14

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 Strata. 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. Data may also be missing owing to death, although the methods described here are generally not appropriate for such situations because such values are not really missing.[1,2]

---

[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

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.[3] 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[1] and Rubin[4] 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 thus is independent of the value of blood pressure at the end of the trial, 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[1] and Rubin[4] 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 attractive option with incomplete data. Little and Rubin[1] offer the following guidelines for creating imputations. They should be: (1) conditional on observed variables; (2) multivariate, to reflect

associations among missing variables; and (3) randomly drawn from predictive distributions rather than set equal to means to ensure that correct variability is reflected.

Unconditional mean imputation, which replaces each missing value with the mean of the observed values of that variable, meets none of the three guidelines listed above. Regression imputation can satisfy the first two guidelines by replacing the missing values for each variable with the values predicted from a regression (e.g., least squares) of that variable on other variables. Replacing missing values of each variable with the mean of that variable calculated within cells defined by categorical variables is a special case of regression imputation. Stochastic regression imputation adds random noise to the value predicted by the regression model, and when done properly, can meet all three guidelines.

Hot deck imputation replaces each missing value with a random draw from a donor pool of observed values of that variable; donor pools are selected, for example, by choosing individuals with complete data who have "similar" observed values to the subject with missing data, e.g., by exact matching or using a distance measure on observed variables to define "similar." Hot deck imputation, when done properly, can also satisfy all three of the guidelines listed above.

Even though analyzing a singly imputed data set with standard techniques can be straightforward, such an analysis will nearly always result in estimated standard errors that are too small, confidence intervals that are too narrow, and *p* values that are too significant, regardless of how the imputations were created. The reason is that imputed data are treated as if they were known with no uncertainty. Thus, single imputation is almost always statistically invalid, although the multiple version of a single imputation method will be valid if the imputation method is "proper." Proper imputations satisfy the three criteria of Little and Rubin.

4.1. Properly Drawn Single Imputations

Let Y represent the complete data, i.e., all the data we would observe in the absence of missing data, and let $Y = \{Y_{obs}, Y_{mis}\}$, where Y_{obs} is the observed data and Y_{mis} is the missing data. For notational simplicity, assume ignorability of the missing data mechanism. Also, let θ represent the (generally multicomponent) parameter associated with an appropriate imputation model, which consists of both a sampling distribution on Y governed by θ, $p(Y|\theta)$, and a prior distribution on θ, $p(\theta)$. A proper imputation is often most easily obtained as a random draw from the "posterior predictive distribution" of the missing data given the observed data, which formally can be written as:

$$p(Y_{mis}|Y_{obs}) = \int p(Y_{mis}, \theta|Y_{obs})d\theta = \int p(Y_{mis}|Y_{obs}, \theta)p(\theta|Y_{obs})d\theta \qquad (1)$$

This expression effectively gives the distribution of the missing values, Y_{mis}, given the observed values, Y_{obs}, under a model governed by θ, $p(Y|\theta)p(\theta)$. This distribution is called "posterior" because it is conditional on the observed Y_{obs}, and it is called "predictive" because it predicts the missing Y_{mis}.

If the missing data follow a monotone pattern (*see* Section 4.1.1.), drawing random samples from this distribution is straightforward. When missing data are not monotone, iterative computational methods are generally necessary, as described shortly.

4.1.1. Theory With Monotone Missingness

A missing data pattern is monotone if the rows and columns of the data matrix can be sorted in such a way that an irregular staircase separates Y_{obs} and Y_{mis}. Figures 1 and 2 illustrate monotone missing data patterns. Missing data in clinical trials are often monotone or nearly monotone when data are missing due to patient dropout, and once a patient drops out, the patient never returns.

Let Y_0 represent fully observed variables, Y_1 the incompletely observed variable with the fewest missing values, Y_2 the variable with the second fewest missing values, and so on. Proper imputation with a monotone missing data pattern begins by fitting an appropriate model to predict Y_1 from Y_0 and then using this model to impute the missing values in Y_1. For example, fit a regression of Y_1 on Y_0 using the units with Y_1 observed, draw the regression parameters from their posterior distribution, and then draw the missing values of Y_1 given these parameters and Y_0. Next, impute the missing values for Y_2 using Y_0 and the observed and imputed values of Y_1. Continue until all missing values have been imputed. The collection of imputed values is a proper imputation of the missing data, Y_{mis}, under this model, and the collection of univariate prediction models is the implied full imputation model. When missing data are not monotone, this method of imputation as described cannot be used directly.

4.1.2. Theory With Nonmonotone Missingness

Creating imputations when the missing data pattern is nonmonotone generally involves iteration because the distribution $p(Y_{mis}|Y_{obs})$ is often difficult to draw from directly. On the other hand, the data augmentation algorithm (DA),[5] a stochastic version of the EM algorithm,[6] is often straightforward to implement.

Briefly, DA involves iterating between randomly sampling missing data given a current draw of the model parameters and randomly sampling model parameters given a current draw of the missing data. The draws of Y_{mis} form a Markov Chain whose stationary distribution is $p(Y_{mis}|Y_{obs})$. Thus, once the Markov Chain has reached approximate convergence, a draw of Y_{mis} obtained by DA is effectively a proper single imputation of the missing data from the correct target distribution $p(Y_{mis}|Y_{obs})$, the posterior predictive distribution of

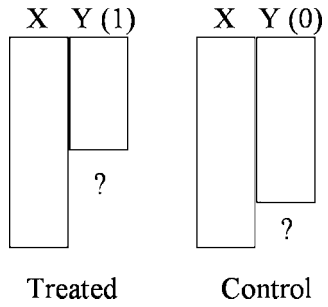

Fig. 1. Pattern of missing data for Intergel® trial.

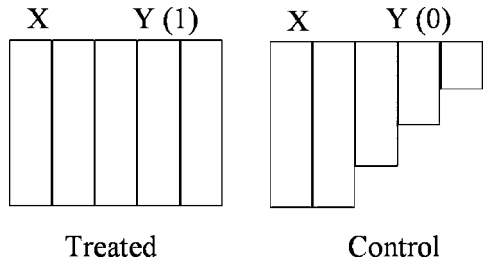

Fig. 2. Pattern of missing data for Genzyme trial.

Y_{mis}. Many of the programs discussed in Section 5.2. use DA to impute missing values. Other algorithms that use Markov Chain Monte Carlo methods for imputing missing values include variations such as Gibbs sampling[7] and Metropolis-Hastings[8–10].

As discussed previously, analyzing a singly imputed data set using complete-data methods usually leads to anticonservative results because imputed values are treated as if they were known, thereby underestimating uncertainty. Multiple imputation corrects this problem, while simultaneously retaining the advantages of single imputation.

5. Multiple Imputation

Described in detail in Rubin,[11] multiple imputation is a Monte Carlo technique that replaces the missing values Y_{mis} with $m > 1$ plausible values, $\{Y_{mis,1}, \ldots, Y_{mis,m}\}$ and therefore reveals and quantifies uncertainty in the imputed values. Each set of imputations creates a completed data set, thereby creating m "completed" data sets: $Y^{(1)}, \ldots, Y^{(l)}, \ldots, Y^{(m)}$, where $Y^{(l)} = \{Y_{obs}, Y_{mis,l}\}$. Typically, m is fairly small; $m = 5$ is a standard number of imputations to use.

Each of the *m* completed data sets is then analyzed as if there were no missing data and the results combined using simple rules described shortly.

Obtaining proper multiple imputations is no more difficult than obtaining a single proper imputation—the process for obtaining a proper single imputation is simply repeated independently *m* times. When the missing data pattern is not monotone, this involves generating *m* sequences of $\{Y^{(t)}_{mis}\}$ each with different starting values. Approximately independent multiple imputations can also be obtained from a single sequence by using only every *p*th draw of Y_{mis}, provided *p* and the length of the sequence are sufficiently large.

5.1. Combining Rules for Proper Multiple Imputations

As in Rubin[11] and Schafer[12], let Q represent the estimand of interest (e.g., the mean of a variable, a relative risk, the intention-to-treat effect, etc.), let Q_{est} represent the complete data estimator of Q (i.e., the quantity calculated treating all imputed values of Y_{mis} as known observed data), and let U represent the estimated variance of $Q_{est} - Q$. Let $Q_{est,l}$ be the estimate of Q based on the lth imputation of $Y_{mis,}$ with associated variance U_l—that is, the estimate of Q and associated variance are based on the complete-data analysis of the lth completed data set, $Y_l = \{Y_{obs}, Y_{mis,l}\}$, $l = 1, \ldots, m$.

The multiple imputation estimate of Q is simply the average of the m estimates: $QMIest = \Sigma^{m}_{l=1} Q_{est,l}/m$. The estimated variance of $Q_{MIest} - Q$ is found by combining between and within imputation variance, as with the analysis of variance: $T = U_{ave} + (1 + m^{-1})B$, where $U_{ave} = \Sigma^{m}_{l=1} U_l/m$ is the within imputation variance, and $B = \Sigma^{m}_{l=1}(Q_{est,l} - Q_{MIest})^2/(m - 1)$ is the between imputation variance. The quantity $T^{-1/2}(Q–Q_{MIest})$ follows an approximate t_v distribution with degrees of freedom $v = (m - 1)(1 + U_{ave}/((1 + m - 1)B))^2$. Rubin and Schenker[13] provide additional methods for combining vector-valued estimates, significance levels, and likelihood ratio statistics, and Barnard and Rubin[14] provide an improved expression for v with small complete data sets (*see also* ref. *1*).

5.2. Computation for Multiple Imputation

Many standard statistical software packages now have built-in or add-on functions for multiple imputation. The S-plus libraries, NORM, CAT, MIX, and PAN, for analyzing continuous, categorical, mixed, and panel data, respectively, are freely available[12] (http://www.stat.psu.edu/_jls/), as is MICE[15] (http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm), which uses regression models to impute all types of data. SAS now has procedures PROC MI and PROC MIANALYZE; in addition IVEwear citeRagu01 is freely available and can be called using SAS (http://support.sas.com/rnd/app/da/new/dami.html). WinBUGS,[16] a stand-alone software package for fitting Bayesian models, imputes

missing values when data are incomplete, assuming missing values are MAR. New software packages have also been developed specifically for multiple imputation. Examples are the commercially available SOLAS (www.statsol.ie/solas/solas.htm), which has been available for years and is most appropriate for data sets with a monotone or nearly monotone pattern of missing data, and the freely available NORM, a stand-alone Windows version of the S-plus function NORM (www.stat.psu.edu/_jls/). Recently, Stata announced that it supports analyses of multiply imputed data sets. (For more information, *see* www.multiple-imputation.com or Horton and Lipsitz.[17])

6. Examples

6.1. Lifecore

Intergel® solution is a medical device developed by Lifecore Biomedical to prevent surgical gynecological adhesions. A double-blind, multicenter randomized trial was designed for the US Food and Drug Administration (FDA) to determine whether Intergel significantly reduces the formation of adhesions after gynecological surgery. The data collection procedure for this study was fairly intrusive: patients had to undergo a minor abdominal surgery (a laparoscopy) weeks after the first surgery in order for doctors to determine the primary endpoint, the number of gynecological adhesions. The trial suffered from missing data because not all women were willing to have another surgery, despite having initially agreed to do so. Medical device trials (and clinical trials in general) often suffer from missing data when data collection methods are invasive.

The original proposal from the FDA for imputing the missing values (the counts of adhesions) was to fill in the worst possible value (defined to be 32 adhesions) for each missing datum, which should lead to conservative results because there were more missing data in the treatment arm than in the placebo arm. This method ignores observed information when creating imputations. For example, most patients with observed data had 10 or fewer adhesions. Furthermore, because the imputed values were so much larger than the observed values, the standard errors based on these worst-possible value imputations were inflated, making it unlikely that significant results would be found, even when the two treatments were substantially different. Figure 1 displays the general pattern of monotone missing data in this case, with X representing covariates, $Y(0)$ outcomes under placebo, and $Y(1)$ outcomes under Intergel. The question marks represent missing values.

Colton, Piantadosi, and Rubin[18] instead used a multiple imputation hot deck procedure to impute missing values. Donor pools for each patient with missing data were defined by treatment group and covariates: treatment center and three

measures of baseline seriousness of adhesions, which were observed for all patients. For each patient whose outcome was missing, the donor pool consisted of the two patients in the same treatment group and treatment center who had the closest baseline adhesion scores. "Closeness" was defined by the Mahalanobis metric, a corrected Euclidean squared distance, calculated for the baseline adhesion measures. For each patient with missing data, the first imputation consisted of a random draw from that patient's donor pool. The remaining value in the donor pool was used for the second imputation. The small number of imputations was deemed acceptable because less than 6% of the outcomes were missing.

Formally, this method is improper, but the limited donor pools should still make the method conservative because the matches are not as close as they would be with bigger sample sizes, or as they could be if a smooth model were used to create the imputations. The donor pool approach also has the advantage that imputations can be created without using any outcome data while remaining blind to treatment group labels, meaning that there is no opportunity to create imputations that influence results in a particular intended way.

6.2. Genzyme

Fabrazyme® is a drug developed by Genzyme Corporation to treat Fabry's disease, a rare and serious X-linked recessive genetic disease that occurs due to an inability to metabolize creatinine. Preliminary results from a phase III FDA trial of Fabrazyme vs placebo showed that the drug appeared to work well in patients in their 30s who were not yet severely ill, in the sense that it lowered their serum creatinine substantially. A similar phase IV trial involved older patients who were more seriously ill. Because there was no other fully competitive drug, it was desired to make Fabrazyme commercially available earlier than initially planned, a decision that would allow patients randomized to placebo to begin taking Fabrazyme but would create missing outcome data among placebo patients after they began taking the drug. The study had staggered enrollment, so that the number of monthly observations of serum creatinine for each placebo patient depended on the time of entry into the study. Figure 2 illustrates the general pattern of monotone missing data with the same length follow-up for each patient. Again, X represents baseline covariates, $Y(0)$ represents repeated measures of serum creatinine for placebo patients, and $Y(1)$ represents repeated measures of serum creatinine for Fabrazyme patients.

In order to impute the missing outcomes under placebo, a complex hierarchical Bayesian model was developed for the progression of serum creatinine in untreated Fabry patients. In this model, inverse serum creatinine varies linearly and quadratically in time, and the prior distribution for the quadratic trend in placebo patients is obtained from the posterior distribution of the qua-

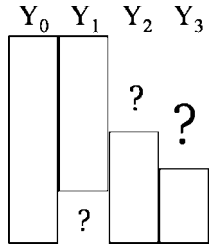

Fig. 3. Illustrative display for type of pattern of missing data in National Medical Expenditure Survey.

dratic trend in an analogous model fit to a historical database of untreated Fabry patients. Thus, the historical patients' data only influence the imputations of the placebo patients' data rather subtly—via the prior distribution on the quadratic trend parameters.

Although the model fitting algorithm is complex, it is straightforward to use the algorithm to obtain draws from $p(\theta|Y_{obs})$ for the placebo patients and then draw Y_{mis} conditional on the drawn value of θ, where, as earlier, θ represents all model parameters. Drawing the missing values in this way creates a sample from $p(Y_{mis}|Y_{obs})$ and thus an imputation for the missing values in the placebo group.

The primary analysis will consider the time to an event, defined as either a clinical event (e.g., kidney dialysis, stroke, death) or a substantial increase in serum creatinine relative to baseline. The analysis will be conducted on each imputed data set and the results combined (as outlined earlier in Section 5.1.) to form a single inference. Although Fabrazyme is not a medical device, the missing data mechanism in this example may be similar to those in medical device trials. Also, the availability of potentially relevant historical data is common in medical device trials.

6.3. National Medical Expenditure Survey

The National Medical Expenditure Survey (NMES) collects data, including hundreds of measurements of medical expenditures, background information, and demographic information on a random sample of approximately 30,000 members of the US population. Again, although NMES does not explicitly deal with medical devices, the general pattern of missing data and corresponding issues arising in this medical database may also arise in medical device databases.

Multiple imputation for NMES was more complicated than in the previous two examples because the missing data pattern was not monotone. Figure 3

shows a tremendous simplification of the missing data pattern for NMES, where, if Y_1 were fully observed, the missing data pattern would be monotone.

Rubin[19] imputed the missing data in NMES by capitalizing on the simplicity of imputation for monotone missing data by first imputing the missing values that destroyed the monotone pattern (the "nonmonotone missing values"), proceeding as if the missing data pattern were in fact monotone, and then iterating this process. More specifically, after choosing starting values for the missing data, iterate between the following two steps: (1) regress each variable with any nonmonotone missing values (i.e., Y_1), on all the other variables (i.e., Y_0, Y_2, Y_3), treating the current imputations as true values, but use this regression to impute only the nonmonotone missing values; and (2) impute the remaining missing values in the monotone pattern; first impute the variable with the fewest missing values (Y_2 in Fig. 3), then the variable with the second fewest missing values (Y_3 in Fig. 3), and so on, treating the nonmonotone missing values filled in during step 1 as known. This process was repeated five times to create five sets of imputations in the NMES example.

7. Summary

MI is a flexible tool for handling incomplete data sets. MIs are often straightforward to create using computational procedures such as data augmentation or using special MI software now widely available. Moreover, the results from imputed data sets are easy to combine into a single MI inference. Although MI is Bayesianly motivated, many MI procedures have been shown to have excellent frequentist properties.[20] In small samples, the impact of the prior distribution on conclusions can be assessed by creating MIs using several different prior specifications, and more generally, the impact of different models on conclusions can be analogously assessed. Furthermore, although only MAR procedures have been considered here, missing data arising from an NMAR mechanism may be multiply imputed by jointly modeling the data and the missingness mechanism; in some cases, results are insensitive to reasonable missingness models and the missing data can then be effectively treated as being MAR.[11] Rubin, Schafer, and Little[1,11,12] provide more detail on the ideas presented here, with some information being less technical and more accessible than others.

References

1. Roderick, J., Little A., and Rubin, D. B. 2002. *Statistical Analysis With Missing Data*. 2nd ed. Wiley Interscience, New Jersey, example 1.7.
2. Zhang, J. L. and Rubin, D. B. 2003. Estimation of causal effects via principal stratification when some outcomes are truncated by "death." *J. Educ. Behav. Statist.* 28:353–368.

3. Rubin, D. B. 1978. Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse (with discussion). *ASA Proceedings of the Section on Survey Research Methods.* 20–34.
4. Rubin, D. B. 1976. Inference and missing data (with discussion). *Biometrika*. 63:581–592.
5. Tanner, M. A. and Wong, W. H. 1987. The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.* 82:528–550.
6. Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion*). J. Royal Statist. Soc., Series B.* 39:1–38.
7. Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intel.* 6:721–741.
8. Metropolis, N. and Ulam, S. 1949. The Monte Carlo method. *J. Am. Statist. Assoc.* 49:335–341.
9. Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* 57:97–109.
10. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*. 2nd ed. Chapman & Hall, London, 2003.
11. Rubin, D. B. 2004. *Multiple Imputation for Nonresponse in Surveys*. 2nd ed. John Wiley & Sons, New York.
12. Schafer, J. L. 1997. *Analysis of Incomplete Data*. Chapman and Hall, London.
13. Rubin, D. B. and Schenker, N. 1991. Multiple imputation in health-care databases: an overview and some applications. *Statist. Med.* 10:585–598.
14. Barnard, J. and Rubin, D. B. 1999. Small-sample degrees of freedom with multiple imputation. *Biometrika.* 86:948–955.
15. van Buuren, S., Boshuizen, H. C., and Knook, D. L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statist. Med.* 18:681–694.
16. Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. 2002. *WinBUGS (Version 1.4)*. Medical Research Council, Biostatistics Unit, Cambridge, England. Available from: www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.
17. Horton, N. J., and Lipsitz, S. R. 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am. Statist.* 55:244–254.
18. Colton, T. Piantadosi, S., and Rubin, D. B. 2001. Multiple imputation for second-look variables based on intergel pivotal trial data. Report submitted to FDA.
19. Rubin, D. B. 2003. Nested multiple imputation of NMES via partially incompatible MCMC. *Statist. Neerlandica,* 57:3–18.
20. Rubin, D. B. 1996. Multiple imputation after 18+ years (with discussion). *J. Am. Statist. Assoc.* 91:473–520.