher. The constant K is used to vary the difference in group means. Note that, when $K = 0$, $\boldsymbol{\mu}_c = \boldsymbol{\mu}_n$ and there is no difference between the cases and controls on levels of the biomarkers. When $K = 1$, mean biomarker levels for each of the three biomarkers exactly match those of Cohort 4, in Table III of Elashoff *et al*. [18].

The power analysis must account for potential missing data. Funds for the study are limited and failed assays will not be rerun, which could result in randomly missing data points. The researcher anticipates a 6% assay failure rate [19]. To account for the missing data, the researcher will use the modified F test recommended by Catellier and Muller [2].

The researcher calculates power for the study using $E\left(N_{m1}\right)$, as suggested in Section 4. Power curves for 0%, 5%, and 10% missing data are shown in Figure 3. At the planned sample size, the study has 90% power for mean differences at least as great as those observed by Elashoff *et al*. [18] and up to 10% missing data.

Example SAS/IML code for the oral cancer biomarker power analysis is given in Appendix D.

7. Discussion

Accurate sample size calculation is a vital component of responsible research. Overestimation of the sample size can expose study participants to unnecessary risk. Underestimation of the sample size can result in studies that cannot be replicated, and which exhaust resources that could have been used for conclusive scientific progress.

The approach presented in the current work gives a general method to approximate power for balanced linear mixed models using a multivariate test. The approach extends power methods for the balanced linear mixed model with complete data [1] to similar models with missing data. The manuscript focuses on an adjusted multivariate Hotelling–Lawley trace test [2] because that test provides accurate Type I error rate, even in small samples. It is important to develop a power method aligned with the planned data analytic approach [8].

The new power approximations can be implemented using the module and example provided in Appendices C and D, respectively. Copies of the free open-source module and example also appear at *www.SampleSizeShop.org*. The implementation has two steps. First, a custom SAS/IML [20] module computes the expected values of N_{m1}, N_{m2} or N_{m9}, as desired. Next, the chosen expected value is passed into POWERLIB [21, Version 2.2]. POWERLIB then provides the chosen power approximation for the Hotelling–Lawley trace statistic.

The power approximations can also be computed using alternative, readily available software packages. The simple module to compute the expected values of N_{m1}, N_{m2} or N_{m9} can be ported easily to other programming languages, such as R [22] or MATLAB [23]. To compute the power, users may choose any power software that provides approximations for multivariate tests. Examples include PASS [24] or GLIMMPSE [25]. If the program does not allow use of fractional sample sizes, the user may need to round the expected values of N_{m1}, N_{m2} or N_{m9} to the nearest whole number.

We provide the following recommendations for researchers who choose to use the power approximations:

(1) For most experimental designs using the balanced linear mixed model, we recommend that researchers use the power approximation with N_{m1} to approximate power. The approximations using N_{m2} and N_{m9} tend to overestimate the power. By contrast, the approximation using N_{m1} is either very accurate, or provides a slight underestimation.

(2) In designs with large numbers of repeated measures and a high anticipated amount of missing data, all of the power approximations deviate substantially from the empirical power. Under these conditions, the choice of power approximation should balance the benefits of the research with the potential harms to study participants. In some cases, the study presents no great harm to participants. Thus, researchers may desire a liberal sample size calculation, one that will enroll more than enough participants. If so, the researchers should use the approximation with N_{m1}. If the study may expose participants to harm, study designers should be conservative in their approach and enroll as few participants as possible. If so, researchers should choose the approximations using N_{m2} or N_{m9}.

(3) Researchers designing studies with small sample sizes and large numbers of repeated measures who expect more than 10% missing data should consider alternatives to the Catellier and Muller [2] data analysis. For many such designs, the Catellier and Muller [2] approach fails more than 9 out of 10 times. If the variance model is known, study investigators should use the Wald test with Kenward–Roger denominator degrees of freedom [4].

