STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:3639–3660
Published online 29 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2827

Statistical tests with accurate size and power for balanced linear mixed models

Keith E. Muller[1,*,†], Lloyd J. Edwards[2], Sean L. Simpson[2] and Douglas J. Taylor[3]

[1]*Department of Epidemiology and Health Policy Research, 1329 SW 16th Street Room 5125, P.O. Box 100177 Gainesville, FL 32610-0177, U.S.A.*
[2]*Department of Biostatistics CB #7420, University of North Carolina, Chapel Hill, NC 27599-7420, U.S.A.*
[3]*Family Health International, Research Triangle Park, NC 27709, U.S.A.*

SUMMARY

The convenience of linear mixed models for Gaussian data has led to their widespread use. Unfortunately, standard mixed model tests often have greatly inflated test size in small samples. Many applications with correlated outcomes in medical imaging and other fields have simple properties which do not require the generality of a mixed model. Alternately, stating the special cases as a general linear multivariate model allows analysing them with either the univariate or multivariate approach to repeated measures (UNIREP, MULTIREP). Even in small samples, an appropriate UNIREP or MULTIREP test always controls test size and has a good power approximation, in sharp contrast to mixed model tests. Hence, mixed model tests should never be used when one of the UNIREP tests (uncorrected, Huynh–Feldt, Geisser–Greenhouse, Box conservative) or MULTIREP tests (Wilks, Hotelling–Lawley, Roy's, Pillai–Bartlett) apply. Convenient methods give exact power for the uncorrected and Box conservative tests. Simulations demonstrate that new power approximations for all four UNIREP tests eliminate most inaccuracy in existing methods. In turn, free software implements the approximations to give a better choice of sample size. Two repeated measures power analyses illustrate the methods. The examples highlight the advantages of examining the entire response surface of power as a function of sample size, mean differences, and variability. Copyright 2007 John Wiley & Sons, Ltd.

KEY WORDS: univariate approach to repeated measures; UNIREP; complete data; Geisser–Greenhouse test; Huynh–Feldt test; sample size

[*]Correspondence to: Keith E. Muller, Department of Epidemiology and Health Policy Research, 1329 SW 16th Street Room 5125, P.O. Box 100177 Gainesville, FL 32610-0177, U.S.A.
[†]E-mail: Keith.Muller@Biostat.ufl.edu

Contract/grant sponsor: NIH NHLBI; contract/grant number: 5 T32 HL007773-09
Contract/grant sponsor: National Center on Minority Health and Health Disparities; contract/grant number: 5 R24 MD000167-01
Contract/grant sponsor: Agency for Healthcare Research and Quality; contract/grant number: 1 R24 HS13353-01
Contract/grant sponsor: Family Health International; contract/grant number: CCP-3079-00-5022-00
Contract/grant numbers: NCI P01 CA47 982-04, NCI RO-1 CA095749-01A1, NIAID 9P30 AI 50410

Received 5 January 2006
Accepted 12 December 2006

Copyright 2007 John Wiley & Sons, Ltd.

1. INTRODUCTION

1.1. Motivation

Reflecting the ever increasing role of imaging in medicine, September 2000 saw the creation of the National Institute of Biomedical Imaging and Bioengineering at the National Institutes of Health in the United States. The tight control typical of medical imaging research often yields repeated measures with little or no missing or mistimed data in the final analysis. Complete and balanced data within independent sampling units allow analysis with either mixed model or multivariate general linear model methods. As discussed in Sections 1.2 and 5.4, mixed model tests can badly inflate test size for such cases, while multivariate model tests do not. A multivariate data analysis also allows selecting a sample size with methods that are well-founded, much more convenient, and have dependable accuracy.

1.2. Choosing a model and a test for analysis

Many biostatisticians choose a mixed model by default for linear models with correlated outcomes and Gaussian errors. However, small sample sizes typically impose a severe penalty. Although estimation of means is well-behaved, test size can be terribly inflated. Catellier and Muller [1] observed simulated test size as high as 0.59 with a target of 0.05 for a mixed model test of time × group (interaction between the repeated measure and the grouping factor) with complete and balanced data. The simulation assumed unstructured covariance, 12 participants in each of four groups, and six repeated measures. Simulations in the present paper (Table III) and by other authors [2] display the same problem. Likelihood ratio tests may be a viable alternative in the mixed model. However, such tests are ignored here due to the lack of general small sample approximations and published simulations.

Linear mixed models allow missing or mistimed data, repeated covariates, and modelling the covariance pattern, while the general linear multivariate model does not. With no missing or mistimed data, multivariate models assume unstructured covariance and apply to repeated measures, profile data, or any set of correlated outcomes. Any multivariate model can be expressed as a special case of a mixed model. In such cases, the mixed and multivariate model mean estimates coincide.

Although mean estimates often coincide between mixed and multivariate models, hypothesis testing and confidence intervals usually differ greatly, except in very special cases. The current variety of mixed model tests generally fail to control test size in small samples because they use what Littel [3] described as ‘approximations piled on approximations.’ For the sake of brevity, we avoid most discussion of the machinery of mixed model tests. Demidenko [4] gave the most recent and comprehensive treatment of the theory. He also discussed a range of applications of mixed models to imaging.

In the absence of missing or mistimed data, the multivariate model, as well as the mixed model, accommodates correlated responses with a common scale (repeated measures data), and correlated responses with a variety of scales (‘multivariate’ data). The hypothesis tests commonly used with the multivariate model fall into two distinct groups: tests for the multivariate approach to repeated measures (MULTIREP: Wilks, Hotelling–Lawley, Roy’s, Pillai–Bartlett) and tests for the univariate approach to repeated measures (UNIREP: uncorrected, Huynh–Feldt (HF), Geisser–Greenhouse (GG), Box conservative). In contrast to mixed model tests, MULTIREP and UNIREP tests always control test size well, even with the smallest sample size. The multivariate model tests also have good power methods, while little is known about power for mixed models.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

The UNIREP approach was originally developed under the assumption of a compound symmetric covariance structure, which leads to the uncorrected test being uniformly most powerful among similarly invariant tests of size α. Violation of compound symmetry can greatly inflate test size of the uncorrected test. However, the HF, GG, and Box conservative tests control test size for any covariance pattern. Although the three UNIREP tests always prove fully robust to misspecification of covariance structure, mixed model tests can badly inflate test size with misspecification.

The choice between a UNIREP test and a MULTIREP test reduces to considering power. Both types appeal because no test has uniformly most power (among similarly invariant tests of size α) except in special cases. However, power software allows determining the best test for any particular combination of covariance and mean structure. A covariance matrix with all equal eigenvalues $\{\lambda_k\}$, which are the variances of the principal components, has spherical principal components. Our focus centres on tests with error covariance patterns somewhat close to sphericity (little variability among λ_k), which leads to preferring the GG or HF corrected tests. Simulation results [1, 5] support the preference because both tests approximately control test size, and give more power than the MULTIREP tests near sphericity. Not surprisingly, the MULTIREP tests have more power than the UNIREP tests for covariance patterns far away from sphericity. In practice, a credible power analysis for the hypothesis of interest will make clear the best choice among the MULTIREP and UNIREP tests for the study being planned. We believe good statistical practice requires using a MULTIREP or UNIREP test rather than a mixed model test, whenever possible, for small to moderate samples. The reader seeking more detailed comparisons of approaches may consult ‘Choosing the Form of a Linear Model for Analysis’ in [6].

The present work was motivated by the need for better approximations of power for GG and HF tests. Their appeal and the appeal of MULTIREP tests lie in robustness to misspecifying the covariance structure, as well as accurate test size in small samples. In contrast, the validity of the mixed model tests depends on correctly specifying the covariance model. Even with correct covariance specification, mixed model tests may still inflate test size.

All widely used software packages that we know provide simple access to the data analysis methods we recommend. However, no current mixed model software that we have seen provides MULTIREP and UNIREP tests. Using the tests requires fitting a multivariate model with a multivariate procedure.

1.3. Previous related work

Muller and Barton [5] described power approximations for the GG and HF tests. Glueck and Muller [7] generalized the method to adjust for a baseline covariate. Muller *et al.* [8] reviewed power approximations for the MULTIREP and UNIREP tests. Muller and Stewart [6] presented the details of the underlying theory and motivating examples for all models and methods discussed in the present paper.

Limitations of mixed model tests led Catellier and Muller [1] to consider approximate MULTIREP and UNIREP tests for repeated measures with missing data. The approximations controlled test size well in simulations even with small sample size. We leave considering power for missing data to future research.

1.4. The need for better UNIREP power approximations

Coffey and Muller [9] reported cases in which the Muller and Barton [5] approximations failed to provide even one digit of accuracy (approximate GG power of 0.80 and a simulated power of

Copyright 2007 John Wiley & Sons, Ltd.

0.98 ± 0.002 standard error). The offending conditions, based on observed data, are part of the simulations in the present paper (simulation 1; row 12 of Table VI).

Given reasonably accurate inputs, a much smaller sample size would be chosen with a more accurate power approximation. With uncertainty about the inputs, we emphasize the value of a thorough sensitivity analysis. In a Gaussian linear model, such an analysis can be especially useful when based on plots relating to mean differences, sample size and power. More accurate power approximation gives a more accurate sample size response surface, as in Figure 2 (discussed in Section 3.1), Figures 5 and 6 (discussed in Section 3.2) and Figure 7 (discussed in Section 6.4).

1.5. Conclusions from the present research

The results reinforce the advantages of using a UNIREP or MULTIREP analysis whenever possible, in lieu of a mixed model analysis. New power approximations (implemented in free software) eliminate accuracy limitations of previous methods. In addition, the methods allow conveniently describing the entire space of plausible designs in terms of sample size, mean differences, and variability.

