METHODS REPORT

A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values

Hrishikesh Chakraborty and Hong Gu

March 2009

RTI Press

RTI INTERNATIONAL

About the Author

Hrishikesh Chakraborty, DrPH, is a senior statistician at RTI International in Research Triangle Park, North Carolina, USA.

Hong Gu, PhD, is an associate professor in the department of mathematics and statistics, Dalhousie University, Halifax, Nova Scotia, Canada.

RTI Press publication MR-0009-0903

Suggested Citation

Chakraborty, H. and Gu, H. (2009). A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values. RTI Press publication No. MR-0009-0903. Research Triangle Park, NC: RTI International. Retrieved [date] from http://www.rti.org/rtipress.

This publication is part of the RTI Press Methods Report series.

RTI International
3040 Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
Fax: +1.919.541.5985
E-mail: rtipress@rti.org
Web site: www.rti.org

doi:10.3768/rtipress.2009.mr.0009.0903

www.rti.org/rtipress

A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values

Hrishikesh Chakraborty and Hong Gu

Contents

Abstract

Missing values and dropouts are common issues in longitudinal studies in all areas of medicine and public health. Intent-to-treat (ITT) analysis has become a widely accepted method for the analysis of controlled clinical trials. In most controlled clinical trials, some patients do not complete their intended follow-up according to the protocol for a variety of reasons; this problem generates missing values. Missing values lead to concern and confusion in identifying the ITT population, which makes the data analysis more complex and challenging. No adequate strategy exists for ITT analyses of longitudinal controlled clinical trial data with missing values. Several ad hoc strategies for dealing with missing values for an ITT analysis are common in the practice of controlled clinical trials. We performed a detailed investigation based on simulation studies to develop recommendations for this situation. We compared sizes (type I errors) and power between some popular ad hoc approaches and the linear mixed model approach under different missing value scenarios. Our results suggest that, for studies with a high percentage of missing values, the mixed model approach without any ad hoc imputation is more powerful than other options.

RTI INTERNATIONAL

Introduction

In most longitudinal studies in all areas of medicine and public health, missing values and dropout are common. For a variety of reasons, some patients do not complete their intended follow-up according to protocol, and they are often described as having "dropped out" before the conclusion of the trial. This situation generates missing values for the study.

Missing data have three important implications for longitudinal studies. First, the dataset becomes unbalanced over time, which complicates choosing the right methods of analyses. Second, because of this missingness, some unavoidable loss in information reduces the efficiency of the study. Finally, the missing values introduce bias that can cause misleading inferences.

The main problem that arises with missing data is that the distribution of the observed data may not be the same as the distribution of the complete data. Some missingness may be unrelated to the observed or unobserved responses, some may be related to the observed data, some related to the unobserved data, and some to both. Little and Rubin[1] classified the missing value mechanisms as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR is a condition in which missing values are randomly distributed across all observations. MAR is a condition in which missing values are not randomly distributed across all observations but are randomly distributed within one or more subsamples. Under MCAR and MAR, the missing data mechanisms are often referred to as being "ignorable"; by contrast, the missing data mechanism NMAR is often referred to as "nonignorable."

Intent-to-treat (ITT) analysis has become a widely accepted method for the analysis of controlled clinical trials. ITT analysis, as suggested by Schwartz and Lellouch,[2] is a pragmatic approach to avoid bias in estimating the effect of treatment assignment in randomized clinical trials. ITT analysis compares the study groups in terms of the treatment to which they were randomly allocated, regardless of protocol deviations and participant compliance or withdrawal. Missing values can lead to problems in identifying the ITT population, which makes the data analysis more complex and challenging. No adequate strategy exists for ITT analyses of longitudinal controlled clinical trial data with missing values.

In this report, we perform a detailed investigation of simulation studies to develop recommendations for ITT analysis of longitudinal controlled clinical trial data with missing values. We compare estimates, sizes (type I errors), and power among several popular ad hoc approaches and the general linear mixed model (GLMM) approach for different proportions of missing values. We also try to answer a fundamental question faced by researchers in the presence of missing values: which method or which combination of imputation and analysis methods will provide the maximum benefits in terms of size and power of the test for different proportion of missing values in a longitudinal design setting? More specifically, does the mixed model provide more powerful tests without inflating the sizes in all missing scenarios in longitudinal design settings?