8. Future work

Development of approximations for power often require consideration of multiple realizations of some random factor. Random factors may include, for example, the pattern of missing data [11, 26, current work] or a random covariate [27–29]. In this manuscript, and in the papers on random covariates, a common approach is described to account for the random factors. The expectation of the power over all possible realizations of the random factor is approximated by the power function, evaluated at the expected value of the random factor. A similar approach should work for other problems with random factors, including random and random time-varying or spatially varying covariates, or covariates that vary within groups or clusters. In addition, the approach could be applied to the Wald statistic for the mixed model with missing data.

In order to calculate the expected value of the random missing data patterns, Ringham *et al*. (in submission) assumed a specific probability process. In future work, additional missing data probability processes will be considered. One process of interest allows for correlation between d_{ij} and $d_{ij'}$. Another process of interest includes a conditional relationship between d_{ij} and $d_{i(j+1)}$, so that if y_{ij} is missing, then $y_{i(j+1)}$ is also likely to be missing. This process produces monotone missing data with a high probability.

Appendix A

Table A.1 reproduces results from Ringham *et al*. (in submission). Using the results in the table, estimate $E\left(N_{m2}\right)$ as:

$$E\left(N_{m2}\right) \approx \begin{cases} N_t & \pi = 0 \\ \sum_{i=1}^{15} \hat{\beta}_i X_i & 0 < \pi < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Table A.1. Regression coefficients for the regression of $E\left(N_{m2}\right)$ on $N_t/10$, p, and π.

i	X_i	$\hat{\beta}_i$	$\mathrm{SE}(\hat{\beta}_i)$	p
1	Intercept	62.7318676	1.39997758	<0.001
2	N_t	−5.5156768	0.48554341	<0.001
3	p	−1.6042196	0.19829228	<0.001
4	$1-\pi$	−147.7255861	2.42449211	<0.001
5	N_t^2	−0.1324363	0.03086649	<0.001
6	p^2	0.0640387	0.00512660	<0.001
7	$(1-\pi)^2$	87.3472243	1.05263141	<0.001
8	$N_t p$	−0.4981166	0.05868018	<0.001
9	$N_t(1-\pi)$	14.8545994	0.56777048	<0.001
10	$p(1-\pi)$	1.2421550	0.23187504	<0.001
11	$N_t p(1-\pi)$	0.4137540	0.06862130	<0.001
12	$N_t^2 p^2$	0.0019218	0.00035756	<0.001
13	$N_t^2(1-\pi)^2$	0.1812043	0.04124270	<0.001
14	$p^2(1-\pi)^2$	−0.0423721	0.00685029	<0.001
15	$N_t^2 p^2(1-\pi)^2$	−0.0013272	0.00047793	0.0055

Appendix B

The numerical evaluations in the current work used variances and correlations as in Barton and Cramer [12] and Catellier and Muller [2], reproduced below. Write $\boldsymbol{\sigma}^2(p, v)$, $v \in \{\text{Equal, Unequal}\}$ to specify the vector of variances used for p repeated measures and i type of variance. Similarly, let $\boldsymbol{\rho}(p, w)$, $w \in \{\text{Low, High}\}$ indicate the correlation matrix used for p repeated measures and w type of correlation. Define

$$\boldsymbol{\sigma}^2(3, \text{Equal})' = \begin{bmatrix} 3 & 3 & 3 \end{bmatrix},$$

$$\boldsymbol{\sigma}^2\,(3, \text{Unequal})' = \begin{bmatrix} 1 & 3 & 5 \end{bmatrix},$$

$$\boldsymbol{\sigma}^2\,(6, \text{Unequal})' = \begin{bmatrix} 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix},$$

$$\boldsymbol{\sigma}^2\,(6, \text{Unequal})' = \begin{bmatrix} 1 & 1 & 3 & 3 & 5 & 5 \end{bmatrix},$$

$$\boldsymbol{\rho}\,(3, \text{Low}) = \begin{bmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{bmatrix},$$