2. FORMULATING A MULTIVARIATE MODEL AND ANALYSIS

2.1. Model notation and estimation

A vector $\mathbf{x}$ is always $n \times 1$, while a matrix, $\mathbf{X}$, has transpose $\mathbf{X}'$. Here, $\mathbf{1}_n$ is an $n \times 1$ vector of 1's, $\mathrm{Dg}(\mathbf{x})$ is diagonal with (i, i) element x_i, $\mathbf{M}^{-1}$ is the inverse, and $\mathbf{M}^{-}$ a generalized inverse. Spectral decomposition of $\mathbf{M}=\mathbf{M}'$ gives $\mathbf{M}=\mathbf{V}\mathrm{Dg}(\boldsymbol{\lambda})\mathbf{V}'$, with $\mathbf{V}$ orthonormal and $\lambda_k \geqslant \lambda_{k+1}$. Detailed treatments of random variables discussed here can be found in [10–14].

For convenience, the independent sampling unit will be described as a person, and the index for repeated measures is referred to as time. In the multivariate model

$$\mathbf{Y}=\mathbf{X}\mathbf{B}+\mathbf{E} \tag{1}$$

rows of $\mathbf{Y}(N \times p)$, $\mathbf{X}(N \times q)$, and $\mathbf{E}$ correspond to persons, columns of $\mathbf{Y}$, $\mathbf{B}(q \times p)$ and $\mathbf{E}$ to time, while columns of fixed, known design matrix $\mathbf{X}$ and rows of fixed, unknown $\mathbf{B}$ correspond to predictors. With $N > \mathrm{rank}(\mathbf{X})$ independent rows of $\mathbf{E}$, $\mathrm{row}_i(\mathbf{E})' \sim \mathscr{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ indicates a Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$.

The general linear hypothesis, H_0: $\boldsymbol{\Theta}=\boldsymbol{\Theta}_0$, considers $\boldsymbol{\Theta}=\mathbf{C}\mathbf{B}\mathbf{U}$ $(a \times b)$, for fixed and known $\mathbf{C}$, $\mathbf{U}$ and $\boldsymbol{\Theta}_0$. Here $\mathbf{C}(a \times q)$ gives contrasts between person, and $\mathbf{U}(p \times b)$ gives contrasts within person. A testable hypothesis has (1) $\mathbf{M}=\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}'$ with full rank of a, (2) $\mathbf{U}$ with full (column) rank of b, and (3) $\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})=\mathbf{C}$.

Least-squares estimates are $\widetilde{\mathbf{B}}=(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ and $\widehat{\boldsymbol{\Theta}}=\mathbf{C}\widetilde{\mathbf{B}}\mathbf{U}$. If $v_e=N-\mathrm{rank}(\mathbf{X})$ then the residuals $\widehat{\mathbf{E}}=\mathbf{Y}-\mathbf{X}\widetilde{\mathbf{B}}$ give $\widehat{\boldsymbol{\Sigma}}=\widehat{\mathbf{E}}'\widehat{\mathbf{E}}/v_e(p \times p)$ for the model and $\widehat{\boldsymbol{\Sigma}}_*=\mathbf{U}'\widehat{\boldsymbol{\Sigma}}\mathbf{U}=\mathbf{S}_e/v_e(b \times b)$ for the test. The $b \times b$ hypothesis sum of squares matrix is $\mathbf{S}_h=\widehat{\boldsymbol{\Delta}}=(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)$. The $b \times b$ matrix $\mathbf{S}_e$ contains the corresponding error sums of squares. Also $\mathbf{S}_h \sim \mathscr{W}_b(a, \boldsymbol{\Sigma}_*, \boldsymbol{\Delta}\boldsymbol{\Sigma}_*^{-1})$, a Wishart with a degrees of freedom, covariance $\boldsymbol{\Sigma}_*$, and non-centrality $\boldsymbol{\Delta}\boldsymbol{\Sigma}_*^{-1}$, while $v_e\widehat{\boldsymbol{\Sigma}}_* \sim \mathscr{W}_b$ $(v_e, \boldsymbol{\Sigma}_*)$.

2.2. *Tests*

As summarized in Table I, each of the MULTIREP and UNIREP test statistics can be expressed in terms of the sums of squares hypothesis and error matrices, $\mathbf{S}_h$ and $\mathbf{S}_e$ [5, 6]. A single test must be chosen *a priori* in order to preserve statistical validity. Except for Roy's largest root, approximate p values and powers may computed with F approximations. Numerator and denominator degrees of freedom are in Table I. Corresponding power approximations use the same degrees of freedom and compute a non-centrality parameter based on using population values to replace sample values in the test statistic. The logic and history of the approach are given in [8].

The eigenvalues of $\mathbf{\Sigma}_*$, indicated $\{\lambda_k\}$, equal the variances of the principal components in the error space of the hypothesis. For UNIREP tests the parameter

$$\varepsilon = \mathrm{tr}^2(\mathbf{\Sigma}_*)/[b\,\mathrm{tr}(\mathbf{\Sigma}_*^2)] = \left(\sum_{k=1}^{b} \lambda_k/b\right)^2 \Bigg/ \left(\sum_{k=1}^{b} \lambda_k^2/b\right) \tag{2}$$

Table I. MULTIREP and UNIREP test statistics based on $\mathbf{S}_h = \widehat{\mathbf{\Delta}}$ and $\mathbf{S}_e = v_e \widehat{\mathbf{\Sigma}}_*$ from a multivariate model.

				Null F approximation d.f.	
Test	Statistic	Principle	Univariate	$v_1(d)$	$v_2(d)$
HLT	$\mathrm{tr}(\mathbf{S}_h\mathbf{S}_e^{-1})$	ANOVA analog	$\dfrac{\mathrm{SSH}}{\mathrm{SSE}}$	ab	$g_1(v_e, a, b)$
PBT	$\mathrm{tr}[\mathbf{S}_h(\mathbf{S}_h+\mathbf{S}_e)^{-1}]$	Substitution	$\dfrac{\mathrm{SSH}}{\mathrm{SSH}+\mathrm{SSE}}$	$ab\dfrac{g_2(v_e, a, b)}{s(v_e+s-b)}$	$g_2(v_e, a, b)$
WLK	$\lvert\mathbf{S}_e(\mathbf{S}_h+\mathbf{S}_e)^{-1}\rvert$	Likelihood ratio	$\dfrac{\mathrm{SSE}}{\mathrm{SSH}+\mathrm{SSE}}$	ab	$g_3(v_e, a, b)$
RLR	max eigenvalue $\mathbf{S}_h(\mathbf{S}_h+\mathbf{S}_e)^{-1}$	Union-intersection	$\dfrac{\mathrm{SSH}}{\mathrm{SSH}+\mathrm{SSE}}$	(none)	(none)
UN	$\mathrm{tr}(\mathbf{S}_h)/\mathrm{tr}(\mathbf{S}_e)$	Most power for sphericity	$\dfrac{\mathrm{SSH}}{\mathrm{SSE}}$	ab	$v_e b$
HF	$\mathrm{tr}(\mathbf{S}_h)/\mathrm{tr}(\mathbf{S}_e)$	$E(\tilde{\varepsilon}) \approx \varepsilon$	$\dfrac{\mathrm{SSH}}{\mathrm{SSE}}$	$ab\tilde{\varepsilon}$	$v_e b\tilde{\varepsilon}$
GG	$\mathrm{tr}(\mathbf{S}_h)/\mathrm{tr}(\mathbf{S}_e)$	$\hat{\varepsilon}$ is MLE	$\dfrac{\mathrm{SSH}}{\mathrm{SSE}}$	$ab\hat{\varepsilon}$	$v_e b\hat{\varepsilon}$
Box	$\mathrm{tr}(\mathbf{S}_h)/\mathrm{tr}(\mathbf{S}_e)$	$\varepsilon \geqslant 1/b$	$\dfrac{\mathrm{SSH}}{\mathrm{SSE}}$	a	v_e

Notes:

$$g_1(v_e, a, b) = \frac{[v_e^2 - v_e(2b+3) + b(b+3)](ab+2)}{v_e(a+b+1) - (a+2b+b^2-1)} + 4$$

$$g_2(v_e, a, b) = \frac{v_e+s-b}{v_e+a}\left[\frac{s(v_e+s-b)(v_e+a+2)(v_e+a-1)}{v_e(v_e+a-b)} - 2\right]$$

$$\frac{g_3(v_e, a, b)}{[v_e-(b-a+1)/2]-(ab-2)/2} = \begin{cases} 1 & a^2b^2 \leqslant 4 \\ [(a^2b^2-4)/(a^2+b^2-5)]^{1/2} & \text{otherwise} \end{cases}$$

Copyright 2007 John Wiley & Sons, Ltd.

indexes the amount of sphericity, and helps simplify any discussion of test size or power. In general $1/b \leqslant \varepsilon \leqslant 1$, while $\varepsilon = 1$ only with sphericity. Here, ε measures sphericity of the hypothesis error variables, with perfect sphericity giving $\varepsilon = 1$ and least sphericity giving $\varepsilon = 1/b$. Although compound symmetry of $\mathbf{\Sigma}$ suffices to guarantee exact test size for the uncorrected test, sphericity of $\mathbf{\Sigma}_* = \mathbf{U}'\mathbf{\Sigma}\mathbf{U}(b \times b)$, namely $\mathbf{\Sigma}_* = \mathbf{I}_b \sigma_*^2$ describes the (weaker) necessary and sufficient condition.