Methods

Many methods have been proposed and developed over the past two decades to deal with missing values and dropout issues. In a review paper, Fitzmaurice[3] summarizes methods for handling dropouts into four categories, namely the complete-case analysis, available-case methods, model-based approaches, and imputation methods. Except for complete-case analysis, the categories are all general terms that refer to a large collection of techniques.

Complete-case analysis refers to the method used to analyze the data that include only the complete cases; it excludes the subjects with missing data. Complete-case analysis, by definition, is not ITT analysis. By omitting all cases with missing values at any measurement occasion, this technique can result in a substantial loss of information, which has a negative impact on precision and power. Complete-case analysis is a valid method only when the dropouts are missing completely at random (MCAR).

Available-case methods refer to a collection of methods that can incorporate the repeated measurements of unequal length in the analysis.

Available-case methods use the available information to estimate means and covariances; thus, it requires a valid assumption about the missing mechanism. A simulation study by Touloumi et al.[4] compared the bias and efficiency of six different methods in this category.

The model-based approaches are the formal statistical methods and the most preferable methods to handle the nonignorable dropouts. However, model-based approaches have to be based on some assumptions that are generally not verifiable. Moreover, they are more difficult and time-consuming to perform than other approaches. For these reasons, model-based approaches are not widely used in clinical trial data analyses.

Imputation methods that fill data in missing values are of two main types: (1) fixed-value imputations and (2) multiple imputations. Both of these types of imputation methods have been applied to fulfill the goal of ITT analysis for longitudinal studies with dropouts. No single strategy is adequate for all combinations of dropouts. However, several ad hoc strategies for dealing with missing values are commonly used in clinical trials.

Fixed-value imputation methods substitute each missing or dropout value with a fixed value that is generated by one of the ad hoc strategies. One widely used ad hoc approach in clinical trial data analyses, which is "last observation carried forward" (LOCF), belongs to the fixed-value class of imputation methods. A simulation-based size and power comparison of various ad hoc strategies under different dropout mechanisms can be found in Unnebrink and Windeler.[5]

LOCF is the most popular method for ad hoc imputations; it fills the missing values with the last available non-missing values of the same subject. Best-value replacement (BVR) and worst-value replacement (WVR) are the other two ad-hoc imputation methods widely used in medical research. In these methods, the missing values are filled in with the best or worst values from that subject. The selection of best or worst value replacement method depends on the scientific knowledge of the variable; in most situations, they are the subject's minimum and maximum values.

Multiple imputation methods impute the missing values using a set of sampled values based on models for the missing data conditional on all relevant observed data. Thus, the multiple imputation methods account for the underlying uncertainty.[6,7] Little and Yau[8] applied multiple imputations in ITT analysis for longitudinal studies under several different assumptions and compared results with those from available-case methods.

In short, available methods to impute missing values are appropriate to use only under some assumptions about the missing data mechanism, and every method has drawbacks. All ad hoc methods require the MCAR assumption; maximum likelihood-based methods require either MCAR or MAR assumptions. If the missing values are NMAR, then standard methods of analysis are not valid and usually a sensitivity analysis is recommended.

The mixed model is the most important available-case method. In it, covariates can be either time-invariant or time-varying. In addition, generalizations exist for non-normal data. The mixed model equation for the kth subject can be written as

$$\mathbf{Y}_k = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \mathbf{d}_k + \mathbf{V}_k$$

with the assumptions that

$$\mathbf{d}_k \sim \text{NID}(\mathbf{0}, \Delta), \mathbf{V}_k \sim \text{NID}(\mathbf{0}, \sigma^2\mathbf{I}).$$

Thus, the covariance matrix is

$$\text{Var}(\mathbf{Y}_k) = \boldsymbol{\Sigma}_k = \mathbf{Z}_k \boldsymbol{\Delta} \mathbf{Z}_k' + \mathbf{V}_k.$$

The model can be constructed using the "usual" linear model method $E(\mathbf{Y}_k) = \mathbf{X}_k \boldsymbol{\beta}$, where:

$\mathbf{Y}_k$ = vector of all available measurements from the kth subject, through all periods of observation

$\mathbf{X}_k$ = fixed effects design matrix for the kth subject corresponding to the available measurements in $\mathbf{Y}_k$

$\boldsymbol{\beta}$ = fixed-effect parameters matrix for all subjects

$\mathbf{Z}_k$ = random effects design matrix for the kth subject

$\mathbf{d}_k$ = random coefficients for the kth subject; $\mathbf{d}_k$ contains *increments* to population intercepts and slopes

$\mathbf{V}_k$ = vector of random measurement errors for the kth subject.

The fixed-effect design matrix $\mathbf{X}_k$ and fixed-effect parameters $\boldsymbol{\beta}$ in the mixed model are similar to the design matrix and the regression parameters in a typical multiple regression, analysis of variance (ANOVA), or analysis of covariance (ANCOVA) model, where $E(\mathbf{Y}_k) = \mathbf{X}_k \boldsymbol{\beta}$. Thus, an element of $\boldsymbol{\beta}$ may represent the "slope" of a regression surface with respect to a covariate, to a treatment effect, or to a similar quantity.

The random-effects design matrix $\mathbf{Z}_k$ and the subject-specific random effects $\mathbf{d}_k$ represent random deviations from $E(\mathbf{Y}_k) = \mathbf{X}_k \boldsymbol{\beta}$ that are associated with the kth subject. $\mathbf{V}_k$ is a vector of random deviations, or "measurement errors" $E(\mathbf{Y}_k|\mathbf{d}_k)= \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \mathbf{d}_k$, the expected value for the kth subject. $\mathbf{V}_k$ is similar to the vector of random errors, usually denoted by ε_k *or* $\boldsymbol{e}_k$, in a multiple regression, ANOVA, or ANCOVA model.

The mixed model can accommodate different numbers of measurements, i.e., different lengths of $\mathbf{Y}_k$, for different subjects. Thus, investigators can base inference on all available measurements. The mixed model has four principal strengths: (1) it can accommodate missing data points often encountered in longitudinal datasets; (2) it does not need to have same number of observations per subject; (3) time can be continuous, rather than a fixed set of points; and (4) specification of the covariance structure among repeated measures can be flexible. These features thus provide a natural way to deal with missing values or dropouts.

This paper focuses on three commonly used ad hoc methods—namely, the LOCF, BVR, and WVR approaches—and mixed model, one of the more widely used available-case analysis methods in longitudinal studies with missing values. We compared the effects when data are imputed by the commonly used ad hoc imputations (i.e., LOCF, BVR and WVR). We use the mixed model approach to analyze the resulting data, based on all the available measurements. We then calculated empirical size and power using simulation techniques.

Analysis

Analyzing missing data requires making assumptions about the missing values as being either ignorable or nonignorable. For simplicity, we assume that all the missing values are ignorable in our simulation. Our simulation is based on a design that compares differences in lumber spine bone mineral density (BMD) values between two groups in a multicenter randomized clinical trial. We adopted this approach because the original drug trial design compared postmenopausal, non-osteoporotic women who received Drug A as treatment with similar women who received a placebo. For confidentiality reasons, we are not specifying the drug name or the manufacturer.

To determine the number of subjects required, we assumed that the variance-covariance matrices would be the same for both the drug A and the placebo groups of subjects. We also assumed that each subject was scheduled to make five visits over a 15-month period. Further, we assumed (a) that the baseline mean value of bone loss is 0.715 gHA/cm^2 for both treatment and placebo groups and (b) that at the end of the five visits, the placebo group's mean value for bone loss will have been the same (0.715) and the treatment group's mean value for bone loss will have increased to 0.745. In other words, we assumed that the *difference* in mean lumber spine BMD loss will be 0.030 at the end of five visits. We estimated that 138 subjects in each group would allow us to detect a lumber spine BMD difference of 0.03 with 90 percent power at 5 percent significance level for a one-sided test.

Our study goal is to simulate a real clinical trial scenario in which missing data occur. For these purposes, data are missing because women in both the treatment and the placebo groups randomly missed some visits at which data would otherwise have been collected. The pattern of the missing data is created in

the simulation following the key assumptions listed below:

1. Women are enrolled into five study centers.
2. Each center randomly assigns subjects to the treatment and placebo arms in a 1:1 ratio.
3. On average, each study center enrolls 55 patients, with a range of variation between 25 and 75 patients.
4. Each randomized patient has five visits over a 15-month period.
5. Twenty percent of the subjects drop out of the trial after the first visit.
6. Of the remaining four visits, between 5 percent and 50 percent are missed by the women remaining in the study.

