Use of resampling to select among alternative error structure specifications for GLMM analyses of repeated measurements*

SCOTT TONIDANDEL,¹ JOHN E. OVERALL,² FRASER SMITH³

¹Department of Psychology, Davidson College, Davidson NC, USA

²Department of Psychiatry, University of Texas Health Sciences Center at Houston, Houston TX, USA.

³The US Food and Drug Administration, Rockville MD, USA

ABSTRACT *Autocorrelated error and missing data due to dropouts have fostered interest in the flexible general linear mixed model (GLMM) procedures for analysis of data from controlled clinical trials. The user of these adaptable statistical tools must, however, choose among alternative structural models to represent the correlated repeated measurements. The fit of the error structure model specification is important for validity of tests for differences in patterns of treatment effects across time, particularly when maximum likelihood procedures are relied upon. Results can be affected significantly by the error specification that is selected, so a principled basis for selecting the specification is important. As no theoretical grounds are usually available to guide this decision, empirical criteria have been developed that focus on model fit. The current report proposes alternative empirical criteria that focus on bootstrap estimates of actual type I error and power of tests for treatment effects. Results for model selection before and after the blind is broken are compared. Goodness-of-fit statistics also compare favourably for models fitted to the blinded or unblinded data, although the correspondence to actual type I error and power depends on the particular fit statistic that is considered.*

Key words: clinical trials, repeated measures, mixed models, goodness of fit, significance testing, type I errors, power

Controlled clinical trials concerned with evaluating the therapeutic efficacy of new drugs often employ a repeated measures design in which patients are randomly assigned between treatments, measured at baseline, and then measured at intervals across a treatment period of fixed total duration. This type of design has historically been characterized as conforming to a simple split plot in the tradition of experimental research in agriculture; however, autocorrelated error and missing data due to dropouts have rendered the use of the ordinary 'repeated measurements ANOVA' questionable. As a consequence, the more complex and adaptable general linear mixed model (GLMM) (Laird and Ware, 1982) has become increasingly proposed as a basis for evaluating the

significance of differences in treatment effects across time in the parallel-groups designs.

Flexibility in GLMM modelling of fixed, random, and error components is, however, provided at the cost of computational complexity and more numerous possibilities for model misspecification. A critical decision when using complex likelihood-based GLMM solutions involves the choice of a structural model to represent the random effects and/or correlated within-subjects error. How should a researcher, with a single data set, decide what error covariance structure is optimal? One solution is to 'try out' a number of alternative specifications and compare the results. Unfortunately, confusion can arise especially if the different specifications lead to

* The views expressed are those of the authors and are not necessarily those of the US Food and Drug Administration.

different substantive conclusions and there is no principled basis for choosing one specification over another. Such a situation is likely given that past research has documented the sensitivity of GLMM to the specification of the error structure. (Overall, Ahn, Shivakumar and Kalburgi, 1999; Ahn, Tonidandel and Overall, 2000). The purpose of this article is to present an empirical approach to selection among alternative error structure models that relies on type I error and power as criteria.

Goodness-of-fit statistics have been the criteria most often recommended in the past as a basis for choosing among alternative variance-covariance structure models (Lindsey, Wang, Byrom and Jones, 1999; Littell, Milliken, Stroup and Wolfinger, 1996; Wolfinger, 1993; 1996; 1997; Singer, 1998). Decisions regarding an appropriate covariance structure for the GLMM have been typically made by evaluating one or both of two information criteria, Akaike (1974) and Schwartz (1978). It is easy to appreciate, however, that goodness of fit in modelling the error structure does not confirm the superiority of the complete mixed model solution as a basis for inferences about treatment effects. Moreover, the same criteria of fit are not necessarily meaningful for comparison across models of differing complexity. In addition, recent Monte Carlo research suggests that results obtained from models selected on the basis of these two criteria may be suspect. For example, Keselman et al. (1998) found that both criteria were highly ineffective in identifying the model that best represented the known structure of simulated data. According to their results, Akaike's information criteria (AIC) correctly identified the 'true' structure only 47% of the time, whereas Schwarz's Bayesian Criteria (BIC) were accurate only 35% of the time. This would not be a problem of concern if Monte Carlo work with simulated data had not documented the sensitivity of the likelihood-based GLMM modelling procedures to variations in the correlational structure of repeated measurements (Overall, Ahn, Shivakumar and Kalburgi, 1999; Ahn, Tonidandel and Overall, 2000).

