ABSTRACT

Title of Document: MISSING DATA ANALYSIS: A CASE STUDY OF A RANDOMIZED CONTROLLED TRIAL

Shaleah Mary Murphy Patzer, MPH, 2009

Directed By: Dr. Guangyu Zhang, Department of Epidemiology and Biostatistics

Missing data is a pervasive problem in the analysis of many clinical trials. In order for the analysis of a study to produce unbiased estimators, the missing data problem must be addressed. First, the missing data pattern must be established; second, the missingness mechanism must be determined; and third, the most appropriate imputation method for imputing the missing values must be found. The purpose of this paper is to explore the imputation methods best suited for the missing data from the Diet and Exercise for Elevated Risk Trial (DEER) in a secondary analysis of the data. The missingness pattern in the data set is arbitrary and the missingness mechanism is MAR. A simulation study suggests that the two best methods for imputation are subject-specific mean imputation and multiple imputation. I conclude that mean imputation is the best method for handling missing data in the DEER data set.

MISSING DATA ANALYSIS:
A CASE STUDY OF A RANDOMIZED CONTROLLED TRIAL

By

Shaleah Mary Murphy Patzer

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Masters of Public Health
2009

Advisory Committee:
Dr. Guangyu Zhang, Chair
Dr. Deborah Young
Dr. Tongtong Wu

Table of Contents

Table of Contents ii
List of Tables iii
List of Figures iv
List of Figures iv
List of Equations v
Chapter 1: Introduction 1
Chapter 2: Methods 7
Chapter 3: Results 18
Chapter 4: Discussion 26
Chapter 5: Conclusion 29
Appendix A 30
Appendix B 31
Appendix C 36
Bibliography 38

List of Tables

Table 1. Missing data in DEER data set....10
Table 2. Missingness Mechanisms....12
Table 3. Missingness Pattern....21
Table 4. Comparison of baseline variable means....22
Table A.1. Between-group parameter estimates....30
Table C.1. Between-group parameter estimates (n = 175, m = 10, 50)....36
Table C.2. Between-group parameter estimates (n = 175, m = 10, 50)....37

List of Figures

Figure 1. Patterns of Non-Response 2
Figure 2. Comparison of RMSE for Change in CRP 18
Figure 3. Comparison of MAD for Change in CRP 19
Figure 4. Comparison of Bias for Change in CRP 20
Figure 5. DEER: Comparison of Models 25
Figure B.1. Comparison of RMSE for Change in CRP (n = 175, m = 10, 50) 31
Figure B.2. Comparison of MAD for Change in CRP (n = 175, m = 10, 50) 32
Figure B.3. Comparison of Bias for Change in CRP (n = 175, m = 10, 50) 33
Figure B.4. Comparison of RMSE for Change in CRP (n = 500, m = 10, 50) 33
Figure B.5. Comparison of MAD for Change in CRP (n = 500, m = 10, 50) 34
Figure B.6. Comparison of Bias for Change in CRP (n = 500, m = 10, 50) 35

List of Equations

Equation 1. Complete data set 3
Equation 2. MCAR 3
Equation 3. MAR 4
Equation 4. MNAR 4
Equation 5. Missing data distribution 4
Equation 6. Likelihood Function 6
Equation 7. MI standard error 14
Equation 8. RMSE 16
Equation 9. MAD 16
Equation 10. Bias 16

Chapter 1: Introduction

Missing data can create problems in statistical analyses for multiple reasons. A major problem is that many statistical procedures depend on complete-case methods of analysis (Allison, 2002; Rubin, 1987). In other words, standard statistical programs require any case being analyzed have a value for every variable in the analysis. Such programs eliminate from analysis any case that contains one or more missing value(s) for any variable of interest, continuing the analysis as though the remaining cases are the complete data set. Inadvertent deletion of cases on the part of the analyst and/or statistical program can lead to two possibly serious problems: non-response bias and reduced analytic power. Both biased and inefficient (reduced analytic power) answers are unreliable (Schafer & Graham, 2002).

Non-response bias occurs when a subset of respondents who fail to answer a particular question, creating missing data, differ in important ways from the subset of respondents who provide the answer (Barnard & Meng, 1999). Potential differences between the two subsets of respondents can cause a bias, or systematic pattern, that characterizes the missing data. The analyst may never know the reason, or reasons, behind the non-response, but simple tests using dummy variables can be conducted to explore for potential differences between the groups.

Compromised analytic power is a function of the percentage of missing information (Allison, 2002; Heitjan, 1997). Incomplete data on only one variable of interest can render a case completely useless in multivariate analysis. Thus a significant proportion of the original sample can be lost in analysis. Such elimination

resulting in a smaller sample not only reduces the analytic power of the study, but it can also introduce systematic selection bias.

1.1. Types of Non-Response

There are two types of non-response that create missing data: item non-response and unit non-response (Rubin, 1987). Item non-response occurs when a respondent fails to answer a particular item or items in a survey. Unit non-response occurs when a respondent fails to answer any items on a survey. The distinction between item and unit non-response is important for determining approaches to handling missing data.

1.2. Patterns of Non-Response

There are three patterns of non-response that are most easily understood in the following figure (Schafer & Graham, 2002).

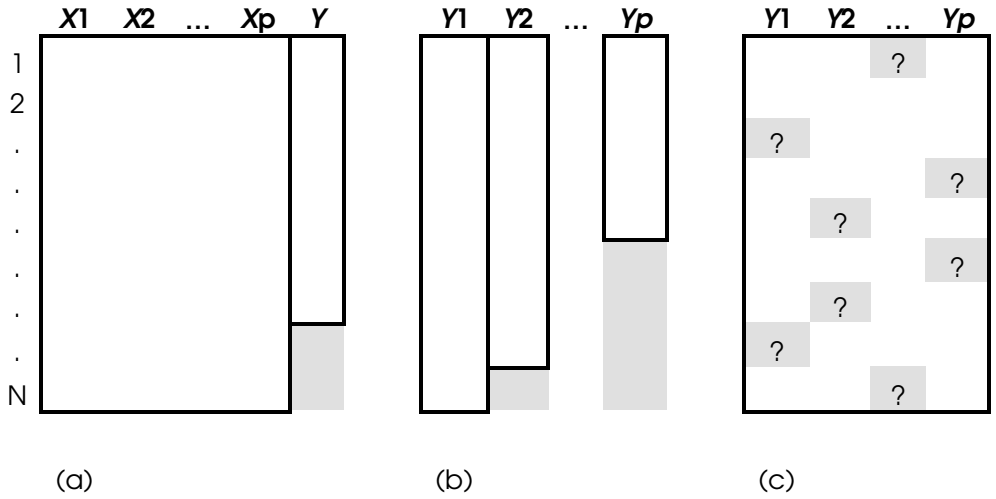

Figure 1. Patterns of Non-Response
(a) univariate non-response; (b) monotone non-response; (c) arbitrary non-response. Missing data represented by shaded squares. Adapted from Little & Rubin, 1989.

Univariate non-response, Figure 1a, occurs when a single variable, *Y*, has missing values but all other variables are completely observed. *Y* can also represent a subset

of variables that are entirely observed or entirely missing for each case. Monotone non-response, Figure 1b, has items or groups of items Y_1 through Y_p that can be ordered so that if Y_j is missing for a case, Y_{j+1} through Y_p are also missing. Finally, arbitrary non-response, Figure 1c, occurs when any variable(s) is missing for any case(s). Arbitrary missingness creates complications in modeling, estimation, and imputation analyses (Rubin, 1987).

1.3. Describing Missing Data

Appropriately handling missing data requires that the missingness mechanism be identified. Data can be missing in three ways: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Schafer & Graham, 2002). In order to describe the missingness mechanism using a generic notation, let Y_{com} represent the complete data set, Y_{obs} represent the observed data set (the subset used in analysis), and Y_{miss} represent the missing cases data set (Rubin, 1976). Therefore,

$$Y_{com} = (Y_{obs}, Y_{miss}).$$

Equation 1. Complete data set

Responses are said to be MCAR when the distribution of missingness does not depend on Y_{miss} or Y_{obs}, such that,

$$P(R| Y_{com}) = P(R),$$

Equation 2. MCAR

where R represents missingness. In other words, a participant's nonresponse does not depend on his or her own values for the observed or missing variables. MAR occurs when the distribution of missingness depends on Y_{obs} , but not Y_{miss},

$$P(R| Y_{com}) = P(R| Y_{obs}).$$

Equation 3. MAR

In other words, a participant's nonresponse may depend on his or her own values for the observed variables, but not the missing variables. MAR is also called ignorable nonresponse. When missingness depends on Y_{miss} as well as Y_{obs},

$$P(R|\ Y_{com}) = P(R|\ Y_{obs}, Y_{miss})$$

Equation 4. MNAR

then the data is said to be MNAR In other words, the probability of a participant missing values depends on the missing variables. MNAR represents nonignorable nonresponse. The definitions of missing data, MCAR, MAR, and MNAR, only describe the relationships between data and missingness: they are not causal.