All four UNIREP tests use the same test statistic

$$f_u = [\mathrm{tr}(\widehat{\mathbf{\Delta}})/a]/\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*) \tag{3}$$

with null approximation

$$\Pr(f_u \leqslant f_0) \approx F_F(f_0; \nu_{*1}, \nu_{*2}) \tag{4}$$

The non-central approximation is a direct generalization. With the spectral decomposition of $\mathbf{\Sigma}_*$ given by $\mathbf{\Sigma}_* = \mathbf{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})\mathbf{\Upsilon}'$ and $\mathbf{\Delta}_* = \mathbf{\Upsilon}'\mathbf{\Delta}\mathbf{\Upsilon}$, power varies directly with the elements of $\boldsymbol{\omega}_*$, namely $\omega_{*k} = \boldsymbol{\upsilon}_k'(\mathbf{\Theta} - \mathbf{\Theta}_0)'\mathbf{M}^{-1}(\mathbf{\Theta} - \mathbf{\Theta}_0)\boldsymbol{\upsilon}_k/\lambda_k$, a diagonal element of $\mathbf{\Omega}_* = \mathbf{\Delta}_*\mathrm{Dg}(\boldsymbol{\lambda})^{-1}$. Coffey and Muller [9] proved that the exact non-central UNIREP test statistic cumulative distribution function (CDF) depends only on $\{a, b, \nu_e, \boldsymbol{\lambda}, \boldsymbol{\omega}_*\}$. The improved non-central approximation may be stated

$$\Pr(f_u \leqslant f_0) \approx F_F\left(f_0 \frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e}; \nu_{*1}, \nu_{*2}, \omega_u\right) \tag{5}$$

All coefficients are defined in the theorem in Appendix A.

Although the non-central distribution approximation applies to all four UNIREP tests, the differences among critical value leads to four distinct tests and hence four distinct power approximations (one for each test). The uncorrected test assumes $\varepsilon = 1$, and the approximations simplify to exact results. The Box test approximation assumes the worst case of $\varepsilon = 1/b$ and reduces the degrees of freedom by the factor $1/b$. The GG test reduces degrees of freedom by the MLE of ε, namely $\widehat{\varepsilon} = b^{-1}\mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*)/\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*^2)$, while the HF test uses $\tilde{\varepsilon} = (Nb\widehat{\varepsilon} - 2)/[b(\nu_e - b\widehat{\varepsilon})]$, truncated to $\tilde{\varepsilon}_t = \min(\tilde{\varepsilon}, 1)$. The uncorrected and Box test critical values are the constants $f_{\mathrm{crit}}(\mathrm{Uncorrected}) = F_F^{-1}(1 - \alpha; ab, b\nu_e)$ and $f_{\mathrm{crit}}(\mathrm{Box}) = F_F^{-1}(1 - \alpha; a, \nu_e)$. For data analysis, random multipliers $\widehat{\varepsilon}$ and $\tilde{\varepsilon}$ give random critical values for the GG and HF tests of $f_{\mathrm{crit}}(\mathrm{GG}) = F_F^{-1}(1 - \alpha; ab\widehat{\varepsilon}, b\nu_e\widehat{\varepsilon})$ and $f_{\mathrm{crit}}(\mathrm{HF}) = F_F^{-1}(1 - \alpha; ab, b\nu_e\tilde{\varepsilon})$.

Muller and Barton [5] used approximate expected values of the degree of freedom multipliers to give approximate expected critical values and power. As an example, with E the approximate expected value of $\widehat{\varepsilon}$, GG power is approximated with $f_0 = F_F^{-1}[1 - \alpha; abE, b\nu_e E]$ in equation (5). The improved method described here (implemented in free software) matches two moments of the numerator and denominator of the test statistic. Most analytic details can be found in [5, 15], while Appendix A contains the logical basis of the new methods.

Specifying $\mathbf{\Sigma}$, $\mathbf{X}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{U}$, $\mathbf{\Theta}_0$, and α completely determines a power analysis. A data analyst starting from a complete and balanced mixed model that corresponds to a multivariate model may compute power as follows. First, specifying the 'fixed effects' determines $\boldsymbol{\beta}$ as a $qp \times 1$ matrix. Reshaping it into a $q \times p$ matrix gives $\mathbf{B}$. Second, specifying the 'random effects' covariance matrix, $\mathbf{\Sigma}_r$, and the 'pure' error covariance $\mathbf{\Sigma}_e$ (often $\sigma^2\mathbf{I}_p$) combine to give the multivariate error covariance, $\mathbf{\Sigma} = \mathbf{\Sigma}_r + \mathbf{\Sigma}_e$. Third, specifying $\mathbf{C}$ and $\boldsymbol{\theta}_0$ in terms of the $qp \times 1$ vector $\boldsymbol{\beta}$, and then converting them to $\mathbf{C}$, $\mathbf{U}$, and $\mathbf{\Theta}_0$ to agree with $\mathbf{B}$ completes the process.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

3. EXAMPLES

3.1. Improving mammography readings: two within-person factors, no between

Digital mammography stores images as numbers, which allows processing to improve quality. Computer scientists developed the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm to improve contrast [16] Pisano *et al.* [17] compared observers' ability to detect breast cancer in mammograms as a function of CLAHE clip and region levels. Region denotes the size of the image (pixels^2) at which contrasts are controlled. Clip level limits the maximum contrast adjustment.

Observers represented the independent sampling units. They were difficult to recruit due to the amount of time required. The variation between observers was known to be far greater than the variation within. Hence, a completely within-person design was chosen for the most power.

All observers served in 10 conditions, $3\times 3=9$ clip × region combinations and an unprocessed condition (CLAHE not applied). Observers were asked to find the lesion appearing in 1 of 4 portions of a mammogram. Counterbalancing was used to compensate for the potential sequence biases. A probit model for proportion correct as a function of contrast (the unitless ratio of target to background image density) was fitted separately for each condition to give the $\log_{10}$(contrast) predicting 88 per cent correct detection. Subtracting the unprocessed condition $\log_{10}$(contrast) from the same measure in the 9 clip × region combinations created 9 response variables.

The primary analysis used a repeated measures test of the clip × region interaction, with a nominal test size of $\alpha=0.04$. The secondary analysis consisted of 9 t tests with a nominal test size of $\alpha=0.01/9$. The t tests were scientifically redundant and likely had less power than the global interaction tests. Power calculations for t tests are well known. Hence, only the interaction test is considered.

The mammography study included only one group. In the multivariate model $\mathbf{X}=\mathbf{1}_N$, while within-person factors clip and region gave $N\times 9\mathbf{Y}$. Also, $\mathbf{B}(1\times 9)$ contained mean $\log_{10}$(contrast) for the unprocessed condition minus the mean for each of the nine combinations of clip and region ($\beta_{\text{cr}}=\mu_{\text{unprocessed}}-\mu_{\text{cr}}$). Choosing a sample size ensuring adequate power for the clip × region interaction illustrates how to use the power methods. Appendix B contains the contrast matrices.

Figure 1 displays the pattern of interaction means predicted from a previous unpublished and similar study. The outcomes of interest are the mean $\log_{10}$(contrast) differences between the unprocessed condition and the 9 combinations of clip and region. Using $\alpha=0.04$ gave corresponding

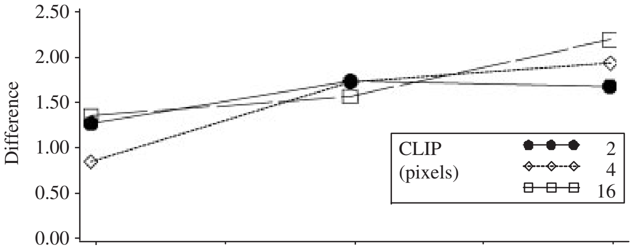

Figure 1. Predicted differences in estimated mean $\log_{10}$(contrast) for previous study of mammogram image processing.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

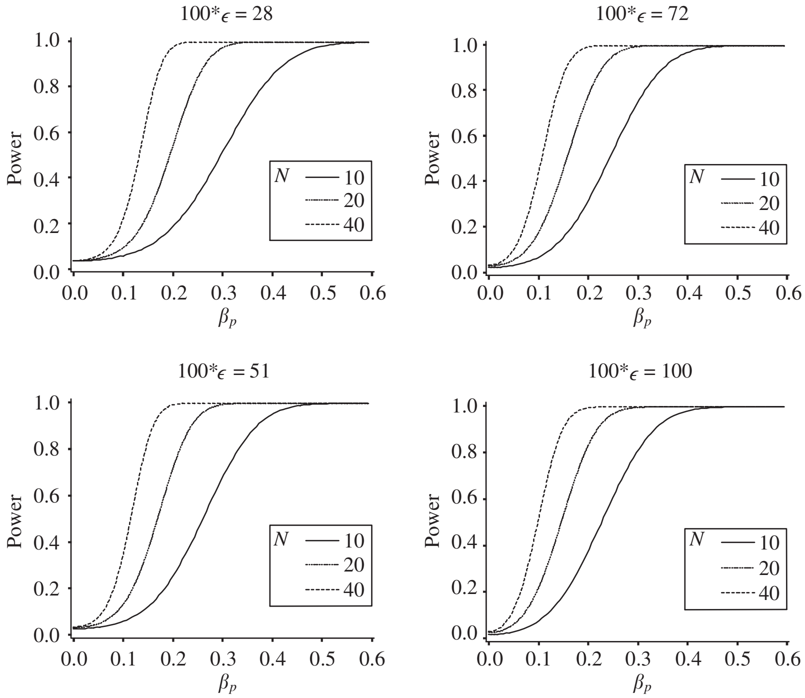

Figure 2. CLAHE approximate Geisser–Greenhouse power (new approximation method, new CDF) for clip × region interaction, with β_P index of effect (in Section 3.1, detailed in Appendix C). $N \in \{10, 20, 40\}$, with ε indexing non-sphericity.

approximate power of 0.828 for the GG test with $N = 15$ and $\varepsilon = 0.28$ (essentially the worst case of least possible sphericity, $\varepsilon = 1/b = 0.25$). In general, $1/b \leqslant \varepsilon \leqslant 1$.