One can easily appreciate that GLMM analyses have little utility if the resulting test statistics have inflated levels of type I error or unsatisfactory power. As a result, Overall and colleagues proposed using actual type I error protection and power of tests of significance of treatment effects as more pragmatic criteria for choosing among alternative simple and

more complex model formulations. Tests of significance based on alternative GLMM model formulations depend not only on the magnitude of treatment effects that are present but also on conformity of the actual variance-covariance structure of the data to the error model specification. The tests of significance assume that the actual variance-covariance structure is the one specified in the model. By eliminating observed treatment effects from consideration, the GLMM analysis itself can be used to test validity of that assumption by examining type I error and power of resulting tests of significance.

When selecting among alternative analytical approaches, a preference is usually given to the procedure with greater statistical power. However, because the power of a procedure is typically correlated with type I error, it is not sufficient to look just at power in isolation. One must also examine the type I error rates to ensure that the analysis is not producing non-conservatively biased results. Although Monte Carlo evaluation of type I error and power of tests of significance based on alternative statistical models can be readily accomplished using artificial sampling data with user-specified error structures, that is of little comfort to a data analyst faced with decision concerning a priori specification of the most appropriate model for the primary analysis of data from a particular study, or when faced with questions concerning the actual error protection afforded by tests that may or may not have already provided 'significant' support for conclusions about treatment efficacy. This article thus examines a strategy of bootstrap sampling-with-replacement from a 'primary data set' to obtain Monte Carlo estimates of actual type I error and power of GLMM tests of significance for purpose of choosing among alternative error structure models.

Bootstrapping involves resampling a given real data set many times in order to generate an empirical estimate of the sampling distribution of a statistic (Chernick, 1999; Efron, 1982; Mooney and Duval, 1993). Bootstrap procedures are commonly employed to make inferences about a population parameter by empirically estimating the sampling distribution of the statistic. More recently, bootstrapping has been applied to a wide variety of other problems, including evaluating error rates for hypothesis tests (Mooney and Duval, 1993). This paper describes an application of bootstrapping for choosing a superior covariance

structure model without taking observed treatment effects into account. Issues that are addressed include:

- evaluation of alternative error model specifications by sampling with replacement across the combined treatment groups prior to breaking the blind,
- utility of sampling with replacement to model the within-groups error distribution after the blind is broken; and
- validities of alternative error structure models selected on the basis of goodness-of-fit criteria.

In contrast with Monte Carlo studies aimed at producing generalizable conclusions, the concern herein is with identification of the most valid error structure model specification for GLMM analysis of treatment effects in a particular study.

Method

The antidepressant drug trial

The data subjected to the bootstrapping procedure for this methodological inquiry came from an international multicentre study sponsored by the Pfizer Company to evaluate the efficacies of two different dose levels of a potential new antidepressant drug against placebo and against an active control. The study employed a 6-week parallel-groups design with a total of 274 mildly to moderately depressed outpatients randomized in approximately equal numbers between the four groups. The total score on the 17-item Hamilton Depression Scale (HAM-D) was the primary outcome measure recorded at baseline and weekly for the following 6 weeks. To avoid prejudicing possible renewed proprietary interest in

clinical development of the particular investigational drug, we chose to use only data for the active control group (N = 81) and the placebo group (N = 77). To provide a more stringent test of the utility of modelling the GLMM error structure by resampling across the combined treatment groups before the blind is broken, we enhanced the original difference between these two treatment groups by adding a small linearly increasing quantity to the observed repeated measurements for subjects in one of the groups but not the other. This produced an endpoint difference of approximately 0.4 SD for the 'primary data set' that was subsequently resampled. Importantly, this left the within-groups correlation structure, within-group variances, and dropout patterns of the original data set unaltered. An overall dropout rate of approximately 20% in those two groups was distributed relatively uniformly across the 6-week study. Group means, standard deviations, and dropout patterns in the primary data set that was resampled for the Monte Carlo model comparisons are shown in Table 1. Although this methodological inquiry was obviously retrospective, using real data that were already available, we will use terminology such as 'prior to breaking the blind' to suggest a chronological applicability that avoids the potential for bias inherent in *post hoc* model selection.