Given the three missingness mechanisms, the implications for analysis are different (Schafer & Graham, 2002). Statistical methods for complete-case analysis (Y_{com}) are generally motivated by the assumption that the data are sampled from a population distribution $P(Y_{com}; \theta)$, where θ represents unknown parameters. The distribution, $P(Y_{com}; \theta)$, can be interpreted in two ways: (1) as a description of the probability of obtaining a particular data set among all possible data sets that could occur over a hypothetical number of samplings and data collections or (2) as the likelihood function for θ. When a data set has missing data, simply basing all statistical analyses on $P(Y_{obs}; \theta)$ and thus discounting the missing data in the distribution of the observed is easily accomplished. The resultant distribution is the definite integral:

$$P(Y_{obs}; \theta) = \int P(Y_{com}; \theta) d\, Y_{miss}.$$

Equation 5. Missing data distribution

However, construction of the $P(Y_{obs}; \theta)$ distribution in this manner does not necessarily yield either a correct sampling distribution or likelihood function (Rubin 1976). For the observed sampling distribution to accurately represent the population, the missing data must be MCAR. For the observed likelihood function to accurately represent the population, the missing data need only be MAR. Based on Rubin's two conditions, the weaker condition, that missing data need only be MAR, implies that statistical procedures based on likelihood functions are more functional than those based solely on repeated-sampling arguments. Such procedures are better suited to handle real-world situations in which MCAR is usually violated and should, therefore, produce more representational and reliable results (Schafer & Graham, 2002).

Equation 5 is also suited to models where the missing values are out of the scope of the universe of interest. In other words, the missing data are not actually missing. This usually occurs in questionnaire surveys (a person with no children leaves blank a question that asks, "How often do you see your children?") and longitudinal studies (participants may die in studies whose outcomes do not include death, ie. cognitive function). In these cases, the hypothetical missing data can be treated as MAR.

Equation 5 cannot be used to define a probability distribution with correct sampling distribution or likelihood function when data is MNAR. In such cases, a joint probability distribution must be calculated that includes the explicit model for missingness, $P(R| Y_{com}; \xi)$ where ξ stands for the unknown parameters of the

missingness distribution. Thus, the joint probability distribution is the product of P(R| Y_{com}; ξ) and P(Y_{com}; θ) and the correct likelihood function is:

$$P(Y_{obs}, R; \theta, \xi) = \int P(R| Y_{com}; \xi) P(Y_{com}; \theta) d Y_{miss}.$$

Equation 6. Likelihood Function

In general, the missingess model is a nuisance because the real questions of interest are usually about the distribution of Y_{com}, not R. However, Equation 6 can offer more and differing information about θ than Equation 5. As a final note, it is impossible to differentiate between MNAR and MAR, only MCAR can be reliably detected (McKnight, et al., 2007). Because of the inability to distinguish between MNAR and MAR, decisions regarding further analyses can only be based on sensible logic.

Chapter 2: Methods

2.1. DEER

The Diet and Exercise for Elevated Risk Trial (DEER) was a year-long, randomized controlled trial conducted at the Stanford Medical School's Center for Research in Disease Prevention (Stefanick, et al., 1998). The trial began in 1992 with the recruitment of 197 men and 180 women into the final cohort. The original objective of DEER was to analyze the effects of (1) low-fat diet, (2) exercise, or (3) low-fat diet plus exercise on lipoprotein levels in individuals at high risk for cardiovascular disease. The three intervention groups were compared to a control group. The original analysis stratified by gender because of the differing inclusion and exclusion criteria for men and women.

This paper is a secondary data analysis of the DEER data set with respect to handling missing data in the ascertaining of the effects of diet, exercise, or diet plus exercise on the change of C-reactive protein (CRP) from baseline to follow-up. The analysis will focus only on the female subject subset because the female subset has a higher percentage of missing data (~26% missing for women versus ~22% missing for men) and because there were no significant between and within group differences for the male subset.

Women were recruited from the Palo Alto, California area. Inclusion criteria included: postmenopausal, 45-64 years of age, BMI $\leq$ 32 kg m^{-2}, LDL 126-209 mg dL^{-1}, HDL < 60 mg dL^{-1}, blood pressure under 160/95 mmHg, fasting glucose below 140 mg dL^{-1}, triglycerides less than 500 mg dL^{-1}, and a normal maximal exercise treadmill test. Exclusion criteria included: history of a life-threatening disease (ie.

stroke, cancer, heart disease), heavy smokers (>9 cigarettes per day), heavy drinkers (>4 alcoholic drinks per day), inability to engage in moderate-intensity physical activity, or taking medications for blood pressure, heart problems, or to lower cholesterol.

A secondary data analysis of the DEER data set was conducted by Camhi (2008). A component of Camhi's research examined the relationship between intervention group (diet, exercise, diet plus exercise, control) and the change in CRP levels from baseline to follow-up. All of the subjects with a baseline or follow-up CRP level greater than ten were removed from the analysis ($n = 5$), so the sample size for the Camhi analysis was $n = 175$. A CRP level greater than ten indicates an acute infection, which is not relevant to the study and can bias the results. Camhi used ANCOVA to determine the between and within group differences in Change in CRP (follow-up CRP minus baseline CRP) in a complete-case analysis ($n = 130$). A total of 45 cases were deleted due to missing data (see Table 1 in Section 2.1.2 for the number of cases missing values for the variables of interest).

This paper also examines between and within group differences for Change in CRP from baseline to follow-up. In a simulation study, the imputed values for Change in CRP are compared with the true values for Change in CRP. The imputation methods are last observation carried forward, last observation carried backward, mean imputation, and multiple imputation. The imputation methods are described in Section 2.3. The most accurate and least biased imputation methods will then be applied to the DEER data set. The final comparison is between the complete case model of the DEER data set and the DEER data sets with imputed data.

2.1.1. Variables of Interest

The analysis controlled for baseline CRP, cohort, baseline body fat percentage, change in body fat, cigarettes per day, alcoholic drinks per day, age, and hormone replacement therapy. Change in body fat was included in the analysis in order to eliminate its effect on CRP. All baseline measurements were taken prior to randomization. All follow-up measurements were taken after one year intervention. The following variables are used in the analysis.

CRP (baseline and follow-up): measured from stored plasma samples using immunoturbidimetric assay on the Hitachi 917 analyzer (Roche Diagnostics – Indianapolis, IN) with reagents and calibrators from DiaSorin (Stillwater, MN); used to compute the dependent variable, Change in CRP.

Intervention status: participants were randomized using the Efron procedure into one of four categories.

- Control: Participants were asked to make no changes to their current lifestyle practices over the intervention period.
- Low-fat Diet: Participants were asked to meet the 1993 Step II dietary guidelines of the National Cholesterol Education Program (total fat < 30% of total calories, total saturated fat < 7% of total calories, dietary cholesterol < 200 mg day^{-1}).
- Exercise: After an initial six week period of one hour aerobics instruction three times a week, participants were asked to perform 20 minutes at 60-85% maximum heart rate three times a week with increased duration over the study

period to 45-60 minutes. If participants were already active, they were asked to increase the duration of their activity by 20 minutes.

- Low-fat Diet plus Exercise: Participants received both the low-fat diet and exercise interventions (separately from the other two groups).

Body fat skinfold (baseline and follow-up): measurements from the right triceps, suprailiac, and thigh were averaged.

Cohort: recruitment cohort.

Age: age at the time intervention began.

Cigarettes per day: a self-report baseline measurement of the average number of cigarettes smoked per day.

Alcoholic drinks per day: a self-report baseline measurement of the average number of alcoholic drinks consumed per day.

Hormone replacement therapy: a binary, self-report baseline measurement of active hormone replacement therapy (Note: randomized women agreed not to change their use of hormone replacement therapy for the intervention period).

2.1.2. Missing Values

As reported in Table 1, all but one of the variables used in the analysis have missing data. The total number of observations removed from analysis due to missing data is 45, which results in approximately a 26 percent reduction in sample size (n = 130).

Table 1. Missing data in DEER data set

Variable	Number of Missing Observations
Baseline CRP	6
Follow-up CRP	11
Baseline Body Fat	4
Follow-up Body Fat	16

Age	1
Cigarettes/day	0
Alcoholic Drinks/day	12
HRT	15
Total Removed	45

2.2. Simulation

The simulation study analyzes 1000 data sets of $n = 175$ (like the DEER data set) and 1000 data sets of $n = 500$. Variable values are generated using parameter estimates (mean and standard deviation for baseline variables, regression coefficients for follow-up variables) of the variables of interest in the DEER data set. In three separate analyses, data are removed so that the missingness mechanism is first MCAR, then MAR, and, finally, MNAR.

Methods are compared not only across missingness mechanisms, but also percentage of missingness. In the DEER data set, approximately ten percent of the values for baseline CRP and follow-up CRP are missing. Imputation methods are compared when the data are missing at ten percent and 50 percent. Table 2 shows how the data are removed from the data set so that the mechanisms are MAR and MNAR. MAR missingness depended on baseline body fat and age values. Baseline CRP values were removed if baseline body fat levels were less than a set value. Follow-up CRP values were removed if age was greater than a set value. The values of baseline body fat and age were chosen so that the rate of missingness was ten or 50 percent. MNAR missingness depended on the value of CRP itself. The value was chosen so that missingness would either be at a ten or 50 percent rate. A random number generator was used to remove data so that the missingness mechanism was

MCAR. The missingness mechanism for all of the other variables in the data set was MCAR.

Table 2. Missingness Mechanisms

	10%		50%	
	CRP1	**CRP2**	**CRP1**	**CRP2**
MAR	body fat < 22	age > 65	body fat < 30	age > 60
MNAR	CRP1 < 0. 25	CRP2 < 0.6	CRP1 < 0.7	CRP2 < 1.3

Summary statistics, described in Section 2.4, of the variable Change in CRP after the four imputations, described in Section 2.3., are compared when the missing data are MCAR, MAR, and MNAR, the sample size is $n = 175$ and the rate of missingness is ten percent. The summary statistics of the other simulations can be found in Appendix A. All parameter estimates can be found in Appendix B.