Figure 2 displays GG power for $N \in \{10, 20, 40\}$, $\varepsilon \in \{0.28, 0.51, 0.72, 1.00\}$ and effect $\beta_P \in [0, 0.6]$, with β_P the scaling factor for **B** corresponding to target power $P \in \{0.20, .0.50, 0.80\}$ for the Muller–Barton GG approximation (Appendix C has details). For $\varepsilon = 0.28$, power essentially asymptotes to 1.0 near $\beta_P = 0.2$ for $N = 40$, and near $\beta_P = 0.3$ for $N = 20$. Therefore, a sample size of $N = 40$ is needed for $\beta_P \in [0.20, 0.30)$ while $N = 20$ suffices for $\beta_P \in [0.30, 0.60]$. In turn, a sample size of $N = 10$ suffices only when $\beta_P \gtrsim 0.60$. Overall, as ε increases, the sample size required to ensure adequate power decreases.

3.2. *Tortuosity study: two between-person factors, one within*

A power analysis for gender and brain region differences in cerebral vessel tortuosity further illustrates the use of the methods. Figure 3 (courtesy of E. Bullitt) displays a vessel map for a normal person, with roughly 25–50 segments in each of four regions of the brain (right middle, left middle, posterior, and anterior). Recent advances allow measuring cerebral vascular tortuosity (bending, twisting, or winding) automatically from magnetic resonance imaging (MRI). The ability to quantify tortuosity holds great promise in diagnosing disease.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

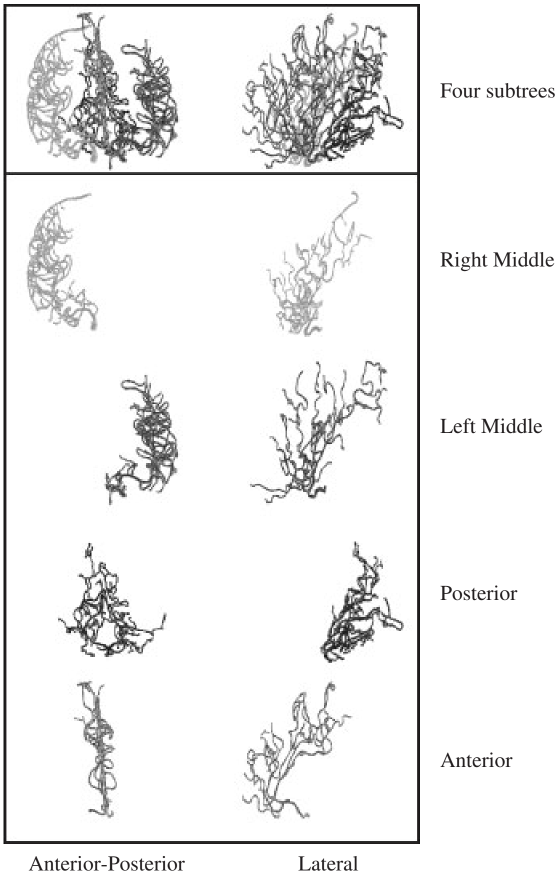

Figure 3. Four regions (right middle, left middle, posterior, anterior) of cerebral vasculature from two views (anterior–posterior, lateral). Vessel map created automatically from MRI. Region of the brain.

The main objective of the study was to create a pool of ‘normal’ brains. Knowledge of brain neuroanatomy led to the expectation of differences within a brain across regions, and to a speculation of a gender difference. Although age seems a natural source of variation, anecdotal knowledge from neurosurgery makes it somewhat uncertain whether strong systematic patterns are present among adults. Overall, the study was designed to have good power for the most complex hypothesis of concern, namely gender × region.

Power analysis focused on a single response variable computed separately in the four regions of the brain. The response variable SOAM1 indicates the sum of all positive angles between successive trios of equally spaced vessel points, divided by total path length (radians/cm), for all vessels in a region. The study included a single within-person factor, region of the brain. The multivariate model had $N \times 4\mathbf{Y}$, with columns anterior, left middle, posterior, and right middle. A factorial design between-persons for gender × age group (20–30, 30–40, 40–50, 50–60, 60+ years old) implied 10 columns in $\mathbf{X}$. The balanced design had $N/10$ participants in each cell and $\mathbf{X}=\mathbf{I}_{10}\otimes\mathbf{1}_{N/10}$. With a cell mean coding, $\mathbf{B}(10\times 4)$ contained mean tortuosity for each combination of age, gender and brain region. Appendix B contains the contrast matrices.

Bullitt *et al.* [18] provided a credible value for $\mathbf{\Sigma}$ of SOAM1 (radians/cm) in four regions (anterior, left middle, posterior, right middle):

$$\mathbf{\Sigma}=\begin{bmatrix}0.0838 & 0.0502 & 0.0356 & 0.0533\\ 0.0502 & 0.0537 & 0.0325 & 0.0333\\ 0.0356 & 0.0325 & 0.0441 & 0.0386\\ 0.0533 & 0.0333 & 0.0386 & 0.0722\end{bmatrix} \tag{6}$$

Diagnostics led to concluding that the data appeared Gaussian with approximately compound symmetric covariance. Here, $\widehat{\varepsilon}=\mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*)/[b\mathrm{tr}(\widehat{\mathbf{\Sigma}}^2_*)]\approx 0.85$ is roughly the expected value under sphericity, from simulations in [5]. The locally best invariant (LBI) test of sphericity statistic is a 1–1 function of $\widehat{\varepsilon}$. Grieve [19] compared the LBI and maximum likelihood (Mauchley) tests. Although the data make the assumption of compound symmetry plausible, we chose the GG test for two reasons. First, GG controls test size for any covariance pattern. Second, the test loses little power relative to the uncorrected test with compound symmetry.

An adequate sample size would provide the desired power for the gender × region interaction hypothesized, which led to proposing

$$\begin{aligned}\mathbf{B}=\mu\cdot\left(\mathbf{1}_5\otimes\begin{bmatrix}1&1&1&1\\1&1&1&1\end{bmatrix}\right)&+\mathbf{0}+\mathbf{0}+\mathbf{0}+\mathbf{0}\\ +\delta_{\mathrm{G}}\cdot\left(\mathbf{1}_5\otimes\begin{bmatrix}0&0&0&0\\1&1&1&1\end{bmatrix}\right)&\\ +\delta_{\mathrm{R}}\cdot\left(\mathbf{1}_5\otimes\begin{bmatrix}-1&0&1&0\\-1&0&1&0\end{bmatrix}\right)&\\ +\delta_{\mathrm{GR}}\cdot\left(\mathbf{1}_5\otimes\begin{bmatrix}0&0&1&0\\0&0&0&0\end{bmatrix}\right)&\end{aligned} \tag{7}$$

Here, μ represents the grand mean while δ_{G}, δ_{R}, and δ_{GR} correspond to the effects of gender, region, and the gender × region interaction, respectively. For the sake of brevity, the $\mathbf{0}$ matrices reflect the assumption that no age, age × gender, age × region, or age × gender × region effect occur.

Particular values were chosen as follows. The previous study gave plausible values of $\mu=3.2$ and $\delta_{\mathrm{R}}=0.30$. The δ_{GR} interaction parameter corresponds to a difference in the posterior region. A more complicated interaction would increase the power. Since no guidance was available about the interaction from the previous study, a conservative form was used. Full rank coding schemes, as in the example, have big advantages in power analysis, perhaps even more than in data analysis [20, Chapters 12–16]. In general, main effects need not be included when computing power for an

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

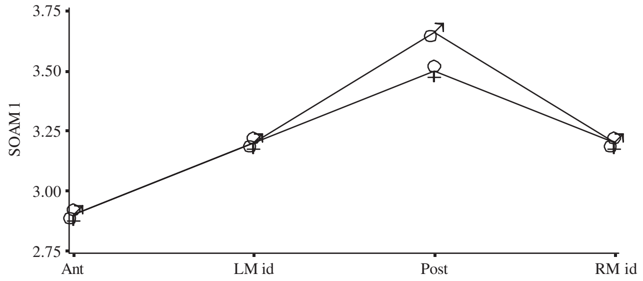

Figure 4. Hypothesized gender × region interaction for cerebral vessel tortuosity (radians/cm) based on means from previous tortuosity study, $\alpha = 0.05/6$.

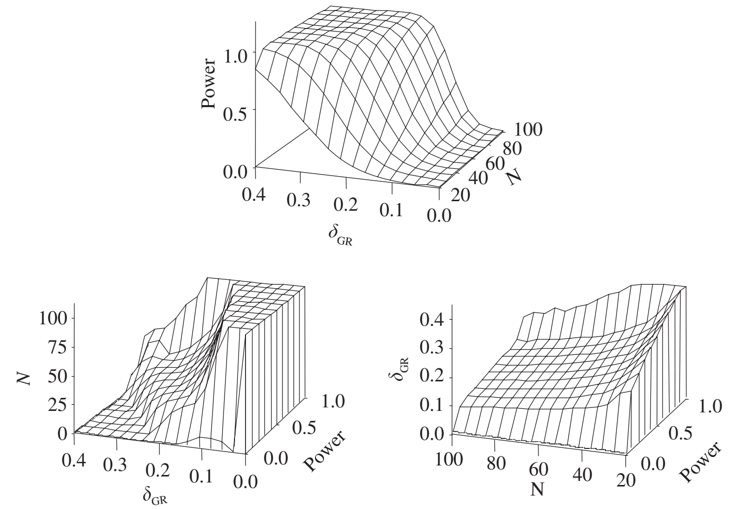

Figure 5. Approximate Geisser–Greenhouse power (new approximation method, new CDF) for gender × brain region effect on cerebral vessel tortuosity, with δ_{GR} size of effect (radians/cm, equation (7)), $\alpha = 0.05/6$. δ_{GR} = gender × region interaction (radians/cm, equation (7)).

interaction since they will cancel in the calculations. Figure 4 displays the pattern of interaction means that results if $\delta_{\mathrm{GR}} = 0.16$ while $\delta_{\mathrm{G}} = 0$. Although we consider only one measure of tortuosity in the power analysis, a total of six different measures were to be studied. Hence, power analysis for the study was conducted with $\alpha = 0.05/6 \approx 0.0083$.