Alternative model formulations

Tests of significance for differences in treatment effects assume that the GLMM error structure is the one specified in the model formulation. With mean differences removed from consideration, Type I error and power of the GLMM tests reflect departure from the assumed error variance-covariance structure. We prefer these indicators of error model misspecification

Table 1. Means, standard deviations, and sample sizes for each group

	Baseline	Score 1	Score 2	Score 3	Score 4	Score 5	Score 6
Group 1							
Mean	25.96	19.62	17.24	15.73	15.38	14.5	13.74
St.Dev.	2.64	4.97	5.44	6.06	6.71	7.21	6.76
N	77	77	75	69	68	65	63
Group 2							
Mean	25.91	19.51	16.47	14.19	13.23	12.38	11.18
St.Dev.	2.69	5.7	7.01	6.8	6.97	7.09	7.27
N	81	81	78	72	69	65	62

over other goodness-of-fit statistics because of their more direct relevance for conclusion that follow from analyses of data from controlled clinical trials. Models providing this type of justification are said to have superior 'inferential validity'.

A primary aim of this empirical investigation was to evaluate the feasibility of using information available prior to breaking the blind in a controlled clinical trial to identify the error structure specification that should have highest inferential validity for a subsequent primary analysis of data from that trial. This strategy is proposed to avoid capitalizing on chance by choosing the preferred results from a number of different models fitted *post hoc* to data in an unblinded study. The importance of avoiding *post hoc* model selection is widely recognized. For example, the Food and Drug Administration (FDA), and regulatory bodies in other countries as well, require complete specification of the proposed primary analyses prior to unblinding data from controlled clinical trials. In a recent NIMH collaborative study of St John's wort (*Hypericum perforatum*) for treatment of major depression, the 'primary statistical analysis' originally specified in the protocol was officially changed with FDA approval on the basis of resampling studies accomplished prior to breaking the blind using a procedure similar to that described herein (Hypericum Depression Trial Study Group, 2002).

For the present inquiry, tests of the 'equal slopes hypothesis' utilizing alternative general linear mixed model (GLMM) formulations on the resampled data sets were accomplished using SAS PROC.MIXED. Seven different formal error model specifications were evaluated for potential representation of the within-subjects (error) variance-covariance structure in testing the significance of treatment effects on means rates of change across repeated measurements in the particular controlled clinical trial. The GLMM model formulations including the alternative error specifications were of the general form illustrated by Littell, Milliken, Stroup and Wolfinger (1996: 103) for modelling time as a regression variable in the repeated measurements design. Singer (1998) and Wolfinger (1997) provide examples of similar GLMM formulations in which a model of the within-subjects covariance structure is specified as error for evaluating the significance of difference in pattern of treatment effects across

time. Random between-subjects variation was not separately modelled for these GLMM analyses of repeated measurements.

The seven GLMM model formulations employed in this first series of analyses differed only by imposing different structural models on the within-subjects error. Four of the seven models were selected for their different representations of the within-subject correlation structure of the repeated measurements. Together with their PROC.MIXED abbreviations, these four models for the error correlation structure were autoregressive moving average (ARMA (1,1)), toeplitz (TOEP), autoregressive, order 1 (AR (1)), and compound symmetry (CS). These correlation models assume model variances to be uniform across the repeated measurements, but GLMM formulations can be amended to accommodate data that violate the typical ANOVA assumption of equal variances over time. The corresponding error covariance models that permit the model variances to change across the repeated measurements, while retaining the designated correlation structure, are indicated in PROC.MIXED abbreviation by appending H to the correlation model designations. The three that we consider herein are TOEPH, ARH (1), and CSH. In addition, we included the unstructured (UN) error model specification, which does not impose any particular mathematical form on the observed within-subject error matrix. In order to appreciate the implications of using actual type I error and power as criteria for choosing a superior error structure model for GLMM analyses, it is not essential to be familiar with the mathematical representation of these correlation and covariance structure models; however, if one wants to examine them in greater detail, useful references are Wolfinger (1993, 1996).