$$\boldsymbol{\rho}\,(3, \text{High}) = \begin{bmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.7 \\ 0.6 & 0.7 & 1 \end{bmatrix},$$

$$\boldsymbol{\rho}\,(6, \text{Low}) = \begin{bmatrix} 1 & 0.3 & 0.2 & 0.2 & 0.4 & 0.4 \\ 0.3 & 1 & 0.2 & 0.2 & 0.4 & 0.4 \\ 0.2 & 0.2 & 1 & 0.3 & 0.3 & 0.3 \\ 0.2 & 0.2 & 0.3 & 1 & 0.3 & 0.3 \\ 0.4 & 0.4 & 0.3 & 0.3 & 1 & 0.3 \\ 0.4 & 0.4 & 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}$$

and

$$\boldsymbol{\rho}\,(6, \text{High}) = \begin{bmatrix} 1 & 0.7 & 0.8 & 0.8 & 0.6 & 0.6 \\ 0.7 & 1 & 0.8 & 0.8 & 0.6 & 0.6 \\ 0.8 & 0.8 & 1 & 0.7 & 0.7 & 0.7 \\ 0.8 & 0.8 & 0.7 & 1 & 0.7 & 0.7 \\ 0.6 & 0.6 & 0.7 & 0.7 & 1 & 0.7 \\ 0.6 & 0.6 & 0.7 & 0.7 & 0.7 & 1 \end{bmatrix}.$$

