Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Name: Mark Bounthavong Date: 20 September 2013

MULTIPLE IMPUTATION WITH MULTIVARIATE MODELS: AN EVALAUTION OF TWO CASE STUDIES

By

Mark Bounthavong
Degree to be awarded: MPH
Career MPH

Kevin M. Sullivan, PhD, MPH, MHA Chair, Thesis Committee	Date
Jonathan H. Watanabe, PharmD, MS, PhD Field Advisor, Committee Member	Date
Melissa Alperin, MPH, MCHES Char, Career MPH Program	Date

MULTIPLE IMPUTATION WITH MULTIVARIATE MODELS: AN EVALAUTION OF TWO CASE STUDIES

By

Mark Bounthavong
Pharm.D., Western University of Health Sciences, 2004
B.S., Bachelor of Science, University of California, Irvine, 1999

Thesis Committee Chair: Dr. Kevin S. Sullivan, PhD, MPH, MHA

An abstract of
A Thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements of the degree of
Master of Public Health in the Career MPH program
2013

Abstract

MULTIPLE IMPUTATION WITH MULTIVARIATE MODELS: AN EVALAUTION OF TWO CASE STUDIES

By
Mark Bounthavong

The purpose of this thesis compared different methods for handling missing data with two observational studies as case studies in order to determine if there were any potential influence on the study results and conclusions.

Both case studies used multivariate models to answer a specific hypothesis. The first retrospective cohort study (Case study 1) constructed logistic regression models to investigate the association between adherence and achievement of lipid panel changes (achieving a ≥25% reduction). The second retrospective cohort study (Case study 2) constructed a multiple linear regression model that investigated the association between drug (exenatide or liraglutide) and change in hemoglobin A1c (HbA1c) level. Multiple imputation (MI) method was compared to complete-case analysis (CCA) to determine the direction and magnitude of the parameter estimates for each case study.

In Case study 1, the regression results for the crude, CCA, and MI methods were similar and did not vary significantly for LDL, HDL, and TC reduction of a ≥25% or greater from baseline. In Case study 2, results for the crude, CCA, and MI methods were similar and did not vary significantly for HbA1c reduction from baseline.

Based on the results of this study, multiple imputation may not be beneficial since the conclusions remained unchanged. Researchers who are involved with multivariate models may consider using multiple imputation to address missing data. Multiple imputation could be presented alongside the results of the complete-case analysis; but this may seem redundant if there are no differences in study conclusions.

MULTIPLE IMPUTATION WITH MULTIVARIATE MODELS: AN EVALAUTION OF TWO CASE STUDIES

By

Mark Bounthavong
Pharm.D., Western University of Health Sciences, 2004
B.S., Bachelor of Science, University of California, Irvine, 1999

Thesis Committee Chair: Dr. Kevin S. Sullivan, PhD, MPH, MHA

A Thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements of the degree of
Master of Public Health in the Career MPH program
2013

Acknowledgements

I would like to thank my Thesis Committee Chair, Dr. Kevin Sullivan, and my field advisor, Dr. Jonathan H. Watanabe, for their assistance with this thesis project. I would also like to thank Dr. Josephine N. Tran for her assistance in extracting and validating the data and the late Dr. Michael Juzba for providing me with access and the initial query for the data, as well as sage advice about using large database analysis at the Veterans Health Administration.

Achieving a Master of Public Health was the first step in my goal to becoming a better researcher and person. I hope that the lessons that I've learned during the development of this thesis instilled in me the discipline and humanity necessary to carry out relevant and important research in the field of pharmacoepidemiology.

TABLE OF CONTENTS

TABLES

FIGURES

CHAPTER 1: INTRODUCTION

1.1. INTRODUCTION AND RATIONALE

In observational studies, epidemiologists are generally concerned about a study's internal and external validity.[1–5] External validity deals with extrapolation of a study's conclusion beyond the sample population to the general population; whereas, internal validity draws upon conclusions from the study population to the source population.[2,3] Internal validity deals with three types of biases: information, confounding, and selection bias. Information bias refers to validity of classification and measurement.[2,3] Classical confounding refers to factors that are associated with both the exposure and its dependent variable relationship, but are not in the causal pathway.[2,3] Selection bias refers to systematic errors in how the study population is distorted from the target population.[2,3,5] Examples of selection bias includes "healthy worker effect,"[5] membership bias,[3] incidence-prevalence bias,[3] and missing data bias.[6,7] Bias arises when patients with missing data are different from the study population; and precision is affected due to a decrease in sample size. This thesis examines the impact of missing data on the conclusions of observational studies.

Depending on the amount, missing data in observational studies have been reported to cause bias and misinterpretation of the study findings. This becomes an inherent problem in regression models which is based on the assumption that data is valid and complete. Most studies in the clinical literature do not address the impact of missing data on conclusions generated from regression models.[8,9] This neglect occurs despite the availability of statistical tools to properly address this problem. Several methods are available to handle missing data's impact on the conclusions of regression models, each associated with advantages and disadvantages. The type

of missingness and the assumptions made by the investigator determine selection of missing data analysis method.

Several guidelines are available for proper handling of missing data;[10–15] but knowledge and understanding of differing circumstances and missing data patterns play a significant role in determining the selection of missing data analysis. According to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, the Methods section should include a strategy on how missing data will be handled.[14,16] In addition, the Results section should have descriptive information on the number of subjects with missing data for each critical variable and, if possible, provide reasons for the missing data. Similarly, the Consolidated Standards of Reporting Trials (CONSORT) statement requires that missing data be addressed in order to maintain an intention-to-treat analysis. The CONSORT statement recommend using imputation methods based on data from other variables over simple imputation methods such as last observation carried forward for missing data that exceed a reasonable amount (e.g., 5% to 10%).[15]

A challenge with missing data analysis is the absence of a "true" or complete dataset often experienced in practice. This lack of a "true" dataset limits the investigator's abilities to validate the results after missing data analysis has been applied. Consequently, this thesis can only make comparisons between missing data methods rather than provide validation.

The purpose of this thesis is to compare different methods for handing missing data with two observational studies[17,18] in order to determine if there were any potential influence on the study

results and conclusions. Although there is a lack of a "true" dataset to compare the results to, side-by-side comparison of different missing data approaches with the published outcomes has the potential to provide different conclusions from what was originally reported.

1.2. PROBLEM STATEMENT

Missing data in observational studies have an elevated potential for bias that may invalidate the results of a study.[6,19–21] Considered a nuisance, mainly because it is not the primary statistical focus, missing data may introduce serious problems of validity and bias that challenge the outcome of a study if left unaddressed.[6,19,20] Oftentimes, this discussion is excluded from the statistical plan or underreported.[8,9]

Specific methods have been developed to address missing data bias; however, limitations based on the data availability and experience with these methods limit their general utilization and integration in published manuscripts. Most studies that evaluate missing data usually perform a descriptive analysis comparing patients with complete data and patients with missing data. However, advanced but complex techniques are available that yield more information while handling missing data. Two studies[8,9] provide a discussion of this issue in their review of the literature. Wood, et al.[8] reviewed randomized trials between July and December 2001 from *BMJ*, *JAMA*, *Lancet* and *New England Journal of Medicine* and reported that 89% of studies (total number of studies, 71) reviewed had partial or missing data and 92% used complete-case analysis to handle missing data. Sterne, et al.[9] evaluated 59 studies that had applied multiple imputation (MI) from 2002 to 2007 in *BMJ*, *JAMA*, *Lancet* and *New England Journal of Medicine*. Sterne, et al.[9] reported that seven studies presented the impact of multiple imputation

and complete-case analysis, 36 studies presented information about the amount of missing data, 22 defined the number of imputations performed, and five studies compared the distribution of key variables with and without missing data. In both these reviews, there were deficiencies on the explanation and influence of missing data on the conclusions of individual studies.

