Use of Multiple Imputation Models in Medical Device Trials[1]

Donald B. Rubin and Samantha R. Cook

1. Abstract

Missing data are a common problem with data sets in most clinical trials, including those dealing with devices. Imputation, or filling in the missing values, is an intuitive and flexible way to handle the incomplete data sets that arise because of such missing data. Here we present several imputation strategies and their theoretical background, as well as some current examples and advice on computation. Our focus is on multiple imputation, which is a statistically valid strategy for handling missing data. The analysis of a multiply imputed data set is now relatively standard, for example in SAS and in Stata. The creation of multiply imputed data sets is more challenging but still straightforward relative to other valid methods of handling missing data. Singly imputed data sets almost always lead to invalid inferences and should be eschewed.

2. Introduction

Missing data are a common problem with large databases in general and with clinical and health care databases in particular. Subjects in clinical trials may fail to provide data at one or more time points or may drop out of a trial altogether, for reasons including lack of interest, untoward side effects, change of geographical location, and success of the procedure with no interest in follow-up assessments, etc. Data may also be “missing” due to death, although the methods described here are generally not appropriate for such situations because such values are not really missing (see Little and Rubin[1], example 1.7, and Zhang and Rubin[2]).

[1] A similar version of this chapter appears in cursory form as an entry in The Encyclopedia of Clinical Trials.

From: Clinical Evaluation of Medical Devices: Principles and Case Studies, Second Edition
Edited by: K. M. Becker and J. J. Whyte © Humana Press Inc., Totowa, NJ

241
An intuitive way to handle missing data is to fill in (i.e., impute) plausible values for the missing values, thereby creating completed data sets that can be analyzed using standard complete-data methods. The past 25 years have seen tremendous improvements in the statistical methodology for handling incomplete data sets using imputation. After briefly discussing missing data mechanisms, we present some common imputation methods, focusing on multiple imputation. We then discuss computational issues and present some examples.

3. Missing Data Mechanisms

A missing data mechanism is a probabilistic rule that governs which data will be observed and which will be missing. Little and Rubin and Rubin distinguish three types of missing data mechanisms. Missing data are missing completely at random (MCAR) if missingness is independent of both observed and missing values of all variables, almost random dart throwing at the data matrix. MCAR is the only missing data mechanism for which “complete-case” analysis (i.e., restricting the analysis to only those subjects with no missing data) is generally acceptable. Missing data are missing at random (MAR) if missingness depends only on observed values of variables and not on any missing values. For example, if the value of blood pressure at the end of a trial is more likely to be missing when some previously observed values of blood pressure are high, and given these the missingness is independent of the value of blood pressure at the end of the trial, then the missingness mechanism is MAR.

If missingness depends on the values that are missing, even after conditioning on all observed quantities, the missing data mechanism is not missing at random (NMAR). Missingness must then be modeled jointly with the data—the missingness mechanism is “nonignorable.” Nonignorable missing data present challenging problems because there is no direct evidence in the observed data about how to model the missing values.

The specific imputation procedures described here are most appropriate when the missing data are MAR and ignorable (see Little and Rubin and Rubin for details). Multiple imputation can still be validly used with nonignorable missing data, although it is more challenging to use it well. Multiple imputation is still more straightforward to use than other valid methods of handling the nonignorable situation.

4. Single Imputation

Single imputation refers to imputing one value for each missing datum. Singly imputed data sets are straightforward to analyze using complete-data methods, which makes single imputation an apparently attractive option with incomplete data. Little and Rubin offer the following guidelines for creating imputations. They should be: (1) conditional on observed variables; (2) multi-