A traditional approach to power analysis computes a single number. Using $\delta_{\mathrm{GR}} = 0.16$, which corresponds to Figure 4, gives approximate GG power of 0.90 with $N = 100$. In contrast, we feel compelled to examine the response surface as sample size and δ_{GR} vary. The three-dimensional GG power curves in Figure 5 illustrate the tradeoffs among sample size, expected size of the effect, and power.

For accuracy, two-dimensional power curves are needed, as in Figure 6 for $N \in \{20, 40, 80\}$ with $\delta_{\mathrm{GR}} \in [0, 0.40]$. With the power reaching 1.0 near $\delta_{\mathrm{GR}} = 0.20$ for $N = 80$, and near $\delta_{\mathrm{GR}} = 0.30$ for

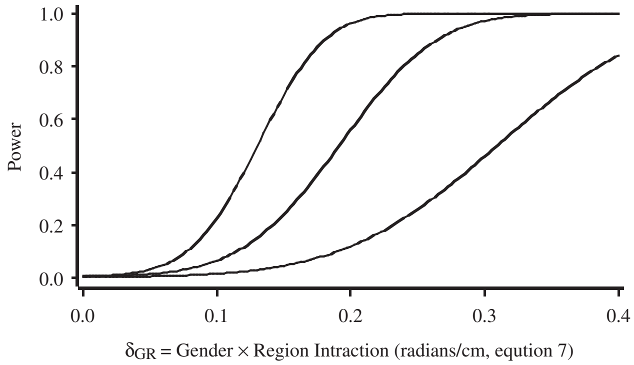

Figure 6. Approximate Geisser–Greenhouse power (new approximation method, new CDF) for gender × brain region effect on cerebral vessel tortuosity, $N \in \{20, 40, 80\}$, $\alpha = 0.05/6$. $\beta_P =$ scale factor for mean differences.

$N = 40$, a sample size of $N = 80$ seems necessary for $\delta_{\text{GR}} \in [0.20, 0.30)$, while $N = 40$ suffices for $\delta_{\text{GR}} \geqslant 0.30$.

4. APPROXIMATIONS

A non-central distribution function (CDF) approximation [15] gives second-order power approximations for the uncorrected and Box tests. Combining the CDF approximation with approximate expected critical values gives approximate power for the GG and HF tests. Appendix A contains details.

5. SIMULATIONS

5.1. *Comparisons to previously published simulations*

The new methods were used to predict $E(\widehat{\varepsilon})$ and $E(\tilde{\varepsilon})$ (degree of freedom adjustments due to non-sphericity), as well as approximate test size and power, for all cases in [5]. All results were consistent with the simulations described in the next section. The new methods provided better approximate power than previous methods for all tests.

5.2. *Overview of new simulations*

All new simulations mimicked the model for the mammography study (described in Section 3.1). Appendix C gives details, including covariance and mean patterns. Results for $\varepsilon \approx 0.72$ were consistent with the remaining values, but were omitted from tables for the sake of brevity, although they were included in summary statistics.

5.3. *Results of new simulations for $E(\widehat{\varepsilon})$ and $E(\tilde{\varepsilon})$*

Table II contains observed and predicted mean values of $\widehat{\varepsilon}$ and ε. Mean/max absolute deviations of observed and expected mean $\widehat{\varepsilon}$ were 0.033/0.143 for MB and 0.018/0.056 for the new

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table II. Observed and predicted mean $\widehat{\varepsilon}$ and mean $\widetilde{\varepsilon}$ (degrees of freedom reduction factors due to non-sphericity indexed by ε) standard error of observed mean $\leqslant 0.0003$.

		Observed	$E(\widehat{\varepsilon})$ Predicted	
N	ε	Mean $\widehat{\varepsilon}$ (s.d.)	MB[†]	New[‡]
10	0.28	0.286 (0.025)	0.285	0.273
	0.51	0.476 (0.103)	0.493	0.420
	1.00	0.691 (0.089)	0.833	0.679
20	0.28	0.284 (0.014)	0.283	0.277
	0.51	0.495 (0.086)	0.500	0.459
	1.00	0.818 (0.065)	0.921	0.813
40	0.28	0.283 (0.009)	0.283	0.279
	0.51	0.502 (0.064)	0.503	0.481
	1.00	0.900 (0.041)	0.962	0.898
			$E(\tilde{\varepsilon})$ Predicted	
		Mean $\tilde{\varepsilon}$ (s.d.)	MB	New
10	0.28	0.301 (0.037)	0.300	0.282
	0.51	0.606 (0.175)	0.651	0.505
	1.00	0.937 (0.100)	1.000*	1.000
20	0.28	0.290 (0.017)	0.289	0.282
	0.51	0.555 (0.114)	0.561	0.505
	1.00	0.965 (0.058)	1.000*	1.000
40	0.28	0.285 (0.010)	0.285	0.282
	0.51	0.529 (0.073)	0.531	0.505
	1.00	0.981 (0.032)	1.000*	1.000

*Result of truncation due to predicted value out of range.

[†] Muller–Barton approximation for expected value.

[‡] New approximation method for expected value.

approximation method. Mean/max absolute deviations for mean $\tilde{\varepsilon}$ were 0.027/0.126 for MB and 0.035/0.100 for the new approximation method. It approximated mean $\widehat{\varepsilon}$ much more accurately for $\varepsilon = 1$ and somewhat worse for $\varepsilon < 1$ than did the MB method. The new approximation method performed somewhat worse or similar to the MB method for $\tilde{\varepsilon}$.

5.4. Results of new simulations for test size

Table III contains simulated test size (target $\alpha = 0.04$) for the GG and HF tests, as well as for a mixed model assuming either unstructured or compound symmetric covariance. The results allow concluding that the UNIREP tests always essentially control test size. In contrast, mixed model tests fail to do so without a correctly specified covariance model and a sufficiently large N. Even with $Np = 360$ observations, incorrectly fitting compound symmetry rather than unstructured (rows with $\varepsilon < 1$) badly inflates test size of all three mixed model tests studied.

The MB and new CDF approximations coincide under the null. The approach gives an extremely good test size approximation for the uncorrected and Box tests.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table III. Simulated interaction test size for target $\alpha = 0.04$ for GG, HF (Standard Error<0.0004) and mixed model (Standard Error<0.003) (Population non-sphericity of $\mathbf{\Sigma}_* = \mathbf{U}'\mathbf{\Sigma}\mathbf{U}$ indexed by ε).

				Mixed model covariance fitted					
		UNIREP		Unstructured			Compound symmetric		
N	ε	GG	HF	Res*	Sat†	KR‡	Res*	Sat†	KR‡
10	0.28	0.042	0.045	0.254	0.138	0.114	0.107	0.106	0.106
	0.51	0.039	0.052	0.263	0.137	0.081	0.074	0.073	0.073
	1.00	0.021	0.036	0.263	0.144	0.053	0.038	0.038	0.038
20	0.28	0.041	0.042	0.116	0.077	0.040	0.099	0.098	0.098
	0.51	0.040	0.046	0.115	0.072	0.036	0.077	0.077	0.077
	1.00	0.029	0.038	0.119	0.075	0.038	0.039	0.038	0.038
40	0.28	0.040	0.041	0.075	0.054	0.040	0.104	0.103	0.103
	0.51	0.041	0.043	0.076	0.061	0.045	0.074	0.073	0.073
	1.00	0.034	0.039	0.074	0.056	0.040	0.041	0.041	0.041

*Residual approximation to denominator degrees of freedom.

†Satterthwaite approximation to denominator degrees of freedom.

‡Kenward–Roger approximation to denominator degrees of freedom.

Table IV contains observed and predicted test size for the GG and HF tests (with target $\alpha = 0.04$). The new method is more accurate for $\varepsilon = 1$, while method MB is more accurate for small ε. The approximate CDF does nearly as well as the exact.

Table IV also contains observed test sizes for mixed models tests, as a function of sample size and true population covariance pattern (as indexed by ε). Either an unstructured or compound symmetric covariance model was fitted, with the test using either the Residual, Satterthwaite, or Kenward–Roger approximation to the denominator degrees of freedom [21]. In contrast to the uniformly good performance of the GG and HF tests, the mixed model tests can all badly inflate test size. The mixed model tests provide correct test size only if the population and the covariance model fitted assume compound symmetry. Correctly fitting an unstructured covariance model does not control test size in small samples.

5.5. *Results of new simulations for power*

Simulation 1, based on the observed means in the CLAHE study, gave the largest deviations between observed and predicted power. For the sake of brevity, only simulation 1 results are reported because all others led to the same conclusions.

If $\varepsilon = 1$ then the uncorrected test should be used (it is then exact size α and most powerful). If $\varepsilon < 1$, uncorrected test power holds no interest due to inflated test size.

Table V contains observed and predicted power for the Box test. Values of β_P (an index of effect detailed in Appendix C) for simulation 1 are reported to eight digits to allow others to use the same conditions in future work. Mean/max absolute deviations for the MB and new CDF approximations are 0.048/0.212 and 0.008/0.025. Hence, the new CDF approximation performed dramatically better.

Copyright 2007 John Wiley & Sons, Ltd.

Table IV. Observed (Standard Error<0.0004) and predicted interaction test size for target $\alpha = 0.04$ with GG and HF (degrees of freedom factors $\widehat{\varepsilon}$ and $\tilde{\varepsilon}$ adjust for non-sphericity indexed by ε).