In epidemiologic or clinical research involving the collection or extraction of data, missing observations are an inherent problem. Even under the best conditions, missing observations are often unavoidable occurrences that require additional investment in time and resources by the investigator(s). Causes of missing observations are numerous and include: study design (longitudinal versus post-test), participant characteristics (refusal to respond and cognitive level), measurement characteristics (length of the questionnaire, instrument failure, and instrument validity), data collection conditions (obstacles such as time constraints and weather/season), data management (transfer, input, and security), and random chance.[6,22]

Strategies that have been developed to handle missing data are grounded on assumptions about the pattern of missing data. There are several types of missing data that should be carefully addressed prior to developing any strategy for handling missing data. The proportion of missing data on critical variables should be evaluated. Variables with 1% missing data should not bias the study; however, variables with 20% missing data would need to be addressed for potential bias. According to the Rubin and Little,[6] 5% missing data on a variable of interest may need to be addressed using some form of missing data analysis method. Once the proportion of missing data has been determined to bias the results, the pattern of missingness should be examined. Determining the type of missing data is based on the knowledge of the pattern of missingness.

Methods for handling missing data are dependent on a balance between complexity of the procedure and its efficiency. Not every study will require a complex method when a single imputation method is efficient. Taken together, the proportion of missing data in a variable of interest, type of data, and preference of the investigator determine the missing data analysis method used for a given project.

1.2.1. Missing data mechanisms:

Missing data mechanism describes the relationship between missingness and values of variables in the data matrix. Assume that a $(n \times K)$ rectangular dataset without missing values where *n* denotes the rows or cases and *K* denotes the variables or columns, with *i*-th row $y_i = (y_{i1}, y_{i2}, y_{i3}, \dots, y_{iK})$. Let $Y = (y_{ij})$, where y_{ij} represents the value of the variable Y_j for subject *i*. In the presence of missing data, the missing data indicator matrix is defined as $M = (m_{ij})$, such that $m_{ij} = 1$ if $y_{ij}$is missing and $m_{ij} = 0$ if $y_{ij}$is not missing. Therefore, *M* represents the pattern of missing data in the dataset (or the missing data indicator matrix).

Let $Y = (y_{ij})$ denote the complete data matrix and $M = (m_{ij})$ denote the missing data indicator matrix. The basic missing data mechanism is based on the conditional distribution of the missing data indicator (*M*) given the data (*Y*) or

$$f(M|Y, \phi),$$

where ϕ denotes unknown parameters or the relationship of *M* to *Y*.

There are three main types of missing data mechanisms in observational studies. A description of each type is provided below according to Rubin's missing data classification system.[7]

1.2.1.1. Missing Completely At Random (MCAR):

Missing completely at random (MCAR) is an extreme example of missing observations that occur through randomness.[6,19] In order words, MCAR represents 'true' randomness of missing data. Missing data under the MCAR assumption underlies that no other variables in the model or the dataset has an influence on the missing observations. This includes the outcome data as well as other potential confounders in the model. Therefore,

$$f(M|Y,\phi) = f(M|\phi) \text{ for all } Y\,(Y_{obs} \text{ and } Y_{mis}) \text{ and } \phi,$$

such that the missingness does not depend on the values of the observed (Y_{obs}) and unobserved values (Y_{mis}) of Y. Of note, MCAR does not result in bias estimates and is considered "ignorable."[6,19]

Identification of MCAR is performed using a statistical test developed by Little.[23] based on a null distribution that is asymptomatically chi-squared. Therefore, if $P<0.05$, then the data is not MCAR.

1.2.1.2. Missing At Random (MAR):

Missing at random (MAR) is based on the assumption that missingness is conditioned on the observed values (Y_{obs}) in the dataset, but not on the unobserved values (Y_{mis}).[6,19] Unlike MCAR where the pattern of missingness is not dependent on the observed values (Y_{obs}) and the unobserved values (Y_{mis}), MAR requires the assumption that the data is dependent only on the observed values (Y_{obs}) and not the unobserved values (Y_{mis}) of Y. Therefore,

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \text{ for all } Y_{mis} \text{ and } \phi,$$

such that missingness only depends on the observed values (Y_{obs}) and not on the unobserved values (Y_{mis}) of Y.

This is the most commonly encountered scenario in practice. The MAR assumption justifies the analysis for several missing data analyses methods (e.g., multiple imputation, Bayesian estimation, and maximum likelihood estimation). MAR can be referred to as "ignorable" missing data because the values that are used to make conclusions for the missing observations 'ignore' the unobserved values (Y_{mis}) of Y.

1.2.1.3. Not Missing At Random (NMAR):

Not missing at random (NMAR) or missing not at random (MNAR) is based on the assumption that it is not possible to make estimations about the missing observations from the known values in the dataset. [6,19] This is also known as "non-ignorable" data. Therefore,

$$f(M|Y,\phi) = f(M|Y_{mis},\phi) \text{ for all } Y_{obs} \text{ and } \phi,$$

such that the missingness depends on the unobserved values (Y_{mis}) and maybe depends on the observed values (Y_{obs}) of Y. In other words, NMAR requires that the assumption of the missingness be conditioned on the unobserved values (Y_{mis}) of Y; however, it does not necessarily need to be conditioned on the observed values (Y_{obs}) of Y. The actual relationship between M and Y_{mis} is unknown because the unobserved values are not available.[6]

The selection of a method for handling missing data is dependent on the type of missing data encountered (MCAR, MAR, and NMAR). Little[23] provides a test to determine if the missing data mechanism is MCAR or not; however, there are no statistical tests to differentiate between MAR

and NMAR. It is also important to note that it is impossible to know the relationship between M and the unobserved values (Y_{mis}) of Y. In these situations, sensitivity analysis would need to be performed.[6,19,22]

1.2.2. Consequences of missing data mechanisms:

If data is assumed to be MCAR, then simple complete-case analysis could be performed. In complete-case analysis (case-wise deletion or list-wise deletion), data with missing values for any variable are dropped from the final analysis. Analysis is limited to only those subjects or data that are not missing. In other words, any missing value for any variable is excluded and the sample size is reduced. This approach is only appropriate if the data is MCAR; however, it is highly unlikely that missing observations follow an MCAR pattern. Therefore, epidemiologists are more likely to assume MAR.

