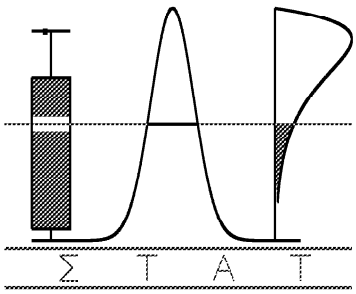
T E C H N I C A L

R E P O R T

0222

Assessing and Interpreting Treatment Effects in Longitudinal Clinical Trials with Missing Data

Craig H. Mallinckrod, Todd M Sanger, Sanjay Dube, David J Debrota, Geert Molenberghs, Raymond J Carroll, WM Zeigler Potter, and Gary D. Tollefson



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Assessing and Interpreting Treatment Effects in Longitudinal Clinical Trials with Missing Data

Craig H. Mallinckrodt[1,4], Todd M Sanger[1], Sanjay Dube[1], David J Debrota[1], Geert Molenberghs[2], Raymond J Carroll[3], WM Zeigler Potter[1], and Gary D. Tollefson[1].

[1]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285.

[2]Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590, Diepenbeck, Belgium

[3]Dept. of Statistics, Texas A&M University, College Station TX.

Running title: Assessing Treatment Effects in Longitudinal Data

Number of tables: 4

Number of figures: 0

Number of text pages: 20

*Correspondence to:

Dr. Craig Mallinckrodt
Lilly Corporate Center
Indianapolis IN 46285
PH: 317-277-2209
FX: 317-651-6269
E-MAIL: Mallinckrodt_craig@lilly.com

Address reprint requests to:

Dr. Craig Mallinckrodt, Lilly Corporate Center, Indianapolis, IN 46285

[4]This work was sponsored by Eli Lilly and Company

ABSTRACT

Treatment effects are often evaluated by comparing change over time in outcome measures. However, valid analyses of longitudinal data can be problematic, particularly if some data are missing. The last observation carried forward (LOCF) approach has for decades been a common method of handling missing data Considerable advances in statistical methodology and our ability to implement those methods have been made in recent years. Thus, it is appropriate to reconsider analytic approaches for longitudinal data. The objectives of this paper are to examine from a clinical perspective the characteristics of missing data that influence analytic choices; 2) the attributes of common methods of handling missing data; and, 3) the use of the data characteristics and the attributes of the various methods, along with empirical evidence, to develop a robust approach for the analysis and interpretation of data from longitudinal clinical trials. We propose that in many settings the primary efficacy analysis should use a likelihood-based mixed-effects modeling approach, with LOCF used as a secondary, composite measure of efficacy, safety, and tolerability. We illustrate how repeated measures analyses can be used to enhance decision-making and what caveats remain in the use of LOCF as a composite measure.

Key words: missing data, longitudinal data, mixed-effects models, repeated measures, depression

INTRODUCTION

Treatment effects are often evaluated by comparing change over time in outcome measures. However, valid analyses of longitudinal data can be problematic, particularly if some data are missing for reasons related to the outcome measure (Milliken et al 1993; Gibbons et al 1993). Missing data is an almost ever-present problem in clinical trials and numerous methods for handling missingness have been proposed, examined, and implemented (Verbeke et al 2000). In fact, there are so many methods that choosing a suitable method and interpreting its results can be difficult.

Perhaps the best place to start in determining how to analyze and interpret longitudinal data, especially in the presence of missingness, is to realize that no universally best method exists. This implies that the analysis must be tailored to the situation at hand. This, in turn, implies that the characteristics of the missing data must be understood. And thus, the objectives of this paper are: 1) to examine the characteristics of missing data that influence analytic choices; 2) to examine the attributes of common methods of handling missing data; and, 3) to use the data characteristics and the attributes of the various analytic methods, along with empirical evidence to develop a robust approach for the analysis and interpretation of data from longitudinal clinical trials. Although certain statistical concepts are inherent to this discussion, we approach it from a clinical perspective and provide references to the more detailed statistical literature when appropriate. Our aim is to translate technical aspects of data analysis to an audience that is familiar with clinical research and clinical trials, but is not expert in statistics. Our focus is on applications to neuropsychiatric disorders with ideas fixed via a specific application to a clinical trial of an antidepressant. However, the concepts covered herein have a broad range of applications.

