JOURNAL OF CLINICAL RESEARCH BEST PRACTICES

Vol. 7, No. 5, May 2011 "Can You Handle the Truth?"

Making Sense of Biostatistics: Missing Data in Clinical Trials

By William Irish

Missing data in clinical trials reduces the power of a study and is a major source of potential bias when interpreting study results, especially if the amount of missing data is substantial. There are many reasons for missing data: subject refusal to continue in the study, treatment failures or successes, adverse events, missed study visits, equipment failures, and loss to follow-up. Some of these reasons may be related to study treatment, but the relationship is often difficult to determine. For example, did a subject miss his study visit because the study drug made him forgetful or because he simply forgot? Not even the subject may know.

Different degrees of data incompleteness can occur. Measurements may be available only at baseline, or they may be missed for one or several follow-up assessments. Even if a subject completes the study, some data might simply be uncollected because the study coordinator forgot to collect it.

Missing data can be classified into two broad categories: If a subject misses a follow-up visit but attends at least one subsequent follow-up visit, then data is considered "interim" missing. However, if a subject is no longer seen after a certain follow-up visit, then all subsequent follow-up data will be missing. This type of data is considered "withdrawal" missing. The terms dropout, attrition and loss to follow-up are often used synonymously with withdrawal.

There is no universally acceptable method for handling missing data. When data are missing, any attempt to draw conclusions from a statistical analysis rests on untestable assumptions concerning the relationship between the missing (unobserved) data and the reasons why the data is missing (the missing data mechanism).

However, the potential for bias can be evaluated to some extent based on the circumstances. The causes for missing data can be grouped into three different classes, depending on the relationship between the unobserved data and the missing data mechanism:

- **Missing Completely at Random (MCAR).** MCAR means that the missing data mechanism is unrelated to the values of any variables, whether missing or observed. Data that are missing because a researcher forgot to perform a test or a subject withdraws because he or she is moving out of the country are likely to be MCAR. Unfortunately, most missing data are not MCAR.
- **Missing at Random (MAR).** MAR means that the missing data mechanism is unrelated to the missing values but is related to either observed covariates or response variables. Data that are missing because a subject is removed from a study based on a pre-defined clinical condition like an elevated blood glucose level are MAR.
- **Non-Ignorable (NI).** NI is also known as Missing Not at Random (MNAR). It means that the missing data mechanism is related to the missing values. For example, an investigator is examining the effect of sleep on pain. Subjects are called daily and asked questions about their last night's sleep and their pain today. Subjects experiencing severe pain are relatively unlikely to come to the phone, leaving the data missing for that particular day. The missing data mechanism for pain is non-ignorable. Whether pain is missing or observed is

related to its value. It may also be related to other variables like concomitant medical conditions and economic level.

A key distinction is whether the mechanism is ignorable (i.e., MCAR or MAR) or non-ignorable. There are excellent techniques for handling ignorable missing data. For example, if a subject's blood pressure is stable in a cholesterol study, it is fair to interpolate a blood pressure value for a single visit. On the other hand, if a subject's blood pressure is gradually rising in a hypertension study and then the subject drops out without any communication, it is impossible to even guess a blood pressure value. Such non-ignorable missing data are more challenging and require a different, more complicated approach. Depending on the data, this approach may include multiple imputation methods, survival analysis techniques, and pattern mixture models.

The reason missing data can be such a problem is best illustrated by an example. Imagine a study in which subjects with known outcomes respond as shown in Table 1. Of the 100 subjects on Treatment A, 50% have a "good" response. Conversely, of the 100 subjects on Treatment B, 70% have a "good" response. The data in this example strongly supports the hypothesis that Treatment B is preferable (p=0.006).

Table 1. Hypothetical Example

Outcome	Treatment A	Treatment B
Good	50 (50%)	70 (70%)
Poor	50	30
Total	100	100

Now suppose there are 12 further subjects with missing outcomes. Depending on the treatment to which these 12 were allocated and their unknown outcomes, combining these with the data in Table 1 could lead to very different conclusions. For example, if six subjects were allocated treatment A and had a "good" response and six subjects were allocated to treatment B and had a "poor" response, then the data would not support the hypothesis that treatment B is preferable (p=0.069).

Given the problems of missing data, two critical elements are needed when conducting a clinical study:

- Careful protocol design and conduct to limit the amount and impact of missing data
- Analysis that makes full use of information on all randomized subjects and is based on careful attention to the assumptions about the nature of the missing data underlying estimates of treatment effects

Author

William Irish, PhD, is Vice President of Outcomes Research and Biostatistics at CTI Clinical Trial and Consulting Services. Contact him at 513.618.4057 or birish@ctifacts.com.