1.3. THEORETICAL FRAMEWORK

This thesis will present two theoretical frameworks for handing missing data: (1) Simple (or ad hoc) methods and (2) Complex methods. Simple methods incorporate single imputation methods that result in an easy and quick approach to missing data. However, they are limited by artificial decrease in variance and; therefore, an artificial increase in precision. Complex methods include multiple imputation (MI) and maximum likelihood (ML) estimation. They provide a valid and robust approach by reducing bias and introducing uncertainty estimators.

1.3.1. Simple methods (ad hoc)

1.3.1.1. Complete-case analysis:

Complete-case analysis has the advantage of being easy to perform and is the default setting for most statistical software (e.g., SPSS and SAS). However, it is can be associated with biased estimates that invariably occur as each subject is excluded from analysis due to missing values for any of the parameters.[6,19]

Subjects with complete data may be different from subjects with incomplete data. Hence, a separate analysis to investigate potential differences between the two is necessary to adjust for missingness. This can be accomplished by creating a missing data indicator (*M*) for complete and incomplete cases or dummy variables. Regression models can control for the presence of missing data and compare estimates to determine the extent of their influence. Although not statistically sound, this method provides differential risk of bias if the parameter estimates deviate from the base-case analysis.

1.3.1.2. Available-case analysis:

Unlike complete-case analysis, available-case analysis excludes data only if the variable of interest has missing values. For example, if a cohort with 100 patients was evaluated and there were 12 missing values for age, 5 missing values for gender, and 7 missing values for ethnicity, the final analysis will use the balance of patients with the complete data for each variable. Hence, 82 patients will be used to analyze the mean age, 95 patients will be used to analyze the gender, and 93 patients will be used to count ethnicity. Ultimately, when presenting the demographic table, the number of patients analyzable for each variable may be different each time an analysis is performed.

In complete-case analysis, subjects with any missing data are eliminated. Conversely, with available-case analysis, subjects are considered for analysis if they have data for any of the variables of interest. As a result, various sample sizes for each variable analyzed are possible. In our above example, age, gender, and ethnicity corresponded with a sample size of 82, 95, and 93, respectively. This may be useful in order to maximize sample size for some of the variables in a univariate analysis resulting in higher efficiency relative to complete-case analysis method. However, in regression models, subjects with missing data will ultimately be excluded through the complete-case analysis method resulting in a reduction in precision and increased potential for bias.

1.3.1.3. Single imputation:

Imputation with a single value in place of a missing value is a basic approach to handling missing data.[6,19] The critical element with this method involves the choice of a plausible value to account for the missing observation. Multiple methods for single imputation exists that include using the mean or median, parameter estimates from regression models, stochastic regression method, hot and cold deck procedures, and missing data indicator variable (*M*).

Single imputation method involving the mean and median require replacement of all missing values with either the mean or median of the variable or parameter of interest. The mean and median are estimated from the observed data (Y_{obs}), which is then inserted for all missing values. This has the potential of underestimating the variance and artificially improving the precision of the estimate. In addition, this will result in an increase in type I error (rejecting the null when the null hypothesis is true).[6,19]

Investigators may use regression analysis to determine a value for single imputation. Regression imputation provides values that may be considered reasonable because of multiple adjustments; however, it underestimates the variance. Similar to the mean and median imputation processes, regression imputation underestimates the variance, and artificially increases precision resulting in an increase in type I error. Moreover, regression imputation also assumes that the between-imputation variance is zero due to the existence of only one value. Stochastic modeling addresses this concern by incorporating residual error with the predicted estimate.[6,19]

Stochastic regression imputation adds variance to the predicted estimate by introducing uncertainty into the equation.[6] However, this method is dependent on assumptions regarding the normality of the distribution. In addition, stochastic regression imputation fails to account for additional variance with other parameters; and thus, limits their utilization for single imputation.

Similar to regression imputation, hot deck procedure replaces missing value with a single imputation that is reflective of a "similar" subject.[6] For example, if a 64-year old white male was a diabetic, then another 64-year old white male with a missing value for diabetes status may also be a diabetic by virtue of similarity. This method assumes that patients are similar to some degree. Variation is underestimated and is limited to the ranges of the observed data. Ranges outside the observed data are not incorporated resulting in an underestimation of variance, an artificial increase in precision, and an increase in type I error. This method has utility when there are a large number of complete cases that is reflective of the source population.

Cold deck procedure requires the use of external data.[19] Data from an external source are used to extract a single value to be used in the single imputation process for the dataset of interest. For example, complete data from another cohort investigation may be used to identify a single value to replace missing observations. The underlying assumption with cold deck procedure is that the external data is exactly the same or similar to the dataset of interest. Realistically, this may not be the case and differences between the datasets limit the use of the external data in the cold deck procedure. This procedure is generally not recommended.[19]

In summary, single imputation methods are limited by an artificial reduction in variance and increase in precision; whereby, type I error is increased. However, complex and valid methods have been developed to address these limitations.

1.3.2. Complex methods

Although single imputation methods offer a simple and easy solution to address missing data, the increased potential for bias, artificial reduction in variance, and artificial increase in precision limit is utilization in missing data analysis. Moreover, practical reasoning implies that missing data requires more variation and complexity in addressing the effect of missing values on the base-case results and, ultimately, conclusions.

Several complex methods have been investigated in the literature and include multiple imputation,[6,12,22,24,25,25–28] maximum likelihood (ML) estimation,[6,29–31] and inverse probability weighting (IPW).[32,33] In this theoretical framework, the focus will be on multiple imputation. ML estimation is similar to logistic regression methods and will not be discussed in this thesis.

Rather, interest has been generated with multiple imputation methods because it requires less computational power to perform and achieves convergence quickly relative to the ML estimation.[6,24,29–31,34] Unlike multiple imputation, ML estimation does not replace or input missing values.[24,34] In addition, it is model specific and is sensitive to different statistical analysis; whereas, multiple imputation is not.[24,34] The results from ML estimation is associated with similar and valid estimates comparable to multiple imputation and is an appropriate alternative missing data method.[24,34]

IPW is another method that reduces bias due to missing data. An advantage of IPW is that it does not need to assume something about the distribution of X values; whereas, MI must make an assumption about the distribution.[32,33] However, in terms of utilizing partially missing data, MI is able to do this to generate values for missing data; whereas, IPW is limited to complete cases.[32,33] Consequently, MI is generally more efficient that IPW.

