1

What’s the problem with missing data?

Michael O’Kelly and Bohdana Ratitch

“For when they reach the scene of crime – Macavity’s not there!”

Macavity the Mystery Cat, TS Eliot*

Key points

• Missing data for the purposes of this book are data that were planned to be recorded during a clinical trial but are not available. Non-monotone or intermediate missing data occur when a subject misses a visit but contributes data at later visits. Monotone missing data, where all data for a subject is missing after a certain time-point due to early withdrawal from the study, is the more serious problem in interpreting the results of a trial.

• The most important thing about missing data is that it is missing: we can never be sure whether the assumptions made about it are true.

• An example illustrates the potential bias of using only observed data in an analysis (a favorable subset of subjects); and of using a subject’s last available

*Reproduced by permission of Faber & Faber and Houghton Mifflin Harcourt.


CLINICAL TRIALS WITH MISSING DATA

observation or baseline observation in place of missing values (bias varies and may be difficult to predict).

- Assuming that data are missing at random (i.e., that given the data and the model, missingness is independent of the unobserved values) allows one to use study data to infer likely values for missing data, but is likely biased in that it assumes that subjects who withdrew from the study have results like similar subjects who remained in the study.

- Given that we can never be sure whether the assumptions made about missingness in the primary analysis are true, sensitivity analyses are needed to stress-test the trial results for robustness to assumptions about missing data: sensitivity analyses will help the reader of the clinical study report to assess the credibility of a trial with missing data.

1.1 What do we mean by missing data?

This book is about missing data in clinical trials. In a clinical trial, missing data are data that were planned to be recorded but are not present in the database. No matter how well designed and conducted a trial is, some missing data can almost always be expected. Missingness may be absolutely unrelated to the subject’s medical condition and study treatment. For example, data could be missing due to a human error in recording data; due to a scheduling conflict that prevented the subject from attending the study visit; or due to a subject’s moving to a region outside of the study’s remit. On the other hand, data may be missing for reasons that are related to subject’s health and the experimental treatment he/she is undergoing. For example, subjects may decide to discontinue from study prematurely if their condition worsens or fails to improve, or if they experience adverse reactions or adverse events (AEs). A contrary situation is also possible, although probably less common, where a subject is cured and observations are missing because the subject is not willing to bother with the rest of the study assessments. Apart from missingness due to missed visits, missing data can arise simply due to the nature of the measurement or the nature of the disease. An example of data that would be missing because not meaningful is a quality-of-life score for a subject who has died. Those cases where missingness is related to the subject’s underlying condition and study treatment have the greatest potential to undermine the credibility of a trial. Sometimes, a subject’s data collected prior to discontinuation reflects the reason for withdrawal (e.g., worsening, improvement or toxicity), but subjects can also discontinue without providing that crucial information that would have enabled us to assess the reason for missingness and thus incorporate it in our analysis. Such cases potentially hide some important information about treatment efficacy and/or safety, without which study conclusions may be biased.

When a subject has provided data over the course of the study, but some assessments, either in the middle of the trial or at the primary time point, are missing for any reason, their data can be referred to as partial subject data. In this book, we explore the implications of this partial data and ways to minimize the potential bias.
WHAT'S THE PROBLEM WITH MISSING DATA?

In many clinical trials, collected data are longitudinal in nature, that is, data about the same clinical parameter is collected on multiple occasions (e.g., during study visits or through subject diaries). In such studies, a primary endpoint (clinical parameter used to evaluate the primary objective of the trial at a specific time point) is typically required to be measured at the end of the treatment period or a period at the end of which the clinical benefit is expected to be attained or maintained, with assessments performed at that point as well as on several prior occasions, thus capturing subject’s progress after the start of the treatment. This is in contrast with another type of trial, where the primary endpoint is event-driven, for example, based on such events as death or disease progression. In this book, we focus primarily on the former type of the trials, and we look at various ways in which partial subject data can be used for analysis.

Most of this book is about ways to handle missing data once it occurs, but it is also important to prevent missing data insofar as this is possible. Chapter 2 discusses this in detail, and describes some ways in which the statistician can contribute to prevention strategies. We now put some of the discussion above somewhat more formally.

1.1.1 Monotone and non-monotone missing data

A subject who completes a clinical trial may have data missing for a measurement because he/she failed to turn up for some visits in the middle of the trial. Such a measurement is said to have “non-monotone missing,” “intermediate missing” or “intermittent missing” data, because the status of the measurement for a subject can switch from missing to non-missing and back as the patient progresses through the trial. In many clinical trials, this kind of missingness is more likely to be unrelated to the study condition or treatment. However, in some trials, it may indicate a temporary but important worsening of the subject’s health (e.g., pulmonary exacerbations in lung diseases).

In contrast, monotone missingness occurs when data for a measurement is not available for a subject after some given time point; in the case of monotone missingness, once a measurement starts being missing, it will be missing for the subsequent visits in the trial, even though it had been planned to be collected. Subjects that discontinue early from the study are the usual source of monotone missing data. In most trials, the amount of monotone missing data is much greater than the amount of non-monotone missing data. In trials where the primary endpoint is based on a measurement at a specific time point, prior intermittent missing data will have a smaller impact on the primary analysis, compared to monotone missing data. Nevertheless, even in these cases, non-monotone missing data can affect study conclusions. This can happen if the intermediate data are utilized in a statistical model for analysis – the absence of such intermediate data may bias the estimates of the statistical model parameters. In this book, however, we will focus mostly on the problem of monotone missing data, because monotone missing data tend to pose more serious problems than non-monotone when estimating and interpreting trial results. For a more detailed discussion of handling non-monotone missing data, see Section 6.2.1. In this chapter, to introduce some of the concepts and problems in handling missing data, we will look at some common methods of handling monotone missing data in clinical trials, and examine the implications of each method.
4 CLINICAL TRIALS WITH MISSING DATA

In Section 4.2.1, we will also briefly discuss situations where subject discontinues study treatment prematurely, but may stay on study and provide data at the time points as planned originally, despite being off study treatment. These cases need special consideration when including data after treatment discontinuation in the analysis, so that the interpretation of results takes into account possible confounding factors incurred after discontinuation (e.g., alternative treatments).

1.1.2 Modeling missingness, modeling the missing value and ignorability

In the missing data methodology, we often use two terms: missing value and missingness (or missingness mechanism). It will be helpful to clarify what these terms refer to as they both play important and distinct roles in the statistical analysis. Missing value refers to a datum that was planned to be collected but is not available. A datum may be missing because, for example, the measurement was not made or was not collected. Missing and non-missing data may also be referred to as unobserved and observed, respectively. Missingness refers to a binary outcome (Yes/No), that of the datum being missing or not missing at a given time point. Missingness mechanism refers to the underlying random process that determines when data may be missing. In other words, missingness mechanism refers to the probability distribution of the binary missingness event(s). The missingness mechanism may depend on a number of variables, which themselves may be observed or not observed. In the analysis, we can use one model (often referred to as a substantive model) for the values of the clinical parameter of interest (some values of which in reality will be missing), and another model for the distribution of a binary missingness indicator variable (datum missing or not). The missingness model may not be of interest in itself, but in some situations it may influence estimation of the substantive model and would need to be taken into account in order to avoid bias. Some analyses make use of both of these models.