For these GLMM analyses, the generalized linear models for regression on time were fitted to the resampled data for post-baseline weeks 1, 2, 3, 4, 5, and 6, with baseline scores entered as a covariate. (We have been counselled (R.D. Wolfinger, personal communication, October, 2000) that the 'customary' way to formulate a GLMM regression model in which baseline scores are to be entered as a covariate to adjust for initial differences is to omit the baseline score from the variable set to which the means model is fitted.) Resampling the primary data set before and after the blind was broken provided Monte Carlo

estimates of type I error and power for test of significance based on the seven alternative error model specifications across runs of 4,000 resampled data sets consisting of 2 groups of size $N = 50$. Two popular goodness-of-fit statistics were also calculated for the alternative models of the GLMM error structure with the results average across the series of resampled data sets. Type I error and power provided by tests of the 'equal slopes hypothesis' accomplished on the unblinded data serve as criteria for judging the utilities of the two popular goodness-of-fit statistics, as well as for evaluating the utility of sampling with replacement from the combined treatment groups prior to breaking the blind.

To illustrate the utility of bootstrap resampling for choice among statistical models for which no comparable goodness-of-fit statistics are available, simple endpoint difference-score analyses and tests on dropout-weighted ordinary least squares (W-OLS) slope coefficients were accomplished on the same resampled data sets for comparison with results from the GLMM analyses. The W-OLS procedure can be conceptualized as OLS regression on a normalized time scale and requires no modelling of the correlation structure (Overall and Shivakumar, 1999).

Modelling the blinded data and unblinded data

To evaluate the type I error rates for tests of the 'equal slope hypothesis' prior to breaking the blind, pairs of independent samples were drawn with replacement completely at random across the combined treatment groups of the primary data set. These duplicate independent samples from the same distribution of correlated repeated measurements were the 'treatment groups' used to obtain Monte Carlo estimates of type I error for alternative GLMM error models without reference to the actual treatment assignments of subjects in the study.

To use the blinded data for evaluating power of tests of significance based on the different statistical models, a 'treatment effect' was introduced by adding a linearly increasing constant to the observed repeated measurements for all of the individuals in one of each pair of treatment groups randomly sampled from the same primary data set. This left the marginal covariance structure and the sampling variability of the resampled means unchanged. The magnitude of the added treatment effect increased

from zero at baseline to approximately 0.4 SD units at endpoint and was empirically scaled to provide power in the minimally acceptable range for two-sided $\alpha = 0.05$ tests of the 'equal slopes hypothesis' with $N = 50$ per resampled group. (Sample sizes of $N = 50$ were chosen because it allows us to compare the results herein with previously published results from Monte Carlo studies.)

Identical procedures were used to evaluate type I error rates and power for analyses accomplished after the blind was broken, except that the respective group means were first subtracted from the individual subject's repeated measurements prior to resampling across the combined treatment groups. This provided a pooled within-groups data matrix with treatment effects removed from which duplicate independent samples were drawn to evaluate Type I error rates. A treatment effect corresponding to the means in Table 1 was reintroduced to compare power under different error model specifications.

The simple endpoint analyses and the tests on means of the dropout-weighted ordinary least squares (W-OLS) slope coefficients were accomplished on the same resampled data sets that were used to compare Type I error rates and power of the GLMM tests under the different error model specifications. The calculation of the ordinary least squares slope coefficients for the W-OLS procedure required no explicit modelling of the within-subjects error structure, although under an independence assumption a diagonal error structure matrix would be implied. An alternative conception is that the W-OLS slope coefficients are simply weighted linear contrasts among the individual subject's repeated measurements. In any event, the W-OLS slope coefficients adjust for dropouts in a manner different from the GLMM analyses. Also in contrast to the GLMM tests of the equal slopes hypothesis, subject-specific W-OLS slopes coefficients were fitted to all of the available repeated measurements including baseline. Baseline scores and the number of repeated measurements available for each subject were then included as covariates for ANCOVA tests on endpoint difference scores and W-OLS slope coefficients. The differences are emphasized to illustrate the advantage of using appropriate test size and power as criteria for selecting among statistical models for which goodness-of-fit statistics cannot be meaningfully compared.

Results

Table 2 contains Monte Carlo estimates of type I error and power obtained by resampling across the combined treatment groups of the original primary data set. Additional analyses obtained by resampling across the combined treatment groups after the original group means (Table 1) were subtracted out are also in Table 2. The latter analyses can be accomplished after the blind is broken and serve here as criteria against which to compare the inferential validity of tests of significance based on error model specifications obtained from bootstrap resampling prior to breaking the blind.

