Move Over ANOVA

Progress in Analyzing Repeated-Measures Data and Its Reflection in Papers Published in the Archives of General Psychiatry

Ralitza Gueorguieva, PhD; John H. Krystal, MD

Background: The analysis of repeated-measures data presents challenges to investigators and is a topic for ongoing discussion in the *Archives of General Psychiatry*. Traditional methods of statistical analysis (end-point analysis and univariate and multivariate repeated-measures analysis of variance [rANOVA and rMANOVA, respectively]) have known disadvantages. More sophisticated mixed-effects models provide flexibility, and recently developed software makes them available to researchers.

Objectives: To review methods for repeated-measures analysis and discuss advantages and potential misuses of mixed-effects models. Also, to assess the extent of the shift from traditional to mixed-effects approaches in published reports in the *Archives of General Psychiatry*.

Data Sources: The *Archives of General Psychiatry* from 1989 through 2001, and the Department of Veterans Affairs Cooperative Study 425.

Study Selection: Studies with a repeated-measures design, at least 2 groups, and a continuous response variable.

Data Extraction: The first author ranked the studies according to the most advanced statistical method used in the following order: mixed-effects model, rMANOVA, rANOVA, and end-point analysis.

Data Synthesis: The use of mixed-effects models has substantially increased during the last 10 years. In 2001, 30% of clinical trials reported in the *Archives of General Psychiatry* used mixed-effects analysis.

Conclusions: Repeated-measures ANOVAs continue to be used widely for the analysis of repeated-measures data, despite risks to interpretation. Mixed-effects models use all available data, can properly account for correlation between repeated measurements on the same subject, have greater flexibility to model time effects, and can handle missing data more appropriately. Their flexibility makes them the preferred choice for the analysis of repeated-measures data.

Arch Gen Psychiatry. 2004;61:310-317

From the Division of Biostatistics, Department of Epidemiology and Public Health (Dr Gueorguieva), and the Department of Psychiatry (Dr Krystal), Yale University School of Medicine, New Haven, Conn.

DURING THE PAST 15 YEARS, the *Archives of General Psychiatry* (ARCHIVES) has published periodic reviews of the status of the analysis of repeated-measures data.[1-3] These reviews have highlighted the importance, challenges, and evolution of analytic methods for these data. In randomized clinical trials and longitudinal follow-up studies, one must use appropriate statistical methods to adjust for multiple measurements of the same individual and model time trends. Repeated-measures data are usually correlated, since sequential observations of the same individual tend to be closer in value to one another than the same number of observations collected from different individuals would be. Missing data also present special challenges for analysis.

The most commonly used approaches for analyzing repeated-measures data shifted over time from end-point analysis to univariate and multivariate repeated-measures analysis of variance (rANOVA and rMANOVA, respectively). Despite the strengths of these approaches, each method has important disadvantages. In the 1990s, studies published in the ARCHIVES began to use mixed-effects regression models.[3] Mixed-effects regression models provide a general framework for the analysis of repeated measures, and recently developed statistical software makes these models accessible to researchers.

The purpose of this article is to take stock of the impact of mixed-effects regression models on the analysis of repeated-measures data in psychiatry research, increase awareness of the advantages and potential pitfalls of these models, and encourage their use over traditional methods. In doing so, this article will briefly review the main methods for repeated-

Table 1. Comparison of Traditional and Mixed-Effects Approaches for the Analysis of Repeated-Measures Data

	End-Point Analysis	rANOVA	rMANOVA	Mixed-Effects Analysis
Complete data required on every subject	Yes	No*	Yes	No
Possible effect of omitting subjects with missing values	Sample bias	Sample bias	Sample bias	Not applicable†
Possible effects of imputation of missing data	Estimation bias	Estimation bias	Estimation bias	Not applicable†
Subjects measured at different time points	Yes	No	No	Yes
Description of time effect	Simple	Flexible	Flexible	Flexible
Estimation of individual trends	No	No	No	Yes
Restrictive assumptions about correlation pattern	Not applicable	Yes	No	No
Time-dependent covariates	No	Yes	No	Yes
Ease of implementation	Very easy	Easy	Easy	Hard
Computational complexity	Low	Low	Medium	High

Abbreviations: rANOVA, univariate repeated-measures analysis of variance; rMANOVA, multivariate repeated-measures analysis of variance.
*Subjects with missing data are often omitted from the analysis.
†It is not necessary to omit subjects with missing values from the analysis or to impute missing values.

measures analysis for continuous data, and will use data examples to illustrate the flexibility and limitations of the mixed-effects models. In addition, we will briefly discuss the extent of the shift from end-point and ANOVA-type methods toward mixed-effects regression models in the ARCHIVES. For clarity of discussion, we consider a randomized clinical trial in which each subject undergoes measurement at baseline (time 0) and then at times t_1, t_2, etc, through t_m, where t_m is the end of the treatment period. A summary of the features and drawbacks of the different models is presented in **Table 1**.

TRADITIONAL METHODS

End-point analysis makes use only of the baseline (time 0) and the final observation on each subject (time t_m). The following 2 approaches are most commonly used: ANOVA and analysis of covariance (ANCOVA). Analysis of variance (or *t* test in the case of 2 treatment groups) is used to compare the final measures or the change scores between the baseline and final measurement between the treatment groups. Analysis of covariance is used to compare the final measures between the groups using the baseline measure as a covariate. Analysis by intention-to-treat means that all randomized subjects contribute data to the analysis regardless of whether they were compliant or dropped out of the study. Such analysis then requires that each subject have a final measure. Thus, missing data on the final measure becomes a serious problem. If the analysis is performed only on these subjects who have undergone measurement at time t_m, then the samples compared may not be representative of the populations, and serious sample bias may occur, despite randomization. If the analysis is performed with some imputation method such as the last observation carried forward (LOCF), there is a serious risk for estimation bias. Thus, end-point analysis can yield misleading results when dropout rates differ between the treatment groups.[3-8]

When observations are made at multiple time points, often a *t* test or ANOVA is performed separately at each time point. If a 5% significance level test is used for each test, then the probability of type I error (that is, of mistakenly declaring treatments to be different) sometimes is considerably greater than 5%. This leads to false-positive conclusions. If this proliferation of type I error is controlled using procedures such as a Bonferroni correction,[1] then power is compromised (ie, treatment differences may not be detected when they exist). Performing separate tests at each time point also ignores trends over time and does not allow for direct comparison between treatment groups over time.

Univariate repeated-measures ANOVA[9] provides a more complete description of the time effect, because it includes data from all time points. This approach allows statistical assessment of whether the treatment groups show different response curves over time (treatment × time interaction effect). The treatment response may take a variety of shapes. For example, there might be little or no improvement in one treatment group and larger improvement in the other, or the same total improvement between the baseline and end point in both treatment groups but faster response in one of the groups. In the latter case, end-point analysis will not find a difference, whereas rANOVA may detect a significant time × treatment effect.

Like end-point analysis, analyses using rANOVA are vulnerable to large effects from missing values or imputation. Researchers frequently drop subjects with even 1 missing observation from analysis or use imputation of missing values without acknowledging it. Omis-

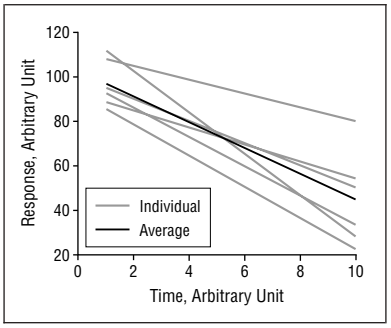

Figure 1. Hypothetical example of a random intercept/random slope model. The 6 gray lines correspond to responses over time of 6 hypothetical subjects. The individuals differ at baseline and in the rate of change over time.

sion of subjects can introduce sample bias, as the group of people with complete data may not be representative of the entire population. Imputation by using the last available observation on each subject in place of all subsequent missing observations (LOCF) usually leads to biased treatment estimates. Another disadvantage of rANOVA is that observations on all individuals need to be made at the same time points.

Univariate repeated-measures ANOVA also requires that correlations among measurements on the same subject satisfy a restrictive condition called *sphericity* or *circularity*. This usually (but not always) amounts to having equal variability of the measurements at each time point and equal correlations between every 2 measurements on the same individual (eg, the correlation between measurements at times t_1 and t_2 is the same as the correlation between measurements at times t_1 and t_m). However, this assumption is infrequently justified, since consecutive observations on the same subject tend to be correlated more highly than observations on the same subject taken further away in time. When this occurs, the type I error rate is inflated, and there is an overestimation of the statistical significance of the treatment × time effect. This problem has long been acknowledged, and the Greenhouse-Geisser[10] and Huynh-Feldt[11] corrections to the significance tests in rANOVA have been proposed. These corrections reduce the numerator degrees of freedom for the statistical tests so that the *P* value is usually adjusted upward and the type I error rate is closer to the target 5% level.[12,13] In addition to serial correlation, the variability of the measurements often increases over time. The Greenhouse-Geisser correction addresses this potential problem, but the correction tends to be very conservative.[14] This means that treatment differences are harder to detect.

Several authors[5,14,15] advocate the general use of rMANOVA[16] rather than rANOVA. This approach is also known as multivariate growth-curve analysis.[4] However, it requires complete data on all subjects and can show significant loss of power (ability to detect treatment differences) if individuals with missing data are dropped from the model. As in rANOVA, individuals with complete data may not be representative of the entire population, and hence results may not be generalized. Imputation of missing data also usually leads to biased estimates, and the multivariate approach has been shown to be less powerful than the univariate approach when the sphericity assumption is satisfied. On the other hand, it does not make any restrictive assumptions about the variances and correlations, it is easily implemented, and, as long as a sufficient sample size is available, its results are valid.[14]

MIXED-EFFECTS MODELS

Random-effects models[17,18] (also called *random-regression models,*[3] *multilevel models,*[19] *hierarchical linear models,*[20] and *empirical Bayes models*[21]) provide a flexible framework for the analysis of repeated measures.[3,4,22] Random-effects models assume that individuals deviate randomly from the overall average response. Consider an example where an individual's response over time is a straight line and individuals differ from one another in their responses at baseline (intercepts) and in their rates of response (slopes). Individuals can start higher or lower and show a higher- or lower-than-average change over time. Such a scenario is illustrated in **Figure 1**, where the gray lines indicate the change in response over time for 6 hypothetical subjects and the black line is the average response over time. The model that generated these data, the random intercept and slope model, is one of the simplest random-effects models. In more complex models, the trajectory over time can be more complicated (eg, there may be an initial fast change and then leveling off of the response), and individuals may differ by the time that the leveling off occurs.

In the random-effects approach, the correlation between repeated observations on the same subject arises from the common random effect(s) for this individual (the random intercept and slope in the example in Figure 1). However, it is also possible to specify directly a pattern for the variances and correlations over time and to use the data to estimate its parameters. Such models are called *covariance-pattern models*[23] and may be used separately or in combination with random effects. The specified structures vary in complexity from a compound symmetry (equal variances at all time points and equal correlations between any 2 measurements on the same subject) to no restrictions at all.[24,25] As an intermediate complexity, one can assume that correlation between observations decreases with increasing time difference. Such a structure is called *autoregressive* and is considered in an example in the "Methods" subsection of the "Simulation Study" section.

ADVANTAGES OF MIXED-EFFECTS MODELS

Mixed-effects (regression) model[23,26] is a general term encompassing models with fixed (eg, treatment) and random effects, covariance pattern models, and combinations of these. The mixed-effects approach has important advantages over traditional methods of repeated-measures analysis. It uses all available data on each subject, it is unaffected by randomly missing data, it can flexibly model time effects, and it allows the use of realistic yet parsimonious variance and correlation patterns for particular applications. For example, in drug trials, variability often increases over time and may differ across subject groups, or variability can be higher

shortly after drug administration than at baseline or later in the study. Traditional methods do not allow the investigator to use that information to achieve more efficient statistical inference and hence greater power. On the other hand, the mixed-effects model allows the researcher to specify several different patterns with varying complexity and to select the best-fitting one using indices of relative goodness of fit such as Akaike information criterion and the Schwarz Bayesian criterion. These indices reflect how well each model agrees with the data and include a penalty for the number of estimated parameters,[23,24] so that overly complicated models are discouraged. Plots of correlation estimates can also help in the choice of the pattern. For a more detailed description, we refer the reader to a tutorial by Littell et al.[24]

Choosing an appropriate pattern of variability over time results in more accurate treatment effect and SE estimates and helps control type I error.[23] Improvements in accuracy may be important because they lead to a decrease of the number of patients required for a particular trial to achieve certain power. Mixed-effects models can handle covariates that change over time (eg, concurrent medication or smoking status) and covariates that do not change with time (eg, sex).

Mixed-effects models allow estimation of average time trends for treatment groups and of the individual's response over time. In contrast to traditional models, the predicted response at each time point is not the same for all subjects in the same treatment group. For example, subjects whose response is larger than the average response at the beginning of the study may tend to have a higher response throughout the study. Mixed-effects models also provide an estimate of the individual variability around the population trend. In the example in Figure 1, this would be the variability of the individual intercepts and slopes and the correlation between them. Mixed-effects models deal seamlessly with unequally spaced observations over time.

Missing data do not present a problem for mixed-effects models as long as data are missing at random, ie, if the chance of a missing value is not related to the unobserved response values. This is the case when an observation is missing because of instrument failure or when the probability of dropping out depends only on covariates included in the model (missing completely at random [MCAR][27]). This is also the case when the likelihood of dropping out depends only on past values of the response variable, but not on future response values (missing at random [MAR][27]). This assumption means that a subject who drops out and a subject who stays in the study with the same response histories have identical probabilities of future response paths. In the MAR and MCAR scenarios, the mixed-effects approach is valid and fully efficient. This is not the case for end-point, rANOVA, and rMANOVA approaches.

However, when data are not randomly missing, the random-effects approach, like all traditional approaches, may yield biased estimates. In psychiatric research, subjects who do not improve are more likely to drop out of the study. This may correspond to random dropout (if the probability of dropping out is related to earlier but not to the current or future response values) or to nonrandom (informative, nonignorable) dropout (if the probability of dropping out is related to the current or future response, or to an unobserved process related to the response). Ascertainment of the type of dropout is difficult and often impossible. When the dropout is nonignorable, one can model the missingness mechanism together with the outcome variable. Different approaches have been proposed in the statistical literature,[5,28-31] but owing to the complexity of such models, the expertise of a statistician is required. Because models for nonignorable missing data rely on assumptions that cannot be verified from the data, they are perhaps most useful in a sensitivity analysis framework to help assess the robustness of a result from a MAR analysis.[32] The MAR assumption is often reasonable in practice,[33] and hence mixed-effects models without modeling the missingness mechanism can be used as a primary analysis tool.

POTENTIAL PROBLEMS AND MISUSES OF MIXED-EFFECTS MODELS

Despite the many advantages of mixed-effects models, their use outside the statistical literature has been limited. This can be attributed to the complexity of these models, and the only recent introduction of reliable software to the general public (eg, SAS PROC MIXED[34] and MIXREG[35]). Applying mixed models to small samples or including unnecessary covariates may bias parameter estimates and statistical tests.[36-39] Assumption violations are also harder to ascertain and may lead to erroneous conclusions. Mixed-effects models occasionally have computational problems when the iterative fitting algorithm fails to converge, so the range of possible models is limited by sample size, number of time points, and correlation structure of the data.

SIMULATION STUDY

Methods

To illustrate some of the advantages of mixed models and especially the less popular covariance-pattern approach, we conducted a small simulation study and analyzed real data from a clinical trial.[40] The simulated data example is motivated by 2 clinical trials in patients with schizophrenia who were given medication after baseline assessments. Outcome measures were collected on several groups of subjects repeatedly over time. Variability right after the drug administration was higher than at the baseline and subsequent assessments, and the correlation between measurements on the same individual decreased with the increasing time difference.

We simulated 500 data sets with such correlation structure. Each data set consisted of 4 repeated measurements on 50 subjects equally divided in 2 groups (a treatment and a control group). Both groups had similar baseline measurements and showed an increase at the second time point and a decrease after the peak. However, the treatment group had a larger increase and

maintained a higher response after the peak (**Figure 2**). The mean scores were 9, 11, 10, and 9 for the treatment group and 9, 10.5, 9, and 8 for the control group. The variances were the same (equal to 1) at all time points except immediately after the drug administration (variance equal to 4), and the correlations followed autoregressive structure with a within-subject correlation of 0.5. This means that the correlation between consecutive observations on the same subject was 0.5, between observations 2 units apart on the same subject was $0.5 \times 0.5 = 0.25$, and between observations 3 units apart on the same subject was $0.5 \times 0.5 \times 0.5 = 0.125$.

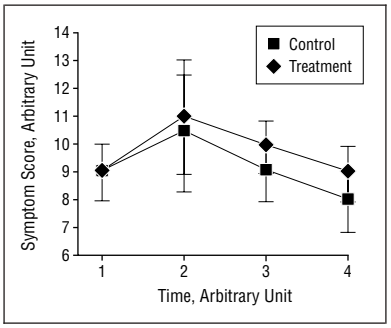

Figure 2. Mean and SD for 2 groups of subjects receiving 2 different treatments in a hypothetical repeated-measures example.

We fit rANOVA, rMANOVA, and 3 different mixed-effects models for the complete data and when 10% to 15% of the data were missing according to 3 different patterns of missingness. The MCAR pattern was created by assigning equal chance to each value to be missing. The MAR and the informative missingness patterns were created by assigning a 40% chance of dropout for patients whose observed value at time 2 was greater than 11 (the average in the treatment group). The only difference was that in the informative dropout pattern, the value that triggered the dropout was treated as unobserved and was deleted from the data set.

Results

The type I error rates and power estimates for the treatment × time interaction tests (the effect of greatest interest) are reported in **Table 2**. Results from the rANOVA approach are shown in 3 columns corresponding to unadjusted F tests and F tests with the Greenhouse-Geisser and Huynh-Feldt adjustments. The 3 mixed models do not have random effects and have compound symmetry and unstructured and heterogeneous autoregression (the correct one) covariance patterns.

For the complete data, the mixed model with the most appropriate correlation structure (ARH) provides the best power (80%) and the lowest type I error rate (0%). The rMANOVA approach and its equivalent mixed model (unstructured pattern [UN]) provide the same power (80%) but have higher type I error rates (10%). The adjusted rANOVAs have the same type I error rate as the ARH mixed model (0%) but have much smaller power (60%). When data are deleted according to all considered missing-data scenarios, all approaches show decrease in power, but the ARH mixed model usually has the best power (≥73%) and some of the lowest type I error rates (2%-5%). The UN mixed model usually has similar power (within 2% difference) but higher type I error rate (4%-9%) than the best-fitting ARH model. Although it is equivalent to the rMANOVA model for complete data, the UN mixed model has more than 10% higher

Table 2. Results From the Tests of the Treatment × Time Interaction in the Simulation Study

	rANOVA				Mixed-Effects Model		
Data Set Effect	Unadjusted	GG	HF	rMANOVA	CS	UN	ARH
Complete data							
Type I error	0.10	0.00	0.00	0.10	0.10	0.10	0.00
Power	0.70	0.60	0.60	0.80	0.70	0.80	0.80
MCAR							
Type I error	0.07	0.05	0.06	0.05	0.06	0.04	0.02
Power	0.50	0.45	0.46	0.60	0.55	0.75	0.73
MCAR, LOCF							
Type I error	0.07	0.04	0.05	0.04	NA	NA	NA
Power	0.48	0.42	0.44	0.52			
MAR							
Type I error	0.07	0.05	0.05	0.08	0.07	0.09	0.05
Power	0.56	0.49	0.51	0.66	0.58	0.78	0.78
MAR, LOCF							
Type I error	0.03	0.01	0.02	0.02	NA	NA	NA
Power	0.53	0.48	0.50	0.43			
Informative missing							
Type I error	0.07	0.05	0.05	0.08	0.07	0.09	0.05
Power	0.56	0.49	0.51	0.66	0.65	0.77	0.77
Informative missing, LOCF							
Type I error	0.07	0.04	0.05	0.09	NA	NA	NA
Power	0.56	0.47	0.49	0.67			

Abbreviations: ARH, autoregressive heterogeneous pattern; CS, compound symmetry pattern of variances and correlations; GG, Greenhouse-Geisser correction; HF, Huynh-Feldt correction; LOCF, missing values filled in with last observation carried forward; MAR, missing at random; MCAR, missing completely at random; NA, imputation with LOCF should not be performed; rANOVA, univariate repeated-measures analysis of variance; rMANOVA, multivariate repeated-measures analysis of variance; UN, unstructured pattern.

power than the rMANOVA model when missing data are present, since it does not drop subjects with incomplete data from analysis. The power deterioration is also evident for the rANOVA models with and without imputation of missing values. The power of the unadjusted rANOVA test in the complete data scenario is 70%, whereas the best power in any of the missing data/imputation scenarios is 56%. The LOCF has differential effect on the results, depending on the missing data scenario and on the method of analysis used. For example, the rMANOVA approach demonstrates a decrease in power (from 66% to 43%) and in type I error rate (from 8% to 2%) after LOCF imputation in the MAR scenario and almost no change in the informative missingness scenario (power and type I error rates are within 1% difference). Although the LOCF is considered by some to be a conservative approach, it may lead to more liberal rather than more conservative tests, depending on the application.

In this example, the mixed-effects approach has the best power to detect treatment differences, especially in the case of missing data. The LOCF has a different effect, depending on the dropout mechanism, whereas the mixed-effects approach was relatively unaffected by the presence and mechanism of missingness. Although examples can be constructed when the mixed model will also give misleading results in the presence of informative dropout simulation, studies suggest that if appropriately used, it may be more robust than the traditional approaches in the presence of informative dropout.[6,7]

A CLINICAL TRIAL DATA SET

Methods

We now illustrate how to use the mixed-model approach and compare it with the standard approaches on a data set from a randomized clinical trial of the effect of naltrexone as an adjunct to standardized psychosocial therapy in the treatment of alcohol dependence.[40] Six hundred twenty-seven veterans with chronic, severe alcohol dependence were assigned to 12 months of naltrexone treatment, 3 months of naltrexone treatment followed by 9 months of placebo treatment, or 12 months of placebo treatment. Herein we consider one of the secondary outcome measures from the original study: the average financial satisfaction score. This score is defined as the average of the 4 financial items of the Lehman Quality of Life Scale. On each item, values ranged from 1 (terrible) to 7 (delighted), the average scores had approximately bell-shaped distribution, and hence the assumption of normality underlying the traditional approaches and the mixed-model approach was at least approximately satisfied. The Lehman Quality of Life Scale was administered at baseline and weeks 4, 26, 52, and 78 of the trial. A graph of means over time is shown in **Figure 3**.

We considered a random intercept-slope mixed-effects model with autocorrelated errors and several different covariance-pattern models with plausible patterns of covariance (compound symmetry, autocorrelated for unequally spaced time points, and unstructured). The unstructured pattern was the best-fitting one according to the Akaike and Bayesian information criteria. Residual plots did not show any obvious patterns, and hence we concluded that the assumptions of the model were reasonably satisfied and that there were no outliers.

Results

Only 211 subjects had complete data, and traditional analyses in the first part of **Table 3** are based on this smaller sample. In the second part of Table 3, results after LOCF imputation are shown. No endpoint differences and no significant group × time effects are observed. The time effect is always highly significant, indicating improvement over time for all subjects staying in the study. The random intercept–random slope model allows for estimation of slopes for each treatment group on average and for each individual. After LOCF imputation, the group main effect is significant using the rANOVA and rMANOVA approaches at the level of .05 ($P = .04$), and hence the LOCF approach is more liberal in this case. The best-fitting mixed model does not show a significant group effect ($P = .07$). Since a large proportion of the data are missing, the results may be significantly affected by the pattern of missingness. For example, the time effect may be overestimated if subjects who drop out of the study get financially more dissatisfied. Because dropout rates do not differ between treatment groups over time, and because the group × time effect is very highly nonsignificant, it is unlikely that taking the dropout mechanism into account will lead to a change in the conclusions of no differences according to treatment regimen over time.

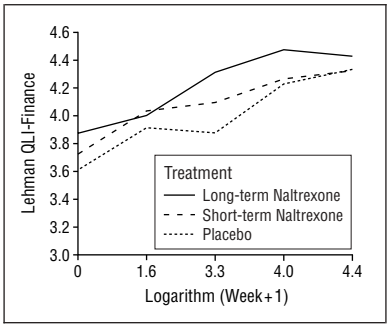

Figure 3. Average Lehman Quality of Life Indices of Finance Satisfaction (QLI-Finance) for the placebo and short- and long-term naltrexone treatment arms of the naltrexone clinical trial. No significant treatment × time effect was observed using traditional and mixed-model approaches.

COMMENT

Univariate repeated-measures ANOVA is still the most commonly used statistical analysis tool for repeated measures in psychiatric research because of its simplicity and familiarity to researchers. However, as psychiatrists become increasingly aware of the advantages of mixed-effects models, the proportion of reports that use the newer methods increases. To detect the extent to which the shift away from traditional methods toward mixed-effects models has taken place within the ARCHIVES, we reviewed all published studies with repeated-measures designs in the past 12 years and classified them according to the most advanced statistical method used (mixed-effects model, rMANOVA, rANOVA, and endpoint analysis, in that order). The

Table 3. Results From Repeated-Measures Analyses of the Financial Satisfaction Outcome From the Naltrexone Randomized Clinical Trial

Effect	ANCOVA on End Point	rANOVA With GG Adjustment	rMANOVA	Mixed-Effects Model With Unstructured Pattern	Random Intercept/Slope Model With Autoregressive Errors
Group	$F_{2,335} = 0.08$ $P = .92$	$F_{2,208} = 0.22$ $P = .80$	$F_{2,208} = 0.22$ $P = .80$	$F_{2,591} = 2.62$ $P = .07$	$F_{2,642} = 1.82$ P = .16
Time		$F_{4,832} = 19.8$ $P<.001$	$F_{4,205} = 16.2$ $P<.001$	$F_{4,402} = 28.1$ $P<.001$	$F_{1,671} = 104.5$ $P<.001$
Group × time		$F_{8,832} = 0.71$ $P = .69$	$F_{8,410} = 0.87$ $P = .54$	$F_{8,566} = 0.77$ $P = .63$	$F_{2,671} = 0.04$ $P = .96$
LOCF					
Group	$F_{2,617} = 0.67$ $P = .51$	$F_{2,618} = 3.15$ $P = .04$	$F_{2,618} = 3.15$ $P = .04$	LOCF not appropriate	LOCF not appropriate
Time		$F_{4,2472} = 38.8$ $P<.001$	$F_{4,615} = 21.6$ $P<.001$		
Group × time		$F_{8,2472} = 0.79$ $P = .58$	$F_{8,1230} = 0.85$ $P = .56$		

Abbreviations: ANCOVA, analysis of covariance; GG, Greenhouse-Geisser correction; LOCF, missing values filled in with last observation carried forward; rANOVA, univariate repeated-measures analysis of variance; rMANOVA, multivariate repeated-measures analysis of variance.

percentage of papers with repeated-measures analysis using mixed-effects models increased from 0% in 1989 to almost 30% in 1999 and 2001.

Mixed-effects models provide a very flexible approach for the analysis of repeated-measures data arising from medical research studies. They allow for assessment of individual and population trends over time, for the use of time-independent and time-dependent covariates and irregular measurement occasions. They use all available data on each individual and provide a choice of appropriate covariance pattern that may lead to more efficient estimation. As the pattern of variability is not known a priori, a comparison of several alternative structures may help the researcher to choose the best-fitting one and hence to obtain additional information about the data. The traditional approaches provide no such flexibility. Even when the rANOVA approach provides a good approximation of the significance level for the treatment × time effect, it tells nothing about the pattern of change in variability over time. The rMANOVA approach provides general estimates of the variances and correlations, but without an alternative model for comparison, it is not possible to see whether there is a more parsimonious structure that describes the data well and may lead to better power.

Mixed-effects models should be used with caution because of their complexity and opportunity for misuse, and because of the requirement of relatively large samples. Mixed-effects models can give biased results in the presence of informative missingness. A number of simulation studies[6,7,36-39,41-43] have investigated the performance of mixed models and documented appropriate uses and misuses of these models. However, specific guidelines for the use of these methods for analysis of data from clinical trials and longitudinal follow-up studies in psychiatry are not yet available.

In general, mixed-effects models are the preferable method of analysis of repeatedly measured outcomes when there are missing data, the repeated measures are irregularly spaced over time, and the sample sizes are modest to large. Mixed models yield unbiased and efficient estimates under MCAR and MAR assumptions, which are often reasonable in clinical trials. Traditional and mixed-effects approaches produce biased results in the presence of informative dropout, but the bias may be smaller in mixed-effects models.[6,7] Mixed models can also be used for sensitivity analysis. Situations in which the traditional methods (end point, rANOVA, and rMANOVA) may be preferred to the mixed-model approach include complete data with observations taken at the same occasions. Univariate repeated-measures ANOVA may also be preferred in small samples ($n<10$) when the implicit assumptions of normality and sphericity are reasonably satisfied.

Although our focus in the present article has been on continuous outcomes, mixed-effects models can also be used in the case of binary, categorical, or other nonnormal data.[44-48] With modern computing power, it is not necessary to transform data to normality, as suitable models can be developed and applied to the raw data. The generalized linear mixed models are more complex than normal mixed-effects models, but such models provide great flexibility in analyzing a wide range of data and should be considered in analysis planning. In 1993, Gibbons et al[3] commented that "methods by which longitudinal studies are analyzed are not commensurate with the level of effort involved in their collection." Although some progress has been made in the psychiatric literature, as demonstrated by the shift from traditional methods to mixed-effects models in the ARCHIVES, more appropriate, well-performed analyses are needed to better understand and use available data.

Submitted for publication June 28, 2002; final revision received August 8, 2003; accepted August 15, 2003.

This study was supported by grants KO2 AA 00261-01, AA-99-005, and AA12870-03 from the National Institute on Alcohol Abuse and Alcoholism, New Haven, Conn, and the Department of Veterans Affairs (the

National Center for PTSD, Alcohol Research Center, and Schizophrenia Biological Research Center), West Haven, Conn, and the Department of Veterans Affairs, Medical Research Service, Cooperative Studies Program, West Haven.

We thank Helena Kraemer, PhD, and C. Neill Epperson, MD, for their comments on the manuscript.

Corresponding author: Ralitza Gueorguieva, PhD, Connecticut Mental Health Center, 34 Park St, Room 329A, New Haven, CT 06519 (e-mail: ralitza.gueorguieva@yale.edu).

REFERENCES

1. Ekstrom D, Quade D, Golden RN. Statistical analysis of repeated measures in psychiatric research. *Arch Gen Psychiatry.* 1990;47:770-772.
2. Lavori P. ANOVA, MANOVA, my black hen: comments on repeated measures. *Arch Gen Psychiatry.* 1990;47:775-778.
3. Gibbons RD, Hedeker D, Elkin I, Waternaux C, Kraemer HC, Greenhouse JB, Shea MT, Imber SD, Sotsky SM, Watkins JT. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Arch Gen Psychiatry.* 1993;50:739-750.
4. Gibbons RD, Hedeker D, Waternaux CM, Davis JM. Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacol Bull.* 1988;24:438-443.
5. Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data. *Psychol Methods.* 1997;2:64-78.
6. Mallinckrodt CH, Clark WS, David S. Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Inf J.* 2001;35:1215-1225.
7. Mallinckrodt CH, Clark WS, David S. Accounting for dropout bias using mixed effects models. *J Biopharm Stat.* 2001;11:9-21.
8. Hennen J. Statistical methods for longitudinal research on bipolar disorders. *Bipolar Disord.* 2003; 5:156-168.
9. Winer BJ. *Statistical Principles in Experimental Design.* New York, NY: McGraw-Hill Co; 1971.
10. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika.* 1959;24: 95-112.
11. Huynh H, Feldt LS. Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *J Educ Stat.* 1976;1:69-82.
12. Jaccard J, Ackerman L. Repeated measures analysis of means in clinical research. *J Consult Clin Psychol.* 1985;3:426-428.
13. Vasey MW, Thayer JF. The continuing problem of false positives in repeated measures ANOVA in psychophysiology: a multivariate solution. *Psychophysiology.* 1987;24:479-486.
14. Park T. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Stat Med.* 1993;12:1723-1732.
15. Hertzog C, Rovine M. Repeated-measures analysis of variance in developmental research: selected issues. *Child Dev.* 1985;56:787-809.
16. Hand DJ, Taylor CC. *Multivariate Analysis of Variance and Repeated Measures.* New York, NY: Chapman & Hall; 1987.
17. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics.* 1982;38:963-974.
18. Ware JH. Linear models for the analysis of longitudinal studies. *Am Stat.* 1985;39:95-101.
19. Goldstein H. *Multilevel Models in Education and Social Research.* New York, NY: Oxford University Press Inc; 1987.
20. Bryk AS, Raudenbush SW. Application of hierarchical linear models to assessing change. *Psychol Bull.* 1987;101:147-158.
21. Casella G. An introduction to empirical Bayes data analysis. *Am Stat.* 1985;39:83-87.
22. Nich C, Carroll K. Now you see it, now you don't: a comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *J Consult Clin Psychol.* 1997;65:252-261.
23. Brown H, Prescott R. *Applied Mixed Models in Medicine.* New York, NY: John Wiley & Sons Inc; 1999.
24. Littell RC, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. *Stat Med.* 2000;19:1793-1819.
25. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics.* 1986;42:805-820.
26. Cnaan A, Laird NM, Slasor P. Tutorial in biostatistics: using the general linear mixed models to analyze unbalanced repeated measures and longitudinal data. *Stat Med.* 1997;16:2349-2380.
27. Little RA, Rubin DB. *Statistical Analysis with Missing Data.* New York, NY: John Wiley & Sons Inc; 1987.
28. Little RA. Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc.* 1995;90:1112-1121.
29. Wu MC, Follmann DA. Use of summary measures to adjust for informative missingness in repeated measures data with random effects. *Biometrics.* 1999;55:75-84.
30. Wu MC, Albert PS, Wu BU. Adjusting for drop-out in clinical trials with repeated measures: design and analysis issues. *Stat Med.* 2001;20:93-108.
31. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *Appl Stat.* 1994;43: 49-93.
32. Mallinckrodt CH, Sanger TM, Dubé S, DeBrote DJ, Molenberghs G, Carroll RJ, Potter WZ, Tollefson GD. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol Psychiatry.* 2003;53:754-760.
33. Mallinckrodt CH, Clark WS, Carroll RJ, Molenberghs G. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J Biopharm Stat.* 2003;13: 179-190.
34. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc; 1987.
35. Hedeker D, Gibbons RD. MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Comput Methods Programs Biomed.* 1996;49:229-252.
36. Guerin L, Stroup WW. A simulation study to evaluate PROC MIXED analysis of repeated measures data. In: Proceedings of the 12th Kansas State University Conference on Applied Statistics in Agriculture; April 30–May 2, 2000; Manhattan, Kan. 2000:170-203.
37. Overall JE, Ahn C, Shivakumar C, Kalburgi Y. Problematic formulations of SAS PROC MIXED models for repeated measurements. *J Biopharm Stat.* 1999;9:189-216.
38. Ahn C, Tonidandel S, Overall JE. Issues in use of SAS PROC MIXED to test the significance of treatment effects in controlled clinical trials. *J Biopharm Stat.* 2000;10:265-286.
39. Delucchi K, Bostrom A. Small sample longitudinal clinical trials with missing data: a comparison of analytic methods. *Psychol Methods.* 1999; 4:158-172.
40. Krystal JH, Cramer JA, Krol WF, Kirk GF, Rosenheck RA. Naltrexone in the treatment of alcohol dependence. *N Engl J Med.* 2001;345:1734-1739.
41. Overall JE. Drop-outs and a random regression model. *J Biopharm Stat.* 1997;7:383-402.
42. Overall JE, Shobaki G, Fiore J. Random regression with imputed values for dropouts. *Psychopharmacol Bull.* 1996;32:377-388.
43. Mazumdar S, Liu KS, Houck PR, Reynolds CF. Intent-to-treat analysis for longitudinal clinical trials coping with the challenge of missing values. *J Psychiatr Res.* 1999;33:87-95.
44. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* Oxford, England: Clarendon Press; 1996.
45. Longford NT. *Random Coefficient Models.* Oxford, England: Clarendon Press; 1993.
46. Lindsey JK. *Models for Repeated Measurements.* Oxford, England: Clarendon Press; 1993.
47. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Comput Methods Programs Biomed.* 1996; 49:157-176.
48. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics.* 1994;50:933-944.