2.3. Imputation Methods

Different approaches to analysis with missing data have been proposed over the years. Traditionally, cases with missing values were removed from the analysis in a deletion process. Another method is single imputation in which missing data are imputed (replaced) by a simple estimate based on the entire data set. The most recent trend in data analysis has been to conduct multiple imputation. More complete descriptions of the methods used in this project, including an analysis of their strengths and weaknesses, follows.

2.3.1. Listwise Deletion

Also known as complete-case analysis, listwise deletion is among the oldest methods of adjusting for missing data. This technique simply deletes all cases with missing value(s) from the analysis. As such, it is the default method used by many statistical programs (Allison, 2002). If the missing data are MCAR, listwise deletion

will yield unbiased parameter estimates, however, the standard errors may be larger because of the smaller sample size. If the proportion of missing data is too large, then bias may be introduced into the parameter estimates and the results may be misleading. When the missing data are MAR, listwise deletion will lead to biased parameter estimates (regression coefficients that are too large or too small). Additionally, the analytic power is reduced when a large portion of the data are removed from the analysis. This method will only be used in the final analysis of the DEER data set; it will not be used for the simulation study.

2.3.2. Single Imputation: LOCF, LOCB, and Population Mean

Single imputation is the ascription of a value to a missing data cell based on a reasonable estimate of the absent data or the values of other variables (Little & Rubin, 1989). Three types of single imputation are used in this analysis: last observation carried forward, last observation carried backward, and mean imputation.

The last observation carried forward (LOCF) method assigns the last known value of a variable to the missing follow-up value. Thus, only follow-up values are imputed. The LOCF method can produce underestimates of variances and covariances (Allison, 2002). The last observation carried backward (LOCB) method assigns the next known value of a variable to the missing previous value. Like LOCF, LOCB can produce underestimates of variances and covariances. In this simulation and in the DEER data set, the values imputed using LOCF and LOCB are the subject mean since there are only baseline and follow-up values for CRP.

Population mean imputation substitutes the mean of the variable (column mean) for missing values. A drawback of this method is that the uniqueness of the

subject is lost—the subject becomes "normal". Mean imputation also ignores non-response bias and can lead to incorrect statistical inferences. Another drawback to mean imputation techniques is that they do not consider the variability between imputations because only one value is imputed, in effect reducing the plausibility of the parameter estimates and error terms (Schafer, 1999). Additionally, single imputation treats the missing values as if they are known when they are not (Rubin, 1987). However, mean imputation performs well when there is a missingness rate of 30 percent or less (Shrive, Stuart, Quan, & Ghali, 2006).

2.3.3. Multiple Imputation

Multiple imputation (MI), a relative newcomer to missing data analysis methods, was first introduced by Rubin in 1977. The basic principles are quite simple: (1) impute the missing values in a data set using an appropriately selected model that includes random variation; (2) impute M times, producing M "complete" data sets (generally accepted number of imputations is 5); (3) conduct the analysis on each data set using complete-data methods; (4) create a single-point estimate by averaging the parameters estimates across the M samples; (5) calculate the standard errors using the following relation:

$$\sqrt{\frac{1}{M}\sum_{k} S_k^2 + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{k}(b_k - \bar{b})^2},$$

Equation 7. MI standard error

where b_k is the estimated regression coefficient in sample k of the M samples, S_k its estimated standard error, and $\bar{b}$ is the mean of b_k.

Appropriate use of MI must meet several requirements (Rubin, 1996). First, the data must be MAR. Second, the model used to impute the data should match the

model that is being used in the complete-data analysis. In other words, the imputation model must preserve the important associations among variables in the data set, including the dependent variable. Finally, the algorithm that generates the imputed values needs to be "correct;" that is, the algorithm must accommodate the included variables and their associations. Good imputation methods use all available information related to missing values (Rubin, 1987).

The benefits of MI are multifold (Allison, 2000). The introduction of random error into the imputation method creates approximately unbiased parameter estimates. The repetition of the imputation makes reliable estimates of the standard error possible. Finally, MI can be used on any type of data, for any type of analysis, without the use of specialized software. For this analysis, PROC MI and PROC MIANALYZE from SAS version 9.1 were used. Multiple imputation was done using the Monte Carlo Markov Chain (MCMC) method for arbitrary missingness. The MCMC method generates pseudo-random draws via Markov chains from multidimensional probability distributions (Schafer, 1997). Markov chains, originating in physics, are a series of random variables in which the distribution of each component depends only on the value of the previous one.

2.4. Comparison of Methods

Three summary measures are used to compare the performance of the four imputation methods. Two are measures of accuracy and the third is a measure of bias.

2.4.1. Root-Mean-Square Error

The root-mean-square error (RMSE) is defined as:

$$RMSE = \sqrt{\sum (\hat{y} - y)^2 / n}$$

Equation 8. RMSE

where $\hat{y}$ is the imputed value of the missing observation, *y* is the true value of the observation, and *n* is the number of observations in the data set. The RMSE is an accuracy measure for how close the estimated values are to the true values. The RMSE penalizes outliers because the difference term, imputed minus true, is squared (Engels & Diehr, 2003). The closer to zero the RMSE is, the more accurate it is.

2.4.2. Mean Absolute Deviation

The mean absolute deviation (MAD) is defined as:

$$MAD = \sum |\hat{y} - y| / n$$

Equation 9. MAD

The MAD is another measure of how close predicted values are to observed values. Similar to the RMSE, the closer the MAD is to zero, the more accurate it is.

2.4.3. Bias

Bias is assessed by computing the mean deviation (MD):

$$Bias = \sum (\hat{y} - y) / n$$

Equation 10. Bias

A MD of zero indicates that no bias exists. A negative bias indicates that the method, on average, underestimates the true value. Alternatively, a positive bias indicates that the method, on average, overestimates the true value.

2.5. Analysis

The simulation study compares the accuracy and bias of each imputation method for the variable Change in CRP when the missingness mechanism is MCAR,

MAR, and MNAR. The missing values of baseline and follow-up CRP are imputed, and then Change in CRP is computed. The analysis of 1000 data sets allows for the computation of confidence intervals. Confidence intervals of 95 percent are computed for each of the summary statistics, for each imputation method. The comparison of summary statistics across methods determines which imputation methods perform better. All comparisons are made relative to each other, although the smaller the summary statistic (the closer to zero), the better the imputation method performs.

Chapter 3: Results

3.1. Comparison of RMSE

When the missingness mechanism is MCAR and ten percent of the values for Change in CRP are missing, mean imputation was the least accurate. LOCB was the most accurate imputation method, and multiple imputation and LOCF performed equally as well. When the missingness mechanism is MAR and ten percent of the values for Change in CRP are missing, mean imputation again was the least accurate. LOCF and LOCB were the most accurate. Multiple imputation also imputed relatively accurate estimates. When the missingness mechanism is MNAR and ten percent of the values for Change in CRP are missing, mean imputation and multiple imputation perform the worst. LOCF was the most accurate.

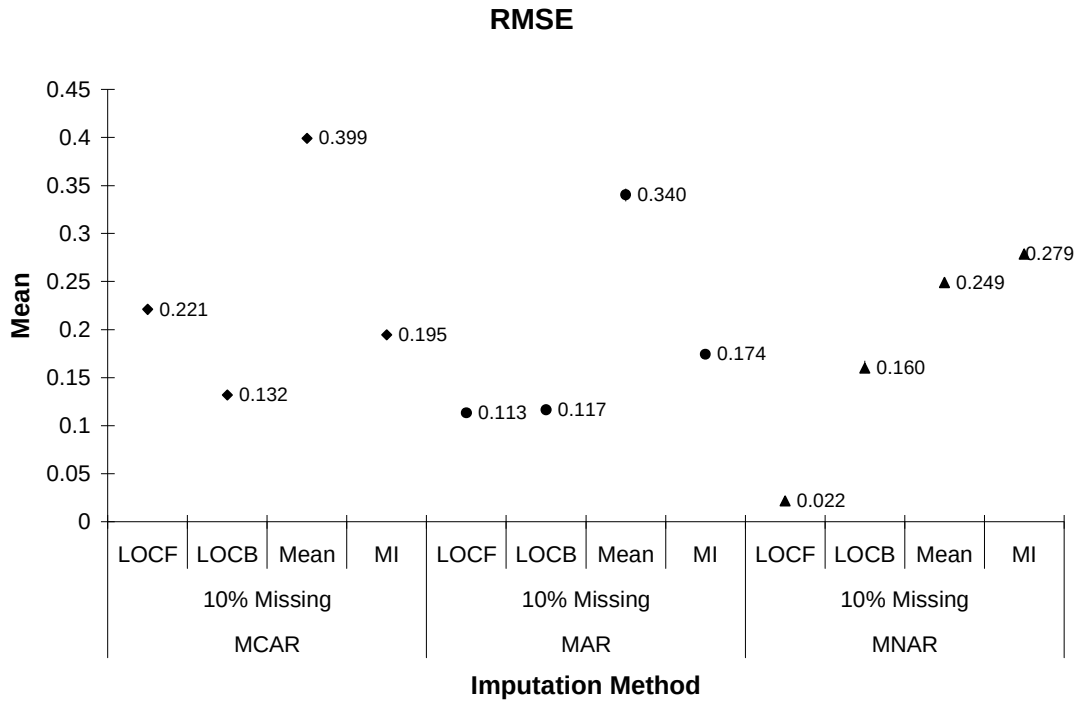

Figure 2. Comparison of RMSE for Change in CRP

3.2. Comparison of MAD

When the missingness mechanism is MCAR and ten percent of the values for Change in CRP are missing, the least accurate method is mean imputation. The most accurate method is LOCB, while multiple imputation and LOCF performed equally as well. When the missingness mechanism is MAR and ten percent of the values for Change in CRP are missing, the least accurate method is mean imputation. Multiple imputation, LOCF, and LOCB are all relatively accurate. When the missingness mechanism is MNAR and ten percent of the values for Change in CRP are missing, the least accurate methods are mean imputation and multiple imputation. LOCF was extremely accurate and LOCB was slightly less accurate.