		Observed (simulation)	Approx. CDF		Exact CDF	
N	ε	GG	$E(\widehat{\varepsilon})$ MB*	$E(\widehat{\varepsilon})$ New†	$E(\widehat{\varepsilon})$ MB*	$E(\widehat{\varepsilon})$ New†
10	0.28	0.042	0.041	0.038	0.040	0.037
	0.51	0.039	0.039	0.030	0.037	0.029
	1.00	0.021	0.030	0.020	0.030	0.020
20	0.28	0.041	0.040	0.039	0.040	0.039
	0.51	0.040	0.039	0.035	0.039	0.035
	1.00	0.029	0.035	0.029	0.035	0.029
40	0.28	0.040	0.040	0.040	0.040	0.040
	0.51	0.041	0.040	0.038	0.040	0.038
	1.00	0.034	0.038	0.034	0.038	0.034
		HF	$E(\widetilde{\varepsilon})$ MB*	$E(\widetilde{\varepsilon})$ New†	$E(\widetilde{\varepsilon})$ MB*	$E(\widetilde{\varepsilon})$ New†
10	0.28	0.045	0.043	0.040	0.042	0.039
	0.51	0.052	0.055	0.040	0.052	0.038
	1.00	0.036	0.040	0.040	0.040	0.040
20	0.28	0.042	0.041	0.040	0.041	0.040
	0.51	0.046	0.046	0.040	0.045	0.039
	1.00	0.038	0.040	0.040	0.040	0.040
40	0.28	0.041	0.041	0.040	0.041	0.040
	0.51	0.043	0.043	0.040	0.042	0.040
	1.00	0.039	0.040	0.040	0.040	0.040

*Muller–Barton approximation method for expected value.

†New approximation method for expected value.

Table VI contains observed and predicted power for the GG and HF tests. The new CDF approximation gave much greater accuracy for both tests. The exact CDF gave no additional accuracy, and even reduced accuracy for some HF cases, which reflects the correlation of the test statistic with $\widehat{\varepsilon}$ and the use of a Taylor series. The new approximation method provides an additional substantial improvement in power approximation.

Figure 7 illustrates the improvement for the GG test with $N = 20$, $\varepsilon = 0.28$, and the mammography study mean pattern (and $\beta_P = 0.295$). The MB approximation suggests $N = 20$ is needed to achieve power of 0.800, while the new method predicts power of 0.828 with $N = 15$, a 25 per cent reduction in sample size.

6. DISCUSSION

6.1. Why use UNIREP (or MULTIREP) for data analysis?

For small to moderate samples, the multivariate model and associated tests, when applicable, guarantee control of test size, while mixed model tests do not. Hence, MULTIREP or UNIREP

Table V. Box test observed and predicted power × 100 of interaction, standard error of observed<0.001.

					CDF		
N	ε	MB GG target	β_P	Observed	MB*	New†	Exact
10	0.28	20	0.18655888	12	18	14	12
		50	0.31625972	54	47	54	54
		80	0.44588762	93	77	92	93
	0.51	20	0.15828381	05	09	06	05
		50	0.25780973	27	30	28	27
		80	0.35468332	69	61	69	69
	1.00	20	0.13933692	02	02	02	02
		50	0.21279863	12	12	12	12
		80	0.28293132	35	35	35	35
20	0.28	20	0.12457780	11	18	13	11
		50	0.21034038	56	47	57	56
		80	0.29558430	98	78	96	98
	0.51	20	0.10614402	06	10	06	06
		50	0.17308635	29	32	31	29
		80	0.23802852	76	65	75	76
	1.00	20	0.09038960	03	03	03	03
		50	0.14067360	14	14	14	14
		80	0.18836995	39	39	39	39
40	0.28	20	0.08580296	11	18	13	11
		50	0.14471410	56	48	59	56
		80	0.20320101	>99	78	98	>99
	0.51	20	0.07326247	06	10	06	06
		50	0.11956019	30	34	32	30
		80	0.16443791	79	66	78	79
	1.00	20	0.06160163	03	03	03	03
		50	0.09666182	15	15	15	15
		80	0.12983560	42	42	42	42

Note: Non-sphericity indexed by ε. Geisser–Greenhouse target power implied by β_P (index of effect, detailed in Appendix C; 8 digits allow others to simulate same conditions).

*Muller–Barton CDF approximation.

†New CDF approximation.

tests should be used whenever possible. With covariance patterns not too deviant from compound symmetry, UNIREP tests seem likely to be more powerful than the MULTIREP tests. The reverse typically holds for less simple covariance patterns.

6.2. *Why not use simulations or simple approximations for power?*

Simulations can be only a last resort because conscientious study design typically involves so many power values. Good planning usually requires many plots like Figure 7. Estimating the

Copyright 2007 John Wiley & Sons, Ltd.

Table VI. GG and HF observed and predicted power × 100 of interaction, standard error of observed<0.001 (degrees of freedom reduction factor $\widehat{\varepsilon}$ adjusts for non-sphericity indexed by ε).

				GG predicted power					HF predicted power			
		MB/GG		$\approx$ CDF		Exact CDF		HF	$\approx$ CDF		Exact CDF	
N	ε	Target	GG Obs.	MB*	New†	MB*	New†	Observed	MB*	New†	MB*	New†
10	0.28	20	15	17	16	15	14	17	18	17	16	14
		50	59	60	58	60	58	60	62	59	62	59
		80	94	94	94	95	95	95	95	94	96	95
	0.51	20	16	17	14	16	13	21	23	18	21	16
		50	52	55	49	55	48	59	65	56	65	56
		80	87	90	87	91	88	90	94	91	95	92
	1.00	20	16	20	16	20	16	22	24	24	24	24
		50	44	50	44	50	44	53	55	55	55	55
		80	75	80	75	80	75	82	84	84	84	84
20	0.28	20	13	15	15	13	12	13	16	15	13	13
		50	61	62	61	61	60	62	63	62	62	61
		80	98	97	97	99	99	98	97	97	99	99
	0.51	20	16	17	15	15	14	18	19	17	17	15
		50	54	56	53	55	52	57	60	56	59	56
		80	91	92	91	93	92	93	93	92	95	93
	1.00	20	18	20	18	20	18	21	22	22	22	22
		50	47	50	47	50	47	51	52	52	52	52
		80	77	80	78	80	78	81	81	81	81	81
40	0.28	20	12	15	15	12	12	12	15	15	12	12
		50	62	63	63	62	62	63	64	63	63	62
		80	>99	98	98	>99	>99	>99	98	98	>99	>99
	0.51	20	15	17	16	15	14	16	18	17	15	15
		50	55	56	55	55	54	56	58	56	57	55
		80	93	93	92	94	94	94	93	93	95	94
	1.00	20	19	20	19	20	19	20	21	21	21	21
		50	48	50	48	50	48	50	51	51	51	51
		80	79	80	79	80	79	80	81	81	81	81

* Muller–Barton approximation for $E(\widehat{\varepsilon})$.

† New approximation method for $E(\widehat{\varepsilon})$.

50–100 power values needed for each curve *via* simulation requires large amounts of computer and programming time for tight confidence limits.

Interest in internal pilot designs [9] makes the problem worse by requiring simulations nested within simulations. An internal pilot design uses a fraction of the data to estimate the variance and adjusts the sample size up or down. Repeated measures make an internal pilot even more appealing by adjusting for a poor choice of planning covariance matrix. An inaccurate covariance matrix presents the biggest hurdle to good power analysis for repeated measures.

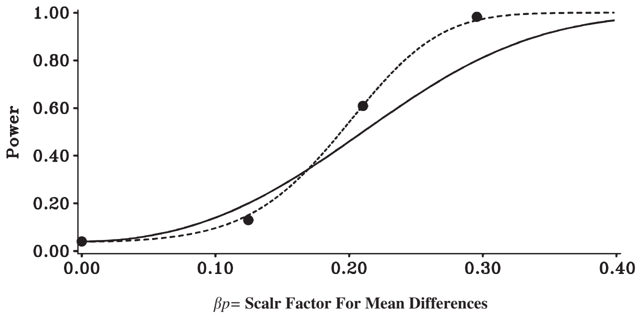

Figure 7. Predicted and observed Geisser–Greenhouse power for $N=20$ and $\varepsilon=0.28$. Dots are simulated values, solid line is the Muller–Barton approximation, and the dashed line is the new approximation method with new approximate CDF.

Simple approximations, such as using t test power to plan a study with a repeated measures analysis, create a substantial risk of an alignment error. Power can be badly overestimated, or badly underestimated by computing power for t tests rather than for repeated measures tests planned for actual studies. Muller *et al.* [8] gave examples of both mistakes.

6.3. Putting UNIREP (and MULTIREP) power calculations into practice

The methods described provide accurate power approximations, even with very small N and extreme covariance patterns. Free software which implements the methods in SAS/IML may be found at http://ehpr.ufl.edu/muller/.

The two examples illustrate the tremendous value of accurate power software for planning new studies. We believe conscientious study planning requires describing the entire response surface (as in Figure 5), not just a single power value.

Representing an effect of interest in terms of a single parameter (β_P for the mammography example, δ_{GR} for the tortuosity example) makes study planning easier. Of course, if the true state of nature differs from the pattern evaluated in the power analysis, then bias can occur. Obviously, the approach risks over-simplification. However, if the actual effect includes the pattern evaluated, in addition to other effects, then the true power will always be larger than the predicted power, and never smaller. Consequently, unless too much power has ethical costs, the approach can be used without worry.

Although our new methods and software make power analysis easier, we do not pretend the task is trivial. We do believe time spent on study planning with credible power analysis yields the highest possible return on the time spent.

