Journal of Biopharmaceutical Statistics, 19: 227–246, 2009

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400802609797

Taylor & Francis
Taylor & Francis Group

MMRM VS. LOCF: A COMPREHENSIVE COMPARISON BASED ON SIMULATION STUDY AND 25 NDA DATASETS

Ohidul Siddiqui, H. M. James Hung, and Robert O'Neill
Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, USA

In recent years, the use of the last observation carried forward (LOCF) approach in imputing missing data in clinical trials has been greatly criticized, and several likelihood-based modeling approaches are proposed to analyze such incomplete data. One of the proposed likelihood-based methods is the Mixed-Effect Model Repeated Measure (MMRM) model. To compare the performance of LOCF and MMRM approaches in analyzing incomplete data, two extensive simulation studies are conducted, and the empirical bias and Type I error rates associated with estimators and tests of treatment effects under three missing data paradigms are evaluated. The simulation studies demonstrate that LOCF analysis can lead to substantial biases in estimators of treatment effects and can greatly inflate Type I error rates of the statistical tests, whereas MMRM analysis on the available data leads to estimators with comparatively small bias, and controls Type I error rates at a nominal level in the presence of missing completely at random (MCAR) or missing at random (MAR) and some possibility of missing not at random (MNAR) data. In a sensitivity analysis of 48 clinical trial datasets obtained from 25 New Drug Applications (NDA) submissions of neurological and psychiatric drug products, MMRM analysis appears to be a superior approach in controlling Type I error rates and minimizing biases, as compared to LOCF ANCOVA analysis. In the exploratory analyses of the datasets, no clear evidence of the presence of MNAR missingness is found.

Key Words: Ignorable missing data; Last observation carried forward; Missing at random; Missing completely at random; Missing not at random.

1. INTRODUCTION

In longitudinal clinical trials of chronic therapies, patients are treated over a period of time and evaluated periodically at a number of time points. That is, for each patient a series of observations is available. In most clinical trials, the treatment period and the number of scheduled visits for efficacy evaluation are predetermined by design. For example, a trial might be designed where each patient would be treated for four weeks with evaluations at baseline and at the end of each

Received October 19, 2007; Accepted May 30, 2008
Address correspondence to Ohidul I. Siddiqui, Food and Drug Administration, 10903 New Hampshire Av., Room 4242, Silver Spring, MD 20993, USA; E-mail: ohidul.siddiqui@fda.hhs.gov

subsequent week. Hence, for each patient completing the entire treatment period, five observations would be available.

Although patients are evaluated at a number of time points, it is a customary, in many longitudinal trials, to define a single "primary end time point" at which efficacy of an experimental drug would be evaluated, for instance, with respect to placebo. Again, in most trials, the primary end time point is taken as the last time point of the predetermined double-blind randomized treatment period. When a patient drops out from the trial before completing the predetermined treatment period, his or her observation will be missing at each subsequent visit. This type of missing data is called monotonic missing data. In contrast, nonmonotonic missing data are also seen in clinical trials due to some patients missing some visits. For example, a patient may miss a visit, but at later visits the patient is available in the trial. In clinical trials, the presence of such nonmonotonic missing data is very minimal. In this manuscript, we focus on monotonic missing data due to patient drop out and refer to this as "missing data" hereafter.

Missing data are commonly grouped into three missing data mechanisms based on reasons why patients drop out. According to Rubin (1976) and Little and Rubin (2002), data are considered *missing completely at random* (MCAR) if, conditional upon the independent variables in the analytic model, missingness is independent of both unobserved and observed outcomes of the variable being analyzed; data are *missing at random* (MAR) if, conditional upon the independent variables in the analytic model, missingness depends on the observed data of the variable being analyzed but is independent of the unobserved outcomes of the variable being analyzed; 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.

Rubin (1976) and Little and Rubin (2002) also define two general classes of missingness mechanisms with respect to likelihood-based analysis. A missingness mechanism is called "ignorable" if a likelihood-based analysis provides valid inferences of the model parameters even when the missingness mechanism is ignored; otherwise, it is called "nonignorable" missingness mechanism. Laird (1988) shows that MCAR and MAR are ignorable missingness, and likelihood-based analyses that ignore the missing data mechanism remain valid. For nonignorable missing data, however, likelihood-based analyses that ignore the missing data mechanism potentially produce biased results.

An intuitive way to explain missingness mechanisms in longitudinal clinical trials is to look at the reasons for dropouts. Heyting et al. (1992) give six common reasons why patients withdraw from clinical longitudinal studies: (i) recovery, (ii) lack of improvement, (iii) treatment-related side-effects, (iv) unpleasant study procedures, (v) intercurrent health problems, and (vi) external factors unrelated to the trial. An informal debate is often encountered in grouping the observed possible drop-out reasons into the three missing data mechanisms defined by Rubin (1976) and Little and Rubin (2002). So far, no consensus is achieved on classifying the reasons for dropouts into the three defined missing data mechanisms or no prospective potential strategies to do so. The lack of consensus might be due to the fact that the three mechanisms are defined based on the outcome measures of interest, and could thus change from one outcome measure to another outcome measure in the same dataset. Consensus also depends on the independent variables

in the model. Therefore, the statistical analysis of incomplete data is solely carried out under the three missing data mechanisms as defined above. Another informal debate of whether or not a particular missing data mechanism, or a mixture of the three missing data mechanisms is present in a clinical trial dataset, often takes place. Moreover, there is no statistical test available to test the presence of nonignorable data. It may be possible to have a mixture of ignorable and nonignorable missing data with certain proportions in the same dataset. However, most of the previous research work on dealing with missing data focused on only one mechanism at a time.

Although some statistical methods (e.g., pattern-mixture models, selection models, etc.) are available to analyze longitudinal data under a nonignorable missing data mechanism, in reality it is not feasible to use such methods in analyzing clinical trial datasets for efficacy evaluation. In fact, there is no clear guidance as to which of the nonignorable methods are preferred under what scenarios. A definitive statistical method does not exist to analyze nonignorable missing data. Therefore, likelihood-based ignorable analyses need to be included in clinical trial data analysis instead of ad hoc methods. Thus, it is important to find out how robust the statistical findings are from ignorable models in analyzing incomplete data that have a mixture of the three missing data mechanisms through simulation studies. In real data analysis, one can explore the impact of deviations from the ignorable missing data assumption on the conclusions based on a sensitivity analysis using models under nonignorable missing data assumption. Several researchers, including Diggle and Kenward (1994) and Rotnitzky et al. (1998), propose different parametric and semiparametric models under a nonignorable missing data framework.

Fitzmaurice et al. (2004) present a brief review of the most often used statistical approaches to handle missing data in longitudinal analysis. Among the approaches, the last value carried forward (LVCF) is widely used for imputing missing data in longitudinal clinical trials, and it is commonly referred to as "last observation carried forward" (LOCF) approach. The LOCF approach is simple, but it makes two strong assumptions that: (i) missing data due to dropouts follow MCAR and (ii) the responses following a patient dropping out remain constant at the last observed value prior to drop out. Both of the assumptions are often unrealistic in clinical trials. Several researchers, including Fitzmaurice et al. (2004), Carpenter et al. (2004), Molenberghs et al. (2004), and Cook et al. (2004) have criticized the use of LOCF approach in imputing missing data due to drop out in clinical trials. Molenberghs et al. (2004) point out that LOCF endpoint analysis typically produces bias, of which the direction and magnitude depend on the true but unknown treatment effects, and the approach is not valid even under MCAR data mechanism. Recently, Shao and Zhong (2003) proposed a LOCF one-way ANCOVA test by averaging the means of efficacy measures of the last observations of all the possible subpopulations, each consisting of patients who dropped out at the same visit time. They claim the test is asymptotically valid where only two treatments are compared and two treatment groups have the same number of patients, regardless of whether the dropout is informative or not. The hypothesis they are testing is not the study endpoint hypothesis of interest in the drug approval process. Several researchers including Carpenter et al. (2004) and Mallinckrodt et al. (2003) have also noted the shortcomings of the hypothesis in clinical terms.

Several authors, including Laird and Ware (1982), propose likelihood-based mixed-effects models to analyze incomplete data from longitudinal clinical trials. Siddiqui and Ali (1998) perform a direct comparison of the likelihood-based mixed-effect regression model analysis with the LOCF analysis on data from a real psychiatric clinical trial. Under an ignorable missing mechanism assumption, likelihood-based models provide likelihood functions for the observed data (i.e., without modeling the dropout process) from which valid inferences on treatment effects and other parameters can be obtained. No explicit imputation of the missing data is required. In general, when dropouts are ignorable, the parameters of dropout and outcome processes are assumed to be distinct, and hence likelihood-based methods can be used on the marginal distribution of the observed data for statistical inferences. One such mixed model is named Mixed-effects Model Repeated Measures (MMRM) analysis by Mallinckrodt et al. (2001). The MMRM analysis is a particular form of a mixed model analysis and is fitted within the ignorable likelihood paradigm. Under a MMRM model the ML estimates can be obtained by maximizing $f(Y_i^O \mid X_i)$, where $f(Y_i^O \mid X_i)$ denotes the ordinary marginal distribution of the particular subset of Y_i determined by Y_i^O, and the missing data are predicted by the observed data via the model for the conditional mean, $E(Y_i^M \mid Y_i^O)$.

A presence of nonignorable missing data in clinical trials is difficult to rule out and is unverifiable with any empirical data. Therefore, when ignorable likelihood-based methods are considered for primary analysis purposes, it is important to explore the impact of the ignorable missing data assumption on the conclusions based on sensitivity analyses using MNAR models. A parametric model of the type of Diggle and Kenward (1994), or a semiparametric approach such as in Rotnitzky et al. (1998), might be used as MNAR models for sensitivity analyses. Molenberghs and Kenward (2007) give a detailed discussion on the importance of sensitivity analyses of incomplete clinical trial datasets within the MNAR modeling framework.

Little (1995) describes the MNAR models in terms of two broad model classes: selection model and pattern-mixture (PM) model. Several authors, including Little (1995), Hogan and Laird (1997), and Michiels et al. (2002), discuss the differences between the selection models and PM models. The basic difference is that selection models depend on distributional assumptions of the missing data that are unverifiable with the observed data (Little, 1995; Little and Rubin, 2002), whereas PM models can be specified without any requirement of missing data mechanism to be ignorable. As an alternative to selection models, a general class of random-effect PM models formulated by Little (1993, 1994, 1995) is often used to analyze incomplete clinical trial data. In the PM models, subjects are divided into groups based on the dropout patterns (e.g., early dropouts, late dropouts, and completers). Then these groups are used to examine the effect of the dropout pattern on the outcome measure of interest. The overall estimate is obtained by averaging the estimates across the dropout patterns. The random-effects PM model is often used in sensitivity analysis to evaluate the robustness of the findings obtained from likelihood-based methods under MAR missingness.

MMRM analyses test the endpoint hypothesis or hypothesis specified at each time point; however, random-effects PM models analyses test either the slope difference (rate of change over time) of treatments groups or an overall treatment mean difference within the study period. That is, there is no straightforward

way to compare the findings obtained from the two approaches. However, it is straightforward to compare the findings obtained from Mixed-Effects Regression (MRM) analysis as proposed by Laird and Ware (1982) and random-effects PM model analysis. A comparison between the two methods might help us to assess any specific reasons for missingness, including the presence of a particular missing data mechanism in the datasets.

Several researchers, including Mallinckrodt et al. (2001), David and Mallinckrodt (2001), Liu and Gould (2002), Lane (2008), and Barnes et al. (2008), have published simulation study results to compare the performance of LOCF analysis with likelihood-based analysis in evaluating treatment efficacy at the study endpoint under a particular missing data mechanism. There is a little evidence of considering a mixture of the three missing data mechanisms in these simulation studies. Therefore, a logical extension of the previous simulation studies is to evaluate performance of the two approaches under a mixture of the three missing data mechanisms. There is also a little evidence of doing sensitivity analyses using the two approaches in the real clinical trial datasets. Therefore, the objective of this paper is to report the results of extensive simulation studies to compare the Type I error rates and biases committed in using the LOCF ANCOVA and MMRM approaches to analyze longitudinal incomplete data under different missing data mechanisms, including a mixture of three missing data mechanisms. In our simulation studies, possible scenarios of treatment efficacy as learned from different clinical trials of neurological and psychiatric drug products are included. In addition, a sensitivity analysis (i.e., comparing parameter estimates and estimated standard errors) is conducted using the above statistical approaches on 48 clinical trial datasets (obtained from 25 NDAs) submitted to the FDA's Division of Neurological and Psychiatric drug products. A brief review of the statistical approaches as stated above is as follows.

2. METHODS

2.1. Last Observation Carried Forward (LOCF) ANCOVA Endpoint Analysis

Let y_{ik} denote change from baseline (y_{i0}) of outcome measurement at the kth time point for the ith subject. Assume that there are $i = 1, \ldots, N$ subjects, and $k = 1, \ldots, K$ (end of study visit) repeated observations per subject. Assume that x_i denotes a dummy coded covariate for subject i, for example, a treatment condition with $x_i = 0$ for control group and $x_i = 1$ for the treatment group. Further, assume that if y_{iK} (end point measurement) is missing, then $y_{iK} = y_{ik}$ (where $k = 1, \ldots, K-1$). That is, if endpoint measurement is missing, it will be filled in by the previous observed measurement. Consider the following regression model for y_{iK}

$$y_{iK} = \beta_0 + \beta_1 y_{i0} + \beta_2 x_i + \varepsilon_{ik}, \tag{1}$$

where β_0 is the intercept, β_1 is the effect of baseline measurement (y_{i0}), and β_2 is the treatment condition difference at time K. Residuals ε_{ik}s are assumed to be independently distributed from a univariate normal distribution. The analysis based on equation (1) is called LOCF ANCOVA analysis.

2.2. MMRM Analysis

The MMRM analysis is a special form of the general Mixed-Effects Regression Model analysis, and hence the MRM model specification is stated first, and then the distinction between the two models is made. A MRM model satisfies followings:

$$Y_i = X_i\beta + Z_i v_i + \varepsilon_i, \tag{2}$$

where

Y_i = the $n_i \times 1$ vector of responses for subject i
X_i = a known $n_i \times p$ design matrix
β = a $p \times 1$ vector of unknown population parameters
Z_i = a known $n_i \times r$ design matrix
v_i = a $r \times 1$ vector of unknown subject effects (random effects) distributed as $N(0, \sum_v)$ and
ε_i = a $n_i \times 1$ vector of random residuals distributed independently as $N(0, \sum_{\varepsilon i})$

v_i and ε_i are independent.

From equation (2), it can be derived that Y_i are distributed as independent normal, with mean $X_i\beta$, and variance–covariance matrix $Z_i \sum_v Z_i' + \sum_{\varepsilon i}$.

The vector Y_i and the matrices X_i and Z_i carry the subscript i, which means that no assumption of complete data (across time points) on the response or covariate measurements is being made. However, it is assumed that for a given time point a subject has complete data on the response variable and all model covariates. The model residuals are assumed, conditional on the random effects, to be uncorrelated and normally distributed. In that case, the simplifying assumption that $\sum_{\varepsilon i} = \sigma^2 I_{ni \times ni}$ is made.

In the above MRM model specification, a saturated treatment group by measurement time (i.e., visit) model for the mean, combined with an unstructured within-subject error covariance, can be included. Such an inclusion in a MRM model leads to a multivariate normal model. In analyzing continuous outcome measures, Mallinckrodt et al. (2001) refer to this model as MMRM analysis. That is, in the MMRM model the *Time* is considered as a factor variable and *Treatment* $*$ *Time* effects is considered as an unstructured interaction effect, instead of considering *Treatment* $*$ *Time* effect as the slope (rate of change) difference of treatment groups over the study time period. The advantage of considering the effect of *Treatment* $*$ *Time* as unstructured is that it provides the direct estimates and statistical test of least square mean (LSMEAN) differences of the treatment groups at the study endpoint, as well as at each scheduled study time point with respect to the primary efficacy measure. Since patients in clinical trials are often evaluated at a fixed number (relatively small) of time points, the MMRM modeling approach facilitates analyzing clinical trial data, considering *Time* as a factor variable in the model.

The MMRM analysis includes a random effect as part of the within-subject error covariance structure by a repeated statement in the model specification. An "unstructured" covariance is often used to model the within-subject errors in the analysis. The advantage of using an "unstructured" (UN) covariance is that no

assumptions are made about the within-subject variability. The UN covariance structure is the most "liberal" in the sense that it allows every term in the matrix to be different. Although a misspecification of wrong covariance structure in MMRM analysis inflates Type I error and alters power, but a specification of unstructured (UN) covariance structure in MMRM analysis, regardless of the true variance–covariance structure, is reasonable and provides better control of Type I error rate and power than LOCF analysis (Mallinckrodt et al., 2004).

3. SIMULATION STUDY

In our simulation studies, 5,000 datasets are generated under multivariate normal assumption with a given mean vector and an unstructured covariance matrix. Two arms (Treatment = 1 and Placebo = 0) are considered in the simulation. In each arm, 200 subjects are included. Seven time points (Baseline and six postbaseline visit time points) are included in the study design.

Two simulation studies are conducted. In the first simulation (Simulation Study 1), each of the three missing data mechanisms is implemented alone. In the second simulation (Simulation Study 2), a mixture of the three mechanisms is implemented. In both simulation studies, differential dropout rates between two groups, as stated in Table 1, are included in creating incomplete data.

Under the null hypothesis (i.e., no difference between the two groups at the study endpoint), two scenarios are considered in the simulations. In scenario 1, the null hypothesis is true at each time point (Fig. 1), and in scenario 2, the null hypothesis is true only at the study endpoint (Fig. 2). Under the alternative hypothesis (i.e., there is a difference between the two groups at the study endpoint), another two scenarios are considered. In scenario 3, the treatment efficacy begins from the early randomized period (Fig. 3), and in scenario 4, the efficacy begins at the later time point of the randomized period (Fig. 4).

In the simulation studies, two sets of unstructured covariance matrices are considered. In one matrix, the observations within each subject are highly correlated, and in another matrix, the observations within each subject are not highly correlated (i.e., moderately correlated).

Simulation Study 1 Under the three missing data mechanisms (MCAR, MAR, MNAR), three incomplete datasets are generated from a full (i.e., no missing) simulated dataset. In generating the incomplete datasets from a full dataset, the following strategies are adopted. For the MCAR mechanism, certain percentages (given in Table 1) of missing data are generated randomly at each visit and all subsequent visits. Similarly, for the MAR mechanism, missing data at visit i and the

Table 1 Cumulative dropout scenario in creating incomplete dataset

	Baseline	Visit 1	Visit 2 (%)	Visit 3 (%)	Visit 4 (%)	Visit 5 (%)	Visit6 (%)
Placebo	—	—	10	20	25	27	30
Treatment	—	—	5	10	20	35	50

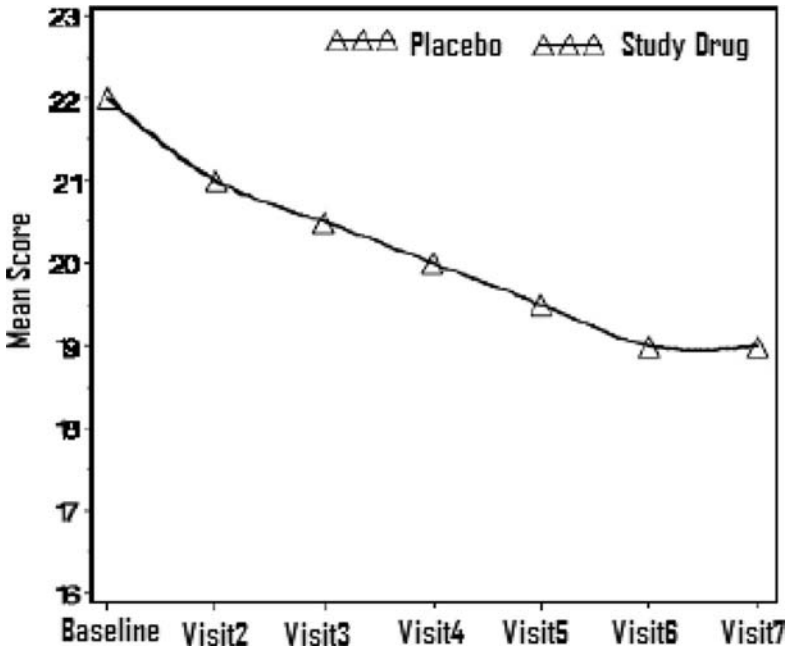

Figure 1 Equal mean score profiles.

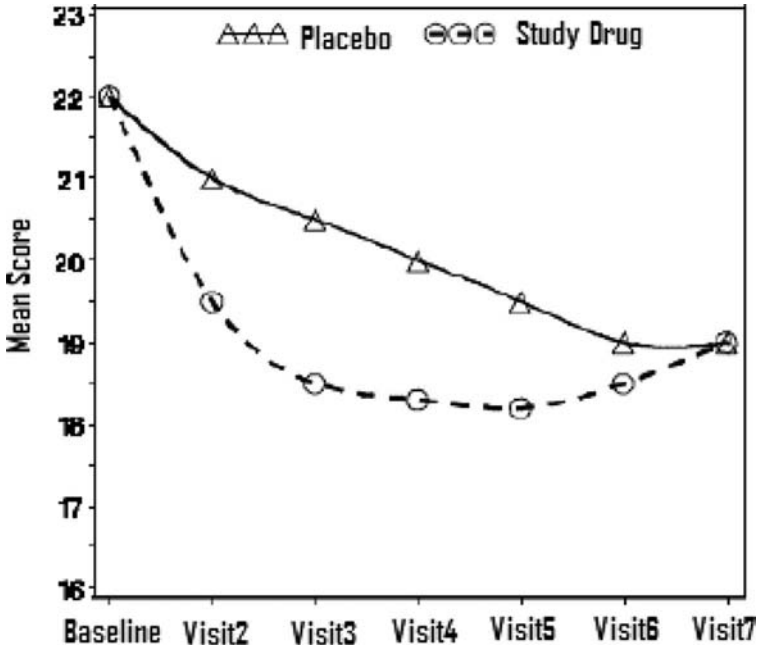

Figure 2 Equal mean scores at study end point.

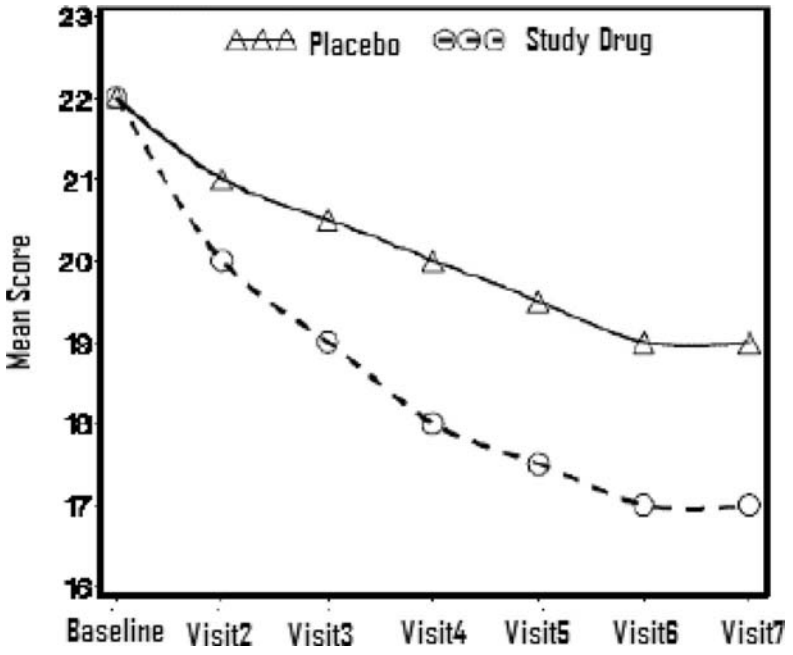

Figure 3 Different mean scores at each post visit.

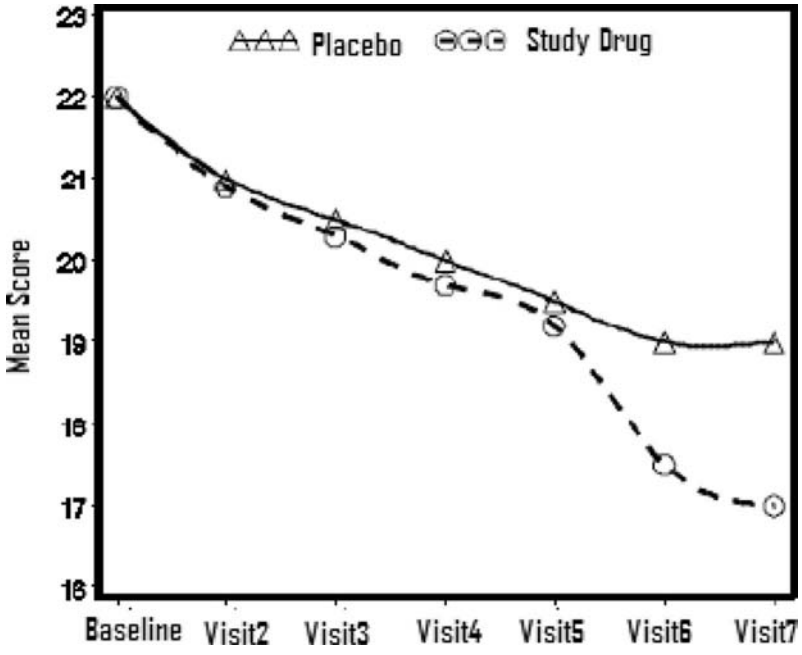

Figure 4 Different mean scores at study endpoint.

subsequent visits are assumed to be dependent on the values of outcome measure at visit $i-1$. For the MNAR mechanism, if the value of the outcome measure is higher at visit i, then the subject will have missing data at ith visit and the subsequent visits.

Simulation Study 2 In generating incomplete datasets having a mixture of the three missing data in the same data set, each incomplete dataset includes 33% of the total missing data at each visit from each of the three missing data mechanisms. For example, in the placebo group, a total of 10% of subjects (as stated in Table 1) will have missing scores from visit 2 and onwards. Among the 10% subjects, 33% subjects' scores will be missing due to each of the three missing data mechanisms.

4. RESULTS

Simulation Study 1 Table 2 lists the findings from Simulation Study 1 (5,000 datasets) under the null and alternative hypotheses in the presence of the three missing data mechanisms and strong covariance structure. When the null hypothesis is true at each time point (i.e., Fig. 1), both the LOCF ANCOVA endpoint analysis and the MMRM analysis are able to reestimate the true treatment difference at the study endpoint with a negligible bias and control Type I error rate at a nominal level (i.e., close to 5%) in presence of MCAR or MAR mechanisms. However, when the null hypothesis is true only at the study endpoint (i.e., Fig. 2), the MMRM analysis is able to reestimate the true treatment difference and control Type I error rate close to a nominal level in the presence of MCAR or MAR mechanisms, but the LOCF ANCOVA analysis fails to reestimate the true treatment difference and it also inflates Type I error rates. Under the MNAR mechanism, both the MMRM and LOCF ANCOVA fail to reestimate the true treatment difference and inflate the Type I error rate severely. The above findings are also true when a moderate covariance structure is assumed in generating the datasets.

Table 2 Simulation study 1 results

		Simulated data[§] analyses					
		Under H_0 – Fig. 1		Under H_0 – Fig. 2		Under	Under
	Method	Mean est.[$]	% Reject H_0	Mean est.[$]	% Reject H_0	H_A – Fig. 3 Bias.[$,¶]	H_A – Fig. 4 Bias.[$,¶]
Full data	ANCOVA	−0.001	5.04	−0.003	5.00	−0.001	−0.003
Incomplete	MMRM	−0.002	5.02	−0.001	5.05	−0.001	−0.001
data (MCAR)	LOCF ANCOVA	−0.063	5.29	−1.025	62.28	0.009	0.528
Incomplete	MMRM	−0.004	5.20	−0.011	5.20	−0.016	−0.011
data (MAR)	LOCF ANCOVA	−0.070	7.86	−1.150	75.18	−0.019	.610
Incomplete	MMRM	−0.218	10.84	−0.198	11.88	−0.216	−0.198
data (MNAR)	LOCF ANCOVA	−0.175	9.20	−1.242	63.84	−0.088	0.518

[§]Under each scenario of treatment efficacy, separate data set was simulated.
[$]Mean of the estimated least square mean differences at the study endpoint of the 5000 datasets.
[¶]The true difference at the endpoint was $\mu_{\text{Treatment}} - \mu_{\text{placebo}} = 2$.

When the alternative hypothesis is true at the study endpoint and the efficacy of a drug starts to show from the early time point (i.e., Fig. 3), both the LOCF ANCOVA endpoint analysis and the MMRM analysis are able to reestimate the true treatment difference at the study endpoint, with a negligible bias in the presence of MCAR or MAR mechanisms. However, when the alternative hypothesis is true at the study endpoint and the efficacy of a drug starts to show at the late time point (i.e., Fig. 4), the MMRM analysis is able to reestimate the true treatment difference at the study endpoint with a negligible bias in the presence of MCAR or MAR mechanisms, but the LOCF ANCOVA analysis fails to reestimate the true treatment difference. Under the MNAR mechanism, both the MMRM and LOCF ANCOVA analyses fail to reestimate the true treatment differences in both cases. The above findings are also true in the presence of a moderate covariance structure.

Simulation Study 2 Table 3 presents the findings of simulation study 2, when the null hypothesis holds in the presence of a mixture of the three missing

Table 3 Simulation study 2 results

		Simulated data[§] analyses					
		Under H_0 – Fig. 1		Under H_0 – Fig. 2		Under	Under
% of Dropout patients	Method	Mean est.[$]	% Reject H_0	Mean est.[$]	% Reject H_0	H_A – Fig. 3 Bias.[$,¶]	H_A – Fig. 4 Bias.[$,¶]
Full data	ANCOVA	−0.001	4.97	−0.001	5.00	−0.001	−0.001
Treat (10%)	MMRM	−0.007	5.40	−0.008	5.13	−0.007	0.001
Placebo (5%)	LOCF ANCOVA	0.040	5.36	−0.210	17.13	0.071	0.328
Treat (15%)	MMRM	0.003	5.58	0.002	5.50	0.001	−0.008
Placebo (5%)	LOCF ANCOVA	0.100	7.76	−0.275	25.77	0.146	0.193
Treat (20%)	MMRM	−0.014	5.98	−0.016	5.80	−0.014	−0.016
Placebo (10%)	LOCF ANCOVA	0.081	7.22	0.420	50.43	0.142	0.384
Treat (25%)	MMRM	−0.002	6.12	−0.003	5.83	−0.002	−0.019
Placebo (10%)	LOCF ANCOVA	0.140	11.08	−0.486	63.07	0.216	0.578
Treat (30%)	MMRM	−0.216	6.26	−0.019	5.90	−0.019	−0.003
Placebo (15%)	LOCF ANCOVA	−0.163	9.32	−0.628	84.47	0.213	0.520
Treat (35%)	MMRM	−0.216	6.24	−0.002	5.60	−0.004	−0.004
Placebo (15%)	LOCF ANCOVA	−0.163	14.82	−0.692	90.74	0.288	0.714
Treat (40%)	MMRM	−0.216	6.90	−0.019	6.50	−0.019	−0.019
Placebo (20%)	LOCF ANCOVA	−0.163	13.42	−0.835	99.37	0.285	0.773
Treat (45%)	MMRM	−0.216	7.33	−0.037	6.33	−0.037	−0.030
Placebo (25%)	LOCF ANCOVA	−0.163	14.00	−0.979	99.87	0.282	0.831

§Under each scenario of treatment efficacy, separate data set was simulated.
[$]Mean of the estimated least square mean differences at the study endpoint of the 5000 datasets.
[¶]The true difference at the endpoint was $\mu_{\text{Treatment}} - \mu_{\text{placebo}} = 2$.

data mechanisms (i.e., consider 1/3 of the total missing data from each mechanism) in a dataset and a strong covariance matrix. The findings indicate that in the presence of a mixture of the three missing data with differential dropout rates between the two treatment groups in a dataset, the MMRM approach is able to reestimate the true treatment difference consistently with a negligible bias and control Type I error rate, at a nominal level when the null hypothesis is true at each time point or null hypothesis is true only at the study endpoint. The LOCF ANCOVA endpoint analysis fails to reestimate the true treatment difference and inflates the Type I error rate in the presence of a small percentage of dropouts. The findings of this simulation indicate that in presence of a mixture of the three missing data mechanisms in a dataset, the MMRM analysis appears to be a superior statistical method, even in the presence of relatively higher dropout rates, compared to the LOCF ANCOVA endpoint analysis to evaluate the treatment efficacy at the study endpoint.

Under alternative hypotheses (i.e., Figs. 3 and 4), the MMRM analysis is able to reestimate the true treatment difference at the study endpoint with a negligible bias in the presence of a mixture of MCAR, MAR, and MNAR mechanisms. The LOCF ANCOVA analysis fails to reestimate the true treatment difference in the presence of even a small percentage of dropouts. The above findings are also true in the presence of both the strong and moderate covariance structures.

5. ANALYSIS OF CASE STUDY

A total of 25 NDA datasets submitted to the division of neurological and psychiatric drug products are reanalyzed to compare the efficacy decisions at the study endpoint based on MMRM and LOCF ANCOVA endpoint analysis. Among the 25 NDA datasets, there are 48 acute phase III double blind randomized trials. There are 108 test comparisons (e.g., Treatment vs. Placebo, and Active Control vs. Placebo comparisons). The study duration at double-blind phase ranges from 6–14 weeks. The number of randomized patients ranges from 100–150 per group. Within each study, the primary hypothesis is to evaluate treatment efficacy at the study endpoint.

The endpoint least square mean (LSMEAN) difference between a study drug vs. placebo in the MMRM analysis appears to be similar (on average) to the corresponding LSMEAN difference estimated in the LOCF ANCOVA endpoint analysis. The estimated standard error of the LSMEAN difference in the MMRM approach appears to be relatively larger compared to the corresponding standard error estimated in the LOCF ANCOVA analysis. Figure 5 plots the T-values of the 108 test comparisons obtained from the two approaches. Both the MMRM and LOCF ANCOVA analyses provide consistent (i.e., similar statistical inference) conclusions of 94% (i.e., 101/108 tests) comparisons with respect to significance or insignificance of the tests. Out of the seven inconsistent comparisons, in five comparisons the LOCF analysis provides statistically significant conclusions, whereas the MMRM analysis provides statistically insignificant conclusions. In general, the T-values obtained from the LOCF ANCOVA analyses look to be consistently larger in the majority of comparisons than the corresponding values obtained from the MMRM analyses.

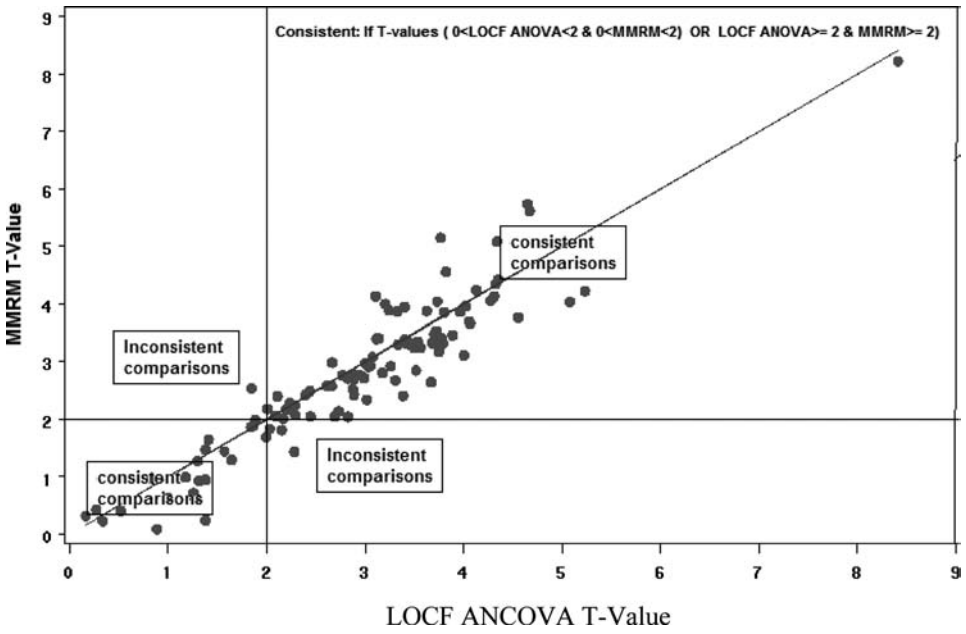

Figure 5 Plot of the *T*-values of 108 test comparisons of LOCF ANCOVA and MMRM analyses.

The overall means of the estimated LSMEAN differences and standard errors of the 108 test comparisons are −4.04 and 1.52 for the LOCF ANCOVA analysis, and −4.06 and 1.74 for the MMRM analysis. The corresponding *T*-values are −2.66 and −2.33. The individual test comparisons, as well as the overall test comparisons, confirm that both the MMRM and LOCF ANCOVA analyses estimate similar LSM EAN difference, but the LOCF ANCOVA analysis underestimates standard errors of the LSMEAN differences. Carpenter et al. (2004) also finds that the estimated standard error of the LSMEAN difference in LOCF ANCOVA analysis is wrong (usually underestimated).

6. SENSITIVITY ANALYSIS

Since the possibility of the presence of a nonignorable dropout mechanism in a real clinical trial incomplete dataset is difficult to role out, it is important to evaluate the robustness of the findings of a nonignorable likelihood-based random-effects PM model analysis, with the findings of an ignorable likelihood-based MRM analysis. Next, the impacts of deviation from nonignorable missing data assumption to ignorable missing data assumption are evaluated based on the real clinical trial incomplete datasets.

We reanalyzed 48 clinical study datasets of neurological and psychiatric drug products to evaluate the performance of the likelihood-based MRM and random-effects PM modeling approaches. In the random-effects PM model analysis of each study data, the subjects are divided into three groups based on their dropout patterns during the study duration (i.e., early dropouts, late dropouts, and

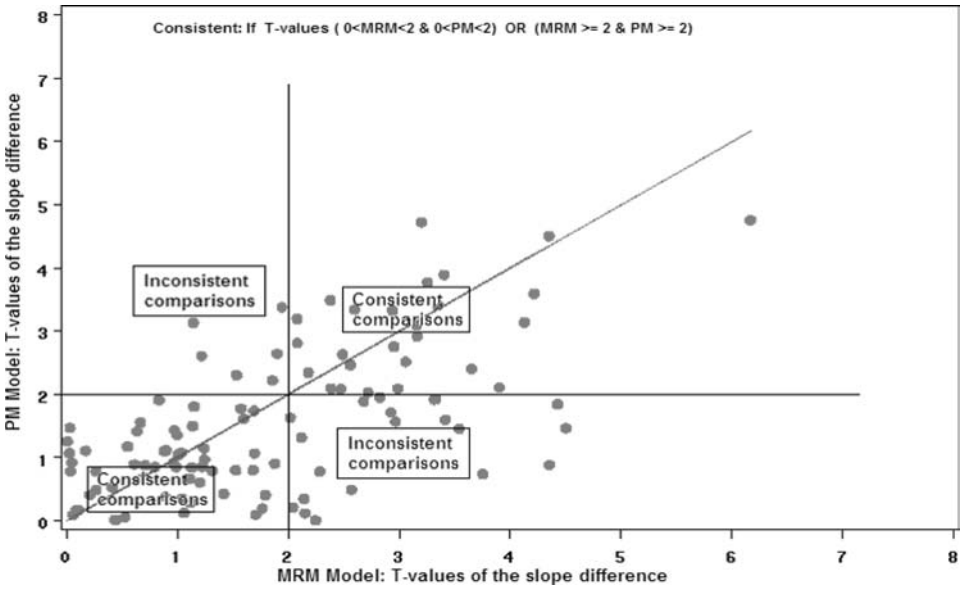

Figure 6 Plot of the T-values (slope comparisons) MRM vs. PM model analyses.

completers). About 80% of the efficacy conclusions of the study results remain the same (i.e., consistent with respect to the significance of the tests at 5% level) across the methods under ignorable and nonignorable missingness mechanisms (Fig. 6). In the remaining 20% of the study datasets where discrepancies in efficacy conclusions are found, the reasons for discrepancy are explained by the observed data (shown in Table 5 and Fig. 8), and hence there is no clear indication of the presence of nonignorable missing data in neurological and psychiatric clinical trials. Several researchers, including Little and Rubin (2002), Verbeke and Molenberghs (2000), and Mallinckrodt et al. (2001, 2003), have concluded that the missing data in clinical trials are mostly MAR. Moreover, Molenberghs et al. (2004) and Shen et al. (2006) find that even if MNAR data exist, it has a small, unimportant impact on treatment contrasts in MAR analysis. The findings based on the 48 study datasets also provide an additional support for the use of likelihood-based ignorable models in analyzing longitudinal, incomplete clinical trial datasets of neurological and psychiatric drug products.

Next, we present two real clinical trial data examples in Table 4 to demonstrate at what circumstances the MRM and random-effects PM models

Table 4 Statistical findings of two psychiatric trial data sets

	Slope diff.		LSMEAN diff.		T-value of slope diff.		T-value of LSMEAN diff.	
TRT vs. PL	MRM	PM	MMRM	LOCF	MRM	PM	MMRM	LOCF
Study 1	−1.15	−1.38	−16.56	−11.25	4.13	3.14	2.99	3.77
Study 2	−0.39	0.05	−3.41	−3.57	2.15	0.12	2.92	3.03

provide similar and dissimilar statistical inferences. In addition, the endpoint LSMEAN comparisons results of the MMRM and LOCF analyses of the two data examples are also reported in the same table.

In study 1, the slope-based analyses (i.e., MRM and PM), as well as the endpoint analyses (i.e., MMRM and LOCF) provide the similar statistical conclusions regarding efficacy of the study drug as compared to the placebo, with respect to the change from the baseline of the primary efficacy measure HAMD total score. However, in study 2, the MRM and PM estimate slopes -0.39 (T-value $= 2.15$) and 0.05 (T-value $= 0.12$), respectively, for the change from the baseline of the primary efficacy measure HAMD total score. The endpoint LSMEAN differences of the change score at the study endpoint obtained from the MMRM and LOCF ANCOVA analyses are -3.41 (T-value $= 2.92$) and -3.57 (T-value $= 3.03$), respectively. That is, based on findings of the MRM, MMRM, and LOCF ANCOVA analyses, the study drug is significantly efficacious compared to placebo; whereas based on the PM analysis it is not efficacious compared to placebo. To understand the reasons for similarity and discrepancy in the conclusions based on the MAR and MNAR models in studies 1 and 2, further exploratory analyses of the dropout patterns and observed mean change scores of the available patients have been done as follows.

Table 5 lists the distributions of dropouts due to different reasons [due to adverse events (AEs), lack of efficacy (LOE), and other reasons]. Within each of the two studies, there were no systematic patterns of dropouts due to different reasons across the early vs. late dropout groups. That is, with respect to dropouts due to different reasons and time to dropout the two studies are comparable.

Figures 7 and 8 list the plots of observed mean change scores in the HAMD total score of the available patients at each visit by treatment groups, as well as the means scores by dropout patterns and treatment groups. In study 1, the observed mean change score profiles of the two treatment groups are separated at the study endpoint and indicate the presence of treatment efficacy for the completers. With respect to the dropout status, the observed mean score profiles of the early dropout patients and completers display the presence of treatment efficacy of the study drug. Based on the analyses of both the MAR and MNAR models, the study drug is statistically significantly efficacious compared to placebo.

Table 5 Distribution of the dropout patients by reasons for dropouts

Study #	Group	Tot drop *N* (%)	Drop time	Dropout *N*	Due to AE *N* (%)	Due to LOE *N* (%)	Due to other[§] *N* (%)
Study 1	PL ($N = 138$)	51 (37)	Early	22	4 (18)	3 (13)	15 (68)
			Late	29	3 (10)	17 (58)	9 (31)
	TRT ($N = 133$)	45 (34)	Early	28	13 (46)	3 (10)	12 (42)
			Late	17	4 (23)	–	13 (76)
Study 2	PL ($N = 119$)	28 (24)	Early	14	–	4 (28)	10 (71)
			Late	14	2 (14)	6 (42)	6 (42)
	TRT ($N = 118$)	23 (19)	Early	11	3 (27)	6 (54)	2 (18)
			Late	12	1 (8)	5 (41)	6 (50)

[§]Other includes unsatisfactory compliance, consent withdrawn, unrelated to therapy etc.

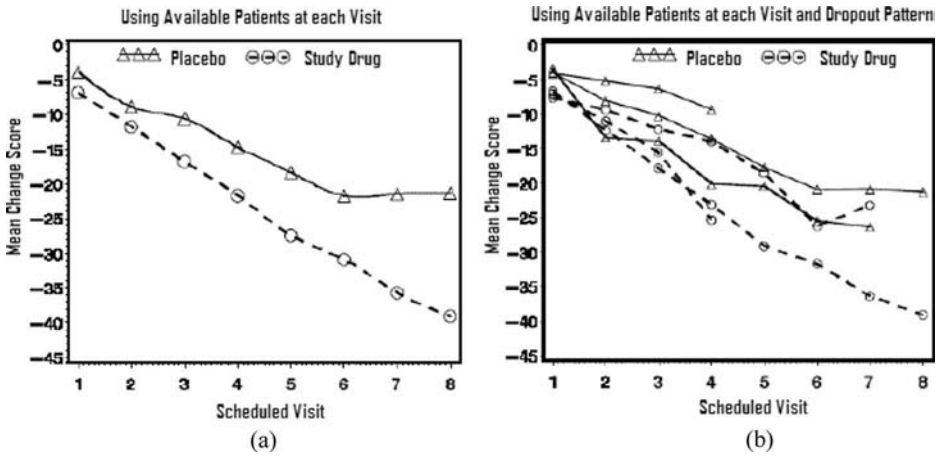

Figure 7 Observed mean profiles by treatment groups and dropout patterns (study 1).

In study 2, the plots of observed mean scores of the two treatment groups (Fig. 8) also suggest the presence of treatment efficacy at the study endpoint for the completers. However, among the late dropout group, the mean change score from baseline score for the study drug group is less, compared to the mean change score for the placebo group. Although the completers (about 80% of the randomized patients) suggest that the study drug is efficacious compared to placebo, the statistical analysis based on the MNAR model (i.e., PM model) demonstrates that the study drug is not, statistically, significantly different from placebo. However, the LOCF ANCOVA analysis and MAR model analyses (MRM and MMRM analyses) consistently demonstrate the efficacy of the study drug. The insignificant efficacy finding in the random-effects PM model is demonstrated by the late dropout patients (only 11% of the randomized patients). In the sensitivity analyses of the

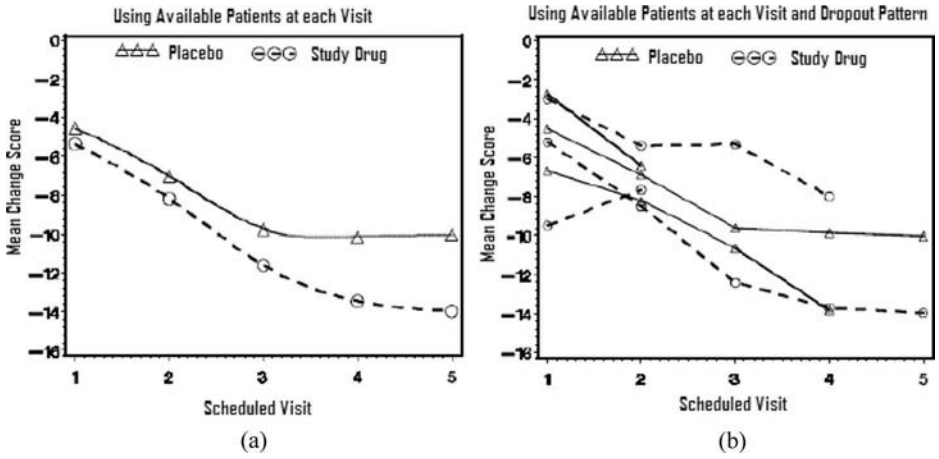

Figure 8 Observed mean profiles by treatment groups and dropout patterns (study 2).

remaining datasets, where discrepancy in the efficacy conclusion is found across the MRM and random-effects PM model results, the reasons for discrepancy are very similar to the reasons as observed in the study 2. That is, a particular group of patients (either early dropout or late dropout groups) that has the opposite direction of efficacy over time, compared to the other groups of patients, can easily mislead the statistical conclusion based on the random-effects PM model analysis. The misleading findings from the random-effects PM analysis might be due to the fact that the defined patterns by the time of their dropout might not be useful. Therefore, the random-effects PM model analysis needs to be used with due caution in analyzing incomplete clinical trial datasets.

7. DISCUSSION AND CONCLUSION

In this paper, an attempt is made to understand the impact of missing data due to dropouts in evaluating the efficacy of study drugs at the study endpoints of clinical trials. Extensive simulation studies under the three missing data mechanisms, as well as under a mixture of these three mechanisms, are carried out to evaluate the relative merits of using MMRM analysis vs. LOCF ANCOVA analysis. In addition, 48 clinical trial datasets (from 25 NDAs submitted to the division of neurological and psychiatric drug products) are reanalyzed using the LOCF ANCOVA model, MAR likelihood-based methods, and MNAR model models to evaluate the consistency of the final conclusions on efficacy of study drugs at the study endpoints.

Our simulation studies suggest that MMRM analysis controls Type I error rate at a nominal level in the presence of MCAR or MAR (i.e., ignorable) missing data mechanisms when the null hypothesis is true at each time point or true only at the study endpoint. LOCF ANCOVA analysis, however, inflates Type I error rates when the null hypothesis is true only at the study endpoint. The simulation study also indicates that in the presence of the MNAR (i.e., nonignorable) mechanism, both the LOCF ANCOVA and MMRM analyses inflate Type I error rates. However, in the presence of a mixture (having $\leq 1/3$ MNAR) of the three missing mechanisms in the same dataset, MMRM analysis is superior in controlling Type I error rate, as compared to LOCF ANCOVA analysis. In general, the simulation studies suggest that MMRM analysis is a better approach in controlling Type I error rates and minimizing biases in treatment differences, as compared to the LOCF ANCOVA approach.

The simulation studies covered in this paper also reveal that assuming unstructured (UN) covariance to explain the within subject covariance in MMRM analysis works reasonably well in protecting Type 1 error rates, and the MMRM model converges within a few iterations in Proc Mixed procedure (using the REML method). The findings of 48 clinical trial data analyses also support that the within-subject covariance follows the UN structure. Therefore, using the UN covariance in the MMRM model to analyze clinical trial data is a reasonable choice for this purpose. A further sensitivity analysis might be important to assess the robustness of the efficacy results under various covariance structures in analyzing clinical trial datasets.

The findings of 48 clinical trial datasets of neurological and psychiatric drug products indicate that both the LOCF ANCOVA and MMRM analyses estimate a

similar treatment difference at the study endpoint. However, the MMRM analysis estimates a larger standard error of the treatment difference, as compared to the corresponding estimate in LOCF ANCOVA analysis; hence, the MMRM approach appears to be a superior approach in evaluating the efficacy of a study drug.

A comparison of the MRM model with the random-effects PM model in reanalyzing the NDA datasets reveals that similar statistical inference regarding the efficacy of a drug at the study endpoint is provided by the two models in a majority of datasets. However, in the comparisons where the two models provide inconsistent statistical conclusions, one of the possible reasons found in the exploratory analysis is that one of the dropout groups (either the early dropout or late dropout groups) has a reverse efficacy effect between the study drug vs. placebo, as compared to the efficacy effect of the study drug for the completers. In such a situation, an average of treatment differences across the groups in the random-effects PM modeling becomes smaller and insignificant. Therefore, random-effects PM models can give biased results if not used with care.

It is often stated that the missing data due to adverse events are nonignorable. Our exploratory data analyses of the neurological and psychiatric clinical datasets reveal that there is a bivariate positive relationship between the number of adverse events and the last observed score of the primary efficacy measure of the dropout patients. The patients who dropped out due to adverse events often had higher scores (i.e., worsening of the disease) of efficacy measures at the last available visit. Since the dropout status of these patients can be explained by their observed score, likelihood-based MMRM analysis appears to be a reasonable statistical approach in such circumstances.

Since the possibility of a nonignorable missingness mechanism in longitudinal clinical trial data cannot be ruled out, it is important to consider some exploratory analyses to understand the missingness mechanism of the dropout patients, as well as some statistical methods that will minimize the impact of missingness on the final findings. Sensitivity analyses using the available statistical methods under various missingness assumptions should be performed routinely to assess the robustness of the findings. These analyses should be planned for in protocol when patient dropouts are expected. Since the statistical finding might be uninterpretable in the presence of high dropout rates, the dropout rates in a trial need be under consideration in interpreting the efficacy findings.

Finally, no universally best statistical method is available for the analysis of longitudinal incomplete clinical trial data. The likelihood-based mixed-effects model repeated measure analysis (i.e., MMRM analysis) under the ignorable missing data framework appears to be a robust approach in estimating the true treatment difference and in controlling Type I error rates. Hence, MMRM analysis is a sensible analytic choice in evaluating the efficacy of a drug, along with a sensitivity analysis framework to assess the robustness of results under various missingness assumptions.

ACKNOWLEDGMENT

We thank the two referees of an earlier version of this paper for valuable comments and suggestions which have been incorporated into this version.

REFERENCES

Barnes, S. A., Kallin, C., David, S. R., Mallinckrodt, C. H. (2008). The impact of missing data and how it is handled on the rate of false positive results in drug development. *Pharmaceut. Statist.* 7:215–225.

Carpenter, J., Kenward, M., Evans, S., White, I. (2004). Last observation carry-forward and last observation analysis. Letter to the Editor. *Stat. Med.* 23(20):3241–3244.

Cook, R. J., Zeng, L., Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics* 60:820–828.

David, S. R., Mallinckrodt, C. H. (2001). Type I error rates from mixed-effects model repeated measures vs. fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal* 35:1215–1225.

Diggle, P. J., Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of Applied Statistics* 43:49–93.

Fitzmaurice, G. M., Laird, N. M., Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley: New York.

Heyting, A., Tolboom, J. T. B. M., Essers, J. G. A. (1992). Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in Medicine* 11:2043–2061.

Hogan, J. W., Laird, N. M. (1997). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* 16:259–272.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine* 7:305–315.

Laird, N. M., Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38:963–974.

Lane, P. W. (2008). Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharmaceutical Statistics* 7:93–106.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 88:125–133.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* 81:471–483.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90:1112–1121.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

Liu, G., Gould, A. L. (2002). Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J. Biopharm. Stat.* 12(2):207–226.

Mallinckrodt, C. H., Clark, W. S., David, S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics* 11:9–21.

Mallinckrodt, C. H., Sanger, T. M., Dube, S., DeBrota, D. J., Molenberghs, G., Carroll, R.J., et al. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol. Psychiatry* 53(8):754–760.

Mallinckrodt, C. H., Kaiser, C. J., Watkin, J. G., Molenberghs, G., Carroll, R. J. (2004). 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. *Clinical Trials* 1(6):477–489.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., Thijs, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine* 21:1023–1041.

Molenberghs, G., Kenward, M. G. (2007). *Missing Data in Clinical Studies*. New York: John Wiley & Sons, Ltd.

Molenberghs, G., Thijs, H., Jansen, I., Beunkens, C., Kenward, M. G., Mallinkrodt, C. H., Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 5:445–464.

Rotnitzky, A., Robins, J. M., Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponses. *Journal of the American Statistical Association* 93(444):1321–1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63:581–592.

Shao, J., Zhong, B. (2003). Last observation carry-forward and last observation analysis. *Statistics in Medicine* 22:3241–3244.

Shen, S., Beunckens, C., Mallinckrodt, C. H., Molenberghs, G. (2006). A local influence sensitivity analysis for incomplete longitudinal depression data. *J. Biopharm. Stat.* 16(3):365–384.

Siddiqui, O., Ali, M. W. (1998). 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.* 8(4):545–563.

Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.