Greater than 20% inflation in actual type I error rate is considered 'significant' non-conservative bias, consistent with the definition we have used in our previous studies. It should not be expected that empirical sampling will result in exact type I error or power values. For tests of the 'equal slopes hypothesis' across repeated measurements provided by sampling with replacement, this particular blinded or unblinded data set produced similar results. The general linear mixed model (GLMM) analyses in which UN, TOEPH, CS or CSH was specified to model the within-subject error structure exceeded the accepted level of non-conservative bias at nominal $\alpha = 0.05$. It is of note that specifying TOEP to model the error structure produced more acceptable type I error than did TOEPH in spite of variances that increased substantially across the repeated measurements (Table 1). Modelling the error structure as AR (1) or ARH (1) resulted in conservatively biased tests of significance for differences in average rates of change between pairs of independent groups that were resampled from this same combined or pooled within-groups distribution. This conservative bias is consistent with previous results from analyses of artificial simulated data in which AR(1) was specified to model repeated measurements with a relatively uniform correlation structure.

Power for detecting the difference in mean rates of change in the resampled data tended, quite naturally, to correlate with bias in the type I error rates for the alternative analyses. Disregarding models that produced unacceptably nonconservative type I error, TOEP (toeplitz) modelling of the error structure provided the largest calculated power values for tests

of the equal slopes hypothesis in resampling this particular data set.

Throughout this discussion, we emphasize that our concern is with choosing the most adequate error model for analysis of a given data set, not with generalizable conclusions about relative merits of the different model formations for use in other applications. Because one cannot verify the actual type I error probability or power from results provided by a single analysis, we propose sampling with replacement to be a superior basis for model comparisons. We are also concerned with justifying the choice of a model prior to breaking the blind. The correspondence of resampling results for models fitted before and after the blind is broken lends support for using resampling for a priori model selection.

Model selection based on goodness of fit

For comparison, we examined the inferential validities of GLMM error model selections based on goodness-of-fit statistics, again using type-I error and power provided by resampling after the blind was broken as the criteria for judging inferential validity. The goodness-of-fit statistics used in this investigation were Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (BIC) as calculated by SAS PROC.MIXED (Littel, Milliken, Stroup and Wolfinger, 1996: 101). In sampling with replacement from the combined treatment groups prior to breaking the blind, the values for each of these two fit statistics were averaged across the resampled data sets for each of the seven GLMM models that were fitted to the resampled data. The same goodness-of-fit statistics were subsequently calculated by applying the seven GLMM models to the 'primary data set' without resampling (see Table 2).

The patterns of fit statistics across the alternative statistical models should be compared, not the absolute values, which differ systematically for the original and resampled data as a function of the different sample sizes considered. For comparison within either the original or resampling conditions, smaller values for goodness-of-fit statistics are considered evidence of superior fit. Both goodness-of-fit measures exact a penalty for the number of parameters estimated, but the Bayesian BIC statistic imposes a greater correction than does the AIC statistic resulting in a preference for more parsimonious

Table 2. Goodness-of-fit criteria, type I error, and power for GLMM model formulations

	Goodness-of-fit criteria ^a				Sampling with replacement			
	Original data		Resampled data		Blinded data		Unblinded data	
	AIC	BIC	AIC	BIC	Type I error	Power	Type I error	Power
UN	4907	4971	3104	3158	6.4	74.7	6.6	78.7
ARMA(1,1)	4911	4921	3118	3126	4.2	64.9	4.0	62.8
TOEP	4914	4933	3119	3135	5.8	69.0	5.6	69.9
TOEPH	4917	4951	3119	3147	6.8	70.3	7.2	72.8
AR(1)	4930	4936	3130	3136	2.1	49.9	1.5	46.9
ARH(1)	4934	4955	3130	3149	3.0	52.1	2.9	49.6
CSH	4963	4984	3153	3171	14.1	83.2	14.1	85.7
CS	4965	4971	3156	3161	14.5	83.7	13.9	85.3

^aValues closer to 0 indicate superior fit of the model covariance structure

structures. To interpret differences between the fits of alternative models, Raftery (1995) offers some general rules of thumb for interpretation of differences in BIC. He suggests that differences smaller than three points are weak evidence of model superiority, while other authors have illustrated choosing the best fitting model regardless of how small are differences in the magnitudes of goodness-of-fit statistics (for example, Wolfinger, 1997).

