HANDLING MISSING DATA IN CLINICAL TRIALS: AN OVERVIEW

WILLIAM R. MYERS, PhD
Senior Statistician, Department of Biometrics and Statistical Sciences, Procter and Gamble Pharmaceuticals, Cincinnati, Ohio

A major problem in the analysis of clinical trials is missing data caused by patients dropping out of the study before completion. This problem can result in biased treatment comparisons and also impact the overall statistical power of the study. This paper discusses some basic issues about missing data as well as potential “watch outs.” The topic of missing data is often not a major concern until it is time for data collection and data analysis. This paper provides potential design considerations that should be considered in order to mitigate patients from dropping out of a clinical study. In addition, the concept of the missing-data mechanism is discussed. Five general strategies of handling missing data are presented: 1. Complete-case analysis, 2. “Weighting methods,” 3. Imputation methods, 4. Analyzing data as incomplete, and 5. “Other” methods. Within each strategy, several methods are presented along with advantages and disadvantages. Also briefly discussed is how the International Conference on Harmonization (ICH) addresses the issue of missing data. Finally, several of the methods that are illustrated in the paper are compared using a simulated data set.

Key Words: Clinical trials; Missing data; Dropouts; Imputation methods; Missing-data mechanism

INTRODUCTION

A PRIMARY CONCERN when conducting a clinical trial is that patients will drop out (or withdraw) before study completion. The reason for withdrawal may be study-related (eg, adverse event, death, unpleasant study procedures, lack of improvement) or unrelated to the study (eg, moving away, unrelated disease). This problem is especially prevalent in clinical trials where a slow-acting treatment or a drug that may be intolerant is being investigated. Dropouts in clinical trials can produce biased treatment comparisons and reduce the overall statistical power. This paper will focus on the case where missing data occur as a result of patients dropping out of the study. More specifically, it focuses on the case in which a patient’s missing response at assessment time t implies it will be missing at all subsequent times. This is termed a monotone pattern of unit-level missing data (1). An example where missing data deviate from the aforementioned pattern is in the case of health related quality-of-life research, where a patient does not answer an item (or question) within a questionnaire, but does not necessarily drop out of the clinical trial.

There are numerous issues one must consider when confronted with a data set where patients have dropped out of the clinical study. First, compliant patients often have a
better response to treatment than noncompliant patients, even in the placebo group. Therefore, the fully compliant subgroup is not a random subsample of the original sample. In addition, the pattern of selection might differ between the placebo group and the active treatment group. It can be problematic if the rate, time to, and reason for withdrawal differ widely among treatment groups. Also, the last observed response for an early drop-out often does not reflect the potential benefit of a drug with a slow onset of action. Furthermore, there are situations where a patient may drop out of a clinical study because of early recovery. He/she may then suffer a relapse. Many of these situations will result in biased treatment comparisons.

Unstated reasons for dropping out of a clinical trial may be associated with the last observed response or other study-related reasons. This will be a primary concern for those methods that incorporate the *reason for dropout* in the statistical analysis. It is imperative to have accurate documentation of the cause of dropout. The method for handling dropouts in clinical studies will often depend on the objective of the study. On the one hand, the goal may be more explanatory in nature, as in the case of a Phase I or early Phase II study, where the true pharmacological properties of a drug are being investigated. On the other hand, it may have a more pragmatic objective, where a pivotal Phase III study is providing an overall evaluation of treatment policy in clinical practice.

**THE MISSING-DATA MECHANISM**

A concept that is often discussed when missing data occur is the missing-data mechanism. Little (2) has also used the term *dropout mechanism* when it relates to patients dropping out of a clinical study prematurely. These two terms will be used synonymously throughout the paper. Little and Rubin (3) classify the missing-data mechanism into three basic categories. They define *missing completely at random* (MCAR) as the process in which the probability of dropout is independent of both observed measurements...
Handling Missing Data in Clinical Trials

(eg, baseline covariates, observed responses) and unobserved measurements (those that would have been observed if the patient had stayed in the study). Under MCAR the observed responses form a random subsample of the sampled responses. When data are MCAR there is no impact on bias and therefore most standard approaches of analysis are valid. A loss of statistical power, however, can still occur. The MCAR assumption can be assessed by comparing the distribution of observed variables between dropouts and nondropouts (4). If no significant differences are found with respect to the variables, then there is no apparent evidence that the data from the clinical trial are representative only of completers. The MCAR assumption is often not plausible in clinical trials (1).