1.3.2.1. Multiple imputations:

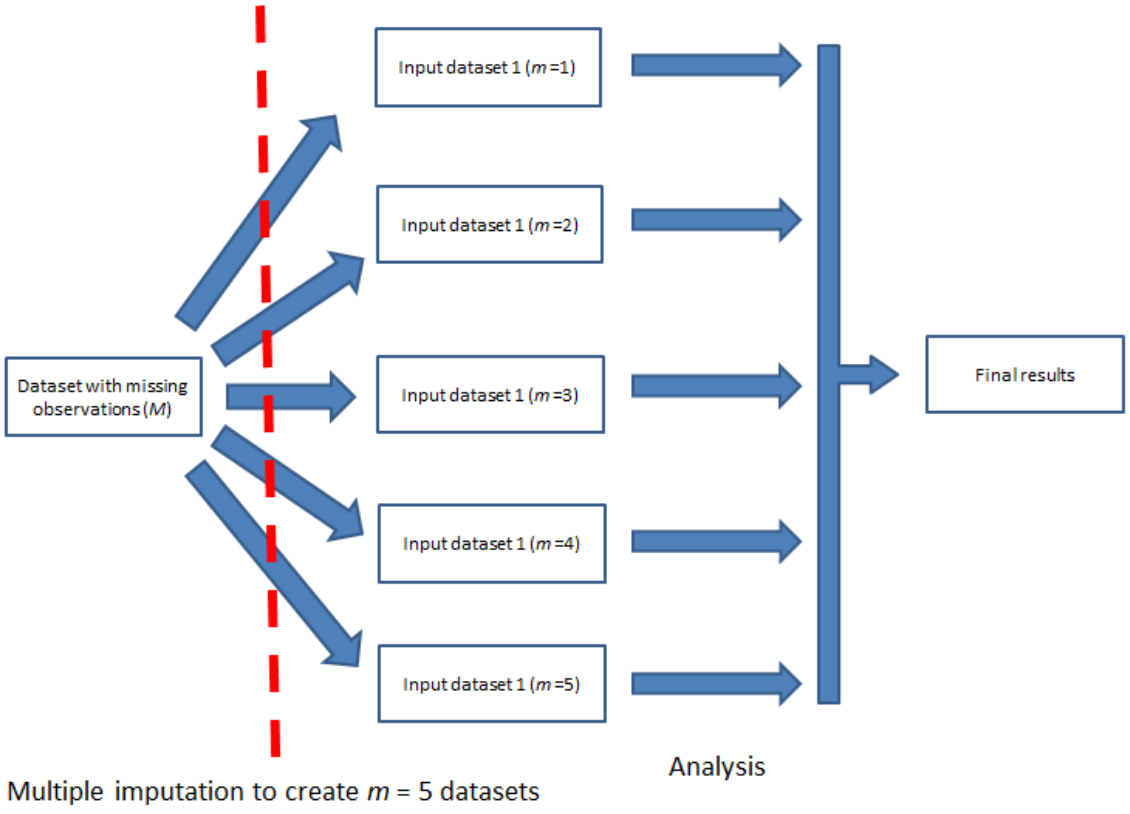
Multiple imputation uses the available observable values (Y_{obs}) in order to generate correlations between observed values (Y_{obs}) and missing values (Y_{mis}) to predict the range of the most probable value for the missing data.[6,22,24] Multiple imputation will generate a complete set of data (m) for the missing values. More than one dataset ($m > 1$) can be generated and then combined to give the investigator a final dataset (Figure 1). Generally, 5 to 10 datasets are created from multiple imputation and then combined for final analysis using Rubin's method.[35] Multiple imputation allows the investigator to draw inference from the missing values and then generate point estimates with confidence intervals for the population rather than the individual.

The efficiency of using $m > 1$ is determined by

$$\left(1+\frac{\gamma}{m}\right)^{-1},$$

where γ is the proportion of missing values for the parameter to be estimated.[20,24,35] Efficiency is inversely related to γ; where the increase in γ results in a decrease in efficiency (Table 1).

Figure 1. Multiple imputation of m = 5 datasets.



At approximately 5 multiple imputations, the efficiency ranges from 85% to 98% with γ ranging from 0.9 to 0.1, respectively. In situations where the proportion of missing data is high, high efficiency can still be achieved with 3 to 5 imputed datasets. Therefore, it is unnecessary to increase the number of imputation datasets beyond $m > 5$.

Table 1. Efficiency of multiple imputations relative to proportion of missing values.

Number of imputations (*m*)	Proportion of missing data (γ)				
	0.1	0.3	0.5	0.7	0.9
3	0.968	0.909	0.857	0.811	0.769
5	0.980	0.943	0.909	0.877	0.847
10	0.990	0.971	0.952	0.935	0.917
20	0.995	0.985	0.976	0.966	0.957

Multiple imputation uses Bayesian methods to generate posterior probability for the parameter estimate using a specified prior distribution with the likelihood function.[6,24,35] The target variable, which contains the unobserved value (Y_{mis}), is dependent on the available observed value (Y_{obs}) on Y. Predictive distribution for Y_{mis} is generated using Markov chain Monte Carlo (MCMC) simulations [$P(Y_{mis} \mid Y_{obs})$], an iterative process that ends when the posterior distribution of Y_{mis} stabilizes and converges.[6,24] This iterative process generates predictive Y_{mis} for each subject resulting in different estimates for the missing values. Each dataset has n number of Y_{mis} that are imputed using MCMC process which results in a single dataset. As m number of dataset is created, they can be combined for use in statistical analysis. The Bayesian process is determined as follows:

$$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis},$$

where $P(Q|Y_{obs})$ is the actual posterior distribution of Q and $P(Y_{mis}|Y_{obs})$ is the posterior predictive distribution of Y_{mis} given Y_{obs}.[20,24] The likelihood function is denoted by $P(Q|Y_{obs}, Y_{mis})$.

Datasets are combined by averaging the parameter estimates (Q_i) over m number of datasets:

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m} Q_i,$$

where Q_i is the point estimate generated from each of the i-th imputed dataset.[20,24,35]

The within-imputation variance or variability (U_m) is determined by the following:

$$\overline{U}_m = \frac{1}{m}\sum_{i=1}^{m} U_i,$$

where $\overline{U}_m$ is the average within-imputation variance for *m* imputations and U_i is the variance for each *i*-th imputed dataset.

The between-imputation variance or variability (B_m) is determined by the following:

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}(Q_i - \bar{Q}_m)^2,$$

where $(Q_i - \bar{Q}_m)^2$ represents the difference between the predicted point estimate from each of the *i*-th imputed dataset (Q_i) and the average predicted point estimate over *m* number of datasets ($\bar{Q}_m$).

The total variance or variability (T_m) is determined by combining the within-imputation variance (U_m) and the between-imputation variance (B_m):

$$T_m = \overline{U}_m + \left(1 + \frac{1}{m}\right) B_m,$$

where $\overline{U}_m$ is the average within-imputation variance for *m* imputations. Overall standard error (SE) is the square root of T_m.

Confidence bands or intervals at the 95% level (95% CI) are determined using the following:

$$\bar{Q}_m \pm t_v(\tfrac{\alpha}{2})T_m^{1/2},$$

where $t_v(\frac{\alpha}{2})$ represents that upper and lower confidence bounds as determined by $100(\frac{\alpha}{2})$. For a 95% confidence interval with lower and upper bounds of 2.5% and 97.5%, respectively, $t_v(\frac{\alpha}{2})$= 1.96.

Several assumptions are required in order for multiple imputation process to be valid. First, the pattern of missing values must be "ignorable" which is achieved when the missing data mechanism is MCAR or MAR.[6,35] The variables in the multivariate model must have a normal distribution; however, multiple imputation is robust to violation of this assumption.[6] The dataset must also follow an item non-response pattern where all subjects contain some observed values (Y_{obs}) of Y.[21] If the dataset has unit nonresponse patterns where the subjects or groups of subjects have none of the observed values (Y_{obs}) of Y, then the multiple imputation procedure will not be suitable.[21]

1.4. PURPOSE STATEMENT

The aim of this investigation was to determine the impact of missing data on published findings by comparing simple (ad hoc) procedures to a complex method, multiple imputation. This will be illustrated in case studies where the methods will be applied to two observational studies by Watanabe, et al.[17] and Bounthavong, et al.[18]

1.5. RESEARCH QUESTION

We focus this discussion on multivariate models as the principal example for performing these methods. The methods discussed will be demonstrated on two cases that were recently published.[17,36] Specific research questions include:

1. Do missing data have an impact on the base-case results of the previously published work by Watanabe, et al?[17]

   H_0: There is no difference in odds ratio between the results of the missing data analyses and published work by Watanabe, et al.[17]

   H_a: There is a difference in odds ratio between the results of the missing data analyses and published work by Watanabe, et al.[17]

2. Do missing data have an impact on the base-case results of the previously published work by Bounthavong, et al?[18]

   H_0: There is no difference in point estimates between the results of the missing data analyses and published work by Bounthavong, et al.[18]

   H_a: There is a difference in point estimates between the results of the missing data analyses and published work by Bounthavong, et al.[18]

1.6. SIGNIFICANCE STATEMENT

Missing data analysis is an integral part of statistical analysis for any epidemiologic study. However, peer-review reports continue to neglect this critical validation step in statistical analysis.[8,9] This research will highlight the importance of missing data analysis and its impact on published reports that did not evaluate the influence of missing observations. The methods demonstrated in this research will demonstrate how investigators can salvage data in order to increase statistical power and confirm their initial findings. The outcome of this study will also illustrate how different missing data methods can report different results while improving precision of the parameter estimates of the multivariate models. More importantly, this research

will also provide SAS codes for other researchers to use in their studies in order to produce research reports that address the problems of missing observations.

1.7. DEFINITION OF TERMS

Table 2. Definition of terms.

Term	Definition
Available-case analysis	Simple method for performing missing data analysis. Cases where data is available are used resulting in different sample sizes.
CI	Confidence Interval
Complete-case analysis	Simple method for performing missing data analysis. Only complete cases are included in any statistical analysis.
Complex methods	Higher order methods that require significant computations such as multiple imputation and maximum likelihood estimation.
Dataset or data matrix (Y)	The dataset that is used for the initial research analysis. This dataset could have complete observations or missing observations.
EMA	European Medicines Agency
IRB	Institutional Review Board
Maximum likelihood (ML) estimation	Complex method for performing missing data analysis. ML relies on the probability model and is model specific.
Missing At Random (MAR)	Missing data mechanism describing a situation where the missing data indicator is conditioned upon the observed values of the dataset.
Missing Completely At Random (MCAR)	Missing data mechanism describing a situation where the missing data indicator is not conditioned upon the observed values of the dataset.
Missing data indicator (M)	Dummy variables used in a dataset with missing values to indicate that there are missing observations.
Missing observations	Missing data, values, or cells denoted as Y_{mis}.
Multiple imputation (MI)	Iterative process using Markov chain Monte Carlo simulations to degenerate values based on the observed data.
Not Missing At Random (NMAR)	Missing data mechanism describing a situation where the missing data indicator is conditioned upon the unobserved values of the dataset; and possibly the observed values of the dataset.
Observed values	Data that is present in a dataset denoted as Y_{obs}.
OR	Odds Ratio
PCORI	Patient Centered Outcomes Research Institute.
Regression model	Statistical model that performs either ordinary least square or maximum likelihood estimation to generate parameters estimates that regress to a particular dependent variable.
RR	Risk Ratio
Simple methods	Ad hoc methods for missing data analysis that include: complete-case analysis, available-case analysis, single imputation, and cold/hot deck methods.
Single imputation	Simple method for performing missing data analysis. Mean, median, or parameter estimation from regression method is inputted for missing values in a dataset.
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology.

CHAPTER 2: REVIEW OF LITERATURE

2.1. SYSTEMATIC REVIEW OF THE LITERATURE

A literature search of the current peer-review publication was performed to identify studies that applied multiple imputation in a clinical or epidemiologic study. Articles were included in the review if there was a comparison of multiple imputation with other missing data analysis methods in a clinical or epidemiologic dataset. Articles were excluded from review if they focused on genetics, simulated missing data as opposed to having a dataset with missing data, used a longitudinal study design, evaluated survey responses or item-response theory, only used a Cox proportional hazards model, did not use multiple imputation, and did not evaluate human patients. A PubMed search using the following combination of key terms was performed for a date range from inception to July 13, 2013: “multiple,” “imputation,” and “regression,” and limited to “HUMAN” and “CLINICAL TRIALS.” The following is the Boolean search strategy used:

(multiple[All Fields] AND imputation[All Fields] AND ("regression (psychology)"[MeSH Terms] OR ("regression"[All Fields] AND "(psychology)"[All Fields]) OR "regression (psychology)"[All Fields] OR "regression"[All Fields])) AND "humans"[MeSH Terms])

The purpose of this review was to evaluate the literature for examples where multiple imputation was compared to an alternative missing data analysis method. In particular, the focus of the review was on epidemiologic studies where multivariate regression models frameworks were applied. Multiple imputation has been used in longitudinal studies where missing data has been a common problem. Longitudinal studies involve repeated measures of specific variables that are considered time-varying. However, this review was interested in examining how multiple

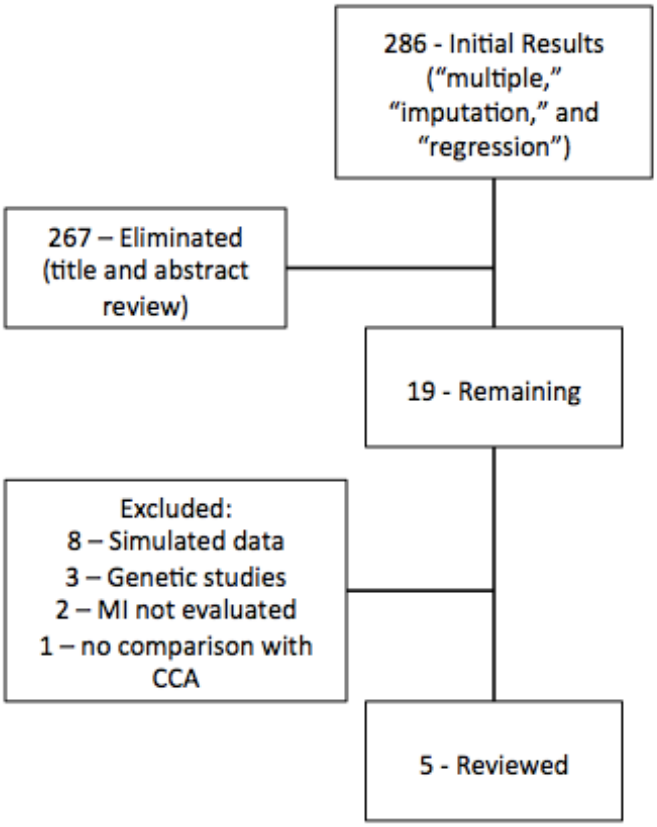
imputation is applied in multivariate regression models that did not involve time-varying predictors or repeated measures. Therefore, the examples that were identified included cohort studies that used a dependent variable that was either continuous or categorical with independent variables that were time-unvarying.