6.4. Limitations and cautions

With unanticipated missing data, the predicted power may overestimate the true power, and the sample size will be too small. Not achieving the recruitment target, and hence target power, may be one of the biggest problem in clinical trials planning.

As the covariance structure deviates from compound symmetry, the UNIREP tests lose power, but remain valid. In such cases a MULTIREP test usually has more power. As mentioned earlier, a credible power analysis gives the best way to choose.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

APPENDIX A: ANALYTIC BASIS OF THE APPROXIMATIONS

Lemma

The UNIREP test statistic CDF can be expressed exactly in terms of the CDF of the sum of b positively and b negatively weighted independent chi-squares.

Proof

$\mathrm{tr}(\widehat{\mathbf{\Delta}}) = \sum_{k=1}^{b} \lambda_k y_{kh}$ and $\nu_e \mathrm{tr}(\widehat{\mathbf{\Sigma}}_*) = \sum_{k=1}^{b} \lambda_k y_{ke}$, with independent $y_{kh} \sim \chi^2(a, \omega_{*k})$ and $y_{ke} \sim \chi^2(\nu_e)$ [5]. Independence of $\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{\Sigma}}_*$ [22, p. 315] gives

$$\Pr\{f_u \leqslant f_0\} = \Pr\left\{\sum_{k=1}^{b} \lambda_k y_{kh} - \left(f_0 \frac{a}{\nu_e}\right) \sum_{k=1}^{b} \lambda_k y_{ke} \leqslant 0\right\} \tag{A1}$$

□

The lemma allows computing exact test size and power for the uncorrected and Box tests with Davies' algorithm [23]. The lemma would also provide exact test size and power for the GG and HF tests, conditional on $\widehat{\varepsilon}$, if f_u were independent of $\widehat{\varepsilon}$.

Theorem

The UNIREP test statistic may be approximated by a non-central F which matches two non-central (and one central) moments of the numerator to a scaled non-central chi-square, and two moments of the denominator to a scaled central chi-square. With $y_{*1} \sim \chi^2(\nu_{*1}, \omega_u)$ non-central chi-square with ν_{*1} degrees of freedom and non-centrality ω_u, and $y_{*2} \sim \chi^2(\nu_{*2})$, $\mathrm{tr}(\widehat{\mathbf{\Delta}}) \approx \lambda_{*1} y_{*1}$, $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*) \approx \lambda_{*2} y_{*2}$

$$\Pr\{f_u \leqslant f_0\} \approx \Pr\left\{\frac{\lambda_{*1} y_{*1}/(ab)}{\lambda_{*2} y_{*2}/(b\nu_e)} \leqslant f_0\right\} = F_F\left(f_0 \frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e}; \nu_{*1}, \nu_{*2}, \omega_u\right) \tag{A2}$$

Using results in [15], if $S_{t1} = \sum_{k=1}^{b} \lambda_k$, $S_{t2} = \sum_{k=1}^{b} \lambda_k \omega_{*k}$, $S_{t3} = \sum_{k=1}^{b} \lambda_k^2$, and $S_{t4} = \sum_{k=1}^{b} \lambda_k^2 \omega_{*k}$ then

$$\lambda_{*1} = (aS_{t3} + 2S_{t4})/(aS_{t1} + 2S_{t2}) \tag{A3}$$

$$\nu_{*1} = aS_{t1}/\lambda_{*1} \tag{A4}$$

$$\omega_u = S_{t2}/\lambda_{*1} \tag{A5}$$

$$\lambda_{*2} = S_{t3}/S_{t1} \tag{A6}$$

$$\nu_{*2} = \nu_e S_{t1}^2/S_{t3} = \nu_e b \varepsilon \tag{A7}$$

Corollary

By separately matching two moments of the numerator and denominator weighted sums of independent chi-squares, the non-central F in the Theorem gives second-order approximations for uncorrected and Box power with $f_0 = f_{\mathrm{crit}}(\mathrm{Uncorrected}) = F_F^{-1}(1-\alpha; ab, b\nu_e)$ or $f_{\mathrm{crit}}(\mathrm{Box}) = F_F^{-1}(1-\alpha; a, \nu_e)$.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Corollary
Values of $E(\widehat{\varepsilon})$ and $E(\tilde{\varepsilon})$ imply approximate expected critical values, $f_0(\mathrm{GG}) = F_F^{-1}[1-\alpha;\, abE(\widehat{\varepsilon}),\, bv_eE(\widehat{\varepsilon})]$ and $f_0(\mathrm{HF}) = F_F^{-1}[1-\alpha;\, abE(\tilde{\varepsilon}),\, bv_eE(\tilde{\varepsilon})]$, and in turn approximate power for GG and HF.

In practice, Taylor series' were used to approximate $E(\widehat{\varepsilon})$ and $E(\tilde{\varepsilon})$. If $t_1 = \mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*)$ and $t_2 = \mathrm{tr}(\widehat{\mathbf{\Sigma}}_*^2)$, then $\widehat{\varepsilon} = b^{-1}t_1t_2^{-1}$. A first-order series for t_2^{-1} about the point t_a gives

$$E(\widehat{\varepsilon}) = b^{-1}E(t_1t_2^{-1}) \approx b^{-1}t_a^{-1}E(t_1) - b^{-1}t_a^{-2}\{E(t_1t_2) - t_aE(t_1)\} \tag{A8}$$

If $m = v_et_2 - t_1$ then a first-order series for m^{-1} about the point m_0 gives

$$E(\tilde{\varepsilon}) \approx b^{-1}E\{(Nt_1 - 2t_2)[m_0^{-1} - m_0^{-2}(m - m_0)]\} \tag{A9}$$

The new approximation method uses only one term in equation (A8) and $t_a = E(t_2)$ to give $E(\widehat{\varepsilon}) \approx b^{-1}E(t_1)/E(t_2)$. The new approximation method uses one term in Equation (A9) and $m_0 = E(m)$ to give $E(\tilde{\varepsilon}) \approx b^{-1}E[NE(t_1) - 2E(t_2)]/[v_eE(t_2) - E(t_1)]$.

The values of $E(t_1)$ and $E(t_2)$ are computed as follows. Here, $\widehat{\varepsilon} = b^{-1}\mathrm{tr}^2(\boldsymbol{\Upsilon}'v_e\widehat{\mathbf{\Sigma}}_*\boldsymbol{\Upsilon})/\mathrm{tr}[(\boldsymbol{\Upsilon}'v_e\widehat{\mathbf{\Sigma}}_*\boldsymbol{\Upsilon})^2]$ and $v_e\boldsymbol{\Upsilon}'\widehat{\mathbf{\Sigma}}_*\boldsymbol{\Upsilon} \sim \mathscr{W}_b[v_e, \mathrm{Dg}(\boldsymbol{\lambda})]$. If z_{ij} i. i. d. $\mathscr{N}(0, 1)$ while $\mathbf{Z} = \{z_{ij}\}v_e \times b$, and $\mathbf{Y}_e = \mathbf{Z}\mathrm{Dg}(\boldsymbol{\lambda})^{1/2}$, then $\mathbf{S} = \mathbf{Y}_e'\mathbf{Y}_e \sim \mathscr{W}_b[v_e, \mathrm{Dg}(\boldsymbol{\lambda})]$, $\widehat{\varepsilon} = b^{-1}\mathrm{tr}^2(\mathbf{S})/\mathrm{tr}(\mathbf{S}^2) = t_1/(bt_2)$. If $\mathbf{z}_{k_j}$ is column k_j of $\mathbf{Z}$, then $s_{k_1k_2} = (\lambda_{k_1}\lambda_{k_2})^{1/2}\mathbf{z}_{k_1}'\mathbf{z}_{k_2}$. Here, $t_1^{1/2} = \mathrm{tr}(\mathbf{S}) \sim Q(\boldsymbol{\lambda}, v_e\mathbf{1}_b, \mathbf{0})$ has cumulant $\kappa_c = 2^{c-1}(c-1)!v_e\sum_{k=1}^{b}\lambda_k^c$, where $Q(\boldsymbol{\lambda}, v_e\mathbf{1}_b, \mathbf{0})$ is a weighted sum of independent, central chi-squares with weights $\{\lambda_j\}$ and df $\{v_e\}$. In turn, $\mu_2' = \kappa_2 + \kappa_1^2 = E(t_1)$. If $k_1 = k_2$ then $E(\mathbf{z}_{k_1}'\mathbf{z}_{k_2})^2 = v_e(v_e + 2)$, and otherwise $E(\mathbf{z}_{k_1}'\mathbf{z}_{k_2})^2 = v_e$ [24]. Hence, $E(t_1) = \kappa_2 + \kappa_1^2 = 2v_e\sum_{k=1}^{b}\lambda_k^2 + v_e^2(\sum_{k=1}^{b}\lambda_k)^2$ and $E(t_2) = \sum_{k_1=1}^{b}\sum_{k_2=1}^{b}\lambda_{k_1}\lambda_{k_2}E[(\mathbf{z}_{k_1}'\mathbf{z}_{k_2})^2] = v_e(v_e+2)\sum_{k_1=1}^{b}\lambda_{k_1}^2 + 2v_e\sum_{k_1=2}^{b}\sum_{k_2=1}^{k_1-1}\lambda_{k_1}\lambda_{k_2}$.