Appendix C

The following SAS/IML module computes the expected value of N_{m1}, N_{m2}, or N_{m9}. An example call is listed in the 'Usage' line of the header comment.

```
/*************************************************************************
Program:      NMK.sas
Date:         6/28/2013
Created By:   Brandy Ringham
Description:  This module calculates the expected value of N_m1, N_m2, and
              N_m9 (Catellier and Muller, 2000; Ringham et al., in
              submission).

      Copyright (C) 2010 Regents of the University of Colorado. This program is
      free software; you can redistribute it and/or modify it under the terms
      of the GNU General Public License as published by the Free Software
      Foundation; either version 2 of the License, or (at your option) any later
      version. This program is distributed in the hope that it will be useful,
      but WITHOUT ANY WARRANTY; without even the implied warranty of
      MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General
      Public License for more details. You should have received a copy of the
      GNU General Public License along with this program; if not, write to the
      Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston,
      MA 02110-1301, USA.


Input:          Nt    number of independent sampling units
                p     number of repeated measures
                pi    probability missing value
```

```
                k     index for missing data summary statistic (see index key)
                print 1 = print on, otherwise printing is turned off

Output:         ENmk    label for mean of the missing data summary statistic
                error   label for error code matrix

Error Code Key: The error matrix contains error indicators for the inputs
                that are checked in the beginning of the module. Error
                indicators are defined as follows.

                0     no error
                1     error

                The error indicator's position in the error matrix defines the
                referent input.

                Row     Input
                1     error state of Nt
                2     error state of p
                3     error state of pi
                4     error state of k

Usage: call NMK( 12, 4, 0.10, 1, 1, ENm1, error );
*************************************************************************/

/*begin module definition*/
start NMK( Nt, p, pi, k, print, ENmk, error );
       /*initialize error indicators to 0*/
       error = j( 4, 1, 0 );
       /*check that Nt is valid*/
       if Nt < 0 then do;
               error[ 1, 1 ] = 1;
               if print = 1 then do;
                        print "invalid number of independent sampling units";
                        print "Nt should be a positive integer";
                        print "currently, Nt =" Nt;
                        print "no expected values or variances were calculated";
               end;
       end;
       /*check that p is valid*/
       if p < 0 then do;
              error[ 2, 1 ] = 1;
              if print = 1 then do;
                       print "invalid number of repeated measures";
                       print "p should be a positive integer";
                       print "currently, p =" p;
                       print "no expected values or variances were calculated";
              end;
       end;
       /*check that pi is valid*/
       if pi < 0 | pi > 1 then do;
               error[ 3, 1 ] = 1;
               if print = 1 then do;
                        print "invalid value for pi";
                        print "pi should be a number between 0 and 1";
                        print "currently, pi =" pi;
                        print "no expected values or variances were calculated";
               end;
       end;
       /*check that k is valid*/
       if k ^= 1 &
          k ^= 2 &
          k ^= 9 then do;
               error[ 4, 1 ] = 1;
               if print = 1 then do;
                        print "invalid index for the missing data summary
                        statistic";
                        print "k should be in the set {1, 2, 9}";
                        print "currently, k = " k;
                        print "No expected values or variances were calculated";
               end;
       end;
       /*only complete the next set of steps if there were no errors*/
       /*otherwise, program skips over the next block and ends with no
       calculations*/ if sum( error ) = 0 then do;
```

```
        /*calculate expected value of N(m1)*/
        if k = 1 then ENmk = Nt * ( 1 - pi )**p;
        /*calculate expected value of for N(m2)*/
        else if k = 2 then do;
                if pi = 0 then ENmk = Nt;
                else if pi = 1 then ENmk = 0;
                else
                /*regression model for expected value of N(m2)*/
                /*see Table A.1, Ringham et al. (in submission)*/
                ENmk =    62.7318676
                          -  5.1567678 * ( Nt / 10 )
                          -  0.1324363 * ( Nt / 10 )**2
                          -   1.6042196 * p
                          +   0.0640387 * p**2
                          - 147.7255861 * ( 1 - pi )
                          +  87.3472243 * ( 1 - pi )**2
                          -   0.4981166 * ( Nt / 10 ) * p
                          +   0.0019218 * ( ( Nt / 10 ) * p )**2
                          +   1.2421550 * p * ( 1 - pi )
                          -   0.0423721 * ( p * ( 1 - pi ) )**2
                          +  14.8545994 * ( Nt / 10 ) * ( 1 - pi )
                          +   0.1812043 * ( ( Nt / 10 ) * ( 1 - pi ) )**2
                          +   0.4137540 * ( Nt / 10 ) * p * ( 1 - pi )
                          -   0.0013272 * ( ( Nt / 10 ) * p * ( 1 - pi ) )**2;
        end;
        /*calculate expected value of Nm9*/
        else if k = 9 then ENmk = Nt * ( 1 - pi );
  end;
/*end module*/
finish;
```

Appendix D

The SAS/IML program below uses the power approximation with $E\left(N_{m1}\right)$ to calculate power for the oral cancer biomarkers example in Section 6. Inputs for the power analysis are taken from Table III in Elashoff *et al*. [18]. The program uses the NMK module in Appendix C to compute $E\left(N_{m1}\right)$. The expectation is passed into POWERLIB, which approximates power for the multivariate hypothesis tests using $E\left(N_{m1}\right)$ as an adjusted sample size. Power is approximated for a range of effect sizes and percentages of missing data. Results are output as a power curve (Figure 3, Section 6).

Note that, should the user wish to approximate power with an alternate statistical package, the program could be truncated prior to the POWERLIB call. Results from the NMK module could be output to a dataset. The dataset could then be imported into a statistical package of the user's choice.

```
/*************************************************************************
Program:        OCBiomarkers.sas
Date:           6/22/13
Created By:     Brandy Ringham
Description:    Create power curve for hypothetical oral cancer biomarkers
                study. Use inputs from Elashoff et al., 2012.

                Fixed predictor = disease status (normal, case)
                Dependent variables = mRNA (molecule1, molecule2, molecule3)
                Hypothesis test = group difference for any mRNA

      Copyright (C) 2010 Regents of the University of Colorado. This program is
      free software; you can redistribute it and/or modify it under the terms
      of the GNU General Public License as published by the Free Software
      Foundation; either version 2 of the License, or (at your option) any later
      version. This program is distributed in the hope that it will be useful,
      but WITHOUT ANY WARRANTY; without even the implied warranty of
      MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General
      Public License for more details. You should have received a copy of the
      GNU General Public License along with this program; if not, write to the
      Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston,
      MA 02110-1301, USA.

*************************************************************************/

title1 h = 12pt "Oral Cancer Biomarkers Demonstration";
footnote1 "&sysdate";
footnote2 "OCBiomarkersDemonstration.sas";
```

```
/*initiate SAS/IML*/
proc iml;
        /*define planned per group sample size*/
        unadjustRepNt = 75;
        /*vector of anticipated proportion of missing data*/
        piVec = { 0, .05, .1 };
        /*correlation between repeated measures*/
        rho = .4;
        /*define 7 inputs of power analysis*/
        /*1.  alpha*/
        alpha = .05;
        /*2.  between subject contrast matrix*/
        /*dimensions 1x2*/
        c = { 1 -1 };
        /*3. within subject contrast matrix*/
        /*dimensions 3x2*/
        u = { 1 0 0,
              0 1 0,
              0 0 1 };
        /*4. sigma*/
        /*see below--values will vary so it resides in the do loop*/
        /*5. beta */
        /*see below--values will vary so it resides in the do loop*/
        /*6. null hypothesis*/
        /*set by POWERLIB as a 1x2 matrix of 0's by default*/
        /*7. essenceX (Muller and Stewart, 2006)*/
        /*dimensions 2x2*/
        essenceX = { 1 0,
               0 1 };
        /*calculate total planned sample size - unadjusted*/
        unadjustNt = unadjustRepNt * nrow( essenceX );
        /*include NMK code*/
        /*NMK module calculate adjusted sample size, or E(Nmk)*/
        %include "NMK.SAS" / nosource2;
        /*include POWERLIB code*/
        /*POWERLIB module calculates power*/
        %include "POWERLIB22.IML" / nosource2;
        /*set options for POWERLIB*/
        opt_off = { HF UN BOX WARN };
        opt_on = { UNIFORCE DS FRACREPN NOPRINT };
        /*loop through rows of proportion missing vector*/
        do piID = 1 to nrow( piVec );
                /*define proportion missing as the value in the current row of
                the vector*/
                pi = piVec[ piID, ];
                /*loop through different scale factors for the beta matrix*/
                /*creates data points for power curve*/
                do k = 0 to 2 by .005;
                  /*variances are from Table 3, Elashoff et al., 2012*/
                       var = 2.9**2;
                       /*define sigma matrix*/
                       oneMat = j( 3, 1, 1 );
                       sigma = var * ( oneMat * oneMat` * rho + I( 3 ) *
                       ( 1 - rho ) );
                       /*cell means are from for
                       1) IL-1B,
                       2) IL-8, and
                       3) SAT, Cohort 4 reported in Table 3, Elashoff et al.,
                          2012*/
                       /*first set means for normals*/
                       mu1n = 20.1;
                       mu2n = 19.8;
                       mu3n = 21.3;
                       /*define difference between normals and cases*/
                       delta1 = -1.3;
                       delta2 = -2.1;
                       delta3 = -1.4;
                       /*calculate mean for cases based on scale factor*/
                       mu1c = mu1n + k * delta1;
                       mu2c = mu2n + k * delta2;
                       mu3c = mu3n + k * delta3;
                       /*define beta matrix from means*/
                       beta = ( mu1c || mu2c || mu3c ) //
                              ( mu1n || mu2n || mu3n );
                       /*calculate sample sizes adjusted for anticipated amount
```

```
        of missing data*/
        /*use Nm1*/
        call NMK( unadjustNt, 2, pi, 1, 0, ENmk, error );
        /*calculated adjusted number of participants per group*/
        /*set repn equal to adjusted repn to pass into POWERLIB*/
        repn = ENmk / nrow( essenceX );
        /*save power results in a dataset*/
        dsname = { work adjustPower };
        /*calculate power adjusted for missing data using POWERLIB*/
        run power;
        use work.adjustPower;
        /*create an IML matrix from the POWERLIB output dataset*/
        read all var{ ALPHA TOTAL_N POWER_MULT EPSILON EXEPS_GG POWER_GG }
        into adjustPower[ colname = name ];
        /*delete dataset after use*/
        call delete( "work", "adjustPower" );
        /*row of output containing experimental conditions and power results*/
        powerRow = rho || k || pi || repn || 2 || adjustPower[ , 1:3 ];
        /*create matrix of results over all proportion missing*/
        power = power // powerRow;
        /*free variables to be reused*/
      free powerRow ENmk repn dsname adjustPower;
    end; /*end k loop*/
  end; /*end pi loop*/
  /*list of column names for output dataset*/
  expNames = { "rho" "k" "PropMiss" "unadjustRepNt" "NumDepVars" };
  allNames = expNames || name[ , 1:3 ];
  /*output matrix as SAS dataset*/
  create out01.appPow from power[ colname = allNames ];
  append from power;
quit;

/*set size of graph*/
goptions hsize = 4in vsize = 4in;
/*plot power curve*/
proc gplot data = out01.appPow;
      title2 h = 12pt "Power Curve";
      /*plot power by scale factor*/
      plot power_mult * k = propMiss / noframe vaxis = axis1 haxis = axis2
                  legend =   legend1 href = 1 vref = .9 cvref = red chref = red;
      /*define linetypes for different proportion missing*/
      symbol1 i = j l = 1 c = black w = 1;
      symbol2 i = j l = 41 c = black w = 1;
      symbol3 i = j l = 35 c = black w = 1;
      /*define axes options*/
      axis1 order = ( 0 to 1 by .2 ) minor = none major = none
            label = ( angle = 90 font = times h = 12pt "Power" )
            value = ( h = 12pt font = times );
      axis2 order = ( 0 to 2 by .5 ) minor = none major = none
            label = ( font = times h = 12pt "k" )
            value = ( h = 12pt font = times );
      /*define legend*/
      legend1     value = ( height = 12pt font = "times" "0%" "5%" "10%" )
                  position = ( inside bottom right )
                  label = ( font = times h = 12pt position = top
                  justify = center "Percent Missing" )
                  mode = protect across = 1 down = 3 frame shape = line( .3in );
run;
quit;
title;
footnote;
```

Acknowledgements

The research presented in this paper was supported by the NIDCR 3R01DE020832-01A1S1, a minority supplement to NIDCR 3R01DE020832-01A1. The parent grant was awarded to the University of Florida, Keith Muller, Principal Investigator, with a subaward to the Colorado School of Public Health. Revisions to the original manuscript were supported by NCI 5R25CA087949-14, a training grant awarded to the University of California, Los Angeles, Roshan Bastani, Principal Investigator. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Dental and Craniofacial Research, the National Cancer Institute nor the National Institutes of Health. This manuscript

was submitted to the Department of Biostatistics and Informatics in the Colorado School of Public Health, University of Colorado Denver, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics for B. M. Ringham.

References

1. Muller KE, Edwards LJ, Simpson SL, Taylor DJ. Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine* 2007; **26**(19):3639–3660.
2. Catellier DJ, Muller KE. Tests for Gaussian repeated measures with missing data in small samples. *Statistics in Medicine* 2000; **19**(8):1101–1114.
3. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd ed). Wiley-Interscience: Hoboken, New Jersey, USA, 2002.
4. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**(3):983–997.
5. Fouladi RT, Shieh Y-Y. A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics - Simulation and Computation* 2004; **33**(3):807–824.
6. Serrano D. *Error of Estimation and Sample Size in the Linear Mixed Model*. ProQuest, UMI Dissertation Publishing: Ann Arbor, Michigan, USA, 2008.
7. Gurka MJ, Edwards LJ, Muller KE. Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine* 2011; **30**(22):2696–2707.
8. 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**(420):1209–1226.
9. Muller KE, Stewart PW. *Linear Model Theory: Univariate, Multivariate, and Mixed Models* (1st ed). Wiley-Interscience: Hoboken, New Jersey, USA, 2006.
10. McKeon JJ. F Approximations to the distribution of Hotelling's T_0^2. *Biometrika* 1974; **61**(2):381–383.
11. 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**(15):2958–2981.
12. Barton CN, Cramer EC. Hypothesis testing in multivariate linear models with randomly missing data. *Communications in Statistics - Simulation and Computation* 1989; **18**(3):875–895.
13. Olson CL. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association* 1974; **69**(348):894–908.
14. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; **39**(1):1–38.
15. Shiboski CH, Schmidt BL, Jordan RCK. Racial disparity in stage at diagnosis and survival among adults with oral cancer in the US. *Community Dentistry and Oral Epidemiology* 2007; **35**(3):233–240.
16. Mashberg A, Samit AM. Early detection, diagnosis, and management of oral and oropharyngeal cancer. *CA: A Cancer Journal for Clinicians* 1989; **39**(2):67–88.
17. Peacock ZS, Pogrel MA, Schmidt BL. Exploring the reasons for delay in treatment of oral cancer. *Journal of the American Dental Association* 2008; **139**(10):1346–1352.
18. Elashoff D, Zhou H, Reiss J, Wang J, Xiao H, Henson B, Hu S, Arellano M, Sinha U, Le A, Messadi D, Wang M, Nabili V, Lingen M, Morris D, Randolph T, Feng Z, Akin D, Kastratovic DA, Chia D, Abemayor E, Wong DT. Prevalidation of salivary biomarkers for oral cancer detection. *Cancer Epidemiology, Biomarkers & Prevention* 2012; **21**(4):664–672.
19. Andreson R, Möls T, Remm M. Predicting failure rate of PCR in large genomes. *Nucleic Acids Research* 2008; **36**(11):e66.
20. SAS Institute Inc. *SAS/IML® 9.22 User's Guide*. SAS Institute Inc: Cary, North Carolina, USA, 2010.
21. Johnson JL, Muller KE, Slaughter JC, Gurka MJ, Gribbin MJ, Simpson SL. POWERLIB: SAS/IML software for computing power in multivariate linear models. *Journal of Statistical Software* 2009; **30**(5):1–27.
22. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2011.
23. *MATLAB and Statistics Toolbox*. The MathWorks, Inc: Natick, Massachusetts, USA, 2012. Available from: http://www.mathworks.com [Accessed 13 November 2015].
24. NCSS, LLC (NCSS Statistical Software). PASS. Kaysville, Utah, USA, 2001. Available from: http://www.ncss.com [Accessed on 13 November 2015].
25. Kreidler SM, Muller KE, Grunwald GK, Ringham BM, Coker-Dukowitz ZT, Sakhadeo UR, Barón AE, Glueck DH. GLIMMPSE: Online power computation for linear models with and without a baseline covariate. *Journal of Statistical Software* 2013; **54**(10):i10. Available at: http://www.glimmpse.org [Accessed on 13 November 2015].
26. Verbeke G, Lesaffre E. The effect of drop-out on the efficiency of longitudinal experiments. *Journal of the Royal Statistical Society, Series C* 1999; **48**(3):363–375.
27. Sampson AR. A tale of two regressions. *Journal of the American Statistical Association* 1974; **69**(347):682–689.
28. Gatsonis C, Sampson AR. Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin* 1989; **106**(3):516–524.
29. Glueck DH, Muller KE. Adjusting power for a baseline covariate in linear models. *Statistics in Medicine* 2003; **22**(16):2535–2551.