The first thing that is apparent in comparing patterns of goodness-of-fit across the alternative GLMM statistical models is a high concordance between patterns for models fitted to the original primary data set and to the data sets resampled from it. The other noteworthy thing is that the pattern of goodness-of-fit across alternative specifications of the GLMM error-structure models depended more on the particular fit statistic (AIC or BIC) than on the blinded or unblinded source of the data. From this we conclude that, to the extent that goodness-of-fit statistics are useful indicators of the inferential validities of alternative statistical models, the present results support their utility for selection of a model to represent the error structure prior to breaking the blind. On the other hand, important differences in validity of the different fit statistics are suggested, as will be considered next.

The difference in patterns for the AIC and BIC goodness-of-fit statistics across alternative modelling of the GLMM error structure bears on their validity as a basis for GLMM model selection both before and after the blind is broken. Again, we consider that type I error and power calculated by fitting alternative models to the unblinded treatment groups is the best available gauge of inferential validity for alternative models applied to data from a particular study. For this particular data set, the AIC statistic indicated TYPE=UN (unstructured) to be the best fitting model for the within-subjects error; however, the GLMM tests of the equal slopes hypothesis based on that model specification provided non-conservative type I error in testing the significance of treatment effects in bootstraps samples from the unblinded study. Reliance on the more conservative Bayesian BIC goodness-of-fit statistic, instead of AIC, applied to either the blinded or unblinded data resulted in identification of a statistical model that provided superior Type I error protection and power for testing the significance of difference in mean rate change.

The fact that neither of the goodness-of-fit statistics resulted in identification of toeplitz (TOEP) as the best error model for the particular data set that was resampled bears on the issue of whether small differences are a meaningful basis for choosing among competing models (for example, Raftery (1995) and Wolfinger (1997) mentioned above).

Discussion

This article is about a procedure for choosing among alternative statistical models in application to any particular data set. The practical concern is with choosing or justifying the best model for analysis of data from a particular controlled clinical trial. The 'inferential validity' of alternative statistical models is determined by the actual type I error protection and power afforded by the associated tests of significance. Those critical inferential parameters cannot be ascertained from the test statistics or p-values customarily reported for the evaluation of treatment effects in a controlled clinical trial, but bootstrapping does provide a basis for Monte Carlo determination of the error protection and power provided by tests of significance based on any statistical model that is considered for analysis of the particular data. Finally, it is proposed that those critical inferential parameters can be usefully evaluated by sampling with replacement before the blind is broken in cases where post hoc model selection is to be avoided.

This article also illustrates the failure of traditional goodness-of-fit criteria to identify the best model in terms of inferential validity. Aside from identifying a more optimal covariance structure, an additional advantage of using type I error and power evaluated by bootstrapping samples with replacement is that it permits one to compare the inferential validities of alternative statistical models for which goodness-of-fit criteria may not be directly comparable. For example, simple two-stage analyses, which rely on ordinary least squares slope coefficients (Overall and Shivakumar, 1999) or simple endpoint difference scores (Overall, 1997) as indices of change, are not amenable to goodness-of-fit statistics comparable to those available for GLMM error structures. Nevertheless, these simple analyses provide tests of significance for meaningful hypotheses concerning differences in rates of change, and the inferential validity represented by actual type I error rates and power ascertained by bootstrap

Table 3. Type I error and power for tests on summary statistics that do not require modelling the correlation structure

	Sampling with replacement			
	Blinded data		Unblinded data	
	Type I error	Power	Type I error	Power
Endpoint Diff	5.4	73.3	5.8	71.2
Weighted OLS	5.1	78.6	5.4	80.2

sampling can be compared with similar validity criteria for tests of significance based on more complex GLMM mixed model procedures. As an example, Table 3 contains results for the simple two-stage endpoint and regression analyses applied to the same resampled data sets used to compare the inferential validities of alternative GLMM model formulations. One might wish to compare the results in Tables 2 and 3 as a basis for choosing among the more and less complex statistical procedures for analysis of the particular illustrative data set considered herein. Other factors such as communicating an understanding of the methodology and clarity of interpretation might also be considered in the decision.

Despite the success of the resampling procedure to choose among alternative error structures, the limitations of the resampling procedure must be acknowledged. Although the assumptions of bootstrapping are weaker than those of parametric methods, resampling can still fail. Nevertheless, the similarity between the resampling results presented here and results previously obtained using simulated data to evaluate various GLMM models (Overall, Ahn, Shivakumar and Kalburgi, 1999; Ahn, Tonidandel and Overall, 2000) lend additional support to the validity of the resampling procedure.

In conclusion, it is important to appreciate the feasibility of routinely employing bootstrap resampling to evaluate inferential validities of models used to test the significance of treatment effects in controlled clinical trials.

To facilitate the use of resampling, a program to accomplish the complete resampling procedure as formatted for SAS 8.1 PROC.MIXED test of the 'equal slope hypothesis' can be downloaded from Web site <http://www.davidson.edu/academic/psychology/Tonidandel/TonidandelProgramsMain.htm>. Using this or a similar computer program, it is entirely

feasible to check and/or compare the inferential validity of alternative model formulations as a routine feature of the data analysis.

Acknowledgement

This work was supported in part by DHHS grant MH32457 and completed when the third author was employed at Pfizer.

References

- Ahn C, Tonidandel S, Overall JE (2000) Issues in use of SAS Proc.Mixed to test the significance of treatment effects in controlled clinical trials. *Journal of Biopharmaceutical Statistics* 10: 265–86.
- Akaike H (1974) A new look at statistical model identification. *IEEE Transaction on Automatic Control*, AC-19, 716–23.
- Chernick MR (1999) *Bootstrap Methods: A Practitioner's Guide*. New York: John Wiley & Sons.
- Efron B (1982) The jackknife, the bootstrap, and other resampling plans. Philadelphia PA: Society for Industrial and Applied Mathematics.
- Frison L, Pocock SJ (1992) Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design. *Statistics in Medicine* 11: 1685–704.
- Hypericum Depression Trial Study Group (2002) Effect of Hypericum perforatum (St John's wort) in major depressive disorder: a randomized, controlled trial. *JAMA* 287: 1807–14.
- Keselman HJ, Algina J, Kowalchuk RK, Wolfinger RD (1998) A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics – Simulation and Computation* 27: 591–604.
- Kirsch I, Saperstein G (1999). Listening to Prozac but hearing placebo: a meta-analysis of antidepressant medication. In Kirsch I (ed.) *How Expectancies Shape Experience*. Washington DC: American Psychological Association.
- Laird NM, Ware JH (1982) Random-effects model for longitudinal data. *Biometrics* 38: 963–74.

- Lindsey JK, Wang J, Byrom WD, Jones B (1999) Modeling the covariance structure in pharmacokinetic crossover trials. *Journal of Biopharmaceutical Statistics* 9: 439–50.
- Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996) *SAS System for Mixed Models*. Carey NC: SAS Institute, Inc.
- Mooney CZ, Duval RD (1993) *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park CA: Sage.
- Overall JE (1997) Drop-outs and a random regression model. *Journal of Biopharmaceutical Statistics* 7: 383–402.
- Overall JE, Ahn C, Shivakumar C, Kalburgi Y (1999) Problematic formulations of SAS PROC.MIXED models for repeated measurements. *Journal of Biopharmaceutical Statistics* 8: 189–216.
- Overall JE, Shivakumar C (1999) Testing differences in response trends across a normalized time domain. *Journal of Clinical Psychology* 55: 857–67.
- Raftery AE (1995) Bayesian model selection in social research. *Sociological Methodology* 25: 111–63.
- Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics* 6: 461–4.
- Singer JD (1998) Using SAS PROC MIXED to fit multi-level models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics* 24: 323–55.
- Wolfinger RD (1993) Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation* 22: 1079–106.
- Wolfinger RD (1996) Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics* 1: 205–30.
- Wolfinger RD (1997) An example of using mixed models and Proc Mixed for longitudinal data. *Journal of Biopharmaceutical Statistics* 7: 481–500.

*Correspondence: Scott Tonidandel, Psychology Department, Davidson College, Box 7061, Davidson NC 28035-7061, USA.
Email: sctonidandel@davidson.edu.*