APPENDIX B: CONTRAST MATRICES FOR EXAMPLES

Mammography example: If $\mathbf{T}_\mathrm{c}$ contains orthonormal linear and quadratic trends for $\log_2(\mathrm{Clip}) \in \{1, 2, 4\}$, and $\mathbf{T}_\mathrm{r}$ does the same for $\log_2(\mathrm{region}) \in \{1, 3, 5\}$, then the 9×4 within-persons contrast matrix, $\mathbf{U}_\mathrm{cr}$ is

$$\mathbf{U}_\mathrm{cr} = \mathbf{T}_\mathrm{c} \otimes \mathbf{T}_\mathrm{r} = \begin{bmatrix} -4/\sqrt{42} & 2/\sqrt{14} \\ -1/\sqrt{42} & -3/\sqrt{14} \\ 5/\sqrt{42} & 1/\sqrt{14} \end{bmatrix} \otimes \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{6} \\ 0 & -2/\sqrt{6} \\ 1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix} \tag{B1}$$

With L for linear and Q for quadratic trend, $\mathbf{U}_\mathrm{cr} = [u_\mathrm{LL}\ u_\mathrm{LQ}\ u_\mathrm{QL}\ u_\mathrm{QQ}]$.

Tortuosity example
H_0: $\mathbf{\Theta} = \mathbf{CBU} = \mathbf{0}$, $\mathbf{C} = [1\ -1] \otimes \mathbf{1}_5'/5$ averages over the age groups separately for each gender. For $\mathbf{U} = [\mathbf{u}_1\ \ \mathbf{u}_2\ \ \mathbf{u}_3]/(2\sqrt{5})$, trend contrasts for $p = 4$ equally spaced points are $\mathbf{u}_1' = [-3\ \ -1\ \ 1\ \ 3]$, $\mathbf{u}_2' = \sqrt{5}[1\ \ -1\ \ -1\ \ 1]\mathbf{u}_1' = [-1\ \ 3\ \ -3\ \ 1]$.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

APPENDIX C: SIMULATION DESIGN AND METHODS

We conducted 11 sets (0–10) of simulations. Each used the covariance conditions 5–8 in Table III of [9]: $\mathbf{\Sigma}_* = \mathrm{Dg}(\boldsymbol{\lambda}_j)$ for $j \in \{1, 2, 3, 4\}$, with $\boldsymbol{\lambda}_1' = [0.47960\ 0.01000\ 0.01000\ 0.01000]$, $\boldsymbol{\lambda}_2' = [0.34555\ \ 0.06123\ \ 0.05561\ \ 0.04721]$, $\boldsymbol{\lambda}_3' = [0.23555\ \ 0.17123\ \ 0.05561\ \ 0.04721]$, $\boldsymbol{\lambda}_4' = [0.12740\ \ 0.12740\ \ 0.12740\ \ 0.12740]$. Hence, $\varepsilon \in \{0.28, 0.51, 0.72, 1.00\}$. Given $\mathbf{\Sigma}_* = \mathrm{Dg}(\boldsymbol{\lambda}_j)$, it follows $\mathbf{\Sigma} = \mathbf{U}_{\mathrm{cr}}\mathbf{\Sigma}_*\mathbf{U}_{\mathrm{cr}}'$. All four covariance patterns were factorially combined with $N \in \{10, 20, 40\}$. Set 0 tabulated test size and ε estimator properties for 500 000 replications with $\mathbf{\Theta}_{\mathrm{cr}} = \mathbf{B}\mathbf{U}_{\mathrm{cr}} = [0\ \ 0\ \ 0\ \ 0]$ and target power $P = \alpha = 0.04$. Mixed model test size was tabulated for 5000 replications with set 0. Set 1 used $\mathbf{\Theta}_{\mathrm{cr}} = \beta_P \cdot [0.5\ \ 1.0\ \ -1.0\ \ 0.5]$, with β_P the scaling factor for $\mathbf{B}$ corresponding to target power $P \in \{0.20, .0.50, 0.80\}$ for the MB GG approximation. For $\omega_{*\min}$ the smallest nonzero element of $\{\omega_{*j}\}$, sets 2–10 used $\mathbf{\Theta}_{\mathrm{cr}} = \beta_P \cdot [b_1\ \ b_2\ \ b_3\ \ b_4]$, with $\{b_j\}$ chosen so that $\omega_{*\min}^{-1} \cdot [\omega_{*1}\ \ \omega_{*2}\ \ \omega_{*3}\ \ \omega_{*4}] \in \{0001, 0010, 0100, 1000, 1111, 1221, 1234, 2122, 4321\}$ and $P \in \{0.70, .0.80, 0.90\}$. Sets 1–10 used $\mathbf{B} = \mathbf{\Theta}_{\mathrm{cr}}\mathbf{U}_{\mathrm{cr}}'$ and $\mathbf{X} = \mathbf{1}_N$ and 500 000 replications per condition.

Computational methods

All power computations were conducted in SAS/IML® [25]. Free software POWERLIB.IML (http://ehpr.ufl.edu/muller/) includes the new methods. A Fortran version (http://lib.stat.cmu.edu) of Davies' algorithm [23] was translated for calculating the exact CDF.

All UNIREP simulations were conducted in SAS/IML, while mixed model simulations used SAS PROC MIXED. The NORMAL function generated an $N \times p$ matrix of pseudo-random Gaussian data. A linear transformation then produced a realization of $\mathbf{Y}$. Free software LINMOD 3.3 (http://ehpr.ufl.edu/muller/) was used for data analysis. Observed means and standard deviations of $\widehat{\varepsilon}$ and $\tilde{\varepsilon}$, counts of $\tilde{\varepsilon}$ truncation, and hypothesis rejection counts for all four tests were stored.

ACKNOWLEDGEMENTS

Muller's work supported in part by NCI P01 CA47 982-04, NCI RO-1 CA095749-01A1 and NIAID 9P30 AI 50410. Edwards' work supported in part by NIH NHLBI Grant 5 T32 HL007773-09, National Center on Minority Health and Health Disparities Grant 5 R24 MD000167-01, the Agency for Healthcare Research and Quality Grant 1 R24 HS13353-01. Taylor's work supported in part by Family Health International with funds from United States Agency for International Development (USAID) cooperative agreement CCP-3079-00-5022-00. The authors thank Dr Mark Weaver for very helpful comments on an early version of the paper.

REFERENCES

1. Catellier DJ, Muller KE. Tests for Gaussian repeated measures with missing data in small samples. *Statistics in Medicine* 2000; **19**:1101–1114.
2. Schaalje GB, McBride JB, Fellingham GW. Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural*, *Biological*, *and Environmental Statistics* 2002; **7**:512–524.
3. Littel R. Analysis of unbalanced mixed models: a case study comparison of ANOVA vs REML/GLS. *Journal of Agricultural*, *Biological*, *and Environmental Statistics* 2003; **7**:472–490.
4. Demidenko E. *Mixed Models Theory and Applications*. Wiley: New York, 2004.
5. Muller KE, Barton CN. Approximate power for repeated-measures ANOVA lacking sphericity. *Journal of the American Statistical Association* 1989; **84**:549–555; *Corrigenda* 1991; **86**:255–256.
6. Muller KE, Stewart PW. *Linear Model Theory*: *Univariate*, *Multivariate*, *and Mixed Models*. Wiley: New York, 2006.

Copyright 2007 John Wiley & Sons, Ltd.

7. Glueck DH, Muller KE. Adjusting power for a baseline covariate in a linear model. *Statistics in Medicine* 2003; **22**:2535–2551.
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**:1209–1226.
9. Coffey CS, Muller KE. Properties of internal pilots with the univariate approach to repeated measures. *Statistics in Medicine* 2003; **22**:2469–2485.
10. Johnson NL, Kotz S. *Continuous Univariate Distributions*, vol. 2. Houghton Mifflin: Boston, 1970.
11. Johnson NL, Kotz S. *Distributions in Statistics*: *Continuous Multivariate Distributions*. Wiley: New York, 1972.
12. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions*, vol. 1 (2nd edn). Wiley: New York, 1994.
13. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions*, vol. 2 (2nd edn). Wiley: New York, 1995.
14. Kotz S, Balakrishnan N, Johnson NL. *Continuous Multivariate Distributions*, vol. 1 (2nd edn). Wiley: New York, 2000.
15. Kim H, Gribbin MJ, Muller KE, Taylor DJ. Analytic, computational and approximate forms for ratios of noncentral and central Gaussian quadratic forms. *Journal of Computational and Graphical Statistics* 2006; **15**:443–459.
16. Pizer SM, Zimmerman JB, Staab EV. Adaptive gray level assignment in CT scan display. *Journal of Computer Assisted Tomography* 1984; **8**:300–305.
17. Pisano ED, Zong S, Hemminger BM, Deluca M, Johnston RE, Muller KE, Brauening MP, Pizer SM. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated speculations in dense mammograms. *Journal of Digital Imaging* 1998; **11**:193–200.
18. Bullitt E, Ewend MG, Aylward S, Lin W, Gerig G, Joshi S, Jung I, Muller KE, Smith JK. Abnormal vessel tortuosity as a marker of treatment response of malignant gliomas: preliminary report. *Technology in Cancer Research and Treatment* 2004; **3**:577–584.
19. Grieve AP. Test of sphericity of normal distributions and the analysis of repeated measures designs. *Psychometrika* 1984; **49**:257–267.
20. Muller KE, Fetterman BA. *Regression and ANOVA*: *An Integrated Approach Using* SAS® *Software*. SAS Institute: Cary, NC, 2002.
21. SAS Institute. *SAS/STAT User's Guide*, *Version 8*, vol. 2 (Chapter 41). SAS Institute, Inc.: Cary, NC, 1999.
22. Arnold SF. *The Theory of Linear Models and Multivariate Analysis*. Wiley: New York, 1981.
23. Davies RB. Algorithm AS 155: the distribution of a linear combination of χ^2 random variables. *Applied Statistics* 1980; **29**:323–333.
24. Wishart J. The generalized product moment distribution in samples from a normal multivariate population. *Biometrika* 1928; **20A**:32–52.
25. SAS Institute. *SAS/IML User's Guide*, *Version 8*. SAS Institute, Inc.: Cary, NC, 1999.

Copyright 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim