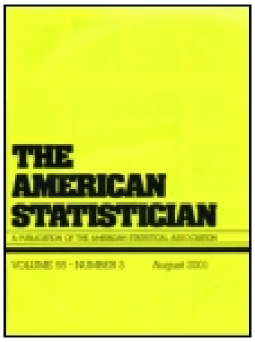

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Power and Sample Size for Fixed-Effects Inference in Reversible Linear Mixed Models

Yueh-Yun Chi, Deborah H. Glueck & Keith E. Muller

THE AMERICAN STATISTICIAN
2018, VOL. 0, NO. 0, 1–10: Statistical Practice
https://doi.org/10.1080/00031305.2017.1415972

Taylor & Francis
Taylor & Francis Group

Power and Sample Size for Fixed-Effects Inference in Reversible Linear Mixed Models

Yueh-Yun Chi[a], Deborah H. Glueck[b], and Keith E. Muller[c]

[a]Department of Biostatistics, University of Florida, Gainesville, FL; [b]Department of Pediatrics, University of Colorado Denver, Denver, CO; [c]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL

ABSTRACT

Despite the popularity of the general linear mixed model for data analysis, power and sample size methods and software are not generally available for commonly used test statistics and reference distributions. Statisticians resort to simulations with homegrown and uncertified programs or rough approximations which are misaligned with the data analysis. For a wide range of designs with longitudinal and clustering features, we provide accurate power and sample size approximations for inference about fixed effects in the linear models we call reversible. We show that under widely applicable conditions, the general linear mixed-model Wald test has noncentral distributions equivalent to well-studied multivariate tests. In turn, exact and approximate power and sample size results for the multivariate Hotelling–Lawley test provide exact and approximate power and sample size results for the mixed-model Wald test. The calculations are easily computed with a free, open-source product that requires only a web browser to use. Commercial software can be used for a smaller range of reversible models. Simple approximations allow accounting for modest amounts of missing data. A real-world example illustrates the methods. Sample size results are presented for a multicenter study on pregnancy. The proposed study, an extension of a funded project, has clustering within clinic. Exchangeability among the participants allows averaging across them to remove the clustering structure. The resulting simplified design is a single-level longitudinal study. Multivariate methods for power provide an approximate sample size. All proofs and inputs for the example are in the supplementary materials (available online).

ARTICLE HISTORY
Received August 2016
Accepted November 2017

KEYWORDS
Cluster design; General linear multivariate model; Longitudinal; MANOVA; Multilevel; Repeated measures

1. Introduction

1.1. Motivation

Despite the widespread popularity of the general linear mixed model for data analysis, corresponding power and sample size methods cover only special cases, and typically use rough approximations. In turn, little or no software is available. In sharp contrast, accurate methods and software for power and sample size are available for general linear multivariate models, which are specials cases of mixed models. In planning health science studies, many power analyses for inference about fixed effects in mixed models use special cases that correspond to multivariate models. Hence, the question arose: Can we determine what conditions a mixed model must satisfy to justify using the accurate power and sample size methods for a multivariate linear model?

Defining the conditions led us to describe a large class of general linear mixed models for which accurate power and sample size methods and software are available for inference about fixed effects. The class includes many popular longitudinal and cluster designs, as well as designs with both longitudinality and clustering. Although all multivariate models can be stated as mixed models, not all mixed models can be stated as multivariate models. Any linear mixed model that can be converted to a multivariate model can also be converted back to a mixed model, which indicates the process is reversible. Hence, we use the term "reversible" to focus attention on models for which it is easy to compute power and sample size for inference about fixed effects. Special cases have been studied (Muller et al. 2007; Ringham et al. 2016). Section 2.1 contains formal definitions and precise methods for deciding whether a mixed model is reversible.

The pregnancy study described in Section 3.1 provides an example design in which both the longitudinal and cluster features are present. Repeated measurements on women across time reflect a longitudinal design feature. Recruiting women within clinics creates clustering. The investigators plan to compare gestational changes in cardiac output across the three pregnancy cohorts by fitting a general linear mixed model to accommodate the complex covariance structure. We show in Section 3.2 that the mixed model is reversible, and we illustrate the process of conducting the power and sample size analysis for the corresponding multivariate model.

Our interest in identifying reversible models is practical. As in the example, in many designs for tests of fixed effects, the test statistic for mixed models corresponds to a multivariate test statistic with a well-developed suite of power and sample size methods. Furthermore, the power and sample size methods are available in both free and commercial software products. Building on the ability to identify a reversible model, we describe sufficient conditions under which a Wald test of fixed effects in a general linear mixed model has power equivalent to a general linear hypothesis test in a general linear multivariate

CONTACT Yueh-Yun Chi ✉ yychi@ufl.edu Department of Biostatistics, University of Florida, 2004 Mowry Road, Room 5238 P.O. Box 117450, Gainesville, FL 32611.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS

model. We describe such a power-equivalent hypothesis as reversible.

The article is organized as follows. The general linear mixed model and its hypothesis for fixed effects appear in the next subsection. The subsequent subsection contains an overview of existing power and sample size methods for the general linear mixed model. The last subsection provides an introduction to the general linear multivariate model and the general linear multivariate hypothesis. The new approach for power and sample size is described in Section 2 for three classes of designs: 1. longitudinal designs, 2. clustering designs with one or more levels, and 3. designs with both longitudinal and clustering features. Section 2 also includes a simulation study demonstrating the accuracy of the new methods. Section 3 contains an example power calculation for a longitudinal study with clusters, and illustrates the steps needed to transform a power calculation for a reversible mixed model to a power calculation for a multivariate model. Section 4 provides a discussion of practical implications and limitations, as well as sketches of future research directions.

1.2. General Linear Mixed Models and Tests of Fixed Effects

Notation follows Muller and Stewart (2006). Throughout, subscripts m and M, respectively, indicate an element of a mixed or multivariate model.

Linear mixed models are commonly used for analyzing clustered and longitudinal data. Specifying a mixed model requires an explicit form for the means (the fixed effects) and an explicit form for the covariance structure (the random effects). Although not correct in all the cases, for the sake of brevity, we use the term subject to indicate the independent sampling unit. The model includes fixed-effect predictors $\boldsymbol{X}_{mi}$ and random-effect predictors $\boldsymbol{Z}_i$ as

$$\underset{(p_i\times 1)}{\boldsymbol{y}_i} = \underset{(p_i\times q_m)(q_m\times 1)}{\boldsymbol{X}_{mi}\boldsymbol{\beta}_m} + \underset{(p_i\times r)(r\times 1)}{\boldsymbol{Z}_i\boldsymbol{d}_i} + \underset{(p_i\times 1)}{\boldsymbol{e}_{mi}} \tag{1}$$

with the subject (independent sampling unit) indexed by $i \in \{1, \ldots, N\}$. The vector $\boldsymbol{\beta}_m$ contains unknown fixed-effect parameters of interest, and the design matrix $\boldsymbol{X}_{mi}$ includes both the between- and within-subject design information. Distributional assumptions are

$$\begin{bmatrix} \boldsymbol{d}_i \\ \boldsymbol{e}_{mi} \end{bmatrix} \sim \mathcal{N}_{r+p_i}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_d & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{e_{mi}} \end{bmatrix} \right), \tag{2}$$

which imply $E(\boldsymbol{y}_i) = \boldsymbol{X}_{mi}\boldsymbol{\beta}_m$ and $\mathcal{V}(\boldsymbol{y}_i) = \boldsymbol{\Sigma}_{mi} = \boldsymbol{Z}_i\boldsymbol{\Sigma}_d\boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{e_{mi}}$. With $\boldsymbol{e}_{mi+} = \boldsymbol{Z}_i\boldsymbol{d}_i + \boldsymbol{e}_{mi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{mi})$, the "population-average" form of a mixed model is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_{mi}\boldsymbol{\beta}_m + \boldsymbol{e}_{mi+}. \tag{3}$$

With $\boldsymbol{\theta}_m = \boldsymbol{C}_m\boldsymbol{\beta}_m$, a general linear mixed model hypothesis about the fixed effects (mean parameters) can be written as

$$H_{m0} : \boldsymbol{\theta}_m = \boldsymbol{\theta}_{m0}. \tag{4}$$

The matrix $\boldsymbol{C}_m$ combines between- and within-subject contrasts. Throughout, we restrict attention to estimable parameters, that is, those with unbiased estimators, and testable hypotheses, that is, tests of full rank transformations of estimable parameters (Muller and Stewart 2006, chap. 11–18). Many tests have been developed for H_{m0}. The Kenward and Roger (1997, 2009) F approximation for the Wald statistic was recommended by Schaalje, McBride, and Fellingham (2002) and Muller and Stewart (2006, sec. 18.5) due to its accuracy in samples of small size. The Kenward and Roger approach uses restricted maximum likelihood (REML) estimators for $\mathcal{V}(\boldsymbol{y}_i)$. The approach uses a series expansion to create an approximate two moment match. The approximation is accurate except for some covariance patterns and few independent sampling units (Park, Park, and Davis 2001; Schaalje, McBride, and Fellingham 2003; Muller et al. 2007).

1.3. Existing Methods for Designing Studies Using Mixed Model Analysis

Existing power and sample size methods for general linear mixed models fall into three classes: large sample approximations, exemplary data approaches, and simulations. The methods differ in the hypothesis, statistical test, and computational format. We describe each class of power and sample size approximations and summarize the weaknesses of each class.

Large sample approximations for fixed-effects inference. Most existing sample size methods for mixed models focus on approximating power by assuming the distribution of the test statistic is Gaussian. Hedeker, Gibbons, and Waternaux (1999), Heo and Leon (2009), Basagaña, Liao, and Spiegelman (2011), and Wang, Hall, and Kim (2015) all considered large sample Gaussian approximations for the Wald statistic using maximum likelihood (ML) estimators. In contrast, Murray et al. (2007) based their power calculation on the t distribution. All of the methods apply only to scalar parameters (i.e., θ_m is 1×1), and are useful only for a restricted class of designs. Hence, the calculations do not apply to the comparisons of more than two groups or analyses of overall trends when more than two repeated measures are present. Subsequently, Tu et al. (2007) suggested another Gaussian approximation allowing tests of multiple parameters and adjustment for predictors important in the model but not part of the hypothesis test for which power is sought.

Table 1 summarizes the large sample methods by their design and hypothesis features, covariance model, applicability to missing data, and software availability. All target the ML-based Wald test for fixed-effects inference with Gaussian or t power approximations. The calculations misalign with the REML-based Wald test recommended for inference about the fixed effects (Kenward and Roger 1997, 2009).

In addition, Gaussian approximations take no account of the fact that the Wald statistic uses the substitution principle for the variance (denominator) parameters, which calls for an F approximation. Most of the methods are limited by the generality of the hypotheses, covariance model, or both. Restricting inference to a scalar parameter disallows comparing either overall time trends or complex interactions involving within- and between-subject effects. Current methods use simple random effect structures (such as a single random intercept) to specify the response covariance. However, results in Gurka, Edwards, and Muller (2011) make it clear that using a simple structure for analyzing longitudinal data can result in liberal inference due to

Table 1. Large sample methods for power and sample size in longitudinal mixed models.

		Article					
Feature	Structure	Hedeker (1999)	Tu (2007)	Murray (2007)	Heo (2009)	Basagaña (2011)	Wang (2012)
Predictor tested	Binary	√	√	√	√	√	√
	Continuous		√				√
	Time-varying					√	
Time effect tested	Linear	√	√	√	√	√	√
	Any trends		√				
Scalar Testing		√		√	√	√	√
Predictors Adjusted			√				
Covariance structure	Rand Intrcpts; iid Errors				√		√
	Rand Intrcpts+Time; iid Errors			√			
	Any Random Effects; iid Errors		√				
	No Random Effects; UN* Errors	√				√	
Missing data		√	√				√
Software		√				√	

*Unstructured covariance.

a greatly inflated Type I error rate. Power methods that assume simple covariance structures do not naturally extend to methods for responses with complex covariance structures, which are typically needed for longitudinal data. Lack of user-friendly software also reduces the practical value of existing methods.

An exemplary data approach. Stroup (2002, 2012) suggested an exemplary data approach to use data analysis software to approximate power for mixed models. The approach assumes an approximate F distribution for the mixed-model Wald statistic for the null case. For power, the reference distribution shifts from a central to a noncentral F distribution (for fixed-effect inference). Stroup (2002) provided example SAS code to construct both the central and noncentral F distributions. He obtained the degrees of freedom and noncentrality parameter by using a mixed model fitting procedure with block and treatment identifications (i.e., $\boldsymbol{X}_{mi}$), treatment differences of interest (i.e., $\boldsymbol{\beta}_m$), and a homogeneous spatial covariance structure (i.e., $\boldsymbol{\Sigma}_{mi}$).

Despite the computational appeal and apparent generality, simulation results in Section 2.4 led us to conclude that the approach is valid only for special cases. In particular, the fixed-effects tests must have degrees of freedom determined by data dimensions N, $\text{rank}(\boldsymbol{X}_{mi})$, and the dimensions of $\boldsymbol{\theta}_m$. Power for the common Kenward–Roger test can not be correctly calculated.

Simulations to estimate power. Simulations can be used for empirical power calculations. Hypothetical data can be repeatedly generated and used to test H_{m0}. Empirical power is reported as the average number of times of rejecting H_{m0}. The approach can be applied to any fixed-effects test, including testing nonscalar parameters (i.e., $\boldsymbol{\theta}_m$ a vector). Siemer and Joormann (2003) used Monte Carlo simulations to iteratively compute estimates of the parameters in the F approximations for the mixed model ANOVA statistic (ratio of mean squared errors). Lu (2012) computed simulation-based empirical power to adjust multiplicity for multiple scalar outcomes with missing data.

Simulations can provide power calculations for a wide range of tests. A minimal standard would require documentation and archiving of reproducible computations, and a verification of reliability (getting the same answer on the second use). However, meeting software industry standards of certification requires extensive and laborious testing and documentation of the specific software for specific use cases. The entire process is required to demonstrate validity (the instrument is measuring accurately what is supposed to be measured). User written simulations do not meet the standard. Simulations may be time consuming for other reasons. Trying a variety of "what if?" analyses is tedious at best, and typically limited by computer time. Deadline pressure often precludes conducting extensive simulations.

Large sample approximations, exemplary data approaches, and simulations approaches for power all have drawbacks. As an alternative, we propose model-based approaches that are more general, and better aligned to recommended data analysis practice. The power and sample size methods we describe apply only to reversible general linear mixed models.

1.4. Borrowing Methods from the General Linear Multivariate Model for the Mixed Model

The ascent to the dominance of the general linear mixed model created a corresponding decline in the use of the general linear multivariate model. The popularity of the mixed model stems from its flexibility. The multivariate model does not allow time varying predictors or missing or mistimed data, which are important for longitudinal and clustering designs.

The situation may be summarized by noting that any general linear multivariate model can be expressed as a general linear mixed model. Although true, the simple statement glosses over important advantages of the multivariate model when it applies: 1. more accurate reference distributions for test statistics used to create confidence intervals and conduct hypothesis tests (Muller et al. 2007); and 2. general and accurate methods for computing power and sample size (Muller et al. 1992). The advantages are especially important in the small to moderate sample sizes often encountered in the longitudinal and clustering designs.

Proving the validity of borrowing multivariate linear model methods for power and sample size requires considering a restricted (but widely useful) class of mixed models. In the notation of Muller and Stewart (2006, chap. 3), the multivariate model of interest may be stated as

$$\underset{N\times p}{\boldsymbol{Y}_M} = \underset{(N\times q_M)(q_M\times p)}{\boldsymbol{X}_M\boldsymbol{B}_M} + \underset{N\times p}{\boldsymbol{E}_M} \ . \tag{5}$$

For subject (independent sampling unit) i, here row i of $\boldsymbol{Y}_M$ is $\boldsymbol{y}_i'$ from Equation (3) and row i of $\boldsymbol{E}_M$ is $\boldsymbol{e}_{mi+}'$ from Equation (3). The number of repeated measures is constant, so $p_i = p$ for all i. Hence, $\boldsymbol{E}_M$ has iid rows, each with $p \times p$ covariance $\boldsymbol{\Sigma}_M$. Every repeated measure (column of $\boldsymbol{Y}_M$) uses the same design matrix, $\boldsymbol{X}_M$. Columns of $\boldsymbol{B}_M$ reflect differences across time (within subject) and rows reflect differences between subjects. Muller and Stewart (2006, sec. 12.1) gave some details used in the next section.

For $a_M \times b_M$ secondary parameter matrix $\boldsymbol{\Theta}_M = \boldsymbol{C}_M\boldsymbol{B}_M\boldsymbol{U}_M$, a general linear multivariate hypothesis about the fixed effects (mean parameters) can be written as

$$H_{M0} : \boldsymbol{\Theta}_M = \boldsymbol{\Theta}_{M0} \ . \tag{6}$$

The matrix $\boldsymbol{C}_M$ contains only between-subject contrasts, of which there are a_M, while $\boldsymbol{U}_M$ contains only within-subject contrasts, of which there are b_M. Throughout, we restrict attention to estimable parameters and testable hypotheses (Muller and Stewart 2006, chap. 11–18).

2. Computing Power for Reversible Mixed Models

2.1. A Class of Reversible Models for Longitudinal Data

We have neither the widest focus of complete generality, nor the narrowest focus of a single model. Instead, we have an intermediate focus and consider a widely used, but not universal, class of models. The focus allows identifying fixed-effect parameters of interest in the multivariate model and therefore computing power and sample size accurately, even for small sample sizes. As generalizations of the results in Muller et al. (2007) and Ringham et al. (2016), we define a class of mixed models called reversible general linear mixed models. We also define the corresponding class of hypothesis tests.

Definition 1. A general linear model mixed model, as in Equation (3), is a reversible model if it can be expressed as a general linear multivariate model, as in Equation (5).

Definition 2. A general linear hypothesis about fixed effects in the general linear model mixed model, as in Equation (4), is a reversible hypothesis if it can be expressed as a multivariate general linear hypothesis, as in Equation (6).

Special cases and sufficient conditions for reversibility are described in four lemmas stated and proven in the supplementary materials. We briefly discuss each.

Lemma 1 guarantees that a model with $p_i = 1$ is reversible. Consequently, any univariate model is reversible, such as an analysis looking at only the last time in a clinical trial.

Lemma 2 gives three sufficient, but not minimally sufficient conditions for model reversibility. The first condition requires complete data, with all subjects observed at the same times for longitudinal studies, the same locations for spatial studies, or the same variables for multivariate responses. The second condition requires covariates to have the same value for all the observations from a subject, which precludes, for example, time-varying covariates. It also requires saturating the model by including all interactions of the baseline covariates with time effects. The third condition requires the errors for each subject to be homogeneous between subjects and share the same covariance pattern within a subject, and is modeled as unstructured. The sufficient condition of an unstructured covariance does not preclude using a structured covariance model to generate the input matrix to a power analysis, which will treat the matrix as unstructured. The approach provides a convenient way to specify a valid and plausible covariance matrix while still using the power methods championed in this article. For example, an autoregressive covariance can be specified as the input matrix, without the need to define a more complex structure. As will be seen in simulation results in the next section, the result can be a somewhat conservative estimate of power.

An example illustrates the second condition in Lemma 2, which stipulates that the design matrix may be written as $\boldsymbol{X}_{mi} = \boldsymbol{T}' \otimes \boldsymbol{x}_{i0}'$, with each row of $\boldsymbol{T}$ $(p_* \times p)$ a polynomial trend $(p_* \leq p)$ and $\boldsymbol{x}_{i0}$ $(q_M \times 1)$ containing subject-specific baseline predictors. For a balanced two-group design with three time points (for all i, $p_i = p = 3$), the linear mixed model with an average and linear time trend has $p_* = 2$ (trends), $\boldsymbol{T} = \left[\begin{smallmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \end{smallmatrix}\right]$, $\boldsymbol{x}_{i0}' = [\,1 \ g_i\,]$ with $g_i = 1$ for the treatment group and $g_i = 0$ for the control group, and $\boldsymbol{\beta}_m' = [\,\mu_C \ \Delta_\mu \ \beta_C \ \Delta_\beta\,]$. The parameters μ_C and β_C are the intercept and slope for the control group, and Δ_μ and Δ_β are differences in intercept and slope for the treatment group. If age is also in the model, then $\boldsymbol{x}_{i0}' = [\,1 \ g_i \ \text{age}_i\,]$ and $\boldsymbol{\beta}_m' = [\,\mu_C \ \Delta_\mu \ \beta_{\text{age}} \ \beta_C \ \Delta_\beta \ \beta_{\text{age}\times\text{time}}\,]$ with β_{age} the effect of age at time 0 and $\beta_{\text{age}\times\text{time}}$ the interaction between age and time (in a linear form).

Lemma 3 provides sufficient conditions for the test reversibility. It has a simple interpretation. Namely, for a reversible model, any test of 1. between-subject effects, 2. within-subject effects, or 3. between-by-within effects has a null hypothesis in the mixed model which coincides with a null hypothesis in a multivariate model.

A counterexample illustrates the value of Lemma 3 by demonstrating that a reversible mixed model can have nonreversible hypotheses about fixed effects. For a model with two groups measured at two time points, using cell-mean coding for the multivariate model leads to $\boldsymbol{B}_M = \{\mu_{g,t}\}$, with group $g \in \{1, 2\}$ and time $t \in \{1, 2\}$. The corresponding mixed model has $\boldsymbol{\beta}_m = \text{vec}(\boldsymbol{B}_M)$ and $\boldsymbol{\beta}_m' = [\,\mu_{1,1} \ \mu_{2,1} \ \mu_{1,2} \ \mu_{2,2}\,]$. In turn, $\boldsymbol{\theta}_m = \boldsymbol{C}_m\boldsymbol{\beta}_m = (\mu_{1,1} - \mu_{2,2})$ is an estimable and testable parameter in the mixed model, but not in the multivariate model. It is worth noting that the parameter rarely holds any scientific interest.