2.2. RESULTS OF THE SYSTEMATIC REVIEW

There were a total of 286 citations identified in the initial key word search. Abstract and title review eliminated 267 results. Of the 19 citations left for review, 14 studies were eliminated: eight were based on simulated missing data,[21,37–43] three focused on genetic studies,[44–46] two did not discuss multiple imputation,[47,48] and one lacked comparison with complete-case analysis.[49] The remaining five studies[24,41–43,49–53] focused on using multiple imputation in actual datasets with missing values that constructed multivariate regression models (Figure 2).

A majority of studies included for review were based on retrospective studies;[50,52–54] however, there was one based on a prospective study.[51] All of the studies used multiple imputation and compared it to complete-case analysis. Two studies included missing indicator methods,[51,53] one study examined regression imputation,[53] and one study examined conditional/unconditional mean imputation.[51]

Figure 2. Flow diagram of the literature search.



MI, multiple imputation
CCA, complete case analysis

2.2.1. Summary of results

Yang, et al. investigated the use of multiple imputation in a retrospective cohort study with a total population of 74 subjects.[52] In this case study, only 47 subjects had complete date available for the multiple linear regression model. The dependent variable was the number of visits a child made to a mental health service provider. Independent variables included age, child behavior score, functional assessment scale, depression inventory index, number of placements, years living with foster parents, number of case-worker home visits, benefits per month from the Department of Child and Family, foster parent education level, race (white or black) and gender. In the regression model using complete-case analysis, being white (T_{44}=2.66, P=0.01) and number of case-worker home visits (T_{44}=2.62, P=0.01) were significant predictors of the number

of visits to a mental health service provider identified using step-wise reverse selection procedures.

Two methods for multiple imputation and variable selection based on the Bayesian framework was applied in this case study.[52] "Impute then select" (ITS) is one method for variable selection that first produces the multiple inputted datasets and subsequently applies Bayesian variable selection to them. Alternatively, "simultaneous impute and select" (SIAS) generates multiple inputted datasets and Bayesian variable selection simultaneously using the Gibbs sampling process. MAR was assumed for these analyses. In both the ITS and SIAS method, being white and the number of case-worker home visits were highly predictive of the number visits for mental health services. In the ITS method, being white and the number of case-worker home visits had a probability of 50% and 89% of being selected as significant predictors in the final multiple linear regression model. Similar results were reported with the SIAS method (being white, 50%; and number of case-worker home visits, 82%). However, the models were sensitive to the selection of prior distribution used for the Bayesian process. In this case study, Yang, et al.[52] provided further information regarding the use of different algorithms for multiple imputation using a Bayesian framework. Although most multiple imputation procedures use the MCMC algorithm, ITS and SIAS offer alternative methods.

Newgard[54] applied multiple imputation to handle missing data in a retrospective cohort study that had matched records between out-of-hospital and ambulance values. Records from 1998 to 2003 were available for four variables (intubation attempt, Glasgow Coma Scale, systolic blood pressure, and respiratory rate) that were used in a logistic regression model to determine

association with mortality. A total of 6,150 matched records were identified with missing values in intubation attempt (9%, missing values), Glasgow Coma Scale (17%), systolic blood pressure (22%), and respiratory rate (17%). The study assumed MAR in order to perform the multiple imputation method and generated ten imputed datasets used for the regression analysis. Results of the complete-case and multiple imputation analyses were compared to the complete matched-record which was defined as the "true" completed data.

In most cases, multiple imputation provided fair to moderate estimates of the missing values for all four independent variables.[54] There were a few instances that deserved notice. In the logistic regression, both complete-case and multiple imputation analysis were biased towards the null when compared to the "true" completed data for the association between having a systolic blood pressure less than 90 mm Hg and mortality. Intubation was associated with mortality using the multiple imputation analysis and "true" complete data; however, complete-case analysis reported a reduction in mortality with intubation. Although the confidence intervals were not explicitly stated, the figures reported confidence bands that were narrower for the multiple imputation analysis compared to the complete-case analysis. Based on the results of Newgard's study,[54] multiple imputations provides an accurate and precise estimation of the point estimates based on a logistic regression model.

Greenland, et al. performed a series of missing data analysis that were compared to complete-case analysis for a retrospective cohort study investigating endometrial cancer among estrogen users.[53] There were 318 endometrial cancer cases and 599 controls; controls were matched to the cases by age and time in the healthcare plan. Proportions of missing data ranged from 0% to 47%

with approximately 59% of the cases missing more than one value in the univariate analyses. The authors assumed MAR and generated ten imputed datasets using multiple imputation methods for the conditional logistic regression model. Three missing data analysis methods were compared to complete-case analysis: ordinary missing indicator, regression imputation, and multiple imputation.

The conditional logistic regression model used two main exposure variables: recent estrogen use and age at menarche (11 versus 16 years) to evaluate their independent association with endometrial cancer.[53] The complete-case analysis was only able to use 140 (44%) of the 318 cases and 248 (41%) of the 599 controls in the regression analysis; whereas, the other methods were able to use all cases and controls. Odds ratios were similar in magnitude and direction for all missing data analyses. There were no major differences in the odds of developing endometrial cancer in subjects who had recent estrogen use relative to those who did not for the complete-case analysis (OR=1.60; 95% CI: 1.10, 2.34), ordinary missing indicator (OR=1.96; 95% CI: 1.44, 2.66), regression imputation (OR=1.82; 95% CI: 1.35, 2.46), and multiple imputation (OR=1.83; 95% CI: 1.36, 2.47). The upper limit of the confidence interval for the ordinary missing indicator was higher than all the other methods; conversely, the lower limit of the confidence interval for the complete-case analysis was the lowest compared to the other methods. The complete-case analysis also demonstrated a wide confidence band which reflects imprecision and uncertainty.

Similar findings were reported for the odds of developing endometrial cancer for subjects experiencing menarche at 11 year versus 16 years old for the complete-case analysis (OR=1.42;

95% CI: 0.91, 2.22), ordinary missing indicator (OR=1.54; 95% CI: 0.70, 3.38), regression imputation (OR=1.07; 95% CI: 0.78, 1.46), and multiple imputation (OR=1.09; 95% CI: 0.72, 1.64).[53] Again, the upper limit of the confidence limit was highest with the ordinary missing indicator method; but the lower confidence limit was also lowest with the same method. This indicates a wide uncertainty and lack of precision with the ordinary missing indicator method. The regression imputation method reported a narrower confidence band that was suggestive of artificial increase in precision. It was recommended by the authors that multiple imputation should be preferred; however, lacking the statistical software, investigators should use the complete-case analysis over the ordinary missing indicator and regression imputation methods where overestimation and large confidence bands reduce their utility.

Mulla, et al.[50] investigated the use of multiple imputation in an population-based study that evaluated the association between hospital mortality and age in patients who were hospitalized for invasive group A streptococcal (GAS) disease. Age was the exposure of interest and transformed into a categorical variable (55 years or older versus 0 to 54 years), but missing data was only present in the confounding variable, serum albumin. Logistic regression was performed to evaluate the association between age and hospital mortality controlling for potential confounders (race, gender, clindamycin use, beta-lactam use, necrotizing fasciitis, and serum albumin). MAR was assumed as the missing data mechanism, and twenty inputted datasets were generated.

Mulla, et al.[50] did not evaluate missing data for the dependent variable which reduced their sample from 257 to 201 (35, deaths; 166, survived). Therefore, a total of 110 (55%) of subjects

had complete data for serum albumin [91 (45%) of patients did not]. No missing values were reported for the other independent variables. In the complete-case analysis, there was no significant association with being 55 years or older and hospital mortality relative to being 0 to 54 years old (OR=2.43; 95% CI: 0.79, 7.53). An increase in serum albumin was significantly associated with lower hospital mortality (OR=0.23; 95% CI: 0.10, 0.55). No other significant associations were reported in the complete-case analysis. In the multiple imputation analysis, being 55 years or older was significantly associated with mortality relative to being 0 to 54 years old (OR=3.08; 95% CI: 1.22, 7.78). Similar to the complete-case analysis, an increase in serum albumin was associated with a reduction in mortality (OR=0.23; 95% CI: 0.10, 0.53).

In the complete-case analysis, there was no significant association between age and mortality which conflicts with the results of the multiple imputation method. Available data supports the latter's conclusion that older age is associated with mortality in GAS disease;[55] therefore, it was concluded by the authors that multiple imputation provided an accurate and precise account of the association between age and hospital mortality. In the complete-case analysis, a total of 110 patients were analyzable, a reduction of 45% from the total data. Type II error is a potential consequence of missing values and should be dealt with appropriately. More importantly, the multiple imputation method allowed the use of all available data; thus, maintaining power, increasing precision and reducing the potential for Type II error. Ultimately, the authors concluded that multiple imputation had a significant contribution in changing their initial conclusions regarding the use of complete-case analysis.

Van der Heijden, et al.[51] took data from a prospective cohort study[56] that investigated the risk of pulmonary embolism (PE) and used it as a case study for missing data methods. The method of analysis was a logistic regression model that investigated possible predictors for a diagnosis of PE. The main dependent variable was the diagnosis of PE, and the independent variables were baseline demographics (age and gender), medical condition at the time of admission (e.g., duration of symptoms, period confined to bed, respiratory rate, cardiac rate, arterial oxygen pressure, arterial carbon-oxygen pressure, Quetelet index, leg paresis, leg pain, family history of deep vein thrombosis (DVT)/PE, fever, dyspnea, pleura rub, wheezing, palpitations, collapse with or without unconsciousness, surgery in the past 3 months, malignancy, signs of DVT, previous history of DVT/PE, leg ultrasound, and chest X-ray). Missing data analysis included the following methods: indicator method, unconditional mean imputation, conditional mean imputation, and multiple imputation.

There were 398 subjects in the study with 246 (61.8%) with no missing data.[51] Forty-two (10.6%) subjects had one missing value, 72 (18.1%) subjects had two missing values, 24 (6.0%) subjects had three missing values, 6 (1.5%) subjects had four missing values, 6 (1.5%) subjects had five missing values, and 2 (0.5%) subjects had 6 missing values. Arterial carbon-oxygen pressure accounted for 86 (21.6%) of the missing values, followed by arterial oxygen pressure with 84 (21.1%) and echography of legs with 55 (13.8%). Dummy variables were used as a missing data indicator and statistical tests were performed to analyze the differences between groups with no missing values and at least one missing value for the potential predictors of PE. Diagnosis of PE (P=0.02), dyspnea (P<0.01), malignancy (P<0.01), surgery in previous 3 months (P=0.04), previous PE (P=0.02), and respiratory rate (P<0.01) were significantly

different between groups with no missing values and at least one missing value. The significant difference in prevalence of PE between the no missing value and at least one missing value groups indicate that the missing data mechanism was not MCAR. Therefore, complete-case analysis was expected to produce biased results.

Variable selections for the final logistic regression model using the backward elimination process were different based on the missing data analysis methods applied.[51] In the complete-case analysis, echography of legs, cardiac rate, chest X-ray, duration of symptoms, age, collapse with or without unconsciousness, wheezing, signs of DVT/PE, and palpitations were part of the final model. For the indicator method, unconditional mean imputation, and conditional mean imputation, the final model also included period confined to bed, fever, and prior DVT/PE; but excluded signs of DVT/PE and palpitations. Unlike the alternative missing data methods, multiple imputation method did not include prior DVT/PE as a predictor in the final logistic regression model. There were no differences in the directionality and outcome of the parameter coefficients when compared to the different missing data analysis methods. Complete-case analysis yielded the largest standard errors for all the parameter coefficients of predictor variables relative to the other methods. This was expected due to a reduction in sample size and decreased precision.

Receiver operating characteristic (ROC) curves were generated to compare the 5 different missing data analyses methods.[51] ROC curves discriminate between the different models by providing probabilities for predicting PE based on the variables included and the parameter estimates generated by the missing data analysis methods.[57] Based on different missing data

analysis methods, missing indicator method produced an overestimation of the ROC curve area (0.813) compared to complete-case analysis (0.794), conditional mean imputation (0.792), multiple imputation (0.787), and unconditional mean imputation (0.755).[51] It was speculated that the lack of any significant differences between the missing data analysis methods was due to the low number of missing values. Therefore, more profound results would have been reported if there were a larger proportion of missing values. Regardless, all methods used produced parameter estimates that were similar in terms of direction, magnitude and variance between each other.

2.3. SUMMARY OF CURRENT PROBLEM AND STUDY RELEVANCE

Multiple imputation methods can reduce bias and overestimation of uncertainty by the complete-case analysis when assuming MAR missing data mechanism. In addition, multiple imputation has influence over the selection of variables in the final regression models. As demonstrated by van der Heijden, et al.,[51] missing data analysis methods yielded different final regression models due to the differences in the variable selection outcomes. Therefore, it is critical that any study constructing multivariate regression models perform missing data analysis to strengthen the study's conclusions or identify the effect that missing data can have.

The studies by Watanabe, et al.[17] and Bounthavong, et al.[18] used complete-case analysis to determine the parameter estimates of their regression models. There were no discussions of the potential effect of missing data on the outcomes. This lack of reporting reflects the common practice in current epidemiologic studies. Fielding, et al.[58] reported that among 61 randomly selected quality of life studies from the New England Journal of Medicine, Journal of the

American Medical Association, BMJ, and Lancet from 2005 to 2006, 36 (59%) had some form of missing data but did not perform any imputation analysis.