MISSING DATA

In many areas of clinical research, the impact of missing data can be profound (Gibbons et al 1993; Laird 1988; Little et al 1987; Lavori 1992). The potential impact of missing data is best understood by considering the process (i.e., mechanisms) leading to the missingness. The following taxonomy of missing data mechanisms is now common in the statistical literature (Little et al 1987)

Data are considered missing completely at random (MCAR) if the missingness does not depend on (is not explained by) either the observed or unobserved outcomes of interest. Data are missing at random (MAR) if the missingness depends on (is explained by) the observed outcomes, but not the unobserved outcomes. Data are missing not at random if the missingness depends on (is explained by) the unobserved outcomes.

The missing data mechanism cannot be deduced with certainty from reasons for patient disposition. However, the following examples illustrate how the various types of missing data might arise. For example, MCAR data may arise from a patient who dropped out because he relocated and was too far away from the investigative site to participate in the trial. Dropout was not in any way related to the outcome of interest. An example of MAR data could be a patient who was observed to be doing poorly ***and then*** the physician and/or the patient decided to discontinue participation. In this case, dropout was related to the outcome of interest, but the observed data explained the dropout. An example of MNAR data could be a patient who had been doing well until midway in a trial, was then lost to follow up because ***after*** the last observed visit the patient relapsed into a worsened condition. Again, dropout was related to the outcome of interest, but in this case the observed data did not explain (predict) the dropout and the unobserved data held information not foreseen by the observed data. Dropouts for adverse

events are difficult to classify as MCAR vs. MAR because the relationship to the observed outcome may vary from situation to situation. However, dropouts due to adverse events are probably not MNAR in many cases because all the relevant data probably were observed.

Frequently, missingness is related to the outcome of interest, and thus the data are not MCAR (Verbeke et al 2000, Molenberghs et al 2002). The MAR assumption is much more plausible than the MCAR assumption (Little et al 1987; Verbeke et al 2000; Molenberghs et al 2002) because the observed data explain much of the missingness in many scenarios. This may be particularly true in well-controlled studies, such as clinical trials where extensive efforts are made to observe all the outcomes and the factors that influence them (Rubin et al 1995; Molenberghs et al 2002). Hence, clinical trials by their very design seek to minimize the amount of MNAR data (missingness explained by non-observed responses).

ANALYTIC APPROACHES for MISSING DATA

Traditional methods

A common choice in many therapeutic areas is to assess mean change from baseline to endpoint via analysis of variance with missing data imputed by carrying the last observation forward (LOCF). The LOCF approach assumes that missing data are MCAR and that subjects' responses would have been constant from the last observed value to the endpoint of the trial. These conditions seldom hold (Verbeke et al 2000). Carrying observations forward may therefore bias estimates of treatment effects and the associated standard errors (Gibbons et al 1993; Verbeke et al 2000, Lavori et al 1995; Siddiqui et al 1998; Heyting et al 1992; Mallinckrodt et al 2001a, Mallinckrodt et al 2001b, Molenberghs et al 2002).

Despite these shortcomings, LOCF has been the long-standing method of choice for the primary analysis in clinical trials because of its simplicity, ease of implementation, and belief that the potential bias from carrying observations forward leads to a "conservative" analysis.

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 to compute the mean change to endpoint for treatment group 1; and, for patient 6, the last observed value, 20, is used to compute the mean change to endpoint for treatment group 2. The imputed data are considered as informative as the actual data because the analysis does not distinguish between the actually observed data and the imputed data

The assertion that LOCF yields conservative results does not appear to have arisen from formal proofs or rigorous empirical study. Recent investigations (detailed in a later section) have demonstrated that LOCF can exaggerate the magnitude of treatment effects and inflate Type I error (falsely conclude a difference exists when in fact the difference is zero).

Furthermore, mean change from baseline to endpoint is only a snapshot view of the response profile of a treatment. Gibbons (Gibbons et al 1993) stated that endpoint analyses are insufficient because the evolution of response over time must be assessed to completely understand a treatment's efficacy profile. By its very design, LOCF change to endpoint cannot assess response profiles over time. Furthermore, advances in statistical theory and in computer hardware and software have made many methods simple and easy to implement.

Alternative approaches

We previously noted that in many settings the MAR assumption (observed responses explain missingness) is more reasonable than the MCAR assumption (missingness not explained by observed or unobserved responses). An MAR method is valid if data are MCAR or MAR,

but MCAR methods are valid only if data are MCAR. Likelihood-based mixed-effects models offer a general framework from which to develop longitudinal analyses under the MAR assumption (Verbeke et al 2000; Cnaan et al 1997). These methods are more robust to potential bias from missing data than LOCF (Gibbons et al 1993; Verbeke et al 2000; Molenberghs et al 2002) and other MCAR methods.

The key general feature of likelihood-based mixed-effects analyses is that they include fixed and random effects, whereas traditional analysis of variance (ANOVA) includes only fixed effects. In clinical trial applications the fixed effect of greatest interest is typically treatment group. Additional fixed effects such as baseline severity, investigative site, or demographic characteristics are also commonly included. The random effect commonly included in mixed-effects analyses that is not included in ANOVA is subject. That is, mixed-effects models consider the unique attributes of each subject and thus account for the fact that responses of individual subjects will vary. In so doing, information from the observed outcomes can be used to provide information about the unobserved outcomes, but missing data are not explicitly imputed.

The following example, using the hypothetical data in Table 1, illustrates the handling of missing data via a mixed-effects model analysis. 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 the fact 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 assumes that patient 3 would have had some additional improvement because the other patients in group 1 all improved after Visit 4. In contrast, LOCF assumes no further improvement. Patient 6 had also been doing 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 the fact that had patient 6 remained in the trial his observations would likely have been worse than the treatment group average. Because all patients in group 2 had been getting worse during Visits 3 – 6, a mixed-effects analysis assumes the rate of worsening for patient 6 would have been greater than the group 2 average. In contrast, LOCF assumes no further worsening.

The magnitudes of these "adjustments" in a mixed-effects analysis are determined mathematically from the data. Additional details may be found in Verbeke et al (2000) Littell et al (1996), and Cnaan et al (1997). While these details go beyond the scope of this paper, the basic principle is easily appreciated. A mixed-effects analysis uses all the available data to compensate for the data missing on a particular patient, whereas LOCF used 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 Visits 5 and 6 will be the same as Visit 4, even though that was not the case for any patient whose data was observed.

Likelihood-based mixed-effects analyses are easy to implement because no additional data manipulation is required to accommodate the missing data; and, the analyses can be implemented using standard software, such as the SAS Procedure Mixed (Littell et al 1996), that has been widely available for a number of years.

Methods that attempt to account for MNAR missingness simultaneously model the measurement process (observed data) and the missingness processes. While important potential advantages of MNAR approaches exist, these methods require assumptions that cannot be validated from the data at hand (Verbeke et al 2000; Molenberghs et al 2002). This, in turn,

argues that for any specific scenario a definitive MNAR analysis does not exist and such analyses are best implemented as sensitivity analyses to assess the robustness of results across different analytic approaches (Molenberghs et al 2002).

Given the implausibility of MCAR, the plausibility of MAR, and the implementation and interpretive difficulties of MNAR, methods developed under the MAR framework are well suited to longitudinal clinical trials. This is the theoretical basis for the shift away from ad hoc methods like LOCF to likelihood-based methods built on the MAR framework. However, the causes of missingness are varied and therefore it is difficult to rule out the possibility of MNAR data in clinical trials – which underscores the benefit from using MNAR analyses in assessing robustness of the results from the likelihood-based MAR analysis. Use of MNAR methods for such sensitivity analysis is beyond the scope of this paper and readers are referred to others sources for a general discussion (Verbeke et al 2000; Molenberghs et al 2002). We instead focus on likelihood-based MAR methods and their application.

Mixed-effects Model Repeated Measures Analyses

Many analytic techniques fall under the general heading of repeated measures analyses. Our focus is on a specific type of repeated measures analysis that we call MMRM because it is a likelihood-based **M**ixed-effects **M**odel **R**epeated **M**easures analysis. The term MMRM refers to a wide array of likelihood-based analyses in which subject specific effects and serial correlation are modeled via the within-subject error correlation structure. The specific implementation of MMRM described herein included an unstructured modeling of time and the within-subject error correlation structure. This version of MMRM was implemented to match the general characteristics of acute phase clinical trials, and illustrations of its use in actual practice are common (Molenberghs et al 2002; Goldstein et al 2002; Detke et al 2002). However, modeling

decisions regarding time and correlation structures are situation dependent and the unstructured approach is not always optimal, or even possible. Cnaan (1997) discussed some of the other useful approaches to modeling time and correlation structures in longitudinal data. While these modeling considerations are important, they do not affect assumptions regarding missing data. Hence, our implementation of MMRM is not a specific solution to the general problem of missing data, but rather one example from the family of likelihood-based analyses developed under the MAR framework.

All details of an MMRM analysis are dictated by the design of the study and can be specified succinctly in the protocol. Molenberghs et al (Molenberghs et al 2002) discussed why likelihood-based MAR methods are consistent with the intent-to-treat principle, and in fact are an improvement over LOCF in this regard, via appropriate use of all available data on all patients.

Mallinckrodt et al (Mallinckrodt et al 2001ab) compared the MMRM analysis as described above with the traditional LOCF ANOVA approach in simulated data. The first study (Mallinckrodt et al 2001a) compared the two methods in simulated scenarios in which a true difference between treatments in mean change from baseline to endpoint existed. The second study (Mallinckrodt et al 2001b) 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 in order to introduce MNAR missingness.

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, important differences in results existed between the methods in analyses of data with missingness.

In the study where treatment differences at endpoint existed, 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 (% of confidence intervals containing the true value) was 94.24% and 86.88% for MMRM and LOCF respectively, compared with the expected coverage rate of 95%. Although LOCF is generally considered a conservative method, it overestimated the treatment effect in some scenarios, typically when there was higher dropout in the inferior (e.g. placebo) group. Other scenarios in which LOCF is likely to overestimate the true treatment effect have been noted (Lavori 1992; Little et al 1996).

In the Type I error rate study, pooled across all scenarios with missingness, the Type I error rates for MMRM and LOCF were 5.85% and 10.36%, respectively, compared with the expected rate of 5.00%. Type I error rates in the 32 scenarios ranged from 5.03% to 7.17% for MMRM, and from 4.43% to 36.30% for LOCF. These results provide empirical justification for the shift away from LOCF to likelihood-based MAR analyses of longitudinal clinical trial data.

EXAMPLE

Methods

A reanalysis of data from a clinical trial is presented to illustrate the use of MMRM. The results were originally reported by Wernicke et al (1987). The study included patients with a baseline $HAMD_{17}$ (Hamilton 1960) total score of 19 or more. Patients were randomized to placebo, fluoxetine 20 mg daily (Flx20, n = 100), fluoxetine 40 mg daily (Flx40, n = 103), fluoxetine 60 mg daily (Flx60, n = 105), and Placebo (n=48) in a 2:2:2:1 ratio.

For these retrospective analyses, mean changes from baseline to endpoint (week 6) were analyzed using three methods. 1) An observed case (OC) analysis that included only those subjects that completed the 6-week acute therapy phase of the trial; 2) an LOCF analysis that included all subjects, with missing values imputed by carrying the last observation forward to week 6; and 3) an MMRM analysis of postbaseline data from all visits. The OC and LOCF analyses were conducted using ANCOVA models that included terms for baseline value, investigator, and treatment. The MMRM analysis included baseline value, investigator, treatment, time, and the treatment-by-time interaction. An unstructured correlation matrix was used to model the within-subject error correlation structure.

Results

Reasons for discontinuation are summarized in Table 2. As the dose of fluoxetine increased, the percentage of completers decreased, the percentage of patients who discontinued for adverse events increased, and the percentage of patients who discontinued for lack of efficacy decreased. The timing of discontinuation also varied across treatments. The Flx60 group had a higher percentage of subjects dropping out at earlier visits.

Results from analyses using the three methods are summarized in Table 3. In the LOCF analysis, only Flx20 was significantly different from placebo, and as dose increased the advantage of fluoxetine over placebo decreased. With MMRM, all doses of fluoxetine yielded similar mean changes and were significantly different from placebo. In the OC analyses, Flx60 had the greatest mean change and was the only dose significantly different from placebo.

Logistic regression analysis showed that rate of improvement had a highly significant influence on probability of missingness ($p<.001$). The dependence of missingness on rate of improvement showed that the observed responses influenced missingness, and thus the MCAR

assumption for LOCF was not valid. In addition, the assumption made by LOCF of no change from last observation to the trial's endpoint was also unrealistic because most patients in all treatments groups tended to improve over time. Molenberghs et al (Molenberghs et al 2002) used a method commonly referred to as the selection model (Diggle et al 1995) to test for the presence of MNAR missingness in these data. Although the authors found some evidence for MNAR data, differences between treatments were not influenced by it and the authors concluded that use of MMRM was sensible.

Interpreting Results

The most important difference in results of our reanalysis was that MMRM yielded significant differences for all doses whereas LOCF yielded significance for only Flx20 and OC yielded significance for only Flx60. The LOCF results were counter-intuitive in that a pronounced inverse dose-response relationship existed, with only the lowest dose being significantly different from placebo. An obvious association existed between the completion rate and the advantage of drug over placebo; namely, as dropout increased the rate of improvement decreased. With MMRM, a more plausible dose-response relationship was found.

Two concepts are central to the decision-making paradigm we propose. 1) Risk and benefit need to be clearly established separately before being combined in order to establish the overall risk-benefit of a drug; 2) Likelihood-based MAR methods such as MMRM (typically) yield appropriate estimates of efficacy (benefit). Because an LOCF result is (typically) influenced by rate and timing of dropout, it is not an efficacy analysis, but rather a composite measure of efficacy and duration on therapy. Several important caveats regarding use of LOCF in this manner are addressed later. However, the long-standing use of LOCF makes this analysis hard to abandon altogether.

In applying this decision-making paradigm in our example, MMRM results suggested that all three doses were efficacious; LOCF suggested that the overall benefit of higher doses was muted by higher dropout rates. Disposition results suggested that higher dropout at higher doses was driven by adverse events. Putting all the pieces together, we concluded that Fluoxetine 20 mg is probably the optimum starting and therapeutic dose, but some patients may benefit from increased dosages. This is a very different picture from what was seen from LOCF or OC alone. One could likely come to the same conclusion using only MMRM and the disposition table, or more preferably MMRM and a formal analysis of missingness.

DISCUSSION

We have noted that no universally best approach to analysis of longitudinal data exists. In general, however, the analyses traditionally used in many longitudinal clinical trials are based on the unrealistic assumption that data are MCAR. The MAR assumption is more plausible than MCAR. Likelihood-based MAR methods can be easily implemented with commercially available software, are consistent with the intent-to-treat principle, all details can be pre-specified, and these methods have been shown to be more robust to biases from missing data than MCAR methods such as LOCF. In fact, likelihood-based repeated measures analyses such as the analysis used in our example are valid in every situation in which LOCF is valid, and in many situations where LOCF is not valid.

Nevertheless, the possibility of MNAR data, and the bias that can result from it, is difficult to rule out. Blindly using likelihood-based MAR methods without consideration of their limitations is dangerous. With high rates of missingness, especially with appreciable losses to follow-up, results may be problematic to interpret regardless of analytic methodology. Results from a likelihood-based MAR method could be misleading. However, MNAR methods can be

complex, do not yield definitive results because assumptions must be made that can not be validated from the data at hand, and MNAR methods are therefore best implemented in a sensitivity analysis framework to assess the validity of the MAR result.

The traditional LOCF approach has been used under the assumption that while potentially biased by non-MCAR data, the bias led to a "conservative" analysis. In this context, conservative is typically thought of as underestimating the magnitude of the treatment effect. However, the simulation studies cited herein (Mallinckrodt et al 2001ab) illustrated, and other authors have shown and noted, that conservative behavior of LOCF is not guaranteed (Lavori et al 1992; Verbeke et al 2000; Little et al 1996; Molenberghs et al 2002).

Even if LOCF is conservative according to the above definition, underestimating the superiority of a superior treatment necessarily results in underestimating the inferiority of the inferior treatment. Thus, such a bias would be conservative in the context of superiority testing, but would be anti-conservative for non-inferiority testing.

Additionally, if a method yielded biased estimates of treatment effects when treatment differences existed, when the true treatment difference was zero, bias would necessarily lead to nonzero estimates of treatment differences and inflation of Type I error. For example, consider Alzheimer's disease, where the therapeutic aim is to delay or slow deterioration of mental status, as compared to situations such as depression where the goal is to improve the condition. If a treatment is in truth no more effective than placebo, but patients drop out due to adverse events, carrying the last observation forward assumes that the patient had no further deterioration in condition. Thus, carrying observations forward could lead to the false conclusion that drug was more effective than placebo. Whether or not the bias from LOCF is conservative may depend on the scenario, the type of test, and on the true difference between treatments. It is these

advantages of likelihood-based MAR methods, and shortcomings of LOCF that motivate the shift to likelihood-based MAR methods.

It is tempting to believe LOCF should be the most valuable analysis because it includes efficacy, safety, and tolerability into a composite measure of benefit. But the use of LOCF in this context must be approached carefully. Although LOCF – in some situations – yields smaller estimates of treatment differences when patients dropout for say adverse events, the reduction is not necessarily proportional to the safety risk. For example, consider the following two patients in an eight-week trial: Patient A dropped out after week 7 due to a dramatically prolonged QT interval; Patient B dropped out during week 1 with nausea. The impact to 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 nausea experienced early in the trial is typically transitory and often resolves with continued therapy and no long term consequences. This non-proportional penalty to individual patients from LOCF may cause misleading inferences regarding the merits of a treatment. For example, consider the following four treatments.

1) Average efficacy and average rate of dropouts due to adverse events
2) Below average efficacy and lower than average rate of dropouts due to adverse events
3) Excellent efficacy and average rate of dropouts due to adverse events, but dropouts typically occur very early in treatment.
4) Average efficacy and average rate of dropouts due to adverse events, but most dropouts are for sustained hypertension and prolonged QT interval.

The four treatments yield similar estimates of mean change from the LOCF analysis, yet they have very different clinical profiles. That is, the four treatments have equal risk-benefit ratios but very different risks and benefits. This is why risk and benefit should first be established individually, and then combined into an overall risk/benefit assessment.

Another shortcoming of using LOCF as a global or composite measure of total benefit is that this assumes the conditions in the clinical trial are reliable indicators of actual practice. However, clinical trials are (typically) designed to delineate causal differences between drug and placebo (or between drugs) – not to mimic actual clinical practice. It is unreasonable to assume doctors and patients make the same decisions regarding continuation of therapy in a double-blind trial where they are unsure if the patient is taking drug or placebo as they would make in actual practice when the drug and its properties are well known.

It is also important to re-emphasize that endpoint analyses of any type provide only a small part of the picture and that the entire longitudinal profile should be considered. The endpoint of a trial has specific, special meaning within the context of the trial; and, well-designed trials choose an endpoint time that is clinically relevant. However, such a time point is not necessarily more meaningful than other time points when extrapolating results from the specific trial to the general patient population. For example, patients may ask how soon until I feel better, or, how soon until I feel well? Endpoint analyses cannot address such questions. However, longitudinal methods such as the likelihood-based MAR analyses are ideally suited to provide such information from the same analysis as that which produces the endpoint contrast.

Change in primary analytic methodology has its drawbacks and warrants careful consideration. Nevertheless, the refinement in statistical theory and in our ability to implement the theory is too compelling to overlook.

CONCLUSION

No universally best approach to analysis of longitudinal data exists. However, likelihood-based mixed-effects analyses developed under the MAR framework are more robust to the bias from missing data than LOCF and are valid in every scenario where LOCF is valid, and in many other scenarios as well. Therefore, likelihood-based repeated measures analyses are a sensible analytic choice in may clinical trial scenarios. Because the possibility of MNAR data cannot be ruled out, MNAR methods can be used in a sensitivity analysis framework to assess the robustness of results from an MAR method. Efficacy results from an MAR method can be combined with results from appropriate safety analyses to assess the overall benefit of a drug. These separate assessments of risk and benefit are more useful than composite assessments of global benefit, such as obtained from an LOCF analysis.

ACKNOWLEDGEMENTS: The authors would like to thank Ms. Renee Bacall for her editorial assistance and the reviews for their details comments. Their contributions greatly enhanced the quality of this manuscript.

REFERENCES

Cnaan A, Laird NM, Slasor P (1997): Using the General Linear Mixed Model to Analyze Unbalanced Repeated Measures and Longitudinal Data. *Stat Med* 16:2349-2380.

Detke MJ, Lu Y, Goldstein DJ, Hayes JR, Demitrack MA (2002): Duloxetine, 60 mg once daily, for major depressive disorder: A randomized double-blind placebo-controlled trial. *J Clin Psychiatry* 63:4 308-315.

Diggle P J, Liang, KY, Zeger, SL (1994): Analysis of Longitudinal Data. Oxford Science Publications. Oxford. Clarendon Press.

Diggle P J and Kenward MG (1995): Informative dropout in longitudinal data analysis (with discussion). *Appl Stat* 43:49-93.

Gibbons RD, Hedeker D, Elkin I, Waternaux C, Kraemer HC, Greenhouse JB et al (1993): Some Conceptual and Statistical Issues in Analysis of Longitudinal Psychiatric Data. *Arch Gen Psych* 50:739-750.

Goldstein DJ, Mallinckrodt C, Lu Y, Demitrack MA (2002): Duloxetine in the treatment of major depressive disorder: A double-blind clinical trial. *J Clin Psychiatry* 63(3):225-231.

Hamilton M (1960): A rating scale for depression. *J Neurol Neurosurg Psychiatr* 23:56-62.

Heyting A, Tolboom J, Essers J (1992): Statistical Handling of Dropouts in Longitudinal Clinical Trials. *Stat Med* 11:2043-2061.

Laird NM (1988): Missing Data in Longitudinal Studies. *Stat Med* 7:305-315.

Lavori PW (1992): Clinical Trials in Psychiatry: Should Protocol Deviation Censor Patient Data. *Neuropsychopharmacol* 6(1):39-48.

Lavori PW, Dawson R, Shera D (1995): A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. *Stat Med* 14:1913-1925.

Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996): The SAS system for mixed models. SAS Institute Inc: Cary, NC, Chap. 1-10.

Little R, Rubin D (1987): *Statistical Analysis with Missing Data*. John Wiley and Sons: New York.

Little R, Yau L (1996): Intent-to-Treat Analysis for Longitudinal Studies with Drop-Outs. *Biometrics* 52:1324-1333.

Mallinckrodt CH, Clark WS, David SR (2001): Accounting for Dropout Bias Using Mixed-Effects Models. *J BioPharm Stat* 11:(1&2) 9-21.

Mallinckrodt CH, Clark WS, David SR (2001): Type I error rates from mixed effects model repeated measures compared with fixed effects ANOVA with missing values imputed via LOCF. *Drug Info J* 35(4):1215-1225.

Milliken GA, Johnson DE (1993): The Analysis of Messy Data. *Designed Experiments: Vol 1*. Chapman & Hall: New York, 326.

Molenberghs G, Thijs H, Carroll RJ, Kenward MG (2002): Analyzing Incomplete Longitudinal Clinical Trial Data. *Biometrics* (Submitted)

Rubin D, Shenker N (1991): Multiple Imputations in Health Care Databases: An Overview and Some Applications. *Stat Med* 10:585-598.

Rubin DB, Stern HS, Vehovar V (1995): Handling “don' t know" survey responses: the case of the Slovenian plebiscite. *J Am Stat Assoc* 90:822-828.

Siddiqui O, Ali MW (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* Springer: New York.

Wernicke JF. Dunlop SR. Dornseif BE. Zerbe RL. Fixed-dose fluoxetine therapy for depression. *Psychopharmacology Bulletin*. 23(1):164-8, 1987.

Table 1. Hypothetical data used to illustrate how various methods handle missing data

Hamilton Depression Rating Scale Total Scores [1]

			Week					
Patient	Treatment	Baseline	1	2	3	4	5	6
1	1	22	20	18	16	14	12	10
2	1	22	21	18	15	12	9	6
3	1	22	22	21	20	19	*	*
4	2	20	20	20	20	21	21	22
5	2	21	22	22	23	24	25	26
6	2	18	19	20	*	*	*	*

Missing values due to patient dropout are marked '*'.

Table 2. Percentages of subjects by reasons for study discontinuation

	Reason for discontinuation			
Treatment	Study Complete	Adverse Event	Lack of Efficacy	Other
Placebo	62%	11%	17%	12%
Fluoxetine 20 mg	64%	8%	14%	14%
Fluoxetine 40 mg	60%	16%	13%	11%
Fluoxetine 60 mg	48%	26%	10%	16%

Table 3. Mean changes from baseline to endpoint (week 6) using Mixed Model Repeated Measures (MMRM), Last Observation Carried Forward (LOCF), and Observed Case (OC) analyses

LOCF

Treatment	Mean Change	Standard Error |	P value contrast with placebo
Placebo	-5.4	0.95	
Fluoxetine 20 mg	-7.9	0.71	.035
Fluoxetine 40 mg	-7.4	0.71	.087
Fluoxetine 60 mg	-5.9	0.69	.654

MMRM

Treatment	Mean Change	Standard Error |	P value contrast with placebo
Placebo	-6.4	1.09	
Fluoxetine 20 mg	-9.8	0.77	.011
Fluoxetine 40 mg	-9.5	0.81	.023
Fluoxetine 60 mg	-9.6	0.85	.024

Observed Case

Treatment	Mean Change	Standard Error |	P value contrast with placebo
Placebo	-8.8	1.10	
Fluoxetine 20 mg	-11.1	0.76	.085
Fluoxetine 40 mg	-10.3	0.80	.284
Fluoxetine 60 mg	-11.7	0.88	.043