We assigned missing visits randomly to each subject. We then simulated data from a multivariate normal distribution in which the variances of lumber spine BMD value is 0.005 for both groups for all visits and the lumber spine BMD values are positively correlated within each patient with a correlation coefficient of 0.7.

Initially, each simulated dataset contained 276 subjects—138 for the placebo group and 138 for the treatment group—with five visits by each subject. We generated dropouts, namely, the random 20 percent of women who dropped out after the first visit. We then generated missing data as the percentages of missed visits, ranging from 0 percent (i.e., no missed visits) to 50 percent (half of the visits were missed); we did this in 5-percentage-point increments (0 percent, 5 percent, …, 45 percent, 50 percent) for the subjects who had not dropped out of the trial after the first visit.

For size calculations, the mean difference between two treatments is 0. For power calculations, we repeated the analysis for 11 different mean differences at the end of five visits: 0.005, 0.010, 0.015, 0.020, 0.025, 0.030, 0.035, 0.040, 0.045, 0.050, and 0.055.

Therefore, initially we have 11 x 12 datasets (N = 132) with different combinations of mean differences and missing percentages. For example, dataset 1 has only 20 percent dropouts, with a mean difference of 0.03; dataset 2 has 10 percent missed visits in addition to 20 percent dropouts, with a mean difference of a 0.03; and so on. For each dataset we used three different ad hoc methods (LOCF, BVR, and WVR) to impute the missing values. For our analyses, for each patient we defined the smallest value as the worst value and the largest as the best value. Thus, from each of the 132 datasets we generated four different datasets: one that kept the missing values as missing and three others that had missing values imputed by LOCF, BVR, or WVR. In all, we generated a total of 528 different simulated datasets for all combinations.

We used a mixed model with an unstructured covariance matrix to compare the mean differences in lumbar spine BMD at the fifth visit between the drug and the placebo groups, for the original datasets with missing values and for all the other versions of imputed datasets. The fixed-effect design matrix includes drug groups (drug A and placebo), centers (1 to 5), and visits (1 to 5). The random-effects design matrix includes a random intercept and a slope for subjects.

We repeated this simulation process 5,000 times and recorded the test results with p values for all four analysis methods during each simulation run. We used SAS 9.2[9] to simulate and analyze the data.

Results

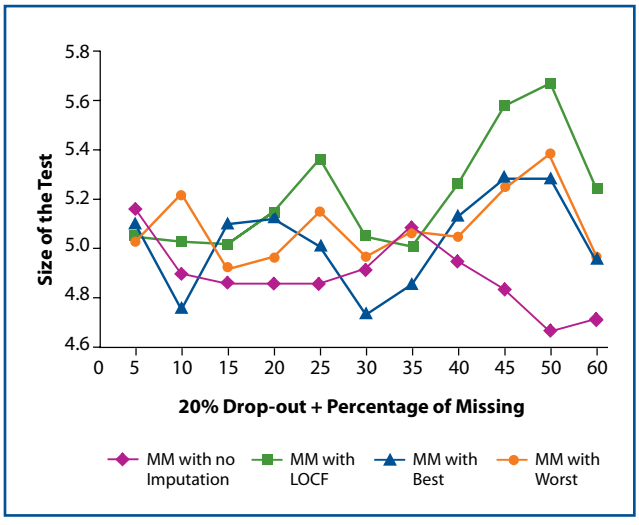
In general, as Figures 1 though 4 indicate, if the study is designed appropriately, a small percentage of missing values does not affect the estimates of size or power. For higher percentages of missing values, different imputation techniques yield various sizes and powers. The simulation results show that the mean BMD estimates do not differ by method.

Figure 1 compares sizes (i.e., type I errors) generated by four different methods for handling missing values for different percentages of missing values (20 percent dropouts plus 0 to 50 percent random missing). We observed no major size fluctuation in size attributable to different analysis methods for relatively low percentages of missing values (up to 30 percent). Conversely, for higher percentages of missing values, from about 35 percent to 50 percent, the size estimates become more unstable.

Figure 2 compares power by four different methods for 0.03 mean differences between two groups for ranges of missing percentages (20 percent dropouts plus 0 to 50 percent random missing). We observed that the power of the trial decreases as the percentage of missing values increases. However, the decrease of power is flat in the no-imputation method, as compared to the other three methods. For example, if the missing percentage is 20 percent dropouts plus 40 percent random missing values, then the power for the no-imputation method is 70 percent and the power of the other method is close to 50 percent.

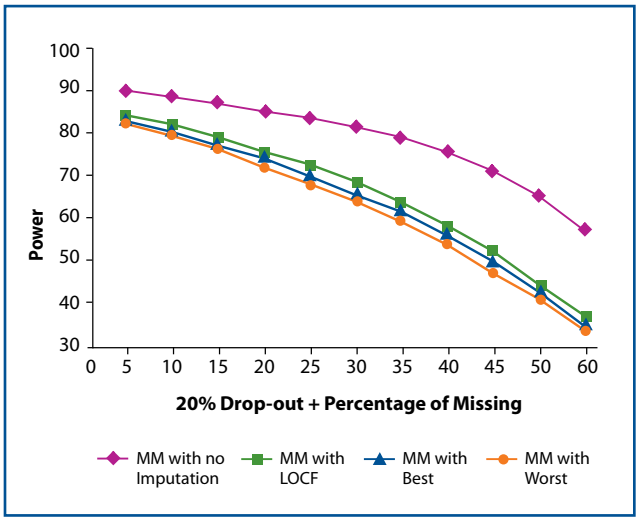
Figure 3 presents a comparison of powers for four different methods and a range of mean differences between two groups from 0.005 to 0.055; missing percentages include 20 percent dropouts, 20 percent dropouts plus 10 percent random missing, 20 percent dropouts plus 30 percent random missing, and

Figure 1: Comparisons of size (type I errors) of a mixed model by different percentages of missing data for different methods of handling missing data, assuming a mean difference of 0 in lumbar bone mineral density between groups



MM = mixed models
LOCF = last observation carried forward
Best = best value replacement for that subject
Worst = worst value replacement for that subject

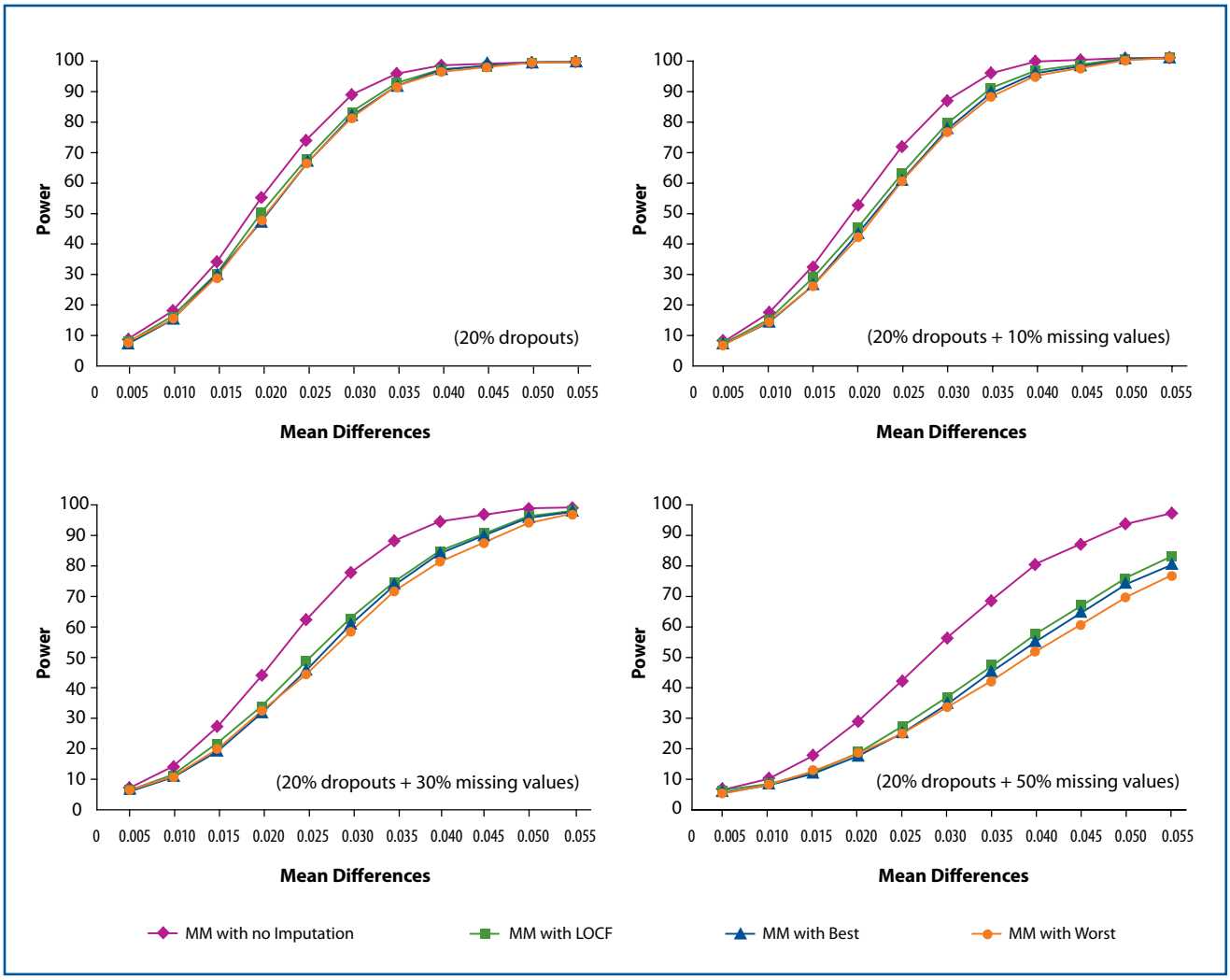
Figure 2: Power comparisons of a mixed model by missing percentages for different methods of handling missing data, assuming a fixed mean difference in lumbar bone mineral density of 0.03 between groups and a 20% dropout rate



MM = mixed models
LOCF = last observation carried forward
Best = best value replacement for that subject
Worst = worst value replacement for that subject

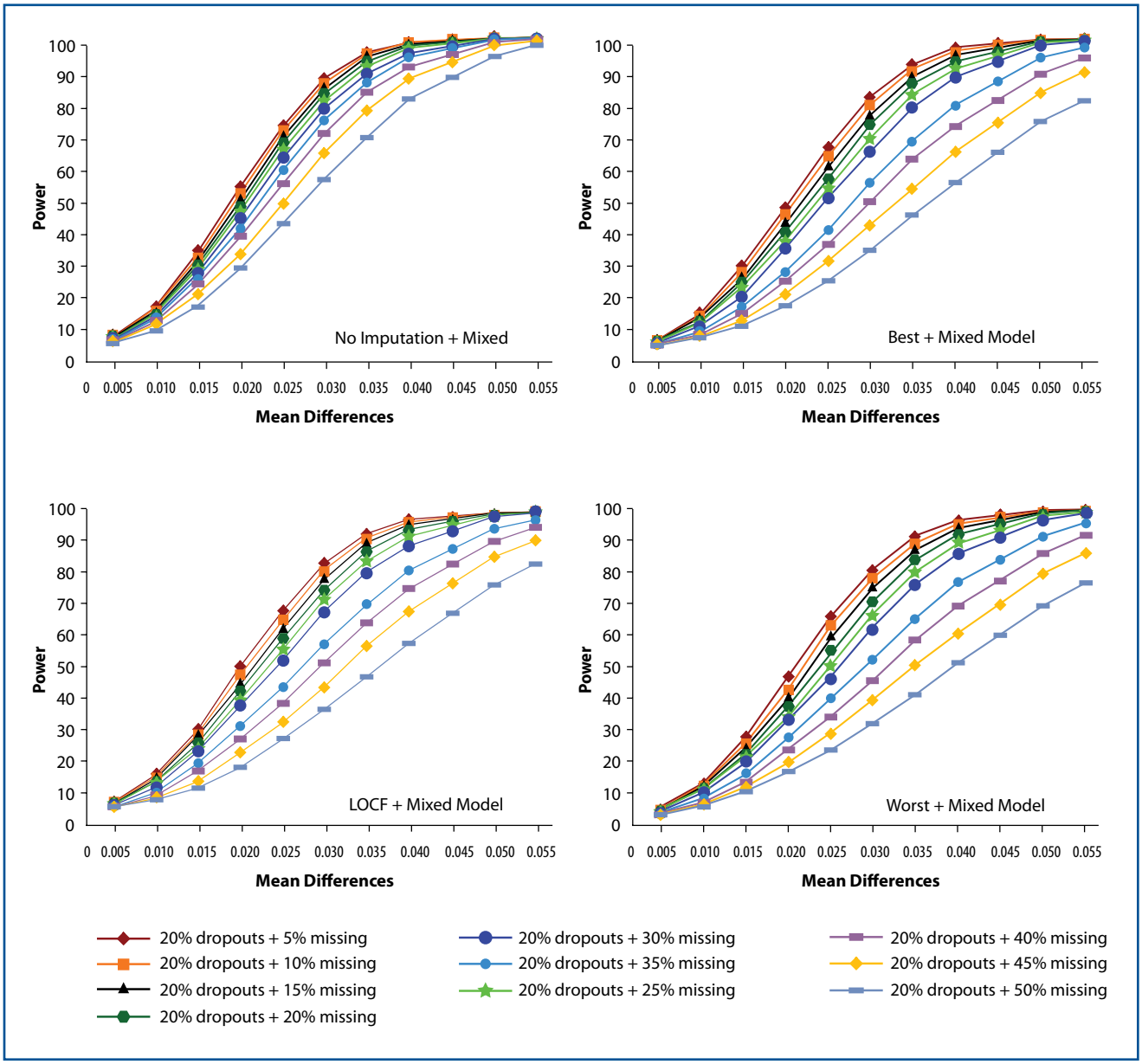
20 percent dropouts plus 50 percent random missing. Figure 4 also compares power for all four different methods and a range of mean differences from 0.005 to 0.055 and missing percentages of 20 percent dropouts plus 0 to 50 percent random missing. In both figures, we observed that if the differences between two groups are very small (say, for example, 0.005), there is no difference between methods irrespective of the percentage of missing values. We do not have adequate samples to detect small differences because we calculated our sample size to detect a difference of 0.3. On the other hand, if the differences between two groups are greater than 0.03, then we have enough power to detect the differences.

Figure 3: Power comparisons of a mixed model by mean differences in lumbar spine bone mineral density for different missing percentages across methods of handling missing data, assuming a 20% dropout rate



MM = mixed models; LOCF = last observation carried forward
Best = best value replacement for that subject; Worst = worst value replacement for that subject

Figure 4: Power comparisons by mean differences in lumbar bone mineral density for different methods of handling missing data across missing percentages, assuming a 20% dropout rate



MM = mixed models; LOCF = last observation carried forward

Best = best value replacement for that subject; Worst = worst value replacement for that subject

Discussion

Our simulations demonstrated that analysis with mixed models without any ad hoc imputation provides more powerful tests than does mixed model analysis with LOCF, BVR, or WVR ad hoc imputations in all missing scenarios. As the missing percentages rise, the power of the test decreases for all four types of analysis. Nevertheless, the rate of power decrease is slower for the analysis with mixed models without any ad hoc imputation than for the mixed model analysis when missing values are filled in by using LOCF, BVR, or WVR. If studies have missing values, then (as is well known) they will lose power. For any percentage of missing values, investigators will not get the power they designed during the planning phase unless they account for the missing values by appropriately increasing the sample sizes. However, in real life, predicting the number of missing values at the beginning of the study is very difficult. Thus, to preserve the power of the study, investigators should seek to strike a balance between cost and power in increasing their sample sizes.

During any study, investigators should also take other measures to minimize missing values by correcting the strategy for dealing with missing values in a blind review stage before final analysis. We propose that when investigators anticipate a good deal of missing values, they use a mixed model without ad hoc imputation as a method of analysis. When investigators expect only a few missing values, meaning that estimates are not expected to differ significantly by imputation methods, investigators can select any method to impute the missing values.

In summary, mixed model analysis without any ad hoc imputation always provides equal or more power than does analysis using mixed models with missing values imputed by LOCF, BVR, or WVR ad hoc imputation methods.

References

1. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987.
2. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis. 1967;20(8):637–48.
3. Fitzmaurice GM. Methods for handling dropouts in longitudinal clinical trials. Stat Neerl. 2003;57:75-99.
4. Touloumi G, Babiker AG, Pocock SJ, Darbyshire JH. Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study. Stat Med. 2001;20:3715-28.
5. Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Stat Med. 2001;20:3931-46.
6. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
7. Rubin DB, Schenker N. Multiple inputation in health-care databases: an overview and some applications. Stat Med. 1991;10:585-98.
8. Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. Biometrics. 1996;52(4):1324-33.
9. SAS Institute, Inc. SAS release 9.2. Cary, North Carolina: SAS Institute, Inc.; 2006.

www.rti.org/rtipress **RTI Press publication MR-0009-0902**