Lemma 4 demonstrates that the Wald statistic for a reversible mixed model coincides with the Hotelling–Lawley test statistic in the corresponding multivariate model. The result holds under both the null and alternative hypothesis.

For testing a reversible hypothesis about the fixed-effects parameters, Theorem 1 provides useful exact results for the power of a special but common case. The four lemmas provide the basis of the proof of Theorem 1. The following Conjecture applies to a more general case which requires approximating power.

Theorem 1. In a general linear mixed model, for any reversible hypothesis about $\boldsymbol{\theta}_m = \text{vec}(\boldsymbol{\Theta}_M)$, with $a_M \times b_M$ secondary parameter $\boldsymbol{\Theta}_M$, if $s = \min(a_M, b_M) = 1$, then null probabilities as well as power and sample size calculations can be computed exactly.

The $s = 1$ case occurs in one of two ways. If $b_M = 1$, then the hypothesis implicitly reduces the model to a univariate model. The other case, with $a_M = 1$, covers many widely used designs including longitudinal comparisons of the two groups.

Conjecture. In a general linear mixed model, for any reversible hypothesis about $\boldsymbol{\theta}_m = \text{vec}(\boldsymbol{\Theta}_M)$, for $a_M \times b_M$ secondary parameter $\boldsymbol{\Theta}_M$, if $s = \min(a_M, b_M) > 1$, then null probabilities as well as power and sample size calculations can be accurately approximated.

Evidence supporting the conjecture. Under the null, an F approximation developed by McKeon (1974) for the HLT statistic matches two moments exactly and reduces to the exact results for $s = 1$. McKeon includes some enumerations to demonstrate accuracy. Catellier and Muller (2000) reported simulations that support the accuracy of the methods even in very small samples ($n = 12$, $p = 6$). Under the alternative, Muller et al. (1992) reviewed power approximations and concluded that a method developed by Muller and Peterson (1984) provides nearly two digits of accuracy in computing power for the HLT statistic. The method automatically reduces to the exact results for $s = 1$. Additional simulation evidence to support the accuracy claim can be found in Kreidler et al. (2013), who reviewed the availability and accuracy of the statistical methods for power and sample size of multivariate models. Extensions to allow a single Gaussian covariate can be found in Glueck and Muller (2003).

The theorem and conjecture have very practical and convenient consequences. Power and sample size for a wide range of common mixed models can be computed with free and open-source software. Johnson et al. (2009) provided an implementation in matrix software (POWERLIB, at *http://SampleSizeShop.org/software-downloads/other/*) and notation. Kreidler et al. (2013) introduced a point-and-click interface (GLIMMPSE, at *http://SampleSizeShop.org/*) that requires only a web browser on a computer, tablet, or smartphone. Many cluster, longitudinal, and repeated measures designs can be handled. Other free and commercial packages can handle many, but not all, of the designs covered by the Johnson et al. (2009) and Kreidler et al. (2013) programs.

2.2. Reversible Models with Single or Multilevel Clustering

As noted earlier, the three conditions in Lemma 2 are sufficient to define a reversible model, but are not minimally sufficient. General linear mixed models for single and multilevel clustering designs with equal cluster sizes as well as a particular and widely used covariance model can be transformed to a reversible model. Cluster designs assume exchangeable sampling, which induces compound symmetric covariance among the responses in a cluster, $\boldsymbol{\Sigma}_c = \sigma^2[\rho \mathbf{1}_p \mathbf{1}_p' + (1-\rho)\boldsymbol{I}_p] = \sigma^2 \boldsymbol{\Gamma}_c$, with σ^2 the common variance and ρ the common (intraclass) correlation, the ICC. A single-level clustering design corresponds to a general linear mixed model with a "random intercept," as the only random effect. Given the focus on inference about fixed effects, for a single level of clustering, and all clusters of equal size, $\boldsymbol{Z}_i = \mathbf{1}_p$, which implies $\boldsymbol{e}_{mi+} = \boldsymbol{Z}_i \boldsymbol{d}_i + \boldsymbol{e}_{mi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_c)$ and $\boldsymbol{\Sigma}_{mi} = \boldsymbol{\Sigma}_c$.

For multilevel clustering designs, direct-product compound symmetry, as in Definition 3, can be used to characterize the exchangeability properties inherent to factorial designs (nesting of cluster effects). Throughout, $\otimes$ denotes a (left) direct product (Muller and Stewart 2006, chap. 1). The model ensures marginal compound symmetry within each level (factor). Other models of multilevel cluster correlations are possible. Heo and Leon (2009) used a different (additive) model for a two-level cluster design with nesting, as considered here.

Definition 3. Compound symmetry generalizes to direct-product compound symmetry of K levels (Gurka, Coffey, and Muller 2007). The matrix $\boldsymbol{\Gamma}_{c,k} = \rho_k \mathbf{1}_{q_k} \mathbf{1}_{q_k}' + (1-\rho_k)\boldsymbol{I}_{q_k}$ is $p_k \times p_k$. If $\boldsymbol{p}_k = [\, p_1 \ldots p_K \,]'$, $\boldsymbol{\rho} = [\, \rho_1 \ldots \rho_K \,]'$, and $\boldsymbol{\sigma} = [\, \sigma_1 \ldots \sigma_K \,]'$, then

$$\boldsymbol{\Sigma}_c\left(\boldsymbol{p}, \boldsymbol{\rho}, \boldsymbol{\sigma}\right) = \bigotimes_{k=1}^{K} \sigma_k^2 \boldsymbol{\Gamma}_{c,k}\,. \tag{7}$$

The matrix $\boldsymbol{\Sigma}_c(\boldsymbol{p}, \boldsymbol{\rho}, \boldsymbol{\sigma})$ is $p_* \times p_*$ with $p_* = \prod_{k=1}^{K} p_k$. Assuming homogeneity across cluster dimensions leads to a constant variance (i.e., $\sigma_k^2 = \sigma^2$) and defining $\boldsymbol{\Sigma}_c(\boldsymbol{p}, \boldsymbol{\rho}, \sigma) = \sigma^2 \bigotimes_{k=1}^{K} \boldsymbol{\Gamma}_{c,k}$.

Theorem 2 and a corollary show that equal cluster sizes with direct product compound symmetry gives a reversible mixed model. Consequently, power analysis is easy to compute.

Theorem 2. Any general linear mixed model with $p_i = p_*$ for all i, homogeneity of covariance across cluster dimensions, and the same direct-product compound symmetric covariance for every subject, that is, $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}_c(\boldsymbol{p}, \boldsymbol{\rho}, \sigma)$, including any single-level and many multilevel clustering designs, can be transformed to one or more univariate models, each of which is reversible. All hypotheses within each model are reversible.

Corollary. Transforming to the average model gives exact power and sample size calculation.

An example with a single level of clustering illustrates the process. Transforming Equation (3) by computing averages gives the scalar model equation $y_{Ai} = \boldsymbol{X}_{Ai}\boldsymbol{\beta}_m + e_{Ai+}$, with $y_{Ai} = p^{-1}\mathbf{1}_p \boldsymbol{y}_i$, $\boldsymbol{X}_{Ai} = p^{-1}\mathbf{1}_p \boldsymbol{X}_{mi}$, and $e_{Ai+} = p^{-1}\mathbf{1}_p \boldsymbol{e}_{mi+}$. Note that $\boldsymbol{\beta}_m$ remains unchanged. For σ^2 the original variance, the univariate model has variance $\sigma_A^2 = \sigma^2[1 + (p-1)\rho]p^{-1}$. The factor $[1 + (p-1)\rho]p^{-1}$ corresponds to the "design effect," as discussed in survey sampling. As the cluster size increases, the variance decreases. For homogeneity and K levels of clustering, the variance becomes

$$\sigma_A^2 = \sigma^2 \prod_{k=1}^{K} [1 + (p_k - 1)\rho_k] p_k^{-1}. \tag{8}$$

After transformation, univariate commercial software can be readily applied. More conveniently, the free software GLIMMPSE (SampleSizeShop.org) directly allows specifying clustering, as does some commercial software.

Not all general linear mixed models based on cluster designs are reversible. Examples that are not reversible include models required for ongoing randomized trials with participants belonging to multiple groups (Andridge et al. 2014).

2.3. Reversible Models with Both Clustering and Longitudinal Features

The results on reversible longitudinal and reversible cluster mixed models can be combined to compute power and sample size for many designs with both longitudinal and clustering dimensions. For three reasons, we restrict attention to longitudinal designs with nested clustering, and hypotheses excluding all cluster effects (although cluster effects are accounted for). First, for the sake of brevity, we omit the more complex calculations needed to include cluster effects in the hypotheses. Second, nested cluster designs are common in behavioral and health research for both observational and experimental studies. Third, in our collaborations, the cluster effects usually hold little interest as a target of inference and are included mainly to improve model accuracy. The hypotheses of primary interest are group (treatment) by time interactions, as well as group and time effects.

Theorem 3 shows how to reduce a longitudinal plus clustering model to a model with only longitudinal features. Doing so allows conveniently computing power and sample size.

Theorem 3. Hypothesis tests of group-by-time (treatment-by-time) interaction, group, and time effects are reversible hypotheses in a longitudinal design with nested clustering if the corresponding general linear mixed model has longitudinal features meeting the conditions of Lemma 2 and clustering features meeting the conditions of Theorem 2. The impact of the clustering on power of the tests is fully accounted for by computing power for a design without clustering and multiplying the (longitudinal) covariance matrix among the repeated measures, $\boldsymbol{\Sigma}_t$, by the factor $\gamma(\boldsymbol{p}, \boldsymbol{\rho}) = \prod_{k=1}^{K}[1 + (p_k - 1)\rho_k]p_k^{-1}$.

An example in Section 3 illustrates the process. After transformation, commercial software can be applied. More conveniently, the free software GLIMMPSE (SampleSizeShop.org) directly allows specifying clustering with longitudinal and multivariate outcomes.

2.4. Missing Data

The sufficient conditions exclude missing data, mistimed observations, and time-varying covariates. Mistimed observations are rare in many studies, but can be common in epidemiologic research. Missing observations are common in many settings. For example, participants may miss a visit or drop out for reasons irrelevant to the study. Such losses are typically addressed by inflating the sample size by the factor $(1 - d)^{-1}$, with d the anticipated probability of drop-outs.

More sophisticated methods are available, at least for some common scenarios. Ringham et al. (2016) described methods for modifying the degrees of freedom to approximate power for a generalization of the Hotelling–Lawley multivariate test in studies with missing data. The power approximations use a noncentral F statistic which is a function of 1. the expected number of complete cases, 2. the expected number of nonmissing pairs of responses, or 3. the trimmed sample size, which is the planned sample size reduced by the anticipated proportion of the missing data. Extensive simulations led to the conclusion that the closest approximation to the empirical power can be obtained from power calculations based on the expected number of complete cases, as given in Equation (7) in Ringham et al. (2016). They provided example code to implement the method using commercially available software.

2.5. Simulations

Numerical simulations were used to 1. compare the performance of the multivariate and exemplary data approaches for estimating power for fixed-effects inference, 2. examine bias in power when the covariance pattern is misspecified, and 3. evaluate bias in the multivariate power approximations as a function of sample size and the number of repeated outcomes. All simulations were conducted using SAS® GLIMMIX (SAS Institute 2012).

The simulations used the following experimental design. An outcome was measured every other week for 10 weeks. Thus, $p = 5$, and data were recorded for weeks 2, 4, 6, 8, and 10. The experiment involved two groups, each with 10 participants. The planned analysis compared overall time trends (linear, quadratic, cubic, and quadratic; four degrees of freedom) between the groups, leading to the group by time interaction hypothesis. The planned analysis used a general linear mixed model, and an F approximation for the REML-based Wald test with Kenward–Roger degrees of freedom. For the power analysis, the pattern of means was chosen as follows. The mean outcome for the first group was β_{Case} at week 2 and 0 otherwise, while the mean outcome for the second group was β_{Case} at week 10 and 0 otherwise. Consequently, the expected value matrix for the corresponding multivariate model was $\boldsymbol{B}_{\text{Case}} = \beta_{\text{Case}}\left[\begin{smallmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{smallmatrix}\right]$. Here, Case $\in \{1, 2, \ldots, 9\}$ indicates a particular simulation scenario. Specific values of β_{Case} are listed in Tables 2 and 3.

The error covariance was assumed to be the same for each independent sampling unit, regardless of group membership. For λ_k, an eigenvalue of $\boldsymbol{U}_M'\boldsymbol{\Sigma}_M\boldsymbol{U}_M$, the parameter $\epsilon = \text{tr}^2(\boldsymbol{U}_M'\boldsymbol{\Sigma}_M\boldsymbol{U}_M)/[b_M\text{tr}(\boldsymbol{U}_M'\boldsymbol{\Sigma}_M^2\boldsymbol{U}_M)] = (\sum_{k=1}^{b_M}\lambda_k)^2/(b_M\sum_{k=1}^{b_M}\lambda_k^2)$, quantifies heterogeneity of the eigenvalues. It is important to recognize that the eigenvalues of $\boldsymbol{U}_M'\boldsymbol{\Sigma}_M\boldsymbol{U}_M$ are the variances of the principal components in the hypothesis subspace based on $\boldsymbol{\Sigma}_M = \boldsymbol{\Sigma}_m$ ($p \times p$ covariance of the longitudinal

Table 2. Empirical and approximate power for the Kenward–Roger test of the group × time interaction. Data were generated for five time points and two groups of 10 subjects.

	Population			Approximate power		Empirical power
Case	$\boldsymbol{\Sigma}_y$	ϵ	β_{Case}	HLT	Exemplary	Fitted UN
1	CS	1.00	1.0	0.510	0.625	0.507
2	AR(0.5)	0.81	1.0	0.705	0.820	0.706
3	AR(0.9)	0.58	0.5	0.699	0.814	0.702
4	IND(H)	0.27	0.2	0.771	0.150	0.773

Table 3. Empirical and approximate power for the Kenward–Roger test of the group×time interaction for two groups. Data were generated with an AR(0.5) covariance pattern.

Case	N	p	ϵ	β_{Case}	Empirical	Predicted (HLT)	Exemplary
5	10	5	0.81	2.0	0.747	0.749	0.982
6	20	5	0.81	1.0	0.706	0.705	0.820
7	100	5	0.81	0.4	0.713	0.723	0.742
8	20	10	0.69	1.5	0.769	0.770	0.981
9	100	10	0.69	0.5	0.781	0.786	0.825

outcomes). Bounded by $1/b_M$ and 1, the lower bound of ϵ is reached when all but one eigenvalue are zero, while the upper bound is reached when eigenvalues are all equal (spherical).

Understanding the two boundary conditions gives insight about the nature of ϵ and values likely to arise in practice. Only compound symmetry (equal variance and equal correlations) gives $\epsilon = 1$. At the other extreme, $\epsilon = 1/b_M$ requires $b_M - 1$ eigenvalues of zero and one nonzero eigenvalue. More generally, low values of ϵ reflect dominance of a small fraction of the eigenvalues, the variances of the principal components in the hypothesis space.

Having $p = 5$ and AR correlation of 0.50 implies $\epsilon = 0.81$, while AR correlation of 0.90 implies $\epsilon = 0.58$. Huynh and Feldt (1976) suggested that educational researchers typically see ϵ values of 0.75 and higher. In a different setting, Littell et al. (2007, chap. 5) considered combining heterogeneous AR and first-order antedependence covariance structures to model respiratory data. Adding heterogeneity of variance will typically (and perhaps always) decrease ϵ. On the other hand, blending two or more patterns seems likely to homogenize variances and correlations, which will increase ϵ.

Additional features of the simulations are as follows. Four covariance structures were used: 1. compound symmetry, CS ($\epsilon = 1$) with $\sigma^2 = 1$ and $\rho = 0.25$; 2. autoregressive, AR (0.5) ($\epsilon = 0.81$) with $\sigma^2 = 1$ and $\rho = 0.5$; 3. autoregressive, AR(0.9) ($\epsilon = 0.58$) with $\sigma^2 = 1$ and $\rho = 0.9$; and 4. independence with heterogeneous variances, IND(H) ($\epsilon = 0.27$) with $\rho = 0$, $\sigma_1^2 = 1$, $\sigma_k^2 = 0.1$, and $k \in \{2, \dots, p\}$. The structures span the range of the sphericity parameter ϵ. Empirical (10,000 replications) and approximate power for REML-based Wald tests with Kenward–Roger degrees of freedom are reported. The group × time interaction (four degrees of freedom) was tested at $\alpha = 0.05$. Empirical power was calculated as the number of rejections, divided by the number of replications of the simulation, here, set to 10,000. A rejection occurred when the p-value was less than the specified $\alpha = 0.05$.

Multivariate power calculation outperforms an exemplary data approach. Table 2 displays empirical power, approximate power for the exemplary approach (exemplary), and approximate power for the McKeon method for multivariate HLT statistic (HLT). In practice, it is difficult to select the true covariance structure. To avoid misspecification of covariance which can inflate the Type I error rate (Gurka, Edwards, and Muller 2011), empirical power was calculated based on fitting an unstructured covariance (Fitted UN). The multivariate method provided accurate power approximations (up to the second decimal place), and remained accurate across a range of population covariance patterns and sphericity (ϵ) values. In contrast, the exemplary approach overestimates power in three of four cases, and underestimates in the fourth (IND(H); $\epsilon = 0.27$). A best case scenario would involve a data analyst choosing to fit the population covariance structure. In the four cases here, empirical powers were 0.624 for CS, 0.818 for AR(0.5), 0.813 for AR(0.9), and 0.853 for IND(H). As expected, power increases with a reduction in the number of covariance parameters that correctly reflects the population structure. The exemplary method predicted the first three cases, but badly missed the fourth case. Using between-within degrees of freedom to calculate exemplary-method power misaligns with the data analysis method (Kenward–Roger degrees of freedom). Approximate powers were in the range of 0.10 for all four cases in Table 2 when Kenward–Roger degrees of freedom were used.

An unstructured covariance model protects the Type I error rate but reduces power. Empirical powers were tabulated for data analysis fitting an unstructured covariance pattern (over-parameterization) as well as for the correctly specified covariance patterns listed in the first column of Table 2. When the covariance matrix was over-parameterized, empirical power decreased by 0.08 to 0.11, due to the need to estimate additional nuisance parameters. Hence, over-specifying the covariance matrix as unstructured predicts a value smaller than the empirical power for a mixed model with the covariance model correctly specified. One cost of the assumption lies in losing the robustness conveyed by fitting an unstructured model.

Bias is a function of sample size and number of outcomes. Table 3 lists empirical and predicted multivariate and exemplary power for comparing two groups with $N \in \{10, 20, 100\}$ and $p \in \{5, 10\}$ at $\alpha = 0.05$. Data were generated with an AR(0.5) covariance structure, leading to $\epsilon = 0.81$ when $p = 5$, and $\epsilon = 0.69$ when $p = 10$. The bias in power approximation (absolute difference between the empirical and predicted powers) remains low for all N and p. The multivariate HLT power provides a good approximation to the mixed model test comparing the overall time trends of the two groups. The exemplary method overestimates power in all cases in Table 3.

3. Example Designs

3.1. *Pregnancy Study with Both Clustering and Longitudinal Features*

Scientists wished to evaluate whether cardiac output in women with spontaneous pregnancies differs from that in women who became pregnant using assisted reproductive technology. Cardiac output may be affected by the number of corpus lutea that exist at implantation. The number differs depending on how the pregnancy occurred. Abnormal early cardiac output seems to increase the risk of pathological pregnancy outcomes such as preeclampsia.

Three forms of conception are compared: spontaneous singleton pregnancy, with one corpus luteum; egg donation, with no corpus luteum; and pregnancy following ovarian stimulation, with multiple corpus lutea possible. Sample size will be selected to ensure good power for testing the primary outcome of cardiac output (L/min). Data for cardiac output will be recorded at four visits: 0, 5–6, 7–9, and 13–15 weeks during pregnancy. The main objective is to compare early gestational changes in cardiac output across the three pregnancy cohorts.

To increase enrollment, pregnant women will be recruited from many clinics. The study design has both longitudinal and cluster features. Results for women within a clinic are correlated. Repeated measures on women over time are correlated. The investigators plan to fit a general linear mixed model to accommodate the complex covariance structure.

The covariance structure can be described in three steps: 1. Correlations among women in the same clinic are assumed constant (i.e., exchangeability among women), which creates a compound symmetric structure. 2. Correlations among longitudinal observations are complex. A safe choice is an unstructured pattern across time, but it requires information from previous research to specify. Structured patterns often provide a credible choice with far fewer parameters to be specified. An AR model has the correlation decaying too rapidly for the data we see. Hence, we prefer a generalization, the LEAR model (linear exponent AR; Simpson et al. 2010), which adds a (third) parameter to slow the decay speed. 3. The two components could be combined in a variety of ways. The nested structure makes a direct-product covariance pattern appealing. Simpson et al. (2014) discussed the approach and its advantages and disadvantages.

3.2. Power for the Pregnancy Study

We focus on early gestational changes (<12 weeks) for women less than 35 years old. Sample size will be selected to control power for cardiac output, one of six primary outcomes. Target changes at 12–16 weeks after the first measurement at 5–6 weeks are 30%, 40%, and 50% increases for egg donation, spontaneous conception, and standard IVF, respectively, relative to a mean cardiac output at 5–6 weeks of 5 L/min. Mean cardiac outputs at 7–9 weeks are predicted assuming linear changes within group. Cardiac output values are log (base 2) transformed to meet the homogeneity and Gaussian assumptions. Times are recorded as interval midpoints: 5.5, 8, and 14 weeks. The primary analysis compares changes in cardiac output among the pregnancy groups, while accounting for baseline values. The two sources of variation (time and clustering) are modeled separately and combined in a direct-product structure. Specifically, the covariance model for all observations in an independent sampling unit (a clinic) is $\mathbf{\Sigma}_m = \sigma^2(\mathbf{\Gamma}_c \otimes \mathbf{\Gamma}_t)$, with $\mathbf{\Gamma}_c = (1 - \rho_c)\boldsymbol{I}_{p_c} + \rho_c \mathbf{1}_{p_c}\mathbf{1}'_{p_c}$ the $p_c \times p_c$ uniform correlation structure for women within a clinic, and $\mathbf{\Gamma}_t$ the $p_t \times p_t$ correlation matrix among longitudinal observations within a woman.

Power analysis requires specifying the common correlation ρ_c among women in the same clinic, the common variance σ^2, and the $p_t \times p_t$ correlation matrix $\mathbf{\Gamma}_t$ among longitudinal responses. The analysis assumed $\rho_c = 0.08$ and $p_c = 5$ (clinics), giving a multiplier of $p_c^{-1}[1 + (p_c - 1)\rho_c] = 0.264$. Here, $\mathbf{\Gamma}_t$ is assumed to have a LEAR structure (Simpson, et al. 2010) with $\rho_t = 0.5$ and decay parameter $\delta_t = 0.5$ for correlations among the repeated cardiac outputs. Results in Van Belle and Martin (1993) allow specifying $\sigma^2 = 0.067 \log_2(\text{L/min})$ from the overall mean and standard deviation of cardiac output (5 L/min and 0.9 L/min, respectively).

Theorem 3 can be applied to the two-level design (time and clustering) to simplify the design by averaging across observations within a cluster. After averaging to remove the clustering structure, seven elements sufficient and required for computing power of the HLT test in a general linear multivariate model can be specified as follows (Muller et al. 1992, sec. 2.4). Subscripts Am and AM denote elements after transformation in the mixed or multivariate models, respectively. The corresponding elements pertaining to the linear mixed model are given in the supplementary materials.

1. The essence matrix of $\boldsymbol{X}_{AM}$ is created from labels defining the three pregnancy groups, with 17, 20, or 24 women per group for each clinic, and a total of $N \in \{255, 300, 360\}$ participants.
2. The Type I error rate $\alpha = 0.0083$ (0.05/6) accounts for multiple testing on the six primary outcome variables.
3. The between-subject contrasts matrix $\boldsymbol{C}_{AM}$ compares the three pregnancy groups.
4. The within-subject contrasts matrix $\boldsymbol{U}_{AM}$ tests linear and quadratic time trends.
5. The null matrix $\mathbf{\Theta}_{AM0} = \mathbf{0}$ (no difference in time trends between the three pregnancy groups).
6. The primary parameters $\boldsymbol{B}_{AM}$ are log-transformed (base 2) values for cardiac outputs, assuming a linear progression.
7. The covariance matrix of the subject-specific errors $\mathbf{\Sigma}_{AM}$ has all diagonal elements of 0.018 $(0.264 \cdot 0.067)$ and an associated LEAR structure to model the slow attenuation of correlation as measurements are taken farther apart in time. Here, $\boldsymbol{C}_{AM}$ and $\boldsymbol{U}_{AM}$ specify testing the pregnancy group $(\boldsymbol{C}_{AM})$ by time $(\boldsymbol{U}_{AM})$ interaction.

Figure 1 displays power as a function of the mean difference in log-transformed (base 2) cardiac output (τ) between egg donation recipients and standard IVF cases at 12–16 weeks. Mean differences between egg donation and spontaneous pregnancies, and between spontaneous pregnancies and standard IVF cases, are assumed half of the difference between egg donation recipients and standard IVF groups. For $t_* = (t - 5.5)/8.5$,

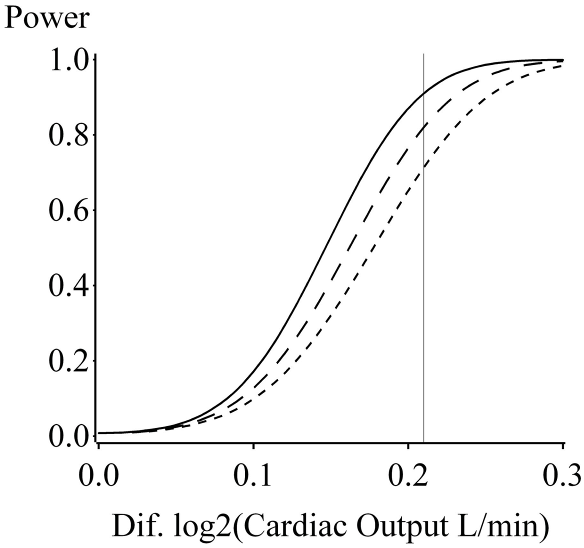

Figure 1. Power as a function of the difference in $\log_2$ cardiac output (τ) between groups at 12–16 weeks for $n = 360$ (solid), 300 (dashed), or 255 (dotted) participants.

$t \in (5.5,\ 8,\ 14)$, and parameter τ, the mean transformed cardiac output is $2.322 + 0.379t_*$ for egg donation reception, $2.322 + (\tau + 0.379)t_*$ for spontaneous conception, and $2.322 + (2\tau + 0.379)t_*$ for standard IVF. For $\tau = 0.21$, the group means increase by 30%, 40%, and 50%, from 5–6 weeks to 12–16 weeks. If $\tau = 0.21$, the predicted power is 0.91 for 24 per pregnancy group, for each of the five clinics. Power reduces to 0.82 (or 0.71), for 20 (or 17) per group for each clinic.

4. Discussion

Designing studies involving humans or animals typically demands considering ethical as well as statistical tradeoffs. Studies with samples that are too small can fail to achieve scientific goals, while studies with samples that are too large waste time and resources. In either scenario, an inaccurate sample size selection creates unnecessary risk to humans or animals.

Casting mixed model tests of fixed effects as multivariate tests provides a convenient way to assess power and select a sample size when planning a study. The analytic results in Section 2 describe model and hypothesis specifications that suffice to ensure the approach is accurate.

The sufficient conditions define common study scenarios in biomedical and behavioral research. Examples include comparing longitudinal trajectories between groups, and testing baseline impacts on changes of the outcome. The approach covers a common class of cluster designs, as well as a class of designs combining longitudinal and clustering features. Convenient adjustments provide good approximations for the impact of missing data in common scenarios.

Monte Carlo simulations provide a widely used alternative. Simulations should be avoided if there is power and sample size software available which follows industry standards. If they must be used, it becomes the responsibility of the programmer to demonstrate both the reliability and the validity of the calculations.

In the context of power analysis, our simulation results demonstrate a modest underestimation of power due to using an unstructured covariance. In many, but not all, study planning scenarios, a conservative power calculation would be acceptable.

Although specifying an unstructured covariance matrix is appealing for data analysis, the approach creates difficulties for power and sample size analysis. Hence, we recommend using a plausible structured covariance pattern for input to power analysis programs, while planning to conduct data analysis assuming an unstructured pattern to protect Type I error rate. The combination can introduce some conservatism in the power analysis.

Many questions about power for mixed models remain unanswered. Missing data and unequal cluster sizes occur in many practical settings. More work is needed to assess the accuracy of adopting the multivariate calculations for mixed model power approximations with missing data and unequal cluster sizes. Although the methods presented here cover a substantial fraction of longitudinal and multilevel designs used in the applied research, the methods are not universally appropriate. Variations and generalizations of the designs considered naturally demand extensions of the mixed model power and sample size tools for fixed-effects inference. For example, trials with participants belonging to more than one group require a mixed model with nonnested correlation structure to achieve valid statistical inference (Andridge et al. 2014).

Funding

The authors thank Dr. Kirk Conrad for permission to use details from research funded by PO1 HD065647-03. Chi's support included NIDCR R01-DE020832, NIDDK R01-DK072398, NIDA R01-DA031017, and NICHD PO1-HD065647-03. Glueck's support included NIGMS R01-GM121081-05, NIDCR R01-DE020832, NIGMS 1R25GM111901-01, and NLM 1G13LM011879-01. Muller's support included NIGMS R01-GM121081-05, NIDCR R01-DE020832, NIDDK R01-DK072398, NIDCR U54-DE019261, NCRR K30-RR022258, NHLBI R01-HL091005, NIAAA R01-AA013458-01, NIDA R01-DA031017, NICHD PO1-HD065647-03, NIGMS 1R25GM111901-01, and NLM 1G13LM011879-01.

References

Andridge, R. R., Shoben, A. B., Muller, K. E., and Murray, D. M. (2014), "Analytic Methods for Individually Randomized Group Treatment Trials and Group-Randomized Trials When Subjects Belong to Multiple Groups," *Statistics in Medicine*, 33, 2178–2190. [6,9]

Basagaña, X., Liao, X., and Spiegelman, D. (2011), "Power and Sample Size Calculations for Longitudinal Studies Estimating a Main Effect of a Time-Varying Exposure," *Statistical Methods in Medical Research*, 20, 471–487. [2]

Catellier, D. J., and Muller, K. E. (2000), "Tests for Gaussian Repeated Measures with Missing Data in Small Samples," *Statistics in Medicine*, 19, 1101–1114. [5]

Glueck, D. H., and Muller, K. E. (2003), "Adjusting Power for a Baseline Covariate in Linear Models," *Statistics in Medicine*, 22, 2535–2551. [5]

Gurka, M. J., Coffey, C. S., and Muller, K. E. (2007), "Internal Pilots for a Class of Linear Mixed Models with Gaussian and Compound Symmetric Data," *Statistics in Medicine*, 26, 4083–4099. [5]

Gurka, M. J., Edwards, L. J., and Muller, K. E. (2011), "Avoiding Bias In Mixed Model Inference for Fixed Effects," *Statistics in Medicine*, 30, 2696–2707. [2,7]

Hedeker, D., Gibbons, R. D., and Waternaux, C. (1999), "Sample Size Estimation for Longitudinal Designs with Attrition: Comparing Time-Related Contrasts Between Two Groups," *Journal of Educational and Behavioral Statistics*, 24, 70–93. [2]

Heo, M., and Leon, A. C. (2009), "Sample Size Requirements to Detect an Intervention by Time Interaction in Longitudinal Cluster Randomized Clinical Trials," *Statistics in Medicine*, 28, 1017–1027. [2,5]

Huynh, H., and Feldt, L. S. (1976), "Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs," *Journal of Educational Statistics*, 1, 69–82. [7]

Johnson, J. L., Muller, K. E., Slaughter, J. C., Gurka, M. J., Gribbin, M. J., and Simpson, S. L. (2009), "POWERLIB: SAS/IML Software for Computing Power in Multivariate Linear Models," *Journal of Statistical Software*, 30, 1–27. [5]

Kenward, M. G., and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, 983–997. [2]

——— (2009), "An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood," *Computational Statistics & Data Analysis*, 53, 2583–2595. [2]

Kreidler, S. M., Muller, K. E., Grunwald, G. K., Ringham, B. M., Coker-Dukowitz, Z. T., Sakhadeo, U. R., Barón, A. E., and Glueck, D. H. (2013), "GLIMMPSE: Online Power Computation for Linear Models With and Without a Baseline Covariate," *Journal of Statistical Software*, 54, 1–26. [5]

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2007), *SAS for Mixed Models*, Cary, NC: SAS institute. [7]

Lu, K. (2012), "Sample Size Calculations with Multiplicity Adjustment for Longitudinal Clinical Trials with Missing Data," *Statistics in Medicine*, 31, 19–28. [3]

McKeon, J. J. (1974), "F Approximations to the Distribution of Hotelling's T2 0," *Biometrika*, 61, 381–383. [5]

Muller, K. E., Edwards, L. J., Simpson, S. L., and Taylor, D. J. (2007), "Statistical Tests with Accurate Size and Power for Balanced Linear Mixed Models," *Statistics in Medicine*, 26, 3639–3660. [1,2,3,4]

Muller, K. E., Lavange, L. M., Ramey, S. L., and Ramey, C. T. (1992), "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications," *Journal of the American Statistical Association*, 87, 1209–1226. [3,5,8]

Muller, K. E., and Peterson, B. L. (1984), "Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis," *Computational Statistics & Data Analysis*, 2, 143–158. [5]

Muller, K. E., and Stewart, P. W. (2006), *Linear Model Theory: Univariate, Multivariate, and Mixed Models*, New York: Wiley. [2,3,4,5]

Murray, D. M., Blitstein, J. L., Hannan, P. J., Baker, W. L., and Lytle, L. A. (2007), "Sizing a Trial to Alter the Trajectory of Health Behaviours: Methods, Parameter Estimates, and Their Application," *Statistics in Medicine*, 26, 2297–2316. [2]

Park, T., Park, J.-K., and Davis, C. S. (2001), "Effects of Covariance Model Assumptions on Hypothesis Tests for Repeated Measurements: Analysis of Ovarian Hormone Data and Pituitary-Pteryomaxillary Distance Data," *Statistics in Medicine*, 20, 2441–2453. [2]

Ringham, B. M., Kreidler, S. M., Muller, K. E., and Glueck, D. H. (2016), "Multivariate Test Power Approximations for Balanced Linear Mixed Models in Studies with Missing Data," *Statistics in Medicine*, 17, 2921–2937. [1,4,6]

SAS Institute (2012), *SAS/STAT 12.1 User's Guide: Survey Data Analysis (book Excerpt)*, SAS Institute Incorporated. [6]

Schaalje, G. B., McBride, J. B., and Fellingham, G. W. (2002), "Adequacy of Approximations to Distributions of Test Statistics in Complex Mixed Linear Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 512–524. [2]

Siemer, M., and Joormann, J. (2003), "Power and Measures of Effect Size in Analysis of Variance with Fixed Versus Random Nested Factors," *Psychological Methods*, 8, 497–517. [3]

Simpson, S. L., Edwards, L. J., Muller, K. E., Sen, P. K., and Styner, M. A. (2010), "A Linear Exponent AR (1) Family of Correlation Structures," *Statistics in Medicine*, 29, 1825–1838. [8]

Simpson, S. L., Edwards, L. J., Styner, M. A., and Muller, K. E. (2014), "Kronecker Product Linear Exponent AR (1) Correlation Structures for Multivariate Repeated Measures," *PloS One*, 9, e88864. [8]

Stroup, W. (2002), "Power Analysis Based on Spatial Effects Mixed Models: A Tool for Comparing Design and Analysis Strategies in the Presence of Spatial Variability," *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 491–511. [3]

Stroup, W. W. (2012), *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*, CRC Press. [3]

Tu, X., Zhang, J., Kowalski, J., Shults, J., Feng, C., Sun, W., and Tang, W. (2007), "Power Analyses for Longitudinal Study Designs with Missing Data," *Statistics in Medicine*, 26, 2958–2981. [2]

Van Belle, G., and Martin, D. C. (1993), "Sample Size as a Function of Coefficient of Variation and Ratio of Means," *The American Statistician*, 47, 165–167. [8]

Wang, C., Hall, C. B., and Kim, M. (2015), "A Comparison of Power Analysis Methods for Evaluating Effects of a Predictor on Slopes in Longitudinal Designs with Missing Data," *Statistical Methods in Medical Research*, 24, 1009–1029. [2]