1.1.3 Types of missingness (MCAR, MAR and MNAR)

The classifications of missing data mechanisms introduced by Rubin (1976; 1987) and Little and Rubin (2002) provide a formal framework that describes how missingness mechanism may affect inferences about the clinical outcome. A value of a clinical outcome variable is said to be missing completely at random (MCAR) when its missingness is independent of observed and unobserved data, that is, when observed outcomes are a simple random sample from complete data; missing at random (MAR) when, given the observed outcomes and the statistical model, missingness is independent of the unobserved outcomes; and missing not at random (MNAR) when missingness is not independent of unobserved data, even after accounting for the observed data. When data are missing for administrative reasons, the missingness mechanism could be MCAR, because the reason for missingness had nothing to do with the outcome model and its covariates. Dropout due to previous lack of efficacy could be MAR, because in some sense predictable from the observed data in the model. It is important to note that MAR is not an intrinsic characteristic of the data
WHAT'S THE PROBLEM WITH MISSING DATA? 5

or missingness mechanism itself, but is closely related to the analysis model: if we include all the factors on which missingness depends in our model, we will be operating under MAR; otherwise, our analysis would not conform to MAR assumptions. Dropout after a sudden unrecorded drop in efficacy could be MNAR, since missingness would be dependent on unobserved data and would not be predictable from the observed data alone. Of these assumptions, MCAR is strongest and least realistic; while MNAR is the least restrictive. However, the very variety of assumptions possible under MNAR may be regarded as a problem: it has been argued that it would be difficult to pre-specify a single definitive MNAR analysis (Mallinckrodt et al., 2008).

We can test for dependence of missingness on observed outcomes, and so test for MAR versus MCAR. However, we cannot test whether the mechanism is MAR versus MNAR, because that would require testing for a relationship between missingness and unobserved data. Unobserved data, we think it is no harm to repeat, is not there, and so the relationship with missingness cannot be tested.

See Appendix 1.A at the end of this chapter for formal definitions of MCAR, MAR and MNAR.

Under some assumptions, missingness can be shown to be ignorable. Missingness is classified as ignorable if a valid estimate of the outcome can be calculated without taking the missingness mechanism into account. In his first paper addressing the problem of missing data, Rubin (1976) showed that, when using Bayesian or direct likelihood methods to estimate any parameter \( \theta \) related to the clinical outcome, missing data are ignorable when the missingness mechanism is MAR and \( \theta \) “is ‘distinct’ from” the parameter of the missing data process (missingness mechanism). Rubin put the word “distinct” in quotation marks because the distinctness condition is a very particular one. The missingness parameter is distinct from \( \theta \) “if there are no a priori ties, via parameter space restrictions or prior distributions, between (the missingness parameter) and \( \theta \).” Thus while we might often expect the same observed data to contribute to the modeling of both missingness and the outcome, \( \theta \) and the missingness parameter will still probably be distinct in such cases, and the missingness ignorable.

1.1.4 Missing data and study objectives

Clinical trial researchers and regulatory authorities are concerned about the effect of missing data on two aspects of the clinical data analysis: the estimates of the difference between experimental and control treatments, and the variance of this estimate. With respect to the difference between treatments, missing data can affect (and can bias) the magnitude of that estimate and make the experimental treatment look more or less clinically efficacious than it is (in the extreme cases even reverse a true comparison) or obscure an important interplay between treatment efficacy and tolerability. With regard to the variance of this estimate, missing data can either compromise the power of the study or, on the contrary, lead to underestimation of the variance, depending on the method chosen for analysis. Regulatory authorities require reasonable assurances that the chosen method of analysis in the presence of missing data is not likely to introduce an important bias in favor of the experimental treatment and will not underestimate the variance.
CLINICAL TRIALS WITH MISSING DATA

1.2 An illustration

To start our exploration of missing data, consider the following illustrative dataset that is patterned after typical Parkinson’s disease clinical data as available, for example, in Emre et al. (2004) and Parkinson Study Group (2004a, 2004b). We suppose our trial had two treatment arms, an experimental arm and a placebo control arm, and that the trial had nine visits, with baseline at Visit 0, and the primary efficacy endpoint at Week 28. Fifty-seven subjects were enrolled in each treatment group. The primary measure of efficacy was a sub-score of the Unified Parkinson’s Disease Rating Scale (UPDRS). For a general description of this illustrative dataset, see Section 1.10.1. A “spaghetti plot” of the complete dataset (Figure 1.1), although showing no strong distinct patterns, allows us to see the mass of data that can be available in a typical longitudinal trial.

A high score here indicates poor subject outcome. Parkinson’s disease is progressive, and for most treatments of the disease, one would expect to see a return to worsening after three to six months treatment seen in Emre et al. (2004) and Parkinson Study Group (2004a, 2004b) just cited. In other words, some transient improvement may be achieved and progression may be delayed for some time by treatment, but progression is not expected to stop completely. The reader may be able to see from Figure 1.1 that indeed, while many subjects in the illustrative dataset improved slightly (lower scores), subjects tended to revert to disease progression towards the end of the trial (higher scores).

In our example dataset, nearly 38% of subjects discontinued early, 18 (32%) and 25 (44%) subjects in the control and experimental arms, respectively, giving rise to substantial amounts of monotone missing data. Figure 1.2 highlights those subjects.

The large proportion of missing data for the primary endpoint in this example (38% of subjects discontinued) is troubling with regard to its impact on the power
WHAT'S THE PROBLEM WITH MISSING DATA?

Figure 1.2 Parkinson’s disease dataset: early discontinuations highlighted.

of the study. Also, the difference between treatment arms in the proportion of withdrawals (12% more in the experimental arm compared to placebo) is large enough to suggest that the reason for discontinuation depends on treatment. Both of these observations should motivate a careful consideration of the impact missing data may have on study conclusions. The statistician will want to consider ways to make inference from available study data while minimizing a possibility of bias that would unfairly favor the experimental treatment.

What options are available to proceed with analysis in the presence of missing data? The most obvious and easiest choice is to use only subjects with data available for the primary endpoint – study completers for whom assessments were performed at the final study visit. A second approach to consider would be to use all available longitudinal data (from all visits), including partial data from study dropouts, with the hope that this partial data could contribute in a meaningful way to the overall statistical analysis. Finally, we can impute missing data of discontinued subjects in some principled way, taking into account the information we have about these subjects prior to their dropout. We will discuss these three basic options in more detail below.

1.3 Why can’t I use only the available primary endpoint data?

Sometimes, only the subjects with available data for the primary endpoint (study completers) are used in the primary study analysis and test. Could we discard the data from subjects who discontinued early, use only available data at Week 28, and still have an unbiased estimate of treatment effect? If we are interested in estimating the treatment effect in the kind of subject who would complete the nine trial visits, then the data available at the ninth visit (Visit 8, Week 28) can be the basis of an unbiased estimate. What is to be estimated – the estimand – is important in assessing
CLINICAL TRIALS WITH MISSING DATA

how to handle missing data. Estimands are discussed at length in the U.S. National Research Council report, *The prevention and treatment of missing data in clinical trials*, commissioned by the U.S. Food and Drug Administration and published in 2010. A variety of estimands are discussed in Section 4.1.1, and US and EU regulatory guidance are discussed in Chapter 3. Usually, however, it is desired to estimate not just the treatment effect among the “elite” selection of subjects that completed the trial, but something more widely applicable such as the treatment effect in all subjects of the type randomized to the clinical trial (including both completers and subjects who discontinued early). The reasons recorded for discontinuation often suggest that many subjects discontinue either because of side effects or because of lack of efficacy. Thus, there is often good reason to believe that the efficacy score would be better in completers than in the full set of randomized subjects. In summary, complete cases (data from study completers) may give an estimate of efficacy that is not representative of all subjects in the study, and likely will be too favorable to the study treatments. An approach that is applicable to all subjects randomized will generally be more useful and more acceptable to the regulator. This approach where results are applicable to all subjects randomized is known as the “intent-to-treat” (ITT) approach. According to the ITT principle, all subjects that were included (randomized) in the trial should be included in the analysis, regardless of their compliance with treatment.

The use of data from completers only has an additional drawback that partial data from subjects that discontinued early, but still provided some information prior to withdrawal, is completely wasted.

In our dataset, Figure 1.3 illustrates the somewhat poorer efficacy scores that can pertain to subjects who discontinue early, taking as an example subjects whose last observation was at Week 6 or 8, (10 discontinuations each in the control and experimental treatment groups). Early withdrawals in the control group had higher (worse) mean efficacy scores from the start, compared to completers in their own

![Figure 1.3 Parkinson’s disease dataset: mean efficacy score at each time point for completers and for subjects whose last observation was at Visit 4 or 5 (Week 6 or 8).](image-url)
WHAT’S THE PROBLEM WITH MISSING DATA?