Current statistical software have the ability to perform multiple imputation methods. Although assumptions about the missing data mechanism are still required, software such as SAS and SPSS provide investigators with user-friendly interfaces to perform complicated missing data analysis. Therefore, investigators should be able to assess for the effects of missing data on their study conclusions.

The key strength of multiple imputation methods is its robustness in generating parameter estimates that are close to the "true" value while addressing the variance. Unlike single imputation methods where parameter estimates have artificially inflated precisions, multiple imputations maintain the variance expected in parameter estimates while reducing bias. Several studies that engineered missing data in order to compare the results of multiple imputation to the "true" data have reported moderate to high correlations and similar confidence levels.[21,41–43,49] These studies provide confidence in the method and support its application in situations where missing data mechanisms are MAR and the unobserved values are unavailable for validation.

The current thesis builds upon the literature by providing further support for using multiple imputation in multivariate regression models where the missing data mechanism is MAR. Similar to the literature, this thesis will compare multiple imputation with complete-data analysis using two case studies where multivariate regression models were used in the primary analysis.

Furthermore, this thesis intends to justify the use of multiple imputation over complete-case analysis when there is missing data.

CHAPTER 3: METHODS

3.1. INTRODUCTION

This project is a secondary analysis of two retrospective cohort studies[17,18] that used multivariate regression models to evaluate the impact of missing data on each of the study outcomes. Each study did not use time-varying predictors and had a dependent variable that was assumed to follow a normal distribution. Moreover, the studies also had a large amount of missing data among the predictors and the outcome variables which may have a meaningful impact on the outcomes of the studies.

The first retrospective cohort study[17] (Case study 1) constructed logistic regression models to investigate the association between adherence and achievement of lipid panel changes (achieving a 25% reduction). The second retrospective cohort study[18] (Case study 2) constructed a multiple linear regression model that investigated the association between drug use (exenatide or liraglutide) and change in hemoglobin A1c (HbA1c) level. Both studies used data from the Veteran population. Case study 1 derived data from the Veteran population in the southwest region of the United States (US); whereas, Case study 2 derived data from the national Veteran population. Both case studies used complete-case analysis methods to generate the parameter estimates of the regression models but did not assess impact of missing data.

Multiple imputation method will be compared to complete-case analysis to determine the direction and magnitude of the parameter estimates for each case study. Five imputed data sets (m=5) will be used to combine the results into the regression methods. Based on these findings,

this project will provide guidance and recommendations on how to handle missing data when a large proportion of data is missing.

3.2. POPULATION AND SAMPLE

This project will use two case studies[17,18] to evaluate the impact of missing data on study results based on multivariate regression models. In the first case study[17] (Case study 1), the study population was based on the Veteran population at the regional level. The study population of the second case study[18] (Case study 2) was based on the Veteran population at the national level. Each study population and their selection process are discussed below.

3.2.1. Case 1 – Regional level

The study population was drawn from the Veterans Integrated Systems Network (VISN) 22 (Desert Pacific Healthcare Network) which includes VA facilities in the Southern California (Los Angeles, Long Beach, Loma Linda, and San Diego) and Nevada (Las Vegas) regions that service approximately 1.4 million veterans.[59] VISN represents the different VA networks across the United States (US) that is responsible for the veterans in their covered networks. There are a total of 21 VISNs across the United States (US) which encompasses 152 VA medical centers and 1,400 community-based outpatient clinics.[60] VISN 22 covers the southwest US and has a total of 5 medical centers with 29 community-based outpatient centers.[59]

Patients were included if they were a new statin user between the periods of November 30, 2006 and December 2, 2007 with a diagnosis of dyslipidemia (or related disorders) based on the *International Classification of Disease, Ninth Revision* (ICD-9), greater than 18 years of age, and

had been continuously enrolled in the VA health plan for at least 2 years. Table 3 provides a list of ICD-9 codes associated with dyslipidemia that were used in Case study 1. Patients were considered new statin users as defined by a 6-month washout period before filling their first statin prescription. Patients were followed for a 1-year observation period after the index date on lipid panel levels [low-density lipoprotein (LDL), total cholesterol (TC), and non-high-density lipoprotein (non-HDL)]. Subjects were required to be eligible for VA medical and pharmacy services 6 months prior to index date and throughout study period and to have complete data for exposure, outcome, and adjustment variables. Patients were excluded if they switched statins during the 12-month follow up period or had an admission for more than 30 consecutive days.

Table 3. ICD-9-CM Diagnosis Codes for dyslipidemia use for Case study 1.

ICD-9-CM Diagnosis Code	Description
272	Disorders of lipid metabolism
272.1	Pure hyperglyceridemia
272	Mixed hyperlipidemia
272.3	Hyperchylomicronemia
272.4	Hyperchylomicronemia

Veterans were subject to prescription drug copayments based on their eligibility status determined by the VA priority category of 2007.[61] Priority categories range from 1 to 8 with 1 being the highest priority. Veterans in priority group 1 were defined as having a service-connected disability that was rated ≥ 50% and determined by the VA to be due to service-related conditions. Veterans in priority group 1 were exempt from prescription drug copayments. Veterans in priority groups 2 to 6 had an $8 prescription drug copayment for each 30-day supply. Exemptions to prescription drug copayment were made for veterans who had a condition that was service-connected and for former prisoners of war. An annual prescription drug copayment cap of $960 was established for veterans enrolled in priority groups 2 to 6; however, no

medication cap was established for veterans in priority groups 7 to 8. Veterans in priority groups 7 to 8 were required to pay a prescription drug copayment. Veterans who did not have a service-related condition for dyslipidemia but were service-connected for another disability were required to pay a prescription drug copayment (labeled as Copayment Service Connected Category). Veterans who did not have a service-related condition for dyslipidemia and were not service-connected for another disability were required to pay a prescription drug copayment (labeled Copayment Non-Service Connected Category).

3.2.2. Case 2 – National level

The study population for Case study 2 was based on the national Veteran population which includes 50 states, the District of Columbia, and all US territories where a VA medical center or community-based outpatient center is present (Guam, Puerto Rico, Samoa, Philippines, and Virgin Islands). The Department of Veterans Affairs is responsible for 22.3 million covered lives at the end of fiscal year 2012.[62]

Patients were included in the second case if they had a diagnosis of type 2 diabetes mellitus (ICD-9, 250.X), greater than or equal to 18 years, and newly initiated on exenatide or liraglutide between January 1, 2006 to December 31, 2011. Patients were considered newly initiated on the study drug if they did not have any active prescriptions to either therapy 6 months prior to the index date. Patients on exenatide were dosed twice daily per the FDA approved labeling. Patients on liraglutide were dose once daily per the FDA approved labeling. Index date was defined as the date when patients filled the study medication. Patients were followed for up to 12 months to measure the change in HgbA1c. Patients who were eligible to receive VA prescription benefits

had to have at least 1 office visit 6 month prior to the index date and another office visit within 12 months post-index date.

Patients were excluded if they had a diagnostic code of type 1 diabetes mellitus or gestational diabetes as determined by ICD-9. Patients who started on one of the study medication and then switched to an alternative study medication were also excluded from analysis.

3.3. RESEARCH DESIGN AND PