Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials

Craig H. Mallinckrodt, PhD
Research Advisor, Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana

Peter W. Lane, MA, CStat
Director of Consultancy and Training, Research Statistics Unit, GlaxoSmithKline, Harlow, United Kingdom

Dan Schnell, PhD
Section Head, Pharmaceutical Statistics, Procter & Gamble Pharmaceuticals, Mason, Ohio

Yahong Peng, PhD
Senior Biometrician, Clinical Biostatistics, Merck Research Lab, Upper Gwynedd, Pennsylvania

James P. Mancuso, PhD
Associate Director, Statistics, Pfizer Inc, Groton, Connecticut

This position paper summarizes relevant theory and current practice regarding the analysis of longitudinal clinical trials intended to support regulatory approval of medicinal products, and it reviews published research regarding methods for handling missing data. It is one strand of the PhRMA initiative to improve efficiency of late-stage clinical research and gives recommendations from a cross-industry team. We concentrate specifically on continuous response measures analyzed using a linear model, when the goal is to estimate and test treatment differences at a given time point. Traditionally, the primary analysis of such trials handled missing data by simple imputation using the last, or baseline, observation carried forward method (LOCF, BOCF) followed by analysis of (co)variance at the chosen time point. However, the general statistical and scientific community has moved away from these simple methods in favor of joint analysis of data from all time points based on a multivariate model (eg, of a mixed-effects type). One such newer method, a likelihood-based mixed-effects model repeated measures (MMRM) approach, has received considerable attention in the clinical trials literature. We discuss specific concerns raised by regulatory agencies with regard to MMRM and review published evidence comparing LOCF and MMRM in terms of validity, bias, power, and type I error. Our main conclusion is that the mixed model approach is more efficient and reliable as a method of primary analysis, and should be preferred to the inherently biased and statistically invalid simple imputation approaches. We also summarize other methods of handling missing data that are useful as sensitivity analyses for assessing the potential effect of data missing not at random.

Key Words
Missing data; Longitudinal data; Primary analysis; Clinical trials

Correspondence Address
Craig Mallinckrodt, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285 (email: cmallinc@lilly.com).

INTRODUCTION

In longitudinal trials, efficacy is often assessed in terms of treatment differences at a specific time point, usually the last time at which observations are planned while patients are under treatment. A major difficulty in analyses of such trials is missing data at the chosen time point, often due to patients withdrawing (or dropping out) from treatment. Inference from the results of a trial can be complicated by the method used to handle the missing data because the inference may depend on the method and its assumptions.

Historically, the simple imputation method, called last observation carried forward (LOCF), has been used for the primary efficacy analysis of clinical trials supporting registration of new medicines (1). This approach is simple to carry out and is generally regarded as conservative in that it tends to under- rather than overestimate treatment effects. Although the appropriateness of LOCF hinges on strong assumptions, it is also generally regarded as less biased than an analysis of completing subjects only, potentially counteracting bias caused by differential timing, rates, and reasons for dropout in the various treatment arms.

Over the past 20 years, statistical methodology and software have been developed that allow for the routine use of alternative approaches with less restrictive assumptions than LOCF. These methods are based on analyzing the observations made at all time points. One such longitudinal approach, which has been extensively studied in regulatory settings, uses a model referred to as *multivariate*, or *mixed*, and is increasingly denoted in the literature by the abbreviation MMRM (mixed model for repeated measures) (2–14).

Drug Information Journal, *Vol. 42, pp. 303–319, 2008 • 0092-8615/2008*

Submitted for Publication: August 6, 2007
Accepted for Publication: November 29, 2007

The MMRM method is from the broader class of direct-likelihood analyses and makes use of fully and partially observed data sequences from individual patients by estimating the covariance between data from different time points (1). As is described in an upcoming section, it is often useful to implement MMRM using an unstructured approach to modeling both the treatment-by-time means and the (co)variances, leading to what is essentially a multivariate normal model wherein treatment group means at the primary time point are adjusted to reflect both the actually observed data and the projected outcomes from the patients with missing data (see, eg, articles by Cnaan et al. [15], Molenberghs and colleagues [5], and Molenberghs and Kenward [1]). Other methods, such as multiple imputation, are also the result of advances in methodology and software but have not been studied as extensively as MMRM in regulatory settings.

Given the strong theoretical and empirical evidence favoring MMRM over LOCF, it is not surprising that use of LOCF as the primary analysis has been questioned by statisticians and clinicians in academic, industry, and regulatory settings. However, regulatory agencies frequently require that primary analyses of efficacy use LOCF. For example, Dr. Linda Yau surveyed statisticians working in phases 2 and 3 from a wide range of therapeutic areas, including neuroscience, antivirals, respiratory, gastrointestinal, urology, and cardiovascular. In her presentation at the DIA Conference in Philadelphia, June 2006, Dr. Yau noted that LOCF was almost universally preferred by regulatory agencies as the primary analysis. However, there was generally no objection to using more recent methods such as MMRM for primary analyses in phase 1, nor for trials on medical devices or diagnostic tests. In addition, plans for some vaccine trials in phase 2 have included MMRM or multiple imputation as the primary analysis.

In our experience, decisions regarding choice of the primary analysis have been hampered by misunderstandings of concepts, some of which stem from inconsistency in terminology. This, in turn, has led to misunderstandings regarding the implications of the research comparing LOCF and MMRM. Additional difficulties may have arisen from differences in the perspectives of pharmaceutical companies and regulators, either real or perceived.

The purpose of this article is to capitalize on the diverse experience of researchers at a number of pharmaceutical companies in order to (1) clarify terminology and concepts regarding use of MMRM and LOCF in regulatory settings, (2) address specific concerns raised by regulatory agencies regarding use of MMRM as the primary analysis, and (3) make specific recommendations for analysis of data from confirmatory longitudinal clinical trials with continuous endpoints.

Regarding our perspective on the choice of primary analysis, this article is the consensus of an expert working team from the Efficiency in Clinical Trials Initiative of the Pharmaceutical Research and Manufacturers of America (PhRMA). We believe there is a compelling public health need to develop drugs using the best possible scientific methods in all disciplines in order to meet patient needs with better and more affordable medicines. We believe regulators share this perspective, as evidenced by the various Critical Path initiatives. Hopefully, this article will help drug developers and regulators achieve their common goal.

TERMINOLOGY AND CONCEPTS REGARDING USE OF MMRM AND LOCF IN REGULATORY SETTINGS

MISSING DATA TERMINOLOGY AND CONCEPTS

In order to understand the potential impact of missing data, the process (ie, mechanisms) leading to the missingness must be considered. The following taxonomy of missing-data mechanisms is now common in the statistical literature (16).

Data are considered *missing completely at random* (MCAR) if, conditional upon the independent variables in the analytic model, the missingness does not depend on either the observed or unobserved outcomes of the variable being analyzed (Y). Data are *missing at random* (MAR) if,

conditional upon the independent variables in the analytic model, the missingness depends on the observed outcomes of the variable being analyzed (Yobs) but does not depend on the unobserved outcomes of the variable being analyzed (Ymiss). Data are *missing not at random* (MNAR) if, conditional upon the independent variables in the analytic model, the missingness depends on the unobserved outcomes of the variable being analyzed.

Several key points arise from these definitions. First, the characterization of the missingness mechanism does not rest on the data alone; it involves both the data and the model used to analyze the data. Consequently, missingness that might be MNAR given one model could be MAR or MCAR given another. In addition, since the relationship between the dependent variable and missingness is a key factor in the missingness mechanism, the mechanism may vary from one outcome to the next within the same data set. Together, these consequences imply that statements about the missingness mechanism without reference to the analytic model and the specific variable being analyzed are problematic to interpret. It also implies that broad statements regarding missingness and validity of particular analytic methods across specific disease states are unwarranted.

Moreover, terms such as *ignorable missingness* or *informative censoring* can be even more problematic to interpret. For example, in the case of likelihood-based estimation, if the parameters defining the measurement process are independent of the parameters defining the missingness process (sometimes referred to as the *separability* or *distinctness* condition), the missingness is ignorable if it arises from an MCAR or MAR process but is nonignorable if it arises from an MNAR process (17). In this context, *ignorable* means the missing-data mechanism can be ignored because unbiased parameter estimates can be obtained from the observed data. Hence, if missing data are described as ignorable or nonignorable, this must be done with reference to both the estimation method and the analytic model. For example, given a certain model, missing data arising from an MAR mechanism might be ignorable if parameters were estimated via maximum likelihood but would not be ignorable if parameters were estimated via a frequentist method that assumes MCAR (18).

These subtleties can be easy to overlook in practice, leading to misunderstandings about missing data and its consequence. For example, when dropout rates differ by treatment group, then it can be said that dropout is not random. But it would be incorrect to conclude that the missingness mechanism giving rise to the dropout is MNAR and that analyses assuming MCAR or MAR would be invalid. Although dropout is not completely random in the simplest sense, if dropout depends only on treatment, and treatment is included in the analytic model, the mechanism giving rise to the dropout would be MCAR. Some authors, such as Little (19), distinguish between *pure* MCAR (missingness depends on nothing at all) and *covariate-dependent* MCAR. The previous example could therefore also be described as being covariate-dependent MCAR.

CONCEPTS AND CHARACTERIZATIONS OF LAST OBSERVATION CARRIED FORWARD

Although this section focuses on LOCF, many of the points also apply to baseline observation carried forward (BOCF). LOCF is not itself an analytic approach, but rather a method for imputing missing values. Therefore, the appropriateness of an analysis using LOCF depends on both the assumptions of LOCF and the assumptions of the method used to analyze the data. When assessing LOCF mean change via analysis of variance (ANOVA), the key assumptions are that missing data arise from an MCAR mechanism and that for subjects with missing endpoint observations, their responses at the endpoint would have been the same as their last observed values.

The following example, using the hypothetical data in Table 1, illustrates the handling of missing data via LOCF: For patient 3, the last observed value, 19, is used in the computation of the mean change to endpoint for treatment group 1; and for patient 6, the last observed value, 20, is used in the computation of the mean

TABLE 1

Hypothetical Data Used to Illustrate How Various Methods Handle Missing Data

			Week					
Patient	Treatment	Baseline	1	2	3	4	5	6
1	1	22	20	18	19	14	12	10
2	1	22	21	18	11	12	11	6
3	1	22	22	21	20	19	*	*
4	2	20	20	20	20	19	21	22
5	2	21	22	22	23	23	25	26
6	2	18	19	20	*	*	*	*

*Missing values due to patient dropout.

change to endpoint for treatment group 2. The analysis does not distinguish between the actually observed data and the imputed data.

Even when the assumptions for LOCF hold, it must also be recognized that because LOCF is a single-imputation method, the uncertainty of imputation is not taken into account. Therefore, the analysis will, in essence, think more data exist than is actually the case (17). This well-known limitation of LOCF results in systematic underestimation of the standard errors (7,8,20).

CONCEPTS AND CHARACTERIZATIONS OF MMRM

Likelihood-based mixed-effects models offer a general framework from which to develop longitudinal analyses under the MAR assumption (15,17). Laird and Ware (21) introduced the general linear mixed-effects model to be any model that satisfies

$$\begin{aligned} Y_i &= X_i\beta + Z_i b_i + \varepsilon_i \\ b_i &\sim N(0, D) \\ \varepsilon_i &\sim N(0, \Sigma_i) \\ b_1 \ldots b_n, &\varepsilon_1 \ldots \varepsilon_n \text{ independent} \end{aligned} \qquad (1)$$

where Y_i is the n_i-dimensional response vector for subject i; β is the p-dimensional vector of fixed effects; b_i is the q-dimensional vector of random (subject-specific) effects; X_i and Z_i are $(n_i \times p)$- and $(n_i \times q)$-dimensional matrices relating the observations to the fixed and random effects, respectively; ε_i is the n_i-dimensional vector of residuals; D is a general $(q \times q)$-dimensional covariance matrix with (i,j) element $d_{ij} = d_{ji}$; and Σ_i is a general $(n_i \times n_i)$-dimensional covariance matrix (usually the same for all i). It follows from this model that, marginally,

$$Y_i \sim N(X_i\beta, V) \text{ and } V = Z_i D Z_i' + \Sigma_i$$

A key general feature of mixed-effects models is that they include fixed and random effects, whereas ANOVA models include only fixed effects (apart from the residuals). In clinical trials, the subject-specific (random) effects are seldom the focus. Rather, the trials are typically designed to assess differences in fixed effects, most notably treatment effects. However, accounting for the random effects is important in order to make the most appropriate inferences regarding the fixed effects. Indeed, not doing so would typically affect the precision of estimates and result in incorrect inferences.

A simple formulation of the general linear mixed model (Eq. 1) can be implemented in which the random effects are not explicitly modeled, but rather are included as part of the marginal covariance matrix V, just defined, leading then to what could alternatively be described as a multivariate normal model. Modeling the random effects as part of the within-patient error correlation structure is the feature that distinguishes MMRM from other implementations of mixed-effects models.

The following example, using the hypothetical data in Table 1, illustrates the handling of missing data via an MMRM analysis: Information

from the observed outcomes is used via the within-patient correlation structure to provide information about the unobserved outcomes, but missing data are not explicitly imputed. Specifically, patient 3 had been doing worse than the average of patients in treatment group 1. Means for treatment group 1 at visits 5 and 6 are adjusted to reflect that had patient 3 stayed in the trial, her observations at visits 5 and 6 would likely have been worse than the treatment group average. But the analysis predicts that patient 3 would have had some additional improvement because the other patients in group 1 all improved. Patient 6 had also been doing marginally worse than the average of patients in his group (treatment group 2). Means for treatment group 2 at visits 3–6 are adjusted to reflect that had patient 6 remained in the trial, his observations would likely have continued to worsen at a rate slightly greater than the treatment group average.

The magnitudes of these adjustments are determined mathematically from the data. Additional details can be found elsewhere (15,17,21). Although these details go beyond the scope of this article, the basic principle is easily appreciated. A mixed-effects analysis uses all the available data (Yobs) to compensate for the data missing on a particular patient, whereas LOCF uses only one data point. Again, using the hypothetical data in Table 1, in dealing with the missing data for patient 3, a mixed-effects analysis considers data from visits 1–4 on patient 3 as well as all the data from patients 1 and 2. In contrast, LOCF uses only the visit 4 value from patient 3, assuming that visit 6 will be the same as visit 4, even though that was not the case for any patient whose data were observed.

SPECIFIC CONCERNS RAISED BY REGULATORY AGENCIES REGARDING USE OF MMRM AS THE PRIMARY ANALYSIS

It is widely recognized that the restrictive assumptions for ANOVA with LOCF seldom hold (1,17). It has also been clearly established that when data fail to conform to these assumptions, use of LOCF can lead to biased estimates of treatment effects, biased tests of the null hypothesis of no treatment effect, underestimates of standard errors, inflated type I error, and coverage probabilities that may be far from the nominal level (1,2,4–8,11,12,17,23–31).

The assumption of MAR is often reasonable because, particularly in longitudinal studies wherein the evolution of treatment effects is assessed by design over time, the observed data and the models used to analyze them can explain much of the missingness (16,17). This point may be especially relevant in well-controlled studies such as clinical trials, in which extensive efforts are made to observe all the outcomes and the factors that influence them while patients are following protocol-defined procedures (32). Hence, longitudinal clinical trials by their very design aim to reduce the amount of MNAR data (missingness explained by unobserved responses), thereby increasing the plausibility of MAR. Further, it is evident that MAR is always more plausible than MCAR because MAR is always valid if MCAR is valid, and MAR can be valid in cases when MCAR is not.

Despite the advantages of MAR methods, LOCF is still favored for use as the primary analysis in many therapeutic areas. The following sections address the concerns cited by regulatory agencies when considering MMRM for the primary analysis, along with responses to those concerns.

LOCF IS CONSERVATIVE

One of the reasons often cited for the continued widespread use of LOCF is that the potential biases in LOCF lead to a conservative analysis. In this context, conservative is typically thought of as underestimating the magnitude of the treatment effect.

It is intuitively obvious that LOCF yields conservative estimates of within-group changes in many scenarios. However, interpretations of treatment effects are based on between-group comparisons. Results from empirical studies (3,7,8,11,12,17,22,23) have clearly shown that conservative behavior of LOCF in regard to between-group comparisons is far from guaranteed and, in fact, is in some scenarios unlikely.

The potential for anticonservative behavior of LOCF has also been confirmed via analytic proof (5,27).

These are not just theoretical concerns. In a summary of all the outcomes from all the placebo-controlled clinical trials included in a new drug application, LOCF yielded a smaller *P* value than MMRM in 34% of the 202 comparisons (13).

When LOCF underestimates the superiority of the superior treatment, it necessarily underestimates the inferiority of the inferior treatment. Thus, such a bias would be anticonservative in noninferiority testing and in a superiority test wherein the test agent is inferior to the control. For safety outcomes, underestimating the magnitude of a treatment effect is certainly not conservative.

The magnitude and direction of the bias from LOCF depends on the average attenuation from flattening out the mean profile in the treatment group compared with the control group. This in turn depends on the rate and timing of the missing data and the rate of change in the trajectory that is being attenuated. It has been shown that the bias from LOCF may involve many factors and can be complex (5). Whether or not the bias from LOCF leads to conservative estimates of treatment benefits is yet another question that further depends on the disease state and scenario. For example, the same direction of bias might be conservative if, on average, patients improve, but would be anticonservative if the treatment goal were the delay of worsening or maintenance of effect (9,12). Hence, it is difficult to anticipate the effects of bias from LOCF in practical situations. However, the general tendencies of LOCF for scenarios in which the overall tendency is for improvement (progressive improvement) include the following:

1. Overestimation of a drug's advantage when dropout is higher or earlier in the comparator, and underestimation of its advantage when dropout is lower or later in the comparator
2. Overestimation of a drug's advantage when the advantage is maximum at intermediate time points, and underestimation of its advantage when the advantage increases over time
3. A greater effect of bias on inferences regarding existence of a treatment effect when a drug's advantage is small and when sample sizes are large

For scenarios in which the overall tendency is for worsening (progressive impairment), the biases in (1) and (2) are reversed.

Additionally, if a method yielded biased estimates of treatment effects when treatment differences truly existed, then when the true treatment difference was zero, bias would necessarily lead to nonzero estimates of treatment differences and inflation of type I error. Moreover, consider Alzheimer disease, wherein the therapeutic aim is to delay or slow deterioration of mental status (as compared to situations such as depression in which the goal is to improve the condition). If a treatment is in truth no more effective than placebo, but a patient drops out early in the treatment arm, carrying the last observation forward assumes that the patient had no further deterioration in condition.

A similar bias can occur in so-called maintenance studies. For example, in a weight maintenance study, patients who lose a substantial amount of body weight through nonpharmacological means begin drug therapy with the goal of maintaining the initial weight loss (33). Any patient who drops out early is likely to have regained little weight. Therefore, applying LOCF in this scenario assumes the patient maintained the weight loss despite having only observed that patient for a short time.

It is reasonable also to question whether conservative analytic approaches are in the best interest of patients. Of course inflation of type I error is never good, but is it necessary to have "extra protection" against type I error when it comes at the expense of losing power, that is, inflated type II error? Moreover, independent of missing-data concerns, using a method that includes only the first and last observation is inherently inefficient. Given the unmet medical needs and the rising costs of health care, the need for new medicines, better medicines, and more affordable medicines is clear. Many factors may influence the success of drug development; however, the reliance on a method such as LOCF

(or BOCF) with known inflation of type I and type II error is an obvious suspect. Therefore, trading reduced power for more than the needed protection against type I error is not in the public interest.

Rather than defining conservatism as underestimating the magnitude of the treatment effect, what if conservatism were defined as being dependable? In the context of a statistical analysis, this might be defined as yielding the type I and type II error rates that were expected. With this dependability, clinical development of individual drugs could be more predictable, while preserving public safety against ineffective drugs. So, by this definition, MMRM is clearly more conservative than LOCF.

EFFECT OF MNAR DATA

As we have previously noted, MMRM assumes data are MAR, which is often a reasonable assumption in longitudinal clinical trials and always at least as plausible as MCAR. However, the possibility of MNAR data can never be ruled out. Therefore, it is not surprising that when MMRM has been proposed as the primary analysis, regulators have asked about the impact of MNAR data on the MMRM results. The more relevant question is, what is the impact of MNAR data on MMRM compared with their impact on LOCF? Research on the comparative impact of MNAR data on MMRM and LOCF is summarized in this section.

Mallinckrodt and colleagues (7,8,11) compared MMRM with LOCF ANOVA in a series of simulation studies with MNAR missingness. The first study included scenarios in which there was a true difference between treatments in mean change from baseline to endpoint. The second study focused on type I error rates by simulating scenarios in which the difference between treatments in mean change from baseline to endpoint was zero. In both studies, comparisons were made in data before introducing missingness (complete data) and in the same data sets after eliminating data via an MNAR mechanism. In analyses of complete data, MMRM and LOCF yielded identical results. Estimates of treatment effects were not biased, and standard errors accurately reflected the uncertainty in the data; however, there were important differences in results between the methods in analyses of incomplete data.

In the study in which there were treatment differences at endpoint, the MMRM estimates were closer to the true value than estimates from LOCF in every scenario simulated. Standard errors from MMRM accurately reflected the uncertainty of the estimates, whereas standard errors from LOCF underestimated uncertainty. Pooled across all scenarios, confidence interval coverage (percentage of confidence intervals containing the true value) was 94% and 87% for MMRM and LOCF, respectively, compared with the expected coverage rate of 95%. Notably, LOCF overestimated the treatment effect in some scenarios, typically when there was higher dropout in the inferior (eg, placebo) group.

In the type I error rate study, pooled across all scenarios with missing data, the type I error rates for MMRM and LOCF were 5.9% and 10.4%, respectively, compared with the expected rate of 5%. Type I error rates in the 32 scenarios ranged from 5.0% to 7.2% for MMRM and from 4.4% to 36% for LOCF.

The third study (11) included a factorial arrangement of scenarios, with four patterns of mean change over time and three true correlation structures (autoregressive, compound symmetry, and unstructured). The mean change patterns included two scenarios in which the null hypothesis of no difference between treatments in mean change from baseline to endpoint was true and another two scenarios in which it was false. Data from each scenario were analyzed using MMRM with each correlation structure and with LOCF. The intent in using these correlation structures was not to advocate their use, but to use very different structures to assess how MMRM would compare with LOCF under extreme conditions of misfitting the correlation structure.

In most cases, the type I error rates from LOCF were greater than or equal to those from any of the corresponding MMRM analyses, indicating that even egregious misfitting of the correlation structure with MMRM was typically less delete-

rious than using LOCF. Importantly, the use of an unstructured covariance matrix in MMRM, regardless of the form of the true covariance matrix, yielded superior control of type I error compared with LOCF in every scenario investigated. In pooled results from scenarios under the true null hypothesis, the type I error rate from MMRM using an unstructured correlation matrix was 6.2% compared with 9.8% for LOCF.

When the true treatment difference was large and dropout rate was higher in the superior arm, MMRM with an unstructured covariance matrix produced an average estimated treatment difference of 12.6 compared with 9.1 from LOCF and the true value of 12. The average power from MMRM was 75% compared to 59% for LOCF and 81% for both methods in the complete data.

In contrast, when the true treatment difference was small and dropout rate was greater in the inferior arm, MMRM with an unstructured covariance matrix produced an average estimated treatment difference of 2.9 compared to 5.2 for LOCF and a true value of 4. The average power from MMRM was 10% compared with 22% from LOCF and 17% with both methods in complete data. This apparent increase in power with LOCF, despite an overall 35% dropout rate, was driven by the bias in its estimates of treatment effect.

Lane (3) conducted simulation studies based on six actual clinical trial data sets. For each trial, multiple sets of data were generated from multivariate normal distributions, with means and covariances set to the estimates obtained from the actual data. Observations were removed from the simulated data to give missing values in accordance with an MNAR mechanism using three different models for probability of dropout depending on the next observation (treated as unobserved). Both equal and differential dropout rates between treatments were investigated.

LOCF led to misinterpretation of results when dropout mechanism was not MCAR (MAR or MNAR), particularly in cases with differential dropout rates. In contrast, MMRM led to misinterpretation only in cases in which data were MNAR with substantial differential dropout. In the majority of comparisons under MNAR data, MMRM resulted in bias that was less than or equal to that obtained with LOCF. Of 63 comparisons, 42 resulted in at least 10% less bias with MMRM, 10 were about the same, and 11 showed more than LOCF. In 6 of the 11 cases in which LOCF was less biased, the treatment difference (on which percentage bias was based) was very small.

Additionally, the perturbations in power caused by MMRM tended to be less than those for LOCF and less subject to extreme differences from the nominal values. Use of MMRM rarely caused a difference in power greater than 20%, whereas use of LOCF caused such a difference in nearly half of the simulations conducted.

Across these studies, the magnitude of bias produced by MNAR data was smaller with MMRM than with LOCF, and MMRM provided more robust control of type I and II error rates than LOCF. Furthermore, in actual clinical trial data, MMRM yielded results similar to those of a selection model (MNAR) approach, and it was determined that MMRM was an appropriate primary analysis for these data (5,14).

DETERMINING AND DEFINING APPROPRIATE MODELING CHOICES FOR MMRM

Regulators have noted that using MMRM entails more explicit modeling choices than using LOCF. While true, the rather modest increase in complexity of MMRM has not been a hindrance in implementation.

When determining a suitable model for a study collecting longitudinal data, it is important to realize that no universally accepted "best" model can be prespecified for the data eventually obtained. However, the main characteristics of the data will be driven by the design of the study. And, to a large degree, an appropriate MMRM model follows logically from the design of the study and thus can be adequately prespecified.

Three important characteristics to consider when specifying a model for data from longitudinal clinical trials are the random effects, the

correlations between the repeated measurements (within-patient errors), and the time trends.

As noted in the earlier section "Terminology and Concepts Regarding Use of MMRM and LOCF in Regulatory Settings," the feature distinguishing MMRM from other mixed-effects analyses is the modeling of the random effects as part of the within-patient error correlation structure. Handling the random effects in this manner simplifies the analysis while having no (or very little) impact on inferences of treatment effects.

Clinical trials often have a common schedule of measurements for all patients, with a large number of patients and a relatively small number of measurement occasions. With such a data structure, MMRM can be implemented using a full multivariate model, featuring an unstructured modeling of time and correlation (10). If the number of patients relative to the number of measurement occasions is not large, more parsimonious approaches are easily implemented. For example, time trends could be modeled using linear and quadratic effects, and some structured form of the V matrix could be fit to the within-patient correlations.

However, the functional form of the longitudinal trends can be difficult to anticipate, and in particular, linear time trends may not adequately describe the response profiles. A parsimonious model using a structured form of the time trends could be more powerful than an unstructured model, but it could also be a poor fit. Therefore in many scenarios, an unstructured modeling of time and the treatment-by-time interaction provides an assumption-free approach, does not require estimation of an inordinate number of parameters, and can be depended upon to yield a useful result—attributes well suited to the primary analysis.

It also is worth noting that an MMRM analysis using the full multivariate approach (unstructured modeling of time, treatment-by-time, and within-patient errors) for analyses of complete data (no missing observations) yields the same inference about the endpoint as an analysis of that endpoint by itself. That is, with fully unstructured treatment-by-time effects (and within-patient errors), MMRM and LOCF yield identical treatment contrasts if no data are missing.

An unstructured modeling of within-patient correlations also removes one layer of assumptions and often provides the best fit to the data. However, overly general correlation structures can lead to an analysis that fails to converge. Although failure to converge often results from improperly preparing the data (eg, two observations on the same patient at the same time point, or poor choice of options in software), a priori specification of the primary analysis must have flexibility to allow alternative models to be fit if an analysis fails to converge because the prespecified correlation structure is too general.

Several approaches can be taken to ensure convergence. First, every attempt should be made to ensure convergence is obtained from a given correlation structure. For example, convergence can be enhanced by using software features such as the inputting of starting values for parameter estimates, or the use in the initial rounds (but not final rounds) of iteration algorithms such as Fisher's scoring rather than the Newton-Raphson algorithm, which is the default algorithm in many software packages.

Rescaling the data is also an option. If outcomes and covariates are made to fall in ranges in the order of magnitude of unity, interpretations and conclusions will not be changed; but avoiding manipulation of large or small numbers from a numerical analysis perspective reduces the risk of ill-conditioned matrices, and ultimately, overflow or underflow. In addition, the protocol can envision one of several model-fitting approaches. One could simply specify a set of structures to be fit, and use as the primary analysis the one yielding the best fit as assessed by standard model-fitting criteria. However, if one does not want to build models from the same data from which hypotheses are to be tested, a series of structures could be specified in a fixed sequence, and the first correlation structure to yield convergence would be considered the primary analysis. For example, *unstructured* could be specified as the structure for the pri-

mary analysis; but if it failed to converge, a series of ever-more parsimonious structures appropriate to the situation at hand could be fit until one converges, which would then be considered the primary analysis.

Although these approaches have always yielded a converged analysis in our experience, it is reasonable to wonder what effect the true correlation structure in the data, and the method of modeling the correlation structure, have on results. One study (11) assessed the effect of correlation structure and how it is modeled on type I error rates and power, and compared results from MMRM with LOCF. Results of this study are detailed in the earlier section, "Effect of MNAR Data." When the correct correlation structure was fit, MMRM provided better control of type I error and power than LOCF. Although misfitting the correlation structure in MMRM inflated type I error and altered power, even egregious misfitting of the structure was typically less deleterious than using LOCF. In fact, simply using an unstructured model in MMRM yielded superior control of type I error than LOCF in every scenario tested.

Therefore, MMRM provides flexibility for modeling the within-patient correlation structure, does so in a manner that can be specified a priori, ensures that analysts following those specifications will independently arrive at exactly the same result, and even in worst-case scenarios provides estimates of treatment effects with less bias than LOCF—all attributes that are well suited to the primary analysis.

Another aspect of time trends that must be considered is in relation to the covariates represented by β in the model (Eq. 1). The treatment effect is clearly the most crucial in pharmaceutical clinical trials, and the interaction of this effect with time has been discussed previously. Other covariates are often also included in the model, and these must also be considered. For example, the effect of a baseline observation may be included, as subjects' responses may be considered dependent on their condition at the start of the trial. In this case, too, it is usually preferable to allow a full interaction of the covariate with time, for if not, a restriction is being imposed that the dependence of response on the baseline measure is the same at all time points. Alternatively, both the baseline and postbaseline measures can be treated as response variables under the assumption that the baseline means are the same across treatment groups to reflect randomization (34). For other covariates, such as age and gender, it may be considered appropriate to include no interaction with time because the effects can be taken to be constant; but such decisions need to be taken and explained at the stage of planning the analysis.

The following example illustrates one way to specify a priori all the details of an MMRM analysis such that independent analysts will arrive at exactly the same results. This particular wording specifies the full multivariate approach, with an unstructured modeling of treatment effects over time and within-patient error correlations.

> Mean changes from baseline will be analyzed using a restricted maximum likelihood (REML)-based repeated measures approach. Analyses will include the fixed, categorical effects of treatment, investigative site, visit, and treatment-by-visit interaction, as well as the continuous, fixed covariates of baseline score and baseline score-by-visit interaction. An unstructured (co)variance structure will be used to model the within-patient errors. If this analysis fails to converge, the following structures will be tested: (insert a list of structures appropriate for the specific application). The (co)variance structure converging to the best fit, as determined by Akaike's information criterion, will be used as the primary analysis. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom. Significance tests will be based on least-squares means using a two-sided $\alpha = .05$ (two-sided 95% confidence intervals). Analyses will be implemented using (insert software package). The primary treatment comparisons will be the contrast between treatments at the endpoint visit.

Note that the primary analysis could be based on contrasts at time points other than endpoint, or could be based on the treatment main effects.

LOCF JUSTIFIED AS A FACTUAL, COMPOSITE, OR EFFECTIVENESS ENDPOINT

Literally taken, the acronym LOCF implies imputation of missing values in a longitudinal context. An alternative interpretation of LOCF is commonly used that might be better termed LO (last observation) or LAV (last available value). In this approach, results are not interpreted as imputations of missing data with changes assessed at a specific time point, but rather as the change that was actually seen at last observation regardless of when it was observed (9). When LOCF is used in this manner, it is said to estimate a factual outcome in that it estimates what was actually (factually) observed at the last assessment, regardless of when that observation was made. In this same context, MMRM is said to be estimating a counterfactual outcome in that it estimates the effect that would have been observed had patients stayed in the trial, contrary to the fact that some patients dropped out (27).

The use of LOCF in the factual context stems from its intuitive appeal as a pragmatic measure of effectiveness, a composite of efficacy, safety, and tolerability (9,12,27). However, the fact that it is easy to understand and calculate an LOCF value should not be confused with its yielding a meaningful measure. If one were to objectively seek a factual or all-encompassing assessment, it seems unlikely that one would arrive at LOCF.

First, the primary purpose of confirmatory clinical trials is typically to delineate causal differences between drug and placebo (or between drugs), not to mimic actual clinical practice. It is unreasonable to assume that doctors and patients make the same decisions regarding continuation of therapy in a double-blind trial—in which they are unsure about whether the patient is taking drug or placebo—as they would make in actual practice, when the drug and its properties are well known. Therefore, the rates and reasons for dropout within the strictly controlled conditions of a confirmatory clinical trial are unlikely to mimic what would happen in general use. If effectiveness were the primary objective, the best place to assess it would be in a general medical (ie, naturalistic) setting. When causal effects are the primary objective, the gold standard design is a double-blind, randomized clinical trial. Hence, using LOCF in a factual context is inconsistent with the design and primary objective of confirmatory clinical trials.

Furthermore, the rate and timing of dropout does not necessarily reflect the true benefit and risk of the drug. While LOCF can in some situations yield smaller estimates of treatment differences when patients drop out due to adverse events, the reduction is not necessarily proportional to the safety risk (9). For example, consider the following two patients in an 8-week trial: patient A dropped out after week 7 because of a dramatically prolonged QT interval; patient B dropped out during week 1 with nausea. The impact on estimates of mean change resulting from patient A's dropout was small because the last observation was close to the trial's endpoint, whereas the impact from patient B's dropout was severe because (in many disease states) little improvement results from one week of treatment. However, patient A developed a potentially life-threatening condition, whereas the nausea experienced by patient B early in the trial is typically transitory and often resolves with continued therapy and no long-term consequences. This nonproportional penalty to individual patients from LOCF may cause misleading inferences regarding the merits of a treatment.

As an even more extreme, but common, example, consider the Alzheimer disease scenario noted in the earlier section, "LOCF Is Conservative," in which the therapeutic aim is to delay or slow deterioration of mental status. Using the last observation from a patient who dropped out early from the treatment arm due to an adverse event (AE) would actually reward the drug for the AE, as the patient would appear to have had no further deterioration in condition.

A related point is that an LOCF result used in this manner does not correspond to a population parameter that can be prespecified. It is essentially a composite whose components (efficacy, safety, tolerability) have unknown, or at least random, weights. Hence, using LOCF in a hypothesis-testing setting violates the funda-

mental approach of statistics wherein we attempt to make inference about population parameters. If a composite measure of efficacy, safety, and tolerability is of primary interest, then it would be better to have a prespecified measure that can capture these facets uniformly for each patient, such as an a priori–defined clinical utility index.

It is sobering to recognize that an LOCF result may be manipulated by design factors and the behavior of investigative site staff who encourage prolonged participation, possibly making the drug look better, but of course not altering the inherent risk-benefit of the drug. This may take place, for example, when an extension period is added to the randomized part of a clinical trial. Minimizing dropout is widely accepted as good scientific practice. However, the concern here is that the amount of dropout should not directly change the measure of interest, parameter being estimated, or the hypothesis that is being tested.

Therefore, while it is true that MMRM estimates a hypothetical parameter in that not all patients stay on medication to the specific time points at which mean changes are estimated, the use of LOCF in the composite or factual context is also fraught with many problems. Importantly, the hypothetical nature of the adjusted means from MMRM is not in practice a hindrance to interpretation. One can take the efficacy results from MMRM and combine them with the various safety and tolerability results in an ad hoc manner, as has traditionally been done, or in a formal clinical utility index to assess the overall benefit-risk of the drug.

In fact, rather than viewing the hypotheses tested by LOCF and MMRM as factual and counterfactual, one might view the hypothesis tested by MMRM as what is expected when patients take the drug as directed, whereas LOCF tests what is expected when the drug is taken as observed. Both are useful; the key is to match the hypothesis with the stage of development and design of clinical trial. The hypothesis tested by MMRM is aligned with confirmatory clinical trials utilizing double-blind, randomized designs, whereas the hypothesis tested by LOCF is best evaluated in naturalistic settings.

It is also important to recognize that an endpoint analysis of any type is able to provide only a small part of the overall picture, and that the entire longitudinal treatment profile should be considered in order to address such questions as "How soon until I feel better?" or "How soon until I feel well?" Longitudinal methods such as MMRM are ideally suited to provide such information from the same analysis as that which produces the endpoint contrast.

HANDLING NONIGNORABLE MISSINGNESS (MNAR)

Although the assumption of MAR is often reasonable in clinical trials, the possibility of data missing not at random (MNAR) is difficult to rule out. Therefore, analyses valid under MNAR are needed. Analyses in the MNAR framework try in some manner to model or otherwise take into account the missingness. Although reasons for (early) discontinuation are routinely collected in clinical trials, they may not reveal much about the missing-data mechanism, and modeling or incorporating information about the missingness into the data analysis may not be straightforward.

The obvious but fundamental problem is that we do not have the missing data, so we cannot definitively know its characteristics; we can only make assumptions. Conclusions from MNAR analyses are therefore conditional on the appropriateness of the assumed model. While dependence on assumptions is not unique to MNAR analyses, a unique feature with MNAR analyses is that (some of) the assumptions are not testable (35) because we do not have the missing data about which the assumptions are made (36).

Importantly, the consequences of model misspecification are more severe with MNAR methods than with other (eg, MAR) methods (19,36–49). Hence, no individual MNAR analysis can be considered definitive. Not surprisingly then, many statistical methodologies have been proposed to analyze data in the MNAR setting.

General classes of MNAR methods have arisen from different factorizations of the likelihood functions for the joint distribution of the out-

come variable and the indicator variable for whether or not a data point is observed. Factorization in this context means that the hypothetical "full" data are split into two parts: the actually observed part and the missing part, which are often described as the measurement process and the missingness process, respectively.

The selection model framework (16,18,41) describes the full data likelihood as the product of the marginal density of the measurement process and the density of the missingness process conditional on the outcomes. Conceptually, a selection model as typically implemented can be thought of as a multivariate analysis. The first outcome variable is the same outcome being analyzed as in an MAR analysis, typically a mean change analysis. The second variable is the indicator variable for dropout, often analyzed via logistic regression. Selection models have been formulated in parametric (41) and semiparametric (42) frameworks.

Pattern-mixture models (43,44) are based on factorization of the full data likelihood as the product of the measurement process conditional on the dropout pattern and the marginal density of the missingness process. Conceptually, pattern mixture models as typically implemented assess the outcome variable separately for different groups (patterns), often defined by time of dropout, and then combine results across groups for final inference.

A third approach, the shared-parameter model (19,45–49), is similar to selection models in that it jointly models the measurement and dropout processes. Shared-parameter models assume that a certain parameter, typically a random effect, influences both the outcome variable and dropout, such that conditional upon this parameter, the measurement and dropout processes are independent.

The conceptual similarity between these different approaches is that they go beyond ignorability by adding something to the analysis to account for the dropout. Another strategy is to add ancillary variables to the analysis of the outcome variable of interest in order to explain the dropout. The basic idea is that data are MAR if, conditional upon the variables in the model, missingness does not depend on the unobserved outcomes of the variable being analyzed. Therefore, if additional (ancillary) variables are added to the model that helps explain missingness, MAR can be valid; whereas if the additional variables were not included, the data would be MNAR.

Collins and colleagues (50) state that multiple imputation (MI), originally proposed by Little and Rubin (16) as an MAR method, is well suited to improving the performance of the missing-data procedure through the use of ancillary variables. They also note that ancillary variables can be included in likelihood-based analyses (such as MMRM). This could be done by adding the ancillary variable either as a covariate or as an additional response to create a multivariate analysis. However, the complexity of multivariate analyses and the features of most of the commercially available software make it easier to use ancillary variables via multiple imputation. Liu and Gould (6) and Lipkovich et al. (51) provided implementations of MI with ancillary variables (including AE information) in clinical trial contexts.

In addition, MI has the added advantage that with separate steps for imputation and analysis, ancillary variables that are postbaseline, time-varying covariates—possibly influenced by treatment—can be included in the imputation step to account for missingness but then not included in the analysis step to avoid confounding with the treatment effects, as might be the case in a likelihood-based analysis.

Although methods to test for the existence and impact of outlier (influential) observations have been around for decades, new methods have been developed for use in MNAR analyses. To this end, interest has grown in local influence approaches (14,52–57), which are often associated with selection models. Local influence provides an objective approach to identifying and examining the impact of influential data points and patients on various aspects of the analysis, including the missing-data mechanisms and treatment effects. Shen and colleagues (14) provide a case study of a longitudinal depression trial showing how local influence can be helpful in conducting sensitivity analysis.

This is by no means an exhaustive list of all the available methods for analyses under MNAR, but rather a brief overview of the fundamental underpinnings that fostered development of many of the methods. See, for example, the study by Ibrahim and colleagues (58) for comparisons of common approaches for ignorable and nonignorable missing-data mechanisms, including maximum likelihood, multiple imputation, fully Bayesian, and semiparametric weighted estimation equations.

Developing an appropriate strategy for analysis under MNAR begins with recognizing that these methods are heavily assumption driven and that the assumptions are not testable. Therefore, no single MNAR approach can be considered definitive. Consequently, a useful and common approach is to fit several MNAR models or methods utilizing different assumptions regarding the data distribution and missingness within a sensitivity analysis framework, thereby allowing assessment of robustness of results to the various assumptions.

RECOMMENDATIONS

Having discussed theoretical and practical considerations, we now turn to specific recommendations.

Our first recommendation is a natural consequence of the inability of any statistical analysis to recoup the loss of information due to missing data. Therefore, whatever methods are to be employed in analysis, they should not detract from efforts to plan a trial that minimizes dropout. In addition, detailed records of the reasons for missing data and data on potential covariates that might help further characterize the probability of dropout should be obtained.

Regarding the primary analysis for confirmatory longitudinal clinical trials, conclusive evidence has demonstrated the need to abandon the simple, ad hoc methods such as LOCF and BOCF. Given that the possibility of MNAR data can never be ruled out, one might be tempted to shift the primary analytic approach to that of an MNAR method. However, MNAR methods are sensitive to untestable assumptions. Also, from a practical standpoint, many MNAR methods require customized programs, may suffer numerical convergence problems, or may be complicated by weakly identified or underidentified models. Therefore, MNAR methods are not well suited for the primary analyses in confirmatory clinical trials wherein a dependable, prespecified method is needed. We conclude, as have others (5,14), that the proper framework for use of MNAR approaches for confirmatory clinical trials is that of sensitivity analyses.

MAR is the most appropriate framework for the primary analysis in confirmatory trials because this assumption is often reasonable and certainly more plausible than MCAR. Use of MAR is further supported in that the consequences of departures from MAR can be evaluated via sensitivity analyses, and MAR methods are often robust to departures from MAR.

Likelihood-based repeated measures approaches, such as MMRM, provide a flexible framework under the MAR assumption from which analyses can be tailored to the specific situation at hand. Flexibility in modeling treatment effects over time and the within-patient error correlation structure are particularly useful in this regard, making MMRM a widely useful analysis in drug development. Specifically, MMRM is an appropriate choice for the primary analysis in many longitudinal confirmatory clinical trials, especially those scenarios in which LOCF has been an acceptable primary analysis in the past. The historical precedent for LOCF makes it a likely choice to include as a sensitivity analysis. However, MNAR-based analyses should be the primary basis upon which sensitivity is assessed because MNAR analyses focus on the assumption key to validity of MMRM, whereas discrepancies between an LOCF and MMRM result could arise for many reasons unrelated to the validity of MMRM.

Our specific recommendation of MMRM for an MAR-based primary analysis needs to be considered in light of the mission of our working group. Our aim was to (a) clarify terminology and concepts regarding use of MMRM and LOCF in regulatory settings, (b) address specific concerns raised by regulatory agencies regarding use of MMRM as the primary analysis, and

(c) make specific recommendations for analysis of data from longitudinal clinical trials.

This degree of focus has the advantage of facilitating a detailed and thorough discussion. However, it has the limitation of not fully considering other analyses. Some readers will wonder why we did not include in our comparisons with LOCF Bayesian analyses, multiple imputation, or weighted generalized estimating equations. All of these methods are intrinsically similar in their underlying assumptions (MAR) and have extensive literature supporting their application. As such, these methods ought to borrow strength from each other, rather than engage in mutual competition.

Our focus on MMRM was driven by extensive experience with this method in the specific situation of relevance—confirmatory clinical trials. The other MAR approaches have not been studied and used as extensively as MMRM in this regard. Therefore, we focused on the area wherein practical experience was greatest so that our recommendations could be implemented immediately and with minimal ambiguity.

Acknowledgments—We benefited from review of the draft article by three prominent academics in the field of missing data and longitudinal analyses, and we thank them for their thoughtful comments and suggestions: Rod Little (University of Michigan School of Public Health), Geert Molenberghs (Hasselt University Centre for Statistics), and Daniel Scharfstein (Johns Hopkins Bloomberg School of Public Health). We are also grateful for review and advice from several colleagues in the pharmaceutical industry: Bruce Binkowitz (Merck), Argyha Chattopadhyay (Johnson & Johnson), David Keller (Pfizer), Frank Liu (Merck), Kaifeng Lu (Merck), Edmund Luo (Merck), Akiko Okamoto (Johnson & Johnson), and James Roger (GSK). Although we incorporated many of the suggestions made by these reviewers, the individuals mentioned above were not asked to specifically endorse the recommendations made in this article.

REFERENCES

1. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Chichester: John Wiley; 2007.
2. Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trials: beyond LOCF. *Obes Rev*. 2003;4:175–184.
3. Lane PW. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharm Stat. (early view)* 2007;DOI: 10.1002/pst.267.
4. Leon AC, Mallinckrodt CH, Chuang-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology. *Biol Psychiatry*. 2006;59:1001–1005.
5. Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004;5:445–464.
6. Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J Biopharm Stat*. 2002;12:207–226.
7. Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. *J Biopharm Stat*. 2001;11(1–2):9–21.
8. Mallinckrodt CH, Clark WS, David SR. 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.
9. Mallinckrodt CH, Sanger TM, Dube S, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol Psychiatry*. 2003;53:754–760.
10. Mallinckrodt CH, Clark SW, Carroll RJ, Molenberghs G. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J Biopharm Stat*. 2003; 13:179–190.
11. Mallinckrodt CH, Kaiser CJ, Watkin JG, Molenberghs G, Carroll RJ. The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. *Clin Trials*. 2004;1:477–489.
12. Mallinckrodt CH, Kaiser CJ, Watkin JG, Detke MJ, Molenberghs G, Carroll RJ. Type I error rates from likelihood-based repeated measures analyses of incomplete longitudinal data. *Pharm Stat*. 2004;3:171–186.
13. Mallinckrodt CH, Raskin J, Wohlreich MM, Watkin JG, Detke MJ. The efficacy of duloxetine: a comprehensive summary of results from MMRM and LOCF_ANOVA in eight clinical trials. *BMC Psychiatry*. 2004;4:26.

14. Shen S, Beunckens C, Mallinckrodt C, Molenberghs G. A local influence sensitivity analysis for incomplete longitudinal depression data. *J Biopharm Stat*. 2006;16:365–384.
15. Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med*. 1997;16:2349–2380.
16. Little R, Rubin D. *Statistical Analysis With Missing Data*. New York: John Wiley; 1987.
17. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer; 2000.
18. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–592.
19. Little RJA. Modeling the drop-out mechanism in repeated measures studies. *J Am Stat Assoc*. 1995;90:1112–1121.
20. Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? *Neuropsychopharmacology*. 1992;6:39–48.
21. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963–974.
22. Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*. 1996; 52:1324–1333.
23. Gibbons RD, Hedeker D, Elkin I, et al. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Application to the NIMH Treatment of Depression Collaborative Research Program dataset. *Arch Gen Psychiatry*. 1993;50:739–750.
24. Heyting A, Tolboom JT, Essers JG. Statistical handling of drop-outs in longitudinal clinical trials. *Stat Med*. 1992;11:2043–2061.
25. Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. *Stat Med*. 1995;14:1913–1925.
26. Siddiqui O, Ali MW. A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *J Biopharm Stat*. 1998;8:545–563.
27. Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Stat Med*. 2003;22: 2429–2441.
28. Carpenter J, Kenward M, Evans S, White I. Last observation carry-forward and last observation analysis. Letter to the Editor. *Stat Med*. 2004;23: 3241–3244.
29. Cook RJ, Zeng L, Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*. 2004;60: 820–828.
30. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. New York: Springer; 2005.
31. Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clin Trials*. 2005;2:379–386.
32. Rubin DB, Stern HS, Vehovar V. Handling "don't know" survey responses: the case of the Slovenian plebiscite. *J Am Stat Assoc*. 1995;90:822–828.
33. Hill JO, Hauptman J, Anderson JW, Fujioka K, O'Neil PM, Smith DK, Zavoral JH, Aronne LJ. Orlistat, a lipase inhibitor, for weight maintenance after conventional dieting: a 1-yr study. *Am J Clin Nutr*. 1999;69:1108–1116.
34. Liang K, Zeger S. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhya: Indian J Stat*. 2000;62(Series B):134–148.
35. Molenberghs G, Kenward MG, Lesaffre E. The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*. 1997;84:33–44.
36. Laird NM. Discussion to Diggle PJ, Kenward MG. Informative dropout in longitudinal data analysis. *Appl Stat*. 1994;43:84.
37. Rubin DB. Discussion to Diggle PJ, Kenward MG. Informative dropout in longitudinal data analysis. *Appl Stat*. 1994;43:80–82.
38. Copas JB, Li HG. Inference for non-random samples (with discussion). *J Royal Stat Soc B*. 1997; 59:55–96.
39. Draper D. Assessment and propagation of model uncertainty (with discussion). *J Royal Stat Soc B*. 1995;57:45–97.
40. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Stat Med*. 1998;17:2723–2732.
41. Diggle PD, Kenward MG. Informative dropout in longitudinal data analysis (with discussion). *Appl Stat*. 1994;43:49–93.
42. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J Am Stat Assoc*. 1998;93:1321–1339.
43. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc*. 1993;88: 125–134.
44. Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika*. 1994;81: 471–483.

45. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988;44:175–188.
46. Ten Have TR, Kunselman AR, Pulkstenis EP, Landis JR. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics*. 1998;54:367–383.
47. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*. 1989;45:939–955.
48. Mori M, Woodworth GG, Woolson RF. Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring. *Stat Med*. 1992;11:621–631.
49. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*. 1995;51:151–168.
50. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6:330–351.
51. Lipkovich I, Duan Y, Ahmed S. Multiple imputation compared with restricted pseudo-likelihood and generalized estimating equations for analysis of binary repeated measures in clinical studies. *Pharm Stat*. 2005;4:267–285.
52. Zhu HT, Lee SY. Local influence for incomplete-data models. *J Royal Stat Soc B*. 2001;63:111–126.
53. Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics*. 2001;57:7–14.
54. Thijs H, Molenberghs G, Verbeke G. The milk protein trial: influence analysis of the dropout process. *Biometric J*. 2000;42:617–646.
55. Molenberghs G, Verbeke G, Thijs H, Lesaffre E, Kenward M. Mastitis in dairy cattle: local influence to assess sensitivity of the dropout process. *Comput Stat Data Anal*. 2001;37:93–113.
56. Troxel AB, Ma G, Heitjan DF. An index of local sensitivity to nonignorability. *Statistica Sinica*. 2004;14:1221–1237.
57. Ma G, Troxel AB, Heitjan DF. An index of local sensitivity to nonignorable drop-out in longitudinal modeling. *Stat Med*. 2005;24:2129–2150.
58. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. *J Am Stat Assoc*. 2005;100:332–346.

Craig Mallinckrodt has disclosed that he is a stock shareholder in Eli Lilly and Co. Peter W. Lane has disclosed that he is a stock shareholder in and has received grants/research support from GlaxoSmithKline. Dan Schnell has disclosed that he is a stock shareholder in Procter & Gamble Co. Yahong Peng and James P. Mancuso report no relationships to disclose.