What’s the problem with using last observation carried forward?

Until recently, the most common method of handling missing data was to estimate the treatment effect using the last available measurement for a subject. This method has been known as “last observation carried forward,” or LOCF. The argument can be made for LOCF, that the observation just before a subject discontinues is likely to give evidence unfavorable to the study treatment — because the subject is likely to have left the study when his/her health was at a low point (unless he/she left for reasons unrelated to the condition under study). Last observation carried forward could thus be regarded as leading to a conservative estimate of efficacy for a treatment. To examine the LOCF method further, we plot some typical trajectories from our example study (Figure 1.4).

Figure 1.4 shows two typical trajectories of the efficacy score for completers and two typical trajectories for subjects who discontinued early. As is common in studies of Parkinson’s disease, the completer on the control arm shows a small
improvement and finishes the study close to his/her baseline value. The completer in the experimental arm also improves, and then reverts to a UPDRS value close to baseline. The two trajectories of early discontinuations are selected to show some implications of the method of handling missing data. Here, the control arm data are of a subject who discontinued very early, and experimental arm data are of a subject who discontinued after the midpoint of the study.

The efficacy score for the subject in the control group had changed little from baseline when he/she discontinued, and so LOCF imputes for Visit 8 (Week 28) a value almost unchanged since baseline (Figure 1.5). The subject in the experimental arm had a somewhat improved (lower) UPDRS score by Visit 6 (Week 12) when he/she discontinued. The values imputed by LOCF here do not seem very unreasonable,

Figure 1.4 Parkinson’s disease dataset: four selected trajectories.

Figure 1.5 Parkinson’s disease dataset: LOCF imputation.
WHAT’S THE PROBLEM WITH MISSING DATA? 11

except that the tendency of subjects to worsen late in the study is not reflected in the
imputation.

For LOCF to provide valid estimates of efficacy at the primary time point (e.g., the
last scheduled visit), a very particular MNAR assumption would need to hold, namely
that with a probability of one, no matter what the general trend of outcomes in the
study, a future outcome is equal to a subject’s last available outcome. Molenberghs and
Kenward (2007, pp. 45–47) point out how strong and unrealistic the LOCF assumption
is. Verbeke and Molenberghs (1997, Chapter 5) show how much at variance LOCF
is with the linear mixed model, with breaches of the usual assumptions about group
differences and evolution over time. In summary, the LOCF assumption is not often
clinically plausible; LOCF is unlikely in general to give a sensible estimate of a
subject’s efficacy at the study endpoint.

It is striking that LOCF makes no use of the information about the likely trajectory
of discontinuations that is available from other subjects in this study. For example,
Figure 1.3 tells us that, among completers, there is a slight worsening (increasing)
in the mean efficacy score from Visit 5 (Week 8) onwards, in both treatment groups,
perhaps reflecting the progressive nature of Parkinson’s disease. No such worsening
(increase) is included in LOCF imputations. Since LOCF fails to take account of the
general worsening observed in subjects with Parkinson’s disease, LOCF is likely to
favor treatment arms that have more discontinuations, especially for subjects who
discontinue mid-study when mean efficacy score is lowest (best). We see this to some
extent in the case of the subject from the experimental arm that discontinued at Week
12. Should efficacy at Week 28 not be somewhat worse than efficacy at Week 12,
given the progressive nature of Parkinson’s disease as seen in the study trend? We
note that our example here is not one that will often be found in a real clinical trial,
as LOCF is generally considered not appropriate for progressive diseases, precisely
because LOCF does not take progression into account. But even for other indications
where the outcomes tend to improve with time or for chronic diseases, there are other
problems with this method that we will discuss later in the book.

Our uncertainty about missing values in the previous paragraph brings us to
another widely expressed objection to LOCF and similar methods such as baseline
observation carried forward (BOCF). LOCF and BOCF are known as single impu-
tation methods. They posit a single imputed value for the missing value, and thereafter
treat the imputed value as though it were “real” data. The objection to single impu-
tation methods is that they fail to reflect uncertainty about missing data. Regulators
have voiced particular concern with regard to this unrealistic lack of variability in
single imputation methods, and this concern is described further in Chapter 3, which
summarizes the regulatory documents; Chapter 6 discusses this issue further in the
context of multiple imputation.

1.5 Can we just assume that data are missing at
random?

Would an MAR assumption give more credible results than LOCF for an estimate
of efficacy at Visit 8 (Week 28)? If we accept the MAR assumption, we take it that
observed data can in some sense account for missing values. Thus, if we assume MAR, missingness of the outcomes \( Y \) is independent of unobserved data conditional on the observed outcomes \( Y_{\text{obs}} \) and other covariables in the statistical model used. The MAR assumption states, as was defined earlier, that probability of missingness does not depend on unobserved data, given observed outcomes. It is helpful to understand that this assumption has an implication for the distribution of unobserved “potential” outcomes, given observed outcomes (or in the repeated measures context, as distribution of future outcomes given earlier outcomes). Informally, MAR can be shown (Verbeke and Molenberghs, 2000, Section 20.2.1, Theorem 20.1, p. 334) to be equivalent to the assumption that the conditional distribution of potential (missing) outcomes for dropouts given their observed outcomes, is the same as the conditional distribution of observed outcomes for patients who continued. As a result, the estimate of treatment effect that we get from the likelihood-based (ignorable) inference under MAR is essentially the estimate of what would have happened, had all the patients who discontinued remained on their respective treatments. See Section 4.2.2.3 for further discussion on what is estimated by MAR, and Section 4.1.1 on what it may be desired to estimate.

Thus under MAR we use all relevant study data, including partial data from discontinued subjects, to infer plausible values for missing data. Chapter 5 shows in detail how to do this in SAS® (SAS Institute Inc., 2011) using a direct likelihood method known as mixed models for repeated measures (MMRM). Direct likelihood approaches can include models of repeated measures for binary and other non-normal outcomes, as well as continuous, normally distributed outcomes. Chapter 6 gives details about how to implement the same MAR assumptions using multiple imputation (MI) When their statistical models are the same, the two methods – MI and MMRM – in theory should give similar results, and in our experience, they usually do. However, we note that with MI, the model used to impute missing values may be distinct from the primary analysis model used to estimate treatment effect. The model used to impute the missing values must have the same explanatory variables as the model of the primary analysis, but can have extra variables in addition to those. MI can even include post-baseline variables to help model the outcome. In contrast, direct likelihood methods such as MMRM, make inferences about missing values and in the same step estimate the treatment effect; inclusion of post-baseline covariables in this single step would almost certainly lead to confounding of the estimate of treatment effect, and so in practice direct likelihood approaches cannot make use of post-baseline covariables other than the outcome being modeled. Thus, MI can use more information than direct likelihood approaches in handling missing data, which may in some circumstances give it an advantage over direct likelihood approaches.

Both methods – direct likelihood modeling and MI – use partially observed study data to make inferences with missing data under MAR and have an advantage over single-imputation LOCF in that they take account of the uncertainty pertaining to missing data. Generally speaking, the confidence intervals provided by MI and by direct likelihood methods such as MMRM depend on the amount of missing information with respect to the estimated parameters, whereas for the single-imputation LOCF estimates this is not the case. Thus, our uncertainty about the missing data is
WHAT’S THE PROBLEM WITH MISSING DATA?

Figure 1.6 Parkinson’s disease dataset: MAR imputation for selected trajectories.

reflected in the summary statistics from MMRM and MI, but we cannot depend upon this being so when we use single imputation methods, such as LOCF or BOCF.

Chapter 6 describes how, in place of the missing data, MI uses a number of draws from the posterior distribution of the missing observation, given the observed data. The variability between the MI draws, calculated and incorporated using Rubin’s rules (Rubin, 1987) reflects the uncertainty about the missing data. Figure 1.6 shows values imputed under the MAR assumption for the trajectories of the two selected subjects that discontinued early, using MI. MI uses all the data in Figure 1.1 to estimate the posterior distribution of the missing data. Imputed values shown on Figure 1.6 represent the mean of 100 draws from the posterior distribution. (For convenience, the variability of the draws is not indicated in the plot, but it would of course be taken into account in subsequent analysis with imputed data). We can see that the slight worsening (increase) in efficacy scores found in the study from Visit 6 (Week 12) onwards is reflected in the MAR estimates for the two subjects who discontinued early. We also see that the study pattern of modest improvement in efficacy scores followed by return to disease progression is reflected in the imputation of the subject from the control group who discontinued shortly after the start of the study. Thus, MAR includes the likely trajectory of symptoms in its inference about missing values.

There is a wide variety of opinion about applicability of the MAR assumption. Many argue that it is “often reasonable” to assume that clinical trial data are MAR, see, for example, Mallinckrodt et al. (2008). Because it assumes that discontinuations follow the pattern of completers, some argue that MAR gives a kind of per-protocol estimate of treatment effect, albeit capable of being applied to the ITT population. In Chapter 3, we will see that the US and EU regulatory guidances point to this limitation of methods that assume MAR. Kenward (2011) similarly talks (broadly speaking) of MAR giving a *de jure* estimate of treatment effect. Section 4.2.3.3 describes
14 CLINICAL TRIALS WITH MISSING DATA

a case where, at the least, an MAR analysis would need to be carefully justified. However, while one acknowledges that MAR uses study data to make inferences for missing data, such use of study data is not always as favorable to study treatments as it sounds. We often find that subjects who discontinue early have efficacy scores that are worse than the study average. Since MAR models missing values using study data, MAR models the missing values of early discontinuations on similar poorly performing subjects who completed the study. Consequently MAR’s use of observed study outcomes to model later missing outcomes may result in acceptably unfavorable estimates of efficacy for subjects who discontinue. See Section 4.2.2.4 for further discussion of these points.

In addition to direct likelihood and MI, there are a number of other approaches that can assume MAR, which we mention briefly here.

Inverse probability weighting (IPW) approaches weight observations by the inverse of the probability of their being observed, and thus “even up” the estimate to give adequate emphasis to the kind of subject associated with missing data. Logistic regression can be used to estimate the probability of an observation being observed, using baseline data or trial data prior to dropout. Since this approach uses study data to take account of missing data, it can give valid estimates assuming MAR. The approach using generalized estimating equations (GEEs) takes account of explanatory variables such as treatment and baseline score, but does not take modeled account of pre-discontinuation outcomes and thus assumes MCAR; but GEEs can use IPW weighting, and thus weighted GEEs can also be valid under the MAR assumption. These weighted GEEs can be further augmented with functions of the observed data and configured so as to be doubly robust in that inference will be valid if at least one model – either the model for missingness or the model for the missing values - is correct, but not necessarily both. Carpenter and Kenward (2006) discuss the theory and practice of this method. Vansteelandt et al. (2010) usefully describe a number of ways of implementing doubly robust estimation. Chapter 8 discusses weighted and doubly robust estimation and shows how to implement Vansteelandt et al.’s version of doubly robust estimation via SAS code and using a SAS macro.

1.6 What can be done if data may be missing not at random?

However, sometimes we may feel we cannot justify the assumption of MAR clinically. In other instances, the regulator may judge that MAR is too favorable an assumption with regard to the experimental arm, and an outcome worse than that assumed by MAR should be imputed.

Selection models take into account the probability of observations being missing by modeling this probability along with the outcome, using a joint likelihood. The missingness of MNAR data is by definition dependent on unobserved outcomes. Selection models can model MNAR assumptions by modeling this dependence. For example, assumptions about the degree of dependence can be modeled by forcing the coefficient for the relationship between the unobserved data and missingness to have a specified value. Shared parameter models posit a latent variable, given which
WHAT’S THE PROBLEM WITH MISSING DATA? 15

the missingness is often assumed to be independent of the unobserved values. The model’s estimate of the magnitude and/or statistical significance of the latent variable is thought by some to give evidence about the strength of a putative MNAR link between missingness and unobserved values of the outcome. We do not give a full treatment of selection models and shared parameter models, but note them among the tools available to the statistician in Chapter 4 and discuss them briefly; while a further description of the two methods, with SAS code and an example, is provided in Appendices 7.I and 7.J.

Both MI and MMRM can also be used to implement analyses based on MNAR assumptions, using pattern-mixture models (PMMs). We have used PMMs with such MNAR assumptions for a primary analysis. An example where this may be appropriate is a trial where the indication is chronic pain. Here, it may be reasonable to assume for purposes of estimating treatment effect that discontinuations do worse than similar subjects who remain in the study, and thus that MAR would not be suitable.

Most sensitivity analyses assume some form of MNAR. We will describe how to use MI for this purpose in detail in the chapter on sensitivity analyses, Chapter 7; use of MMRM-type estimates for PMMs is also noted in Chapter 7. Our next section discusses in more detail how PMMs can be used to facilitate sensitivity analyses.

1.7 Stress-testing study results for robustness to missing data

One of the things we know for certain about missing data is that it’s not there – we cannot be sure of what the missing data might have been. Ultimately, we do not know the level of bias that may result from a particular assumption about missing data. Could a few subjects have discontinued due to sudden worsening of disease that was not recorded, and thus have data MNAR? Even if a small proportion of data is MNAR, the assumptions of a standard MAR analysis such as MMRM will be incorrect. We would argue that the unknown bias due to missing values is not just a nuisance parameter or a crack in the evidence that may be papered over. The fact that a subject discontinues early from a study is in itself an important piece of evidence about the study treatment. The US and EU guidance (see Chapter 3) strongly advises that results from MAR analyses should be “stress tested” for robustness to the MAR assumption.

Chapter 7 describes approaches for testing study conclusions for robustness to missing data. We think such tests of robustness are an essential part of handling missing data, whether MCAR, MAR or indeed MNAR is assumed.

1.8 How the pattern of dropouts can bias the outcome

For the subject in Figure 1.5 who discontinued at Week 12 in the experimental arm, LOCF assumed no worsening – even though on average efficacy scores did worsen after Week 12 in the study. For progressive disease, LOCF thus tends inappropriately to favor early discontinuations. Does this mean that LOCF is “anti-conservative?” Not necessarily. If for a trial in progressive disease we expect more or earlier
16 CLINICAL TRIALS WITH MISSING DATA

discontinuations in the control arm than in the experimental arm, LOCF could be
regarded as “conservative” in estimating the difference between treatment groups
in that LOCF will be expected to favor the control group. However, LOCF-based
estimates would still paint an overly optimistic picture of disease progression within
each arm. We may begin to see that, for a given estimand, bias due to missing data
will depend upon the interplay between

- the assumptions we make (LOCF, MAR, etc.)
- the expected trajectory of outcome over time
- relative timing and proportion of missing data in experimental arm and control
  arm
- reasons for discontinuation in experimental arm and control arm
- objective of the study (superiority, non-inferiority)

Among these factors that contribute to our assessment of bias, the MAR
assumption can often, by its nature, take account of the second factor – the expected
trajectory of outcome over time. In Figure 1.6, we see how the disease trajectory can
usually be modeled well from the observed study data. Thus if, for example, subjects
with a particular indication tend to worsen over time (whatever their treatment),
MAR will model this in its imputation of their missing values. In general, then, MAR
methods allow for disease trajectory in their inference about missing data. Because
of this, disease trajectory is often less important in determining the bias of MAR
methods. This does not take away from the weakness of the MAR assumption noted
in Section 1.5, that it infers that subjects who stop treatment will have symptoms
like similar subjects who remain in the study and on treatment. Thus, while MAR
will often correctly model the “natural” disease trajectory, it may well model a
continuing effect of treatment after the subject has stopped taking the treatment in
a way that is not clinically realistic. As noted, this could be a particular problem,
for example, in studies of chronic diseases, where symptoms are known to return to
baseline when treatment stops. In such cases, MAR may erroneously infer that the
symptoms of subjects who stop treatment will maintain a smooth trajectory similar
to subjects who remain on treatment.

1.9 How do we formulate a strategy for missing data?

It is not possible to predict bias accurately, and no approach for handling missing data
will be perfect. This fact is recognized in the wording of the EU guidance: the primary
analysis should be chosen that “would … provide a point estimate that is unlikely to
be biased in favor of experimental treatment to an important degree (under reasonable
assumptions)” (European Medicines Agency, 2010). Because we cannot be sure that
the assumptions of the primary analysis are true, a comprehensive report of study
results must include sensitivity analyses that stress-test the primary study result,
to assess its robustness when those assumptions do not hold. Sensitivity analyses
usually implement pre-specified MNAR assumptions that can be demonstrated to be
likely to be more unfavorable to the experimental treatment than those of the primary analysis. We have noted that statistical frameworks for implementing sensitivity analyses include selection models, shared parameter models and PPMs. Our Chapter 7 concentrates on the use of PMMs to implement a variety of sensitivity analyses.

The strategy for missing data must be chosen and justified individually for each trial. Chapter 2 shows how steps can be taken when both designing and implementing a trial to prevent missing data in the first place. Chapter 4 maps the process of how to identify justifiable approaches to missing data. The choice of both main analysis and sensitivity analyses can take into account the likely patterns of outcome and of missingness in the trial data, using historic data about the efficacy and safety trajectory of the indication, and the likely effects of treatment. Those historic data, if they exist, will be needed to support a scientific justification of the chosen strategy in the study protocol. Historic data can also suggest sensitivity analyses that stress-test the study result by including assumptions that are likely to be unfavorable to the experimental arm, with regard to missing data, while remaining clinically plausible.

There is no one strategy that is best for missing data. Nevertheless, we propose the following pointers:

- a plan should be devised to prevent or minimize missing data in the design and implementation of a trial;
- the MAR assumption can in many, but not all, cases offer an estimate that is “unlikely to be biased to an important degree”;
- but the MAR approach is likely to have some bias;
- the likely extent of bias due to missing values would need to be assessed on a study by study basis;
- an MNAR approach may be warranted for the primary analysis in some cases;
- sensitivity analyses to assess robustness to the assumptions of the primary analysis are always warranted;
- the study design and analyses should make assumptions about missing data that are clinically meaningful, transparent and easily understood;
- in a growing number of cases, regulators have in their advice and their approvals been following the latest regulatory documents issued in the United States or the EU; on occasion, however, particular regulatory views, sometimes linked to a strong tradition for a therapeutic area, can dictate primary and even sensitivity analyses that do not follow the current guidances or the reasoning we have presented in this chapter; in this case, of course, agreement must be reached on the best strategy taking regulatory views into account.

A detailed guide to planning a strategy for handling missing data will be presented in Chapter 4.

The examples in this book are drawn from phase II and phase III trials, but we note that it will be helpful to pay attention to missing data in early-phase trials not
only for the sake of those trials, but also because this will make for a more informed strategy for missing data in later confirmatory trials. Exploration of missing data patterns should be included in the clinical study report of early-phase trials.

Finally, missing data is not only the statistician’s problem – the whole team needs to take part in discussions pertaining to assumptions and prevention.

1.10 Description of example datasets

This section describes three example datasets that are used throughout the book to illustrate various approaches for dealing with missing data.

1.10.1 Example dataset in Parkinson’s disease treatment

The illustrative example described in this section has been patterned after typical Parkinson’s disease data, for example, in Emre et al. (2004) and Parkinson Study Group (2004a, 2004b). The dataset represents a randomized, double-blind trial designed to compare two treatment arms – experimental and placebo, with 57 subjects per arm. The primary efficacy endpoint is a sub-score of the UPDRS with lower values representing less severe symptoms. Time points included in our example dataset represent study visits scheduled to occur at baseline (Visit 0), and then at post-baseline Weeks 2, 3, 4, 6, 8, 12, 20 and 28 (a total of eight post-baseline visits).

In our example, 68% of subjects in the placebo arm completed the study versus 56% of subjects in the experimental arm. Table 1.1 summarizes (cumulative) percentages of subjects discontinued from the trial by visit. For example, this table shows that 19% of subjects in the placebo arm discontinued at or before Visit 4.

Table 1.1 Parkinson’s disease dataset, subjects completing and discontinuing from the trial by visit of discontinuation.

<table>
<thead>
<tr>
<th>Visit (Time Point)</th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Week 2)</td>
<td>4%</td>
<td>7%</td>
</tr>
<tr>
<td>2 (Week 3)</td>
<td>9%</td>
<td>16%</td>
</tr>
<tr>
<td>3 (Week 4)</td>
<td>9%</td>
<td>21%</td>
</tr>
<tr>
<td>4 (Week 6)</td>
<td>19%</td>
<td>32%</td>
</tr>
<tr>
<td>5 (Week 8)</td>
<td>26%</td>
<td>39%</td>
</tr>
<tr>
<td>6 (Week 12)</td>
<td>30%</td>
<td>44%</td>
</tr>
<tr>
<td>7 (Week 20)</td>
<td>32%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Study Completers

<table>
<thead>
<tr>
<th></th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>68%</td>
<td>56%</td>
</tr>
</tbody>
</table>
WHAT'S THE PROBLEM WITH MISSING DATA? 19

Figure 1.7 Parkinson’s disease dataset. Kaplan-Meier plot of time to discontinuation from the study in each treatment arm. The two rows of numbers within the plot at the bottom are counts of those “at risk” of discontinuation at each time point.

(Week 6). Overall, there were considerably more subjects who discontinued from the experimental arm compared to the placebo arm.

Figure 1.7 presents a Kaplan-Meier plot of probability of continuing in the study in both treatment arms. It further highlights in a visual manner the difference in the proportion of discontinuations between the arms, which started to become quite apparent from around Week 4 and was sustained until the end of the study. Larger proportions of subjects withdrew up to Week 6 from the experimental arm; after Week 6 the probability curves remained mostly parallel throughout the rest of the study. This suggests that there may have been a group of subjects who were not able to tolerate the experimental treatment relatively early in the study and discontinued for this reason, while the withdrawals at the later stages were similar in both arms.

Figure 1.8 summarizes mean change from baseline (CFB) in UPDRS sub-score values by time point for study completers and a number of dropout cohorts (corresponding to the time point of discontinuation) for each treatment arm. Most discontinuation cohorts contain a relatively small number of subjects, so one can expect the means to have more diverse trajectories, compared to the trajectory for study completers. In the placebo arm, we can see that subjects who discontinued at Week 6 and 12 showed on average some improvement prior to discontinuation, while, for example, those who discontinued at Week 8 and 20 experienced both improvement and worsening. In the experimental arm, the slope of the trajectory of most discontinued subjects was initially close to the slope of completers in their arm, although those that discontinued at Weeks 6 and 8 had worsening of symptoms prior to discontinuation after some initial improvement.

Figure 1.9 summarizes the CFB by grouping all subjects who discontinued at any point versus completers in each treatment arm. It appears that, on average, withdrawals from the experimental arm are very similar to placebo completers prior
20 CLINICAL TRIALS WITH MISSING DATA

Figure 1.8 Parkinson's disease dataset, summary of mean change from baseline (CFB) in UPDRS sub-score by time point across dropout cohorts (grouped by time of discontinuation) and study completers.

Table 1.2 also clearly suggests that withdrawals differed between treatment arms: 38% of subjects discontinued from the experimental treatment due to an adverse event (AE) versus 9% in the placebo arm, indicating that the experimental treatment toxicity was an important factor in subjects’ inability to continue on treatment. In the placebo arm, 16% of subjects discontinued due to lack of efficacy, while no subjects discontinued due to this reason from the experimental treatment. The proportion to discontinuation, while placebo dropouts were on all but one visit doing worse compared to placebo completers.
WHAT’S THE PROBLEM WITH MISSING DATA?

Figure 1.9  Parkinson’s disease dataset, summary of mean change from baseline (CFB) in UPDRS sub-score by time point for study dropouts and completers in each treatment arm.

of subjects discontinuing due to other reasons (such as loss to follow-up, protocol violation, withdrawal of consent) is similar in the two arms.

Figure 1.10 summarizes mean CFB across time by discontinuation status/reason in each treatment arm. In the experimental arm, subjects who withdrew for reasons other than AE did so quite early in the study. Those who discontinued due to AEs stayed on study longer and showed some modest improvement (comparable to that of placebo subjects) before adverse reactions led to the decision to discontinue treatment. In the placebo arm, these data show a tendency towards worsening among subjects who discontinued due to AEs and other reasons. Interestingly, in the group of placebo subjects who had a primary reason for discontinuation recorded as lack of efficacy, CFB during the interval just prior to discontinuation at Week 12 appeared as favorable. A situation like this may indicate that there is another (e.g., secondary) efficacy parameter that played an important role in the decision to discontinue, while the primary efficacy endpoint does not directly correlate with the recorded primary reason.

In SAS code examples provided throughout this book, we will refer to variables as contained in this dataset and used for analysis. Depending on the analysis performed,

Table 1.2  Parkinson’s disease dataset, reasons for discontinuing from the trial.

<table>
<thead>
<tr>
<th>Reason for Discontinuation</th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse Event</td>
<td>9%</td>
<td>39%</td>
</tr>
<tr>
<td>Lack of Efficacy</td>
<td>16%</td>
<td>0%</td>
</tr>
<tr>
<td>Other Reasons</td>
<td>7%</td>
<td>5%</td>
</tr>
</tbody>
</table>
22 CLINICAL TRIALS WITH MISSING DATA

![Graph showing mean CFB by time point for study completers and dropouts by reason for discontinuation.]

Figure 1.10 Parkinson’s disease dataset, summary of mean CFB in UPDRS subscore by time point for study completers and dropouts by reason for discontinuation.

we may utilize either a vertically structured or a horizontally structured dataset. Variables that are common in both of these structures are as follows:

- *subj* – subject number;
- *region* – geographic region;
- *trt* – treatment arm (with 1 representing placebo arm, and 2 – the experimental arm);
- *lastvis* – last study visit subject attended (taking the value of 0 for baseline, and values 1 through 8 for post-baseline Weeks 2, 3, 4, 6, 8, 12, 20 and 28, respectively);
- *reason* – primary reason for discontinuation (taking values Adverse Event; Lack of Efficacy; and Other Reasons).
WHAT’S THE PROBLEM WITH MISSING DATA? 23

A dataset with a vertical structure contains one record per subject and visit, and has the following additional variables:

- \( \text{visit} \) – study visit (taking the value of 0 for baseline, and values 1 through 8 for post-baseline Weeks 2, 3, 4, 6, 8, 12, 20 and 28);
- \( \text{upd} \) – subjects’ UPDRS sub-score for a given visit;
- \( \text{base upd} \) - subjects’ UPDRS sub-score at baseline;
- \( \text{inv} \) – investigator site.

A dataset with a horizontal structure contains one record per subject with the following variables representing analysis values at different visits:

- \( \text{upd}_0, \text{upd}_1, \ldots, \text{upd}_8 \) – subjects’ UPDRS sub-scores for baseline and Visit 1 through 8 (or Weeks 2, 3, 4, 6, 8, 12, 20, and 28) respectively.

### 1.10.2 Example dataset in insomnia treatment

Data introduced in this section have been patterned after typical clinical trials in an adult population with an insomnia indication, for example, in Krystal et al. (2008); Roth et al. (2005); Purdue Pharma LP (2007); Eli Lilly and Company (2008). Our dataset represents a randomized, double-blind trial with 320 subjects in each of the two arms – experimental and placebo. The primary efficacy endpoint is the mean total sleep time (TST) per night measured in minutes. We also consider another efficacy parameter, morning sleepiness, commonly measured in insomnia trials. This parameter, which we refer to as the morning sleepiness score (MSS), is evaluated on a 0 to 10 scale, with lower values representing less sleepiness and thus a more favorable outcome. Additionally, this dataset contains another sleep-related parameter – sleep quality score (SQS), which we will make use of in some analyses. Typically, TST, MSS and SQS are recorded by subjects daily, while the analysis endpoints represent averages of these daily values over the period of time between clinic visits. Time points included in our example dataset represent study visits scheduled to occur prior to the commencement of the study treatment (referred to as baseline or Week 0), and then at post-baseline Weeks 1, 2, 4, 6, and 8. In other words, an analysis value for the TST endpoint at Visit \( k \) for a given subject in our dataset is the average of daily TST values recorded between Visits \( k - 1 \) and \( k \). For a baseline time point (\( k = 0 \)), it is assumed that subjects recorded their daily values between a screening visit and a Week 0 visit and the average over the days between these visits is used as baseline. Analysis values for the morning sleepiness score represent similar between-visit averages of daily scores for each subject.

In our example, 82% of subjects in the placebo arm completed the study versus 80% of subjects in the experimental arm. Table 1.3 summarizes (cumulative) percentages of subjects discontinued from the trial by visit. For example, this table shows that 15% of subjects in each arm discontinued at or before Visit 3 (Week 4). Percentages of dropouts are fairly similar across the two treatment arms for all time points.

Table 1.4 provides a summary of reasons for discontinuation. In this example dataset, there is not much difference between the treatment arms in this respect.
24 CLINICAL TRIALS WITH MISSING DATA

Table 1.3 Insomnia dataset, subjects completing and discontinuing from the trial by visit.

<table>
<thead>
<tr>
<th>Visit (Time Point)</th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Week 1)</td>
<td>6%</td>
<td>6%</td>
</tr>
<tr>
<td>2 (Week 2)</td>
<td>9%</td>
<td>10%</td>
</tr>
<tr>
<td>3 (Week 4)</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>4 (Week 6)</td>
<td>18%</td>
<td>20%</td>
</tr>
<tr>
<td>Study Completers</td>
<td>82%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Figure 1.11 contains a Kaplan-Meier plot of probability of remaining in the study by treatment arm. The probability curves are very similar in both treatment arms and show only a very small difference towards the end of the study. The rate of discontinuation was very similar at different stages of the trial, with approximately 5% of subjects dropping out after each visit.

Figure 1.12 and 1.13 summarize mean CFB in TST and MSS values by time point for study completers and different dropout cohorts (corresponding to the time point of discontinuation) for each treatment arm. Upon examining these figures, we can see that while subjects that discontinued from the experimental arm look similar to completers in their TST and MSS, subjects that discontinued from the placebo arm are noticeably different from placebo completers, in terms of the CFB in TST and MSS. Placebo subjects discontinued after Weeks 2 and 6 had little improvement or worsening in TST. Those who discontinued after Week 4, seem to have improved in terms of TST, but less so in terms of MSS. If we examine baseline values (see Table 1.5), we can see that mean baseline TST seemed substantially worse in this group of dropouts compared to other subjects (220 min vs. 337 min for completers).

Table 1.4 Insomnia dataset, reasons for discontinuing from the trial.

<table>
<thead>
<tr>
<th>Reason for Discontinuation</th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse Event</td>
<td>4%</td>
<td>5%</td>
</tr>
<tr>
<td>Consent Withdrawn</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Lack of Efficacy</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Lost to Follow-up</td>
<td>5%</td>
<td>6%</td>
</tr>
<tr>
<td>Protocol Violated</td>
<td>4%</td>
<td>2%</td>
</tr>
<tr>
<td>Other Reasons</td>
<td>1%</td>
<td>3%</td>
</tr>
</tbody>
</table>
WHAT’S THE PROBLEM WITH MISSING DATA?

Figure 1.11 Insomnia dataset, Kaplan-Meier plot of time to discontinuation from the study in each treatment arm. The two rows of numbers within the plot at the bottom are counts of those “at risk” of discontinuation at each time point.

Similarly, the mean baseline morning sleepiness score in this group of subjects was the highest (equal to 5.9), while, for example, completers had a baseline mean of 5.1. This seems to indicate that even though these subjects showed improvement in terms of TST and, to some extent, MSS, it might not have been a satisfactory one given their more severe symptoms at baseline. Overall, the trajectory of TST and MSS values prior to withdrawal as well as baseline values seem to be correlated with dropout in the placebo arm, and may suggest that an MAR assumption would be plausible, at least for the placebo arm.

However, as already noted, compared to the placebo arm, discontinuations from the experimental arm appear to be somewhat more similar to completers in their arm, in terms of their trajectory prior to dropout. Nevertheless, experimental arm subjects last observed at Weeks 1, 2, or 6 showed little improvement from baseline in term of TST. Subjects last observed at Weeks 1 and 6 had changes in MSS that showed trends similar to those in TST. For experimental arm subjects who were last observed at Week 4, the picture is less clear. They seem to have improved both in terms of TST and especially MSS. Their baseline TST and MSS values are, on average, somewhat worse than those of other subjects but to a less important degree than in the case of placebo subjects.

Primary reasons for discontinuation do not seem to shed any more light: reasons are distributed similarly for the two treatment groups. Discontinuation after an apparent improvement in the primary and key secondary efficacy parameter may either indicate that subjects were reasonably satisfied with their achieved sleep pattern and did not consider that further treatment was necessary, or it may be suggestive of an MNAR mechanism.

In SAS code examples provided throughout this book, we will refer to variables as contained in the datasets we used for analysis. Depending on the analysis performed,
26 CLINICAL TRIALS WITH MISSING DATA

Figure 1.12  Insomnia dataset, summary of mean CFB in total sleep time (minutes) by time point across dropout cohorts (grouped by time of discontinuation) and study completers.

we utilize either a vertically structured or a horizontally structured dataset. Variables that are common in both of these structures are as follows:

- *subj* – subject number;
- *agegrp* – subject’s age group at study entry (taking value of 1 through 5 for age groups of <30 years, 30–34 years, 35–44 years, 45–54 years and ≥55 years);
- *sex* – subject’s sex;
WHAT’S THE PROBLEM WITH MISSING DATA?  27

Figure 1.13 Insomnia dataset, summary of mean change from baseline (CFB) in morning sleepiness score (0 to 10) by time point across dropout cohorts (grouped by time of discontinuation) and study completers.

Table 1.5 Insomnia dataset, summary of baseline values of efficacy endpoints.

<table>
<thead>
<tr>
<th>Subjects discontinued at visit (Time point)</th>
<th>Baseline TST Mean (±STD)</th>
<th>Baseline MSS Mean (±STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Placebo</td>
<td>Experimental</td>
</tr>
<tr>
<td>1 (Week 1)</td>
<td>345 (±113)</td>
<td>336 (±38)</td>
</tr>
<tr>
<td>2 (Week 2)</td>
<td>345 (±115)</td>
<td>363 (±77)</td>
</tr>
<tr>
<td>3 (Week 4)</td>
<td>220 (±142)</td>
<td>298 (±96)</td>
</tr>
<tr>
<td>4 (Week 6)</td>
<td>383 (±68)</td>
<td>295 (±145)</td>
</tr>
<tr>
<td>Completers</td>
<td>337 (±79)</td>
<td>322 (±80)</td>
</tr>
</tbody>
</table>
CLINICAL TRIALS WITH MISSING DATA

- \textit{trt} – treatment arm (with 1 representing placebo arm, and 2 – the experimental arm);
- \textit{lastvis} – last study visit subject attended (taking the value of 0 for baseline, and values 1 through 5 for post-baseline Weeks 1, 2, 4, 6 and 8 respectively);
- \textit{reasond} – primary reason for discontinuation (taking values Adverse Event; Consent Withdrawn; Lack of Efficacy; Lost to Follow-up; Protocol Violated; and Other Reasons).

A dataset with a vertical structure contains one record per subject and visit, and has the following additional variables:

- \textit{visit} – study visit (taking the value of 0 for baseline, and values 1 through 5 for post-baseline Weeks 1, 2, 4, 6 and 8);
- \textit{tst} – subjects’ average total sleep time value for a given visit;
- \textit{base\_tst} - subjects’ average total sleep time value at baseline;
- \textit{mss} – subjects’ average morning sleepiness score for a given visit;
- \textit{base\_mss} – subjects’ average morning sleepiness score at baseline;
- \textit{sqs} – subject’s average sleep quality score for a given visit;
- \textit{base\_sqs} – subject’s average sleep quality score at baseline.

A dataset with a horizontal structure contains one record per subject with the following variables representing analysis values at different visits:

- \textit{tst\_0}, \textit{tst\_1}, …, \textit{tst\_5} – subjects’ average total sleep time values for baseline and Visit 1 through 5 (or Weeks 0, 1, 2, 4, 6, and 8) respectively;
- \textit{mss\_0}, \textit{mss\_1}, …, \textit{mss\_5} – subjects’ average morning sleepiness scores for baseline and Visit 1 through 5 (or Weeks 0, 1, 2, 4, 6, and 8) respectively;
- \textit{sqs\_0}, \textit{sqs\_1}, …, \textit{sqs\_5} – subjects’ average sleep quality scores for baseline and Visit 1 through 5 (or Weeks 0, 1, 2, 4, 6, and 8) respectively.

1.10.3 Example dataset in mania treatment

The example dataset described in this section has been patterned after typical trials in mania treatment, for example, Bowden \textit{et al.} (2010); Lipkovich \textit{et al.} (2008); Post \textit{et al.} (2005). Our dataset represents a randomized, double-blind, two-arm trial comparing an experimental treatment to placebo, with 550 subjects randomized to each arm. The primary efficacy endpoint is a total score from the Young Mania Rating Scale (YMRS) based on the subject’s evaluation of his/her clinical condition over the past 48 hours using 11 items evaluating severity of a variety of symptoms. The YMRS score can take values between 0 and 88, with higher scores representing more severe symptoms. Time points included in our example dataset represent study visits
Table 1.6  Mania dataset, subjects completing and discontinuing from the trial by visit of discontinuation.

<table>
<thead>
<tr>
<th>Visit (Time Point)</th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Baseline)</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>1 (Day 4)</td>
<td>5%</td>
<td>4%</td>
</tr>
<tr>
<td>2 (Day 7)</td>
<td>8%</td>
<td>10%</td>
</tr>
<tr>
<td>3 (Day 14)</td>
<td>21%</td>
<td>18%</td>
</tr>
<tr>
<td>4 (Day 21)</td>
<td>30%</td>
<td>25%</td>
</tr>
<tr>
<td>5 (Day 28)</td>
<td>70%</td>
<td>75%</td>
</tr>
</tbody>
</table>

scheduled to occur at baseline (Visit 0), and then at post-baseline Days 4, 7, 14, 21, and 28 (a total of five post-baseline visits).

In this example dataset, 70% of subjects in the placebo arm completed the study versus 75% of subjects in the experimental arm. Table 1.6 summarizes (cumulative) percentages of subjects discontinued from the trial by visit. For example, this table shows that 21% of subjects in the placebo arm discontinued at or before Visit 3 (Day 14). The overall proportion of subjects who discontinued from the placebo arm was somewhat larger (by 5%), and the difference in the proportion of discontinuations between the two arms increased towards the end of the trial.

Figure 1.14 shows a Kaplan-Meier plot of probability of not withdrawing from the study for each treatment arm. The probability curves of the two arms are very

![Kaplan-Meier plot](image)

Figure 1.14  Mania dataset, Kaplan-Meier plot of time to discontinuation from the study in each treatment arm. The two rows of numbers within the plot at the bottom are counts of those “at risk” of discontinuation at each time point.
Figure 1.15 Mania dataset, summary of mean change from baseline (CFB) in YMRS score by time point across dropout cohorts (grouped by time of discontinuation) and study completers.

close until a midpoint of the study (Day 14), with the probability of discontinuation increasing more rapidly in the placebo arm from that point on.

Figure 1.15 summarizes mean CFB in YMRS score by time point for study completers and a number of dropout cohorts (corresponding to the time point of discontinuation for each treatment arm). In this dataset, 1% of subjects in each arm discontinued before providing any post-baseline assessments; this group is not depicted on the plots of mean CFB. In the placebo arm, we can see that subjects who discontinued tended to experience worsening or small improvement compared to placebo completers. Mean improvement in YMRS score up to Day 28 is quite pronounced among placebo completers (strong placebo effect). In the experimental arm, subjects who discontinued at or before Day 14 also experienced either a very small improvement or worsening. Subjects that discontinued after Day 21 (7% of subjects), on the contrary, showed an important improvement, similar to completers in their arm. Only a small number of subjects in this cohort discontinued due to an adverse event, suggesting that they may have had deterioration on efficacy parameters other than the primary YMRS score, or discontinued for other reasons.

The left panel of Figure 1.16 summarizes observed CFB by grouping all subjects who discontinued at any point versus completers in each treatment arm. From this plot, for example, it is apparent that using an approach such as last observation carried forward could favor the experimental arm because of the cohort of experimental arm withdrawals that had a significant improvement prior to discontinuing after Day 21.
WHAT’S THE PROBLEM WITH MISSING DATA?

Placebo Arm Completers: 70%
Placebo Arm Dropouts: 30%
Experimental Arm Completers: 75%
Experimental Arm Dropouts: 25%

Figure 1.16 Mania dataset, summary of mean CFB in YMRS score by time point for study dropouts and completers in each treatment arm, based on observed values (left panel) and LOCF-imputed values (right panel).

This is illustrated on the right panel of Figure 1.15 which depicts mean CFB using LOCF-imputed values. In this case, LOCF imputation makes placebo withdrawals look bad and some experimental withdrawals look good. Thus, LOCF may be viewed as anti-conservative, especially if a rapid deterioration of mania symptoms after premature treatment discontinuation is clinically plausible between Day 21 and 28. Nevertheless, we shall see in Section 4.2.3.2 that it is possible to implement the assumption that LOCF applies only to the experimental arm while assuming MAR for the control arm, and that this could be a sufficiently conservative approach with this illustrative dataset (based on the fact that there are more withdrawals in the placebo group, so more of them will be favored by MAR, and not all experimental improved before discontinuation, so LOCF will not favor all in the experimental arm).

Table 1.7 summarizes discontinuations by reason for discontinuation. In this dataset, percentage of subjects discontinuing due to an AE is very small in both treatment arms (1% in placebo and 3% in experimental), indicating that the experimental

<table>
<thead>
<tr>
<th>Reason for Discontinuation</th>
<th>Placebo arm</th>
<th>Experimental arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse Event</td>
<td>1%</td>
<td>3%</td>
</tr>
<tr>
<td>Consent Withdrawn or Decision of Investigator</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>Other Reasons</td>
<td>11%</td>
<td>6%</td>
</tr>
</tbody>
</table>
treatment was, in general, well tolerated. The majority of subjects discontinued due
to an investigator decision or withdrawal of consent, in proportions that are similar
in both arms (18% in placebo and 16% in experimental). Approximately 5% more
of subjects withdrew from the placebo arm due to other reasons compared to the
experimental arm.

Figure 1.17 summarizes mean CFB across time in each treatment arm by reason
for discontinuation. In both arms, subjects who withdrew consent or discontinued
due to an investigator decision did not show adequate improvement in the YMRS
score. Subjects who withdrew due to other reasons did improve, especially those in
the experimental arm.

In SAS code examples provided throughout this book, we will refer to variables as
contained in this dataset and used for analysis. Depending on the analysis performed,
we may utilize either a vertically structured or a horizontally structured dataset.
Variables that are common in both of these structures are as follows:

- **subj** - subject number;
- **trt** - treatment arm (with 1 representing placebo arm, and 2 - the experimental
  arm);
- **lastvis** - last study visit subject attended (taking the value of 0 for baseline, and
  values 1 through 5 for post-baseline Days 4, 7, 14, 21, and 28 respectively);
- **reason** - primary reason for discontinuation (taking values Adverse Event;
  Consent Withdrawn or Investigator Decision; and Other Reasons).
WHAT’S THE PROBLEM WITH MISSING DATA?

A dataset with a vertical structure contains one record per subject and visit, and has the following additional variables:

- \( visit \) – study visit (taking the value of 0 for baseline, and values 1 through 5 for post-baseline Days 4, 7, 14, 21, and 28);
- \( ymrs \) – subjects’ YMRS score for a given visit;
- \( base_{ymrs} \) – subjects’ YMRS score at baseline;
- \( inv \) – investigator site.

A dataset with a horizontal structure contains one record per subject with the following variables representing analysis values at different visits:

- \( ymrs_0, ymrs_1, \ldots, ymrs_5 \) – subjects’ YMRS scores for baseline and Visit 1 through 5 (or Days 4, 7, 14, 21, and 28) respectively.

Appendix 1.A: Formal definitions of MCAR, MAR and MNAR

Missing data mechanisms are often classified as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) as per the terminology introduced by Rubin (1976) and Little and Rubin (2002). These types of missingness can be formally defined as follows, using notation similar to that of Molenberghs and Kenward (2007). Consider the full data density of the study outcome (including both missing and observed outcomes):

\[
f(y, r | W, F, \theta, \psi) = f(y | W, \theta)f(r | F, \psi) \tag{1.1}
\]

where \( y \) is a matrix of outcomes, \( r \) is the matrix of missingness indicators, \( W \) and \( F \) are the design matrices for \( y \) and \( r \), respectively, and the corresponding parameter vectors are \( \theta \) and \( \psi \). Under MCAR, missingness is independent of outcomes, and we can transform (1.1) as follows:

\[
f(y, r | W, F, \theta, \psi) = f(y | W, \theta)f(r | F, \psi) \tag{1.2}
\]

and consequently

\[
f(y_{obs}, r | W, F, \theta, \psi) = f(y_{obs} | W, \theta)f(r | F, \psi) \tag{1.3}
\]

where \( y_{obs} \) is the observed subset of outcomes \( y \). In this case, we can treat \( y \) as independent of the missingness \( r \).

Under MAR, \( y \) is independent of missingness \( r \), given the observed subset of outcomes \( y_{obs} \):

\[
f(y, r | W, F, \theta, \psi) = f(y | W, \theta)f(r | y_{obs}, F, \psi) \tag{1.4}
\]

and consequently

\[
f(y_{obs}, r | W, F, \theta, \psi) = f(y_{obs} | W, \theta)f(r | y_{obs}, F, \psi) \tag{1.5}
\]
CLINICAL TRIALS WITH MISSING DATA

Under MNAR, missingness depends on unobserved outcomes, and we cannot transform (1.1) using independent factors for observed outcomes and missingness as we can under the MCAR and MAR assumptions. In this case, the joint distribution has to be treated as follows:

$$f(y_{obs}, r|W, F, \theta, \psi) = \int f(y|W, \theta)f(r|y, F, \psi) dy_{mis}$$ (1.4)

These definitions of the types of missingness are conditional on the statistical model used. Some authors (including Little and Rubin as cited) do not make the definitions conditional on the statistical model, but since the correctness of the model is important in practice, we follow Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007) in including it in the definition.

References


WHAT’S THE PROBLEM WITH MISSING DATA? 35