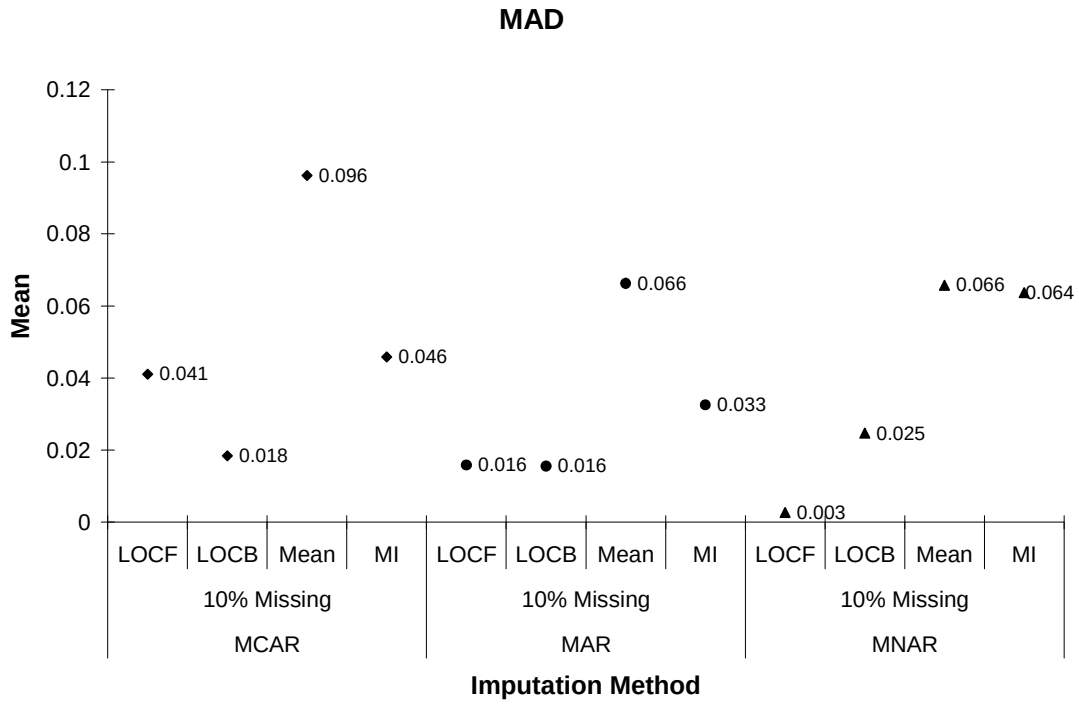

Figure 3. Comparison of MAD for Change in CRP

3.3. Comparison of Bias

When the missingness mechanism is MCAR and ten percent of the values for Change in CRP are missing, multiple imputation is the least biased imputation method, it slightly underestimates the true value. Mean imputation is the most biased, also underestimating the true vale. Both LOCF and LOCB overestimate the true value of Change in CRP, but are not as biased as mean imputation. When the missingness mechanism is MAR and ten percent of the values for Change in CRP are missing, multiple imputation is the least biased imputation method, however the other three methods are also relatively unbiased. When the missingness mechanism is MNAR and ten percent of the values for Change in CRP are missing, the least biased method is LOCF. LOCB, mean imputation, and multiple imputation are all extremely biased.

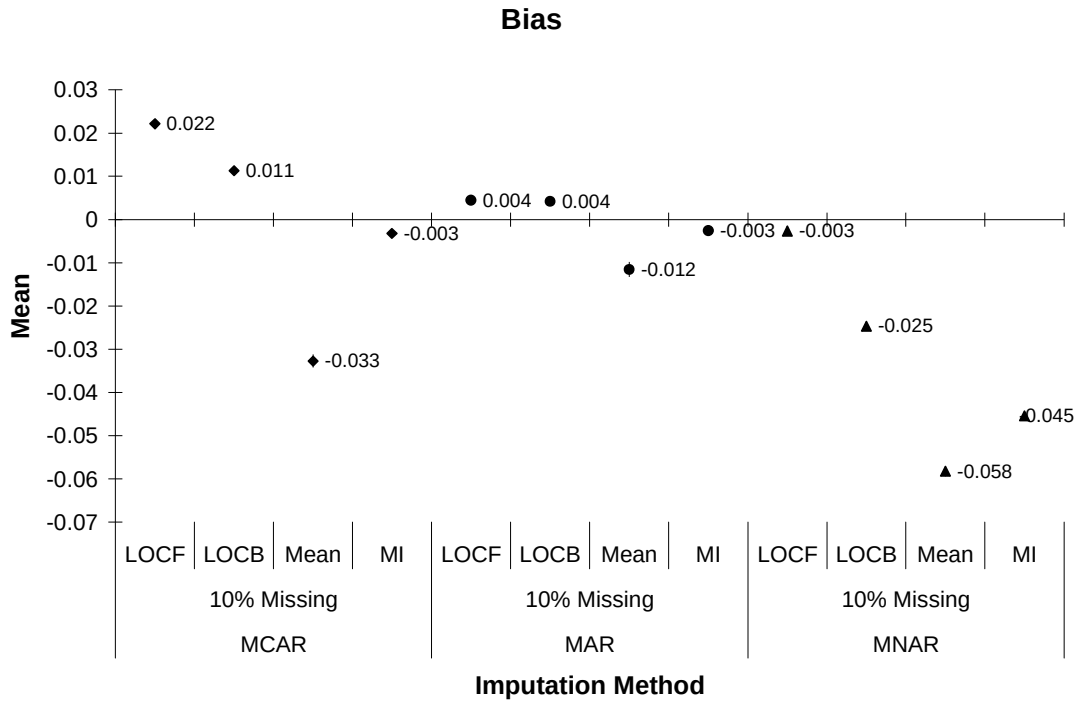

Figure 4. Comparison of Bias for Change in CRP

3.4. DEER

3.4.1. Missingness Pattern

The missingness pattern for the DEER data set is arbitrary (Table 2). Each pattern describes the missing values and the number of subjects that fall into that particular pattern. For example, subjects in Pattern 1 (n = 130) are the subjects without any missing data. Subjects in Pattern 10 are missing values for follow-up CRP, follow-up body fat, and hormone replacement therapy (n = 3). Since the missingness pattern is arbitrary (non-monotone), we expect that the best method for multiple imputation is the Markov chain Monte Carlo (MCMC), a Monte Carlo integration method using Markov chains (Zhang, 2003).

Table 3. Missingness Pattern

Pattern	Baseline CRP	Follow-up CRP2	Baseline Body Fat	Follow-up Body Fat	Age	HRT	Cig/Day	Alc/Day	Number of Subjects
1	X	X	X	X	X	X	X	X	130
2	X	X	X	X	X	X	X	.	7
3	X	X	X	X	X	.	X	X	8
4	X	X	X	.	X	X	X	X	8
5	X	X	X	.	X	X	X	.	1
6	X	X	X	.	X	.	X	X	2
7	X	X	.	X	X	X	X	X	2
8	X	X	.	X	.	X	X	X	1
9	X	.	X	X	X	X	X	X	5
10	X	.	X	.	X	.	X	X	3
11	X	.	X	.	X	.	X	.	2
12	.	X	X	X	X	X	X	X	2
13	.	X	X	X	X	X	X	.	2
14	.	X	.	X	X	X	X	X	1
15	.	.	X	X	X	X	X	X	1

3.4.2. Missingness Mechanism

In order to eliminate the missing data as MCAR, t-tests were conducted to compare the means of baseline variables for the group missing baseline CRP values with the group not missing baseline CRP values and to compare the group missing follow-up CRP values with the not missing follow-up CRP values. If the missingness

mechanism is MCAR, then there should not be any significant differences between groups. However, as can be seen in Table 3, there are significant differences between the group missing baseline CRP and not missing baseline CRP on three variables: BMI (P = .029), body fat (P = .009), and weight (P = .003). There is a significant difference in mean age (P = .046) between the group missing follow-up CRP and not missing follow-up CRP.

Table 4. Comparison of baseline variable means

Variable	Baseline Covariate	Group with Missing Value Mean (Std. Dev.)	Group without Missing Value Mean (Std. Dev.)	t-Value	P-Value
Baseline CRP	Age	55.15 (6.94)	56.46 (5.04)	0.609	0.543
	BMI	28.96 (1.02) *	26.11 (3.15)	-2.207	0.029
	Body Fat	38.12 (3.31) §	32 (5.19)	-2.619	0.009
	Weight	74.38 (2.73) §	69.21 (10.58)	-3.75	0.003
	Cholesterol	252.83 (28.96)	240.43 (26.08)	-1.141	0.256
	HDL	46.5 (4.04)	45.29 (7.19)	-0.409	0.683
	LDL	174.67 (25.26)	164.05 (21.27)	-1.193	0.234
Follow-up CRP	Age	53.45 (4.44) *	56.61 (5.09)	2.007	0.046
	BMI	26.42 (2.99)	26.2 (3.16)	-0.227	0.82
	Body Fat	33.7 (5.68)	32.08 (5.21)	-0.998	0.32
	Weight	70.36 (9.32)	69.32 (10.54)	-0.321	0.75
	Cholesterol	229.36 (26.45)	241.64 (26.07)	1.51	0.133
	HDL	45.91 (8.81)	45.29 (6.99)	-0.277	0.782
	LDL	161.36 (21.95)	164.63 (21.44)	0.489	0.626

* Significant mean difference at $P<0.05$.
§ Significant mean difference at $P<0.01$

Because there were significant differences between groups with missing data and groups without missing data, the missingness mechanism cannot be MCAR for the variables baseline and follow-up CRP. There is no way to mathematically determine if the missing data are MAR or MNAR. However, the missingness mechanism for this data set is most likely MAR because of what we know about why some of the data are missing. In the original trial, three women were lost to follow-up (Stefanick, et al., 1998), so there are no follow-up data available for them. We also know that when the plasma samples were later transported for analysis of CRP levels,

some samples were broken in transit (Camhi, 2008). Both these reasons suggest that the missing data are MAR; they are not missing because of the value of CRP itself.

3.4.3. Imputations

The simulation study identified three imputation methods whose imputed values of Change in CRP were the least biased and most accurate: LOCF, LOCB, and multiple imputation. For the DEER data set, LOCF and LOCB will be used together to increase sample size. The combination of these methods is known as subject mean imputation (as opposed to population mean imputation). Multiple imputation will also be used to impute missing values in the DEER data set.

The model for multiple imputation included all of the variables in the final analysis, as well as the other baseline covariates listed in Table 3. The other baseline covariates were included because the intent of the original study was to reduce cholesterol in people at high risk for cardiovascular disease. The imputation model incorporates the missingness mechanisms because of the inclusion of baseline covariates that are significantly different between missing and non-missing groups. Since the imputation model is more restrictive (has more assumptions) than the analysis model, the MI model leads to valid, more efficient estimates than the estimates from the observed data alone.

3.4.5. ANCOVA Results

There were slight differences between the three models. Figure 5 displays the within-group parameter estimates and standard deviations and the significant between-group differences. Parameter estimates for the between-group differences can be found in Appendix A. In the complete case analysis (n = 130), there were

significant between-group differences between the control group and the diet plus exercise group ($P = 0.04$) and between the exercise group and the diet plus exercise group ($P = 0.005$). The diet plus exercise group had a significant within-group change in CRP ($P = 0.009$).

The model after subject mean imputation (LOCF/LOCB) ($n = 160$) had the same results, although the between-group difference between the control group and the diet plus exercise group was significant at $P = 0.002$, the between-group difference between the exercise group and the diet plus exercise group was significant at $P = 0.003$, and the within-group significance for the diet plus exercise group was $P = 0.003$. For the model after multiple imputation ($n = 175$), the between-group significance for control versus diet plus exercise was $P = 0.007$ and for exercise versus diet plus exercise the significance was 0.003. There was no significant within-group change for the diet plus exercise group.

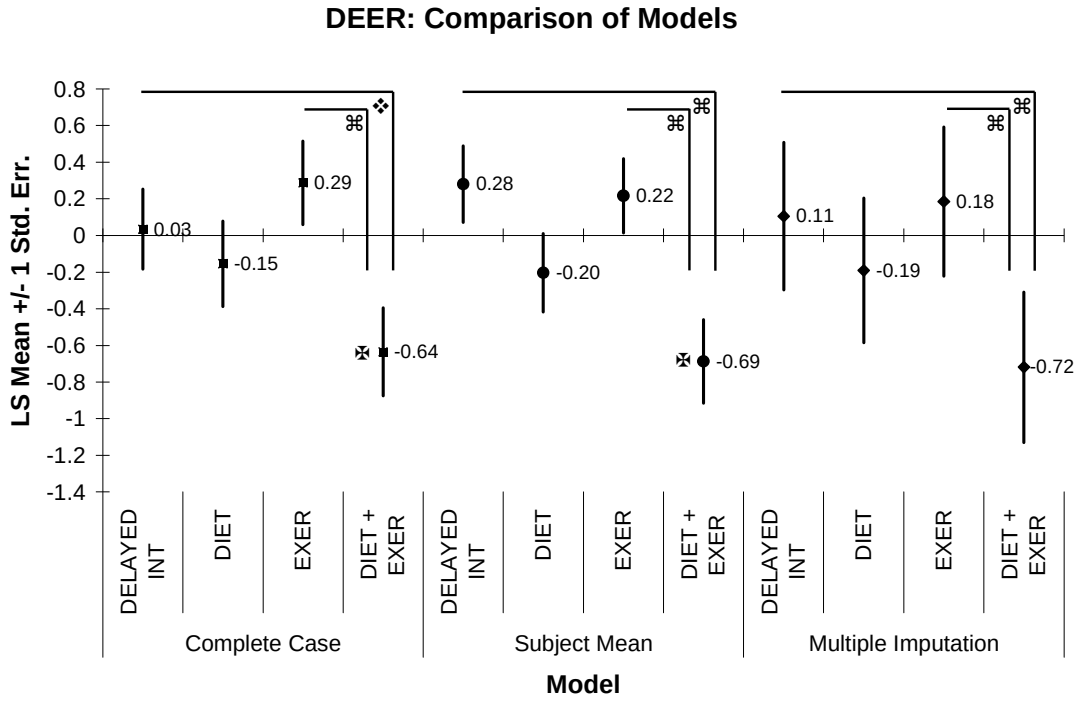

Figure 5. DEER: Comparison of Models
⌘ Between-group significance ($P < 0.01$)
❖ Between-group significance ($P < 0.05$)
✠ Within-group significance ($P < 0.01$)

Chapter 4: Discussion

Section 4.1. Simulation

In all cases, population mean imputation was the least accurate. This may be because it eliminates the relationship between baseline and follow-up CRP by assigning the column mean to the missing values. LOCF, LOCB, and multiple imputation retain the relationship between baseline and follow-up CRP. Other studies have found that subject-specific imputations are more accurate than population imputations (Shrive, Stuart, Quan, & Ghali, 2006).

When the missingness mechanism is MNAR and ten percent of the values for Change in CRP are missing, multiple imputation and population mean imputation performed the worst. Multiple imputation most likely performed the worst because the assumption of MAR was violated—the missing data could not be imputed based reliably on the values of other observed variables. Mean imputation performed the worst because it imputed the column mean for the missing values when, in reality, all of the values that were removed were much smaller than the column mean.

Section 4.2. DEER

The results of the three analyses of the DEER data set raised some questions. There was not one method that performed best in the simulation study for MAR missing data—both subject mean imputation and multiple imputation performed well. Therefore, both methods were used to impute the data in the DEER data set.

The larger standard errors for the estimates after multiple imputation reflect the missing data uncertainty as well as the ordinary sampling variation. There are

three explanations for the loss of the significant within-group difference for the diet plus exercise group. First, the significance could be lost due to the increased variance due to multiple imputation. Second, the MAR assumption may not hold, in which case MI is not an appropriate method to use. Because there are multiple variables in the DEER data set that have missing values, the missingness mechanism for each variable may not be the same. In multiple imputation, every variable included in the final model must also be used in the imputation model, so variables other than baseline and follow-up CRP that were missing values also had values imputed. The MAR assumption necessary for MI may not hold for these variables. Finally, there could also be systematic differences between the observed data and the missing data, so there is not actually a significant within-group difference. All three explanations are plausible and there is no way to know the truth.

Although mean imputation can produce biased estimates, the results of the simulation study and the final analysis after the mean imputation suggest that it is a plausible method in this situation. This may be because of the low percentage (26 percent) of missing data. Other studies have shown that mean imputation performs well when there is a missingness rate of 30 percent (Shrive, Stuart, Quan, & Ghali, 2006).

One limitation to the study is in regards to multiple imputation. Currently, PROC MI in SAS is only able to include continuous variables in the MCMC model (Horton & Kleinman, 2007). This means that the categorical variables used in the analysis could not be included in the imputation model. The strength of the study is the comparison of imputation methods that compared subject-specific, population,

and multiple imputation methods. Future work could include the use of indicator variables for the categorical variables that currently were not used in the multiple imputation using PROC MI in SAS.

Chapter 5: Conclusion

The results of the analyses of the DEER data set may have raised more questions than were answered. Although it is impossible to know the truth, the results of the simulation study suggested that mean imputation and multiple imputation were the two best methods to use to impute the missing data for the DEER data set. While the two between-group differences remained significant, the within-group difference for diet plus exercise was not significant after multiple imputation.

The results of simulation study highlight the importance of exploring multiple methods of imputation to impute values for missing data. Although most of the literature suggests that multiple imputation is the best method for imputing missing values when missing data is MAR, the application of multiple imputation in a real data set with an arbitrary missingness pattern may not be appropriate. All of the variables with missing data may not have the same missingness mechanism, in which case multiple imputation is not be appropriate to use.