The second classification, which is more restrictive than MCAR, is missing at random (MAR). Under MAR, the probability of dropout depends on the observed data, but not the unobserved data. When the response is MAR the observed responses are a random subsample of the sampled values within a subclass defined by the observed data (eg, age). In this case, the missing-data mechanism can be considered ignorable (5,6). Laird (6) points out that ignorable missingness is plausible in longitudinal studies, when patient withdrawal is related to a previous response. In addition, Murray and Findlay (7) point out that if a protocol removes patients from a study because they reach a threshold for the response (eg, diastolic blood pressure exceeding 110 mmHg) then the data are MAR. The rationale for the MAR assumption here is traced to the fact that the lack of response can be deduced from the recorded value. Little (2) also discusses a special case of MAR dropout which is not discussed extensively in the literature. For this case the dropout mechanism depends only on the covariates X and is classified as covariate-dependent dropout. For this missing-data mechanism Diggle and Kenward (8) used the term completely random dropout. Murray and Findlay (7) state that an observation is MAR if the fact that it is missing is not in itself informative. Lavori et al. (1) point out that the MAR assumption is inherently untestable. Therefore, one can never truly achieve complete certainty that conditioning on observed variables achieves ignorable dropout.

If the dropout mechanism is neither MCAR or MAR then it is nonignorable. In this case, the dropout mechanism needs to be incorporated into the analysis. This paper will briefly discuss, in the following section, available methods in the case of nonignorable dropout. Diggle (9) proposed a method for testing for “random dropouts” in repeated measures data. Diggle uses the term “random dropout” in a more restrictive sense than Rubin’s “missing at random.” In addition, Park and Davis (10) proposed a method for testing the missing-data mechanism for repeated categorical data.

**GENERAL STRATEGIES FOR STATISTICAL ANALYSIS**

Much of the literature involving missing data (or dropouts) in clinical trials pertain to the various methods developed to handle the problem. This paper will divide those methods into the following five basic classifications:

1. Complete-case analysis,
2. “Weighting methods,”
3. Imputation methods,
4. Analyzing data as incomplete, and
5. “Other” methods.

The category defined as “other” methods includes those methods that do not logically fit into the other four classifications.

Complete-case methods use only those patients with complete data. For example, in a longitudinal study, a complete-case analysis will use only those patients who have observed responses at each scheduled time point. Another example is the case of a single-endpoint study where a regression analysis is used. Only those patients who have the observed endpoint and observed values for all relevant covariates are involved in the analysis. An obvious advantage of this type
of analysis is ease in implementation. In addition, it provides valid results in the case of MCAR. There are numerous disadvantages, however, to excluding patients with incomplete data. First, the complete-case method provides inefficient estimates, that is, loss of statistical power. If the dropout mechanism is not MCAR, then the analysis can produce biased treatment comparisons. In the case of a clinical trial with longitudinal measurements, it is typically not good practice to “throw out” data. Maybe the most important concern with this type of analysis is that it does not follow the “Intention-to-Treat” paradigm. The following simple example demonstrates the limitations of the complete-case analysis. Suppose Treatment X is modestly effective for patients regardless of the baseline severity of their condition. On the other hand, Treatment Y provides a benefit for the less severe patients, while providing no improvement for the more severely ill patients. If the patients have a tendency to drop out before completion because of lack of efficacy, then the complete-case analysis may unduly favor Treatment Y.

A second strategy to handle missing data is weighting methods. Some may actually consider this a form of imputation. The general idea is to construct weights for complete cases in order to reduce or remove bias. Little and Schenker (11) discuss the basic concept of weighting adjustments in the sample survey setting. Heyting et al. (12) describe the heuristic appeal of this method as follows. Each patient belongs to a subgroup in the patient population in which all patients have a similar baseline and response profile. A proportion within each subgroup are destined to complete the clinical trial, while the remainder are destined to drop out early. Those “completer” patients with a very low probability of completing can certainly have an overly strong influence on the results. Heyting et al. (12) provide a particular weighting method where the evaluation of the mean treatment differences at the end of the study is of primary interest. The authors’ primary objective was explanatory in nature. Robins et al. (13) introduced a weighting method which allows generalized estimating equation (GEE) analyses to be correct under the MAR assumption.

A third classification of statistical analyses to handle missing data is imputation methods. Imputation is any method whereby missing values in the data set are filled with plausible estimates. The choice of plausible estimates is what differentiates the various imputation methods. The objective of any imputation method is to produce a complete data set which can then be analyzed using standard statistical methods.

The Last Observation Carried Forward (LOCF) method is a commonly used imputation procedure. This method is implemented when longitudinal measurements are observed for each patient. LOCF takes the last available response and substitutes the value into all subsequent missing values. The LOCF can be problematic if early dropouts occur and if the response variable has expected changes over time. It can provide biased treatment comparisons if there are different rates of dropout or different time to dropout between the treatment groups. For example, LOCF can provide conservative results with respect to active treatment if placebo patients drop out early because of lack of efficacy. In this case, the mean placebo response is biased upward. On the other hand, when the active treatment is slowing down the severity of an illness and if those patients in the active treatment group are dropping out early due to intolerability, the LOCF method can render anti-conservative results with respect to active treatment.

Another type of imputation method is a “worst case” analysis. This analysis imputes the worst response observed among the active treatment group for those missing values within the active treatment group. For the placebo group, the method imputes the best observed response among the placebo group for those missing values within the placebo group. This particular analysis can be viewed as a type of sensitivity analysis (14). From a purely statistical perspective this type of method can increase the overall variability, bias the active treatment mean downward, and
bias the placebo mean upward. This could potentially limit a promising treatment from demonstrating efficacy. In many cases, however, this method is usually not the planned primary analysis, but rather a secondary analysis. It can be used to assess the robustness of the results and provide a so-called "lower bound" on treatment efficacy. If the "worst case" analysis demonstrates a treatment benefit it can be a very powerful result. For example, one could state either that the treatment efficacy is so strong that even imputing the worst case scenario does not alter the positive results or the missing response rate is so low that it does not alter the conclusions.

Brown (15) proposed a slightly different method than that discussed above. A predetermined percentile (e.g., median) of the placebo group is assigned to all those patients who dropped out from either the placebo group or the active treatment group. This predetermined score is assigned to all patients in both groups with values worse than the assigned score. A Mann-Whitney statistic is used to test the equality of the distribution of the two groups and thus provides a bound on the test of the efficacy for the treatment. There are other single imputation methods such as mean imputation, conditional mean imputation, and the Hot deck method. Little and Rubin (3) discuss these methods as well as other single imputation methods. Paik (16) proposed a mean imputation method as well as a multiple imputation method for handling missing data in GEE analysis. Each method provides valid estimates when data are MAR.

One particular imputation method that has received a significant amount of attention in the literature recently is multiple imputation. The idea of multiple imputation, first proposed by Rubin (17), is to impute more than one value for the missing item. The advantage of multiple imputation is that it represents the uncertainty about which value to impute. This is as opposed to imputing the mean response which does not incorporate the degree of uncertainty about which value to impute. Therefore, the analyses that treat singly imputed values just like observed values generally underestimate the variability (18). Multiple imputation can be implemented for either longitudinal measurements or a single response.

The general strategy for multiple imputation is to replace each missing value with two or more values from an appropriate distribution for the missing values. This produces two or more complete data sets. Repeated draws are made from the posterior predictive distribution of the missing values, \( \hat{Y}_{\text{miss}} \). As Rubin and Schenker (18) point out, in practice, implicit models can be used in place of explicit models. Lavori et al. (1) discuss a propensity-based imputation where one models the probability of remaining in the study given a vector of observed covariates. A logistic regression model is typically used. This method stratifies patients into groups based on propensity scores, that is, the propensity to drop out of the study. Imputations are made by the approximate Bayesian bootstrap. One first draws a potential set of observed responses at random with replacement from the observed responses in the propensity quintile. Then the imputed values are chosen at random from the potential observed sample. This process is performed \( m \) times in order to produce \( m \) complete data sets. Analyses for each complete data set are then combined in a way that reflects the extra variability. The total variability consists of both within and between imputation variability. Multiple imputation methodology relies on the MAR assumption.

Little and Schenker (11) and Rubin (19) indicate that typically only a few imputations (\( m = 3-5 \)) are necessary for a modest amount of missing information (e.g., < 30%). The previous authors do point out, however, that as the percentage of missing data increases more imputations will be necessary. The software SOLAS for Missing Data Analysis (20) can implement the previously discussed multiple imputation methods as well as other single imputation methods. A good list of references on this topic include Lavori et al. (1), Little and Schenker (11), Rubin and Schenker (18), and Rubin (19).

A fourth strategy to handle missing data is to analyze data as incomplete. This is typi-
cally performed in longitudinal studies. One option is to use summary statistics. In longitudinal studies a slope estimate could be used to summarize each patient’s response. Early dropouts could obviously be problematic with this type of analysis. Likelihood methods that ignore the drop-out mechanism are also a popular analysis of choice. Most software packages (eg, PROC MIXED in SAS®) assume MAR and ignore the missing-data mechanism. In the case of nonignorable dropout, inferences based on likelihood methods that ignore the dropout mechanism may produce biased results. Little and Rubin (2) and Little (3) discuss nonignorable missing-data models. Implementing a nonignorable missing-data model is not trivial. Little (3) indicates that results can be sensitive to misspecification of the missing-data mechanism and if little knowledge is known about the mechanism then sensitivity analysis should be performed. In longitudinal studies where non-Gaussian responses are measured (eg, binary or count data) GEE is often used. GEE generally requires MCAR or covariate-dependent dropout in order to yield consistent estimates (2).

The fifth and final category of analyses is defined as “other” methods. As stated earlier, these are methods that do not fall into any of the four previously discussed classifications. One of the methods, which was proposed by Gould (21), converts all information on an outcome variable into ranking of patients in terms of desirability outcome. These desirability outcomes are ordered from the least desirable (eg, early withdrawal due to lack of efficacy or intolerance) to the most desirable (eg, early withdrawal due to efficacy). Between the two ends of the desirability spectrum are the ranks of the scores for those patients who complete the study. A standard two-sample test based on ranks could then be implemented. This particular method excludes those patients whose reason for dropout is unrelated to the study. The utility of this approach certainly requires a clear understanding of the reason for dropout. This method does not directly address the issue of estimation. Cornell (22) proposed a modification of Gould’s method by taking into account the time to dropout in order to provide a finer measure for the desirability outcome. For example, patients who drop out very early in the study due to intolerability would be assigned a lower desirability outcome than those patients who dropped out later in the study due to intolerability.

Shih and Quan (23) proposed a method that they defined as a “composite approach.” They consider the situation where the outcome measure is a continuous response variable and the final outcome is of main interest. Shih and Quan (23) indicate that the relevant clinical question when dropouts occur is what is the chance that a patient completes the prescribed treatment course and if he/she does complete the therapy, what is the expected response? This results in two hypotheses. The first is the probability of dropping out of the clinical study and the second is the expected response for those patients who completed the clinical study. A logistic regression model could be used to model the probability of dropping out of the study, while an analysis of variance could be implemented for the expected response of completers. Shih and Quan (23) point out that the alternative hypotheses need to be in the same direction. For example, the placebo has a greater percent of patients dropping out for unfavorable reasons compared to active treatment and the mean response for completers is favorable for active treatment. The proposed method first tests the joint hypothesis. If that is rejected the individual hypotheses are subsequently tested. A closed testing procedure is used in order to control the overall type I error rate. This particular method makes no MAR type assumption. The following are a few papers that discuss some additional methods: Dawson and Lagakos (24) and Shih and Quan (25).

ICH GUIDELINES
The International Conference for Harmonization (ICH) guidelines “E9 Statistical Principles for Clinical Trials” address the issue of missing data. The guidelines indicate that
methods for dealing with missing data should be predefined in the protocol. The guidelines also point out that methods for dealing with missing data can be refined in the statistical analysis plan during the blind review of the data. This is a very important step to consider, given that it can be difficult to anticipate all potential missing-data problems that could occur. Probably the most important suggestion the ICH guidelines make, however, is to investigate the sensitivity of the results of the analysis to the method of handling missing data, that is, sensitivity analysis.

EXAMPLES BASED ON A SIMULATED DATA SET

The following examples are not intended to be an extensive simulation study, but rather some simple scenarios based on simulated data sets. They are presented to demonstrate the potential advantages or disadvantages of some of the methods under certain dropout scenarios. The examples will consider a study where a single endpoint is of primary interest. The following two treatment scenarios are considered:

1. The active treatment is known to be significantly superior to placebo \( \mu_T = 15 \) units and \( \mu_P = 10.5 \) units, and
2. The active treatment is known not to be superior to placebo \( \mu_T = 16 \) units and \( \mu_P = 15 \) units.

Three hundred patients were initially randomized to each treatment group with a known standard deviation of 19 units. The first scenario provides approximately 80% statistical power, while the second scenario provides approximately 10% statistical power.

After the complete data sets were created, patient observations were deleted based on the following two dropout scenarios. First, patients in the active treatment group who were poor responders and who had significant side effects have a higher dropout rate (60%) than the rest of those in either the active treatment group or the placebo group (20%). Side effects were classified into three categories (mild, moderate, severe). The “high dropout group” was those with severe side effects and a response 0.5 standard deviations below the mean response for the active treatment group or those patients with moderate side effects and a response 1.0 standard deviation below the mean response for the active treatment group. The second scenario is where low responders, irrespective of treatment, have a higher dropout rate (60%) than the rest of the patients (20%). For this case, the “high dropout group” was those patients who have a response that is 1.0 standard deviation below the mean response of the placebo group.

Of the four possible treatment/dropout scenarios, this paper focuses on two of the more interesting cases. The first is where the active treatment is known not to be significantly superior to placebo and where there is a higher dropout rate for the active treatment group (Scenario 1). The second is where the active treatment is known to be superior to placebo and there is a higher dropout rate among the placebo group (Scenario 2). The three methods that were applied were the complete-case analysis, multiple imputation method, and the “composite approach.”

The first row of Table 1 provides the results when analyzing the complete data set. The other rows provide the results for the three methods. For Scenario 1, the complete-case analysis, as expected, provided a treatment mean that was biased upward and thereby erroneously demonstrated efficacy of a nonefficacious treatment. On the other hand, the multiple imputation method provided mean values and results that more closely resembled the truth. For the “composite approach” the differences for each hypothesis were in different directions (the active treatment group has a higher dropout rate for unfavorable reasons, while the mean response for the completers favors the active treatment group), therefore, it was not reasonable to combine the tests and perform the analysis. For Scenario 2, the complete-case analysis, as expected, provided a placebo mean that was biased upward and thereby
was unable to demonstrate that an efficacious treatment was effective. The multiple imputation method provided results that more closely mimicked the complete data set, even though it was not significant at the 0.05 significance level. The “composite approach” was unable to demonstrate a statistically significant result with the joint hypothesis, thus, the individual hypotheses were not tested.

**SUMMARY/DISCUSSION**

The objective of the paper was to present an overview of issues, concerns, and available methodology in the case of missing data as a result of patients dropping out of a clinical trial. It should be emphasized that sophisticated statistical analysis is no substitute for a good clinical plan in order to mitigate patients dropping out of a study. It is important to continue following patients even after they have dropped out of a clinical study. In addition, understanding both the disease and the therapy being studied can be helpful in selecting an appropriate statistical method.

The choice of a particular method for handling missing data depends on whether one is considering a more pragmatic or a more explanatory perspective. There is often the question of whether there are too many missing data. Spriet and Dupin-Spriet (14) point out that the tolerable amount of missing data is that which would not conceal an effect in the opposite direction. They go on to add that in order to detect whether this level of missing data has been reached one can perform what was earlier called the “worst case” analysis.

A major part of the paper was dedicated to the method of multiple imputation. It appears that regulatory agencies are still uncertain about the degree of usefulness of multiple imputation methods. Shih and Quan (23) provide several examples by which making inferences on the complete-data parameter may not always be of practical interest. For example, when a patient dies before the end of the study, the outcome measure for the end of the study simply does not exist. Another example is when the glomerular filtration rate in a renal disease study is not legitimate for patients who drop out due to renal failure and who are referred for kidney dialysis. They continued to point out that one should not estimate their nonexisting/missing values that would have been observed if the censoring event had not occurred. As Lavori et al. (1) point out, further work should be performed to better understand how well multiple imputation works under various types of dropout mechanisms, most specifically the nonignorable case. Obviously, it is very important to clearly understand the limitations of the various methods. This paper attempted to outline many of these limitations. This directly leads to the utility of performing some form of sensitivity analysis and how necessary and valuable it is. In conclusion,
it is important not to necessarily consider the various methods that handle dropouts (missing data) as rivals but rather consider them as methods that can compliment one another.

REFERENCES