There are also problems with subject mean imputation. It can produce biased estimates and, when the mean is imputed, reduce the true variability of the data because estimates are regressed to the mean. However, in this situation it may be the most appropriate method because the percent of missing data is low.

Although it is impossible to state with absolute certainty, I believe that the results of the analysis after subject mean imputation are more accurate than the results of the analysis after multiple imputation. However, the most important conclusion is that handling missing data can be very complicated and it is important to compare multiple methods to find the best fit for the data set that is being analyzed.

Appendix A

Table A.1. Between-group parameter estimates

	control vs. diet β (std.dev.)	**control vs. ex** β (std.dev.)	**control vs. diet+ex** β (std.dev.)	**diet vs. diet+ex** β (std.dev.)	**ex vs. diet+ex** β (std.dev.)
Complete Case	0.19 (0.31)	-0.25 (0.31)	0.67* (0.32)	0.48 (0.31)	-0.29* (0.31)
Subject Mean	0.48 (0.29)	0.06 (0.28)	0.97* (0.31)	0.48 (0.29)	-0.9* (0.3)
Multiple Imputation	0.3 (0.29)	-0.08 (0.29)	0.82* (0.3)	0.53 (0.3)	-0.9* (0.3)

*$P < 0.05$

Appendix B

Simulation 1: *n* = 175, *m* = 10, 50

RMSE:

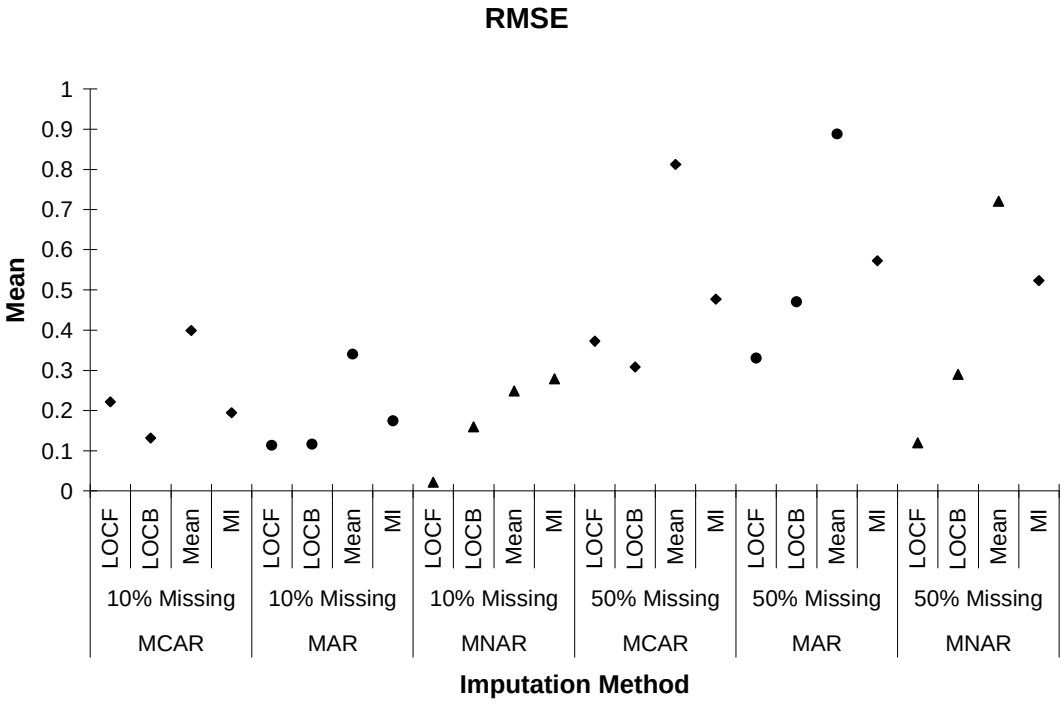

Figure B.1. Comparison of RMSE for Change in CRP (*n* = 175, *m* = 10, 50)

MAD:

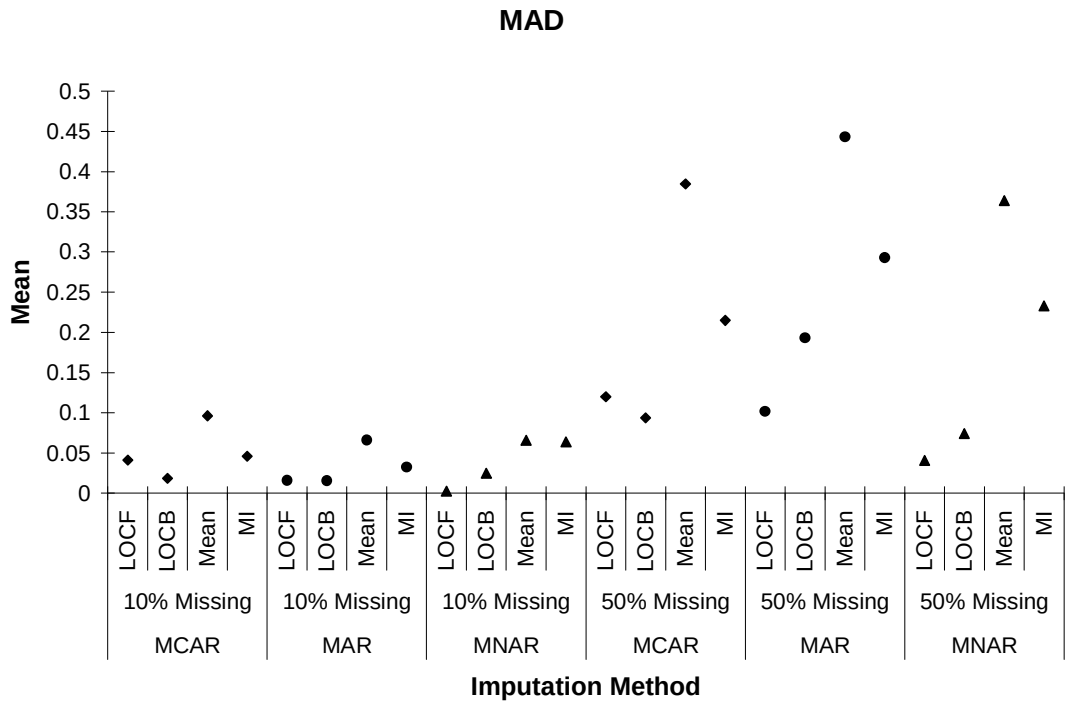

Figure B.2. Comparison of MAD for Change in CRP (n = 175, m = 10, 50)

Bias:

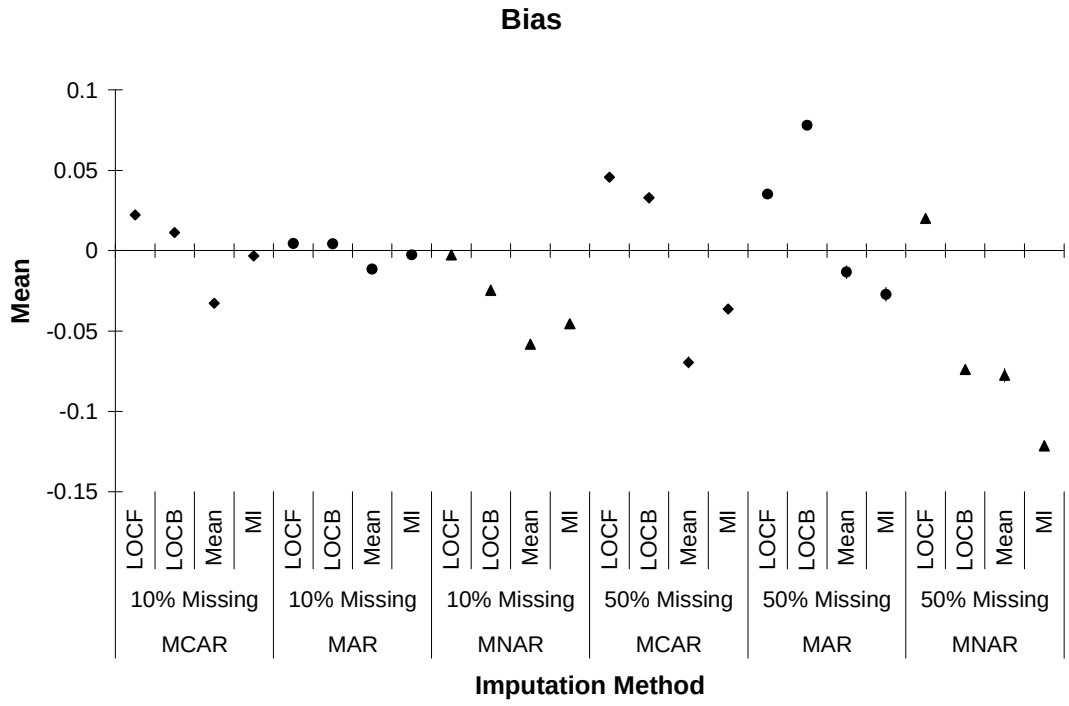

Figure B.3. Comparison of Bias for Change in CRP (n = 175, m = 10, 50)

Simulation 2: n = 500, m = 10, 50

RMSE:

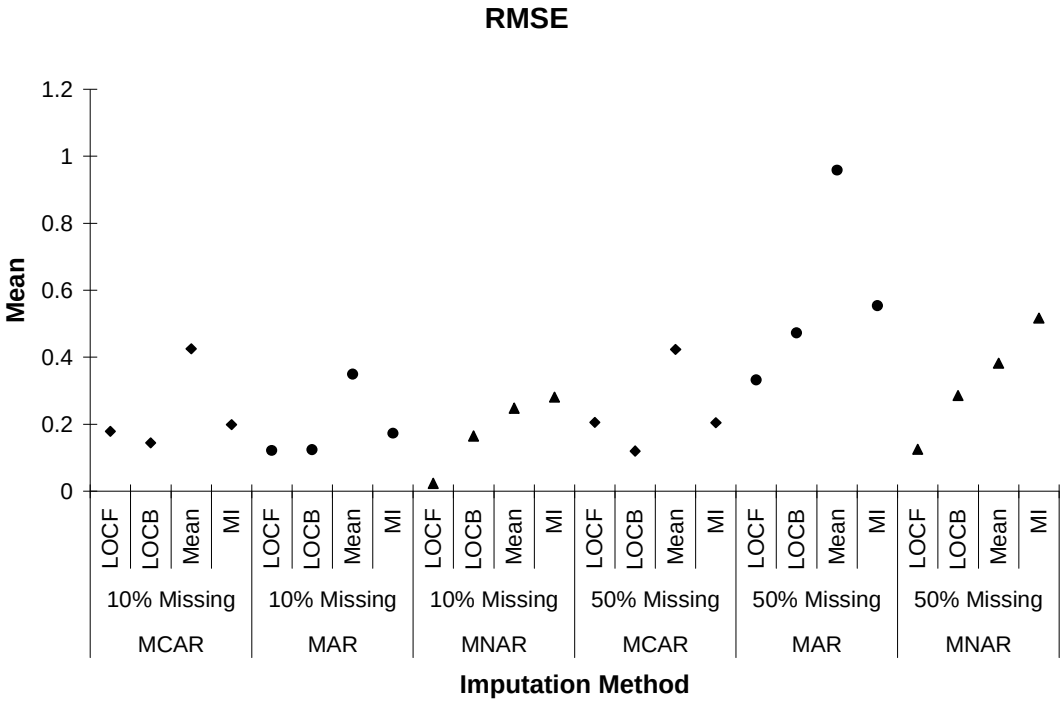

Figure B.4. Comparison of RMSE for Change in CRP (n = 500, m = 10, 50)

MAD:

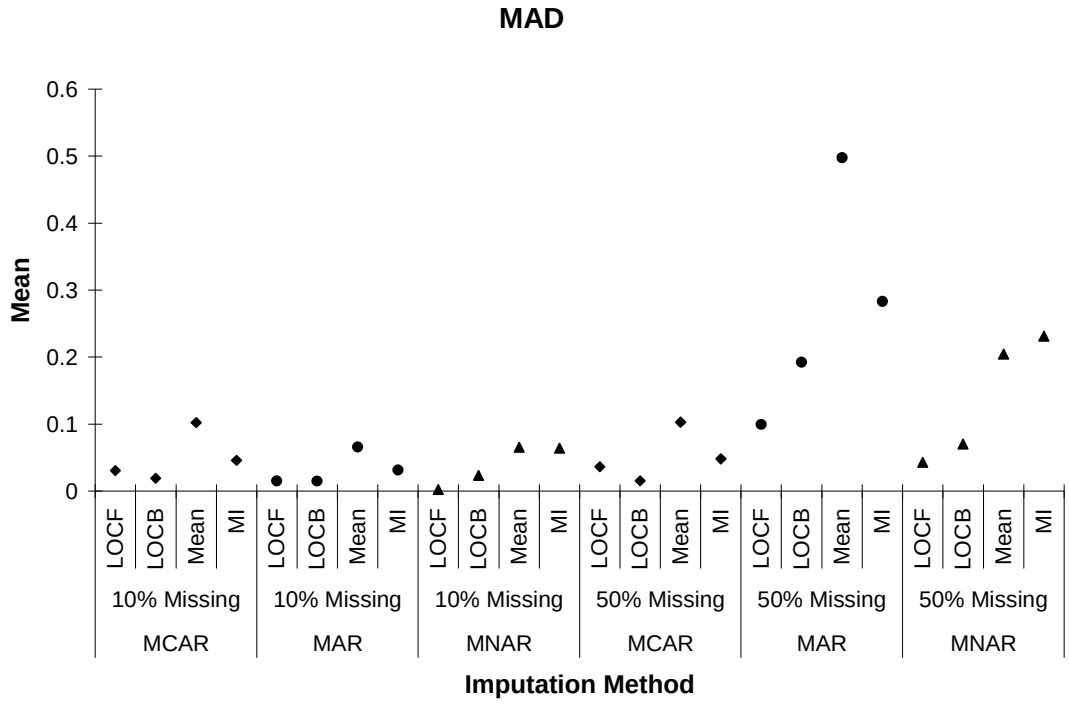

Figure B.5. Comparison of MAD for Change in CRP (n = 500, m = 10, 50)

Bias:

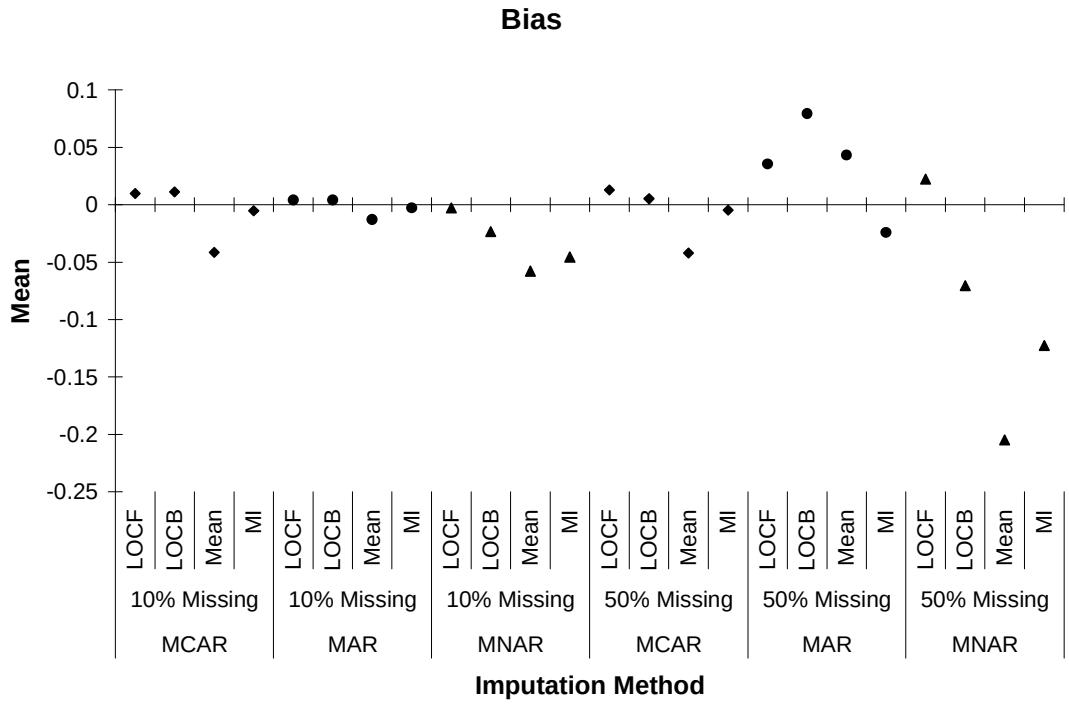

Figure B.6. Comparison of Bias for Change in CRP (n = 500, m = 10, 50)

Appendix C

Simulation 1: $n = 175$, $m = 10, 50$

Table C.1. Between-group parameter estimates (n = 175, m = 10, 50)

n=175			control vs. diet β (std. dev.)	control vs. ex β (std. dev.)	control vs. diet+ex β (std. dev.)	diet vs. diet+ex β (std. dev.)	ex vs. diet+ex β (std. dev.)
Full			0.39 (0.32)	0.75 (0.43)	-0.08 (0.15)	0.35 (0.12)	-0.83 (0.33)
10%	**MCAR**	**LOCF**	0.41 (0.13)	0.68* (0.19)	-0.07§ (0.12)	0.27§ (0.09)	-0.74§ (0.17)
		LOCB	0.41 (0.38)	0.75 (0.51)	-0.07 (0.18)	0.34§ (0.15)	-0.82 (0.39)
		Mean	0.35§ (0.07)	0.73§ (0.09)	-0.05 (0.12)	0.38§ (0.09)	-0.78§ (0.13)
		MI	0.11§ (0.13)	0.32§ (0.18)	-0.16§ (0.1)	0.21§ (0.07)	-0.47§ (0.16)
	MAR	**LOCF**	0.38 (0.14)	0.73 (0.18)	-0.08 (0.12)	0.35 (0.08)	-0.81 (0.16)
		LOCB	0.37 (0.38)	0.71 (0.51)	-0.09* (0.18)	0.34 (0.14)	-0.82 (0.39)
		Mean	0.41 (0.08)	0.73§ (0.09)	-0.06 (0.13)	0.33 (0.07)	-0.9 (0.12)
		MI	0.09§ (0.12)	0.31§ (0.17)	-0.17§ (0.1)	0.23§ (0.07)	-0.49§ (0.16)
	MNAR	**LOCF**	0.36§ (0.13)	0.74§ (0.16)	-0.15 (0.12)	0.38§ (0.07)	-0.89§ (0.15)
		LOCB	0.36§ (0.32)	0.74§ (0.43)	-0.21 (0.17)	0.38§ (0.13)	-0.95§ (0.33)
		Mean	0.37§ (0.06)	0.73§ (0.07)	-0.13 (0.11)	0.36* (0.06)	-0.86§ (0.11)
		MI	0.08§ (0.13)	0.35§ (0.17)	-0.27§ (0.1)	0.27§ (0.07)	-0.62§ (0.16)
50%	**MCAR**	**LOCF**	0.39 (0.21)	0.68 (0.27)	-0.04 (0.14)	0.29 (0.13)	-0.72 (0.22)
		LOCB	0.36§ (0.36)	0.72§ (0.47)	-0.06 (0.18)	0.36 (0.14)	-0.78§ (0.34)
		Mean	0.25§ (0.1)	0.52§ (0.11)	-0.16§ (0.16)	0.27§ (0.11)	-0.67§ (0.17)
		MI	0.07§ (0.09)	0.24§ (0.12)	-0.08 (0.08)	0.17§ (0.06)	-0.32§ (0.11)
	MAR	**LOCF**	0.3§ (0.25)	0.62§ (0.31)	-0.06§ (0.16)	0.32§ (0.14)	-0.68§ (0.25)
		LOCB	0.03§ (0.4)	0.22§ (0.52)	-0.16§ (0.18)	0.19§ (0.15)	-0.38§ (0.39)
		Mean	0.39 (0.14)	0.73§ (0.14)	0.04 (0.19)	0.34§ (0.13)	-0.69§ (0.18)
		MI	0.07§ (0.11)	0.24§ (0.14)	-0.06§ (0.09)	0.16§ (0.07)	-0.3§ (0.13)
	MNAR	**LOCF**	0.29§ (0.16)	0.6§ (0.21)	-0.3§ (0.13)	0.31§ (0.11)	-0.9§ (0.19)
		LOCB	0.22§ (0.33)	0.82§ (0.44)	-0.46§ (0.16)	0.6§ (0.14)	-1.28§ (0.33)
		Mean	0.41 (0.08)	0.82§ (0.1)	-0.11§ (0.12)	0.4§ (0.07)	-0.93§ (0.11)
		MI	-0.03§ (0.12)	0.22§ (0.16)	-0.4§ (0.1)	0.25§ (0.08)	-0.62§ (0.16)

* $P<0.05$
§ $P < 0.01$

Simulation 2: $n = 500$, $m = 10, 50$

Table C.2. Between-group parameter estimates (n = 175, m = 10, 50)

n=500			control vs. diet	control vs. ex	control vs. diet+ex	diet vs. diet+ex	ex vs. diet+ex
			β (std. dev.)	β (std. dev.)	β (std. dev.)	β (std. dev.)	β (std. dev.)
Full			0.39 (0.19)	0.74 (0.25)	-0.09 (0.09)	0.35 (0.07)	-0.83 (0.19)
10%	**MCAR**	**LOCF**	0.38 (0.09)	0.72§ (0.12)	-0.07* (0.07)	0.35 (0.05)	-0.8§ (0.1)
		LOCB	0.38 (0.16)	0.72§ (0.21)	-0.08§ (0.09)	0.34§ (0.06)	-0.8§ (0.17)
		Mean	0.36* (0.04)	0.69§ (0.05)	-0.04§ (0.07)	0.31§ (0.04)	-0.73§ (0.07)
		MI	0.14§ (0.08)	0.39§ (0.11)	-0.14§ (0.06)	0.24§ (0.04)	-0.53§ (0.1)
	MAR	**LOCF**	0.38 (0.09)	0.73* (0.12)	-0.08 (0.07)	0.35 (0.05)	-0.81§ (0.1)
		LOCB	0.35§ (0.16)	0.69 (0.21)	-0.09§ (0.09)	0.34§ (0.06)	-0.78§ (0.17)
		Mean	0.39 (0.05)	0.73§ (0.05)	-0.06 (0.07)	0.34§ (0.04)	-0.79§ (0.07)
		MI	0.14§ (0.08)	0.4§ (0.11)	-0.15§ (0.06)	0.25§ (0.04)	-0.54§ (0.1)
	MNAR	**LOCF**	0.36§ (0.08)	0.76§ (0.11)	-0.16 (0.07)	0.38§ (0.04)	-0.9§ (0.09)
		LOCB	0.34§ (0.14)	0.71§ (0.19)	-0.22§ (0.08)	0.38§ (0.06)	-0.93§ (0.15)
		Mean	0.36§ (0.03)	0.73§ (0.04)	-0.13 (0.06)	0.36§ (0.03)	-0.86§ (0.06)
		MI	0.12§ (0.08)	0.41§ (0.11)	-0.26§ (0.06)	0.29§ (0.04)	-0.66§ (0.1)
50%	**MCAR**	**LOCF**	0.33§ (0.08)	0.65§ (0.1)	-0.09 (0.07)	0.31§ (0.05)	-0.74§ (0.09)
		LOCB	0.39§ (0.11)	0.74§ (0.15)	-0.08 (0.08)	0.35§ (0.05)	-0.82§ (0.12)
		Mean	0.35§ (0.04)	0.69§ (0.05)	-0.09 (0.07)	0.34§ (0.04)	-0.77§ (0.07)
		MI	0.17§ (0.07)	0.43§ (0.09)	-0.13§ (0.06)	0.25§ (0.04)	-0.56§ (0.09)
	MAR	**LOCF**	0.32§ (0.12)	0.63§ (0.15)	-0.06§ (0.09)	0.31§ (0.07)	-0.68§ (0.12)
		LOCB	0.2§ (0.16)	0.44§ (0.21)	-0.09§ (0.09)	0.24§ (0.07)	-0.53§ (0.17)
		Mean	0.41§ (0.08)	0.69§ (0.08)	0.02§ (0.1)	0.28§ (0.07)	-0.67§ (0.1)
		MI	0.1§ (0.06)	0.27§ (0.08)	-0.05§ (0.05)	0.17§ (0.03)	-0.32§ (0.07)
	MNAR	**LOCF**	0.2§ (0.07)	0.61§ (0.11)	-0.29§ (0.07)	0.31§ (0.05)	-0.9§ (0.1)
		LOCB	0.23§ (0.11)	0.82§ (0.15)	-0.46§ (0.07)	0.59§ (0.05)	-1.27§ (0.13)
		Mean	0.42§ (0.04)	0.82§ (0.05)	-0.11§ (0.07)	0.4§ (0.04)	-0.93§ (0.06)
		MI	-0.01§(0.06)	0.24§ (0.08)	-0.4§ (0.05)	0.26§ (0.04)	-0.64§ (0.08)

* $P<0.05$
§ $P < 0.01$

Bibliography

Allison, PD. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods & Research,* 28(3): 301-309.

Allison, PD. (2002). Missing Data. Thousand Oaks, CA: Sage Publications.

Barnard, J & Meng, X. (1999). Application of Multiple Imputation in Medical Studies: From AIDS to NHANES. *Statistical Methods in Medical Research,* 8(1): 17-36.

Camhi, SM. (2008). The effects of low-fat diet and exercise on C-reactive protein and metabolic syndrome: Findings from a randomized controlled trial. Ph.D. dissertation, University of Maryland, College Park, United States -- Maryland. Retrieved December 12, 2008, from Dissertations & Theses: Full Text database. (Publication No. AAT 3321975).

Heitjan, DF. (1997). What Can Be Done about Missing Data? Approaches to Imputation. *American Journal of Public Health,* 87: 548-550.

Horton, NJ & Kleinman, KP. (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Journal of the American Statistical Association,* 61(1): 79-90.

Little, RJ. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association,* 83: 1198-1202.

Little, RJ & Rubin, DB. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research,* 18: 292-326.

McKnight, PE; Figueredo, AJ & Sidani, S. (2007). Missing Data: A Gentle Introduction. New York, NY: Guilford Press.

Padilla, MA & Algina, J. (2006). Type I Error Rates for a One Factor Within-Subjects Design with Missing Values. *Journal of Modern Applied Statistical Methods,* 3(2): 406-416.

Rubin, DB. (1976). Inference and Missing Data. *Biometrika,* 63: 581-592.

Rubin, DB. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association,* 91: 473-515.

Rubin, DB & Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association,* 81: 366-374.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York, NY: Chapman and Hall.

Schafer, JL. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, 8: 3-15.

Schafer, JL & Graham, JW. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2): 147-177.

Schafer, JL & Olsen, MK. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, 33(4): 545-571.

Shrive, FM; Stuart, H; Quan, H; & Ghali, WA. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Medical Research Methodology*, 6: 57-67.

Stefanick, ML; Mackey, S; Sheehan, M; Ellsworth, N; Haskell, WL & Wood, PD. (1998). Effects of Diet and Exercise in Men and Postmenopausal Women with Low Levels of HDL Cholesterol and High Levels of LDL Cholesterol. *The New England Journal of Medicine*, 339(1): 12-20.

Wittes, JT. (Oct. 2008). Proceedings from Temple-Merck Conference 2008: *Missing Inaction: Why Do So Many People Ignore Missing Data in RCTs?* Philadelphia, PA.

Zhang, P. (2003). Multiple Imputation: Theory and Method. *International Statistical Review*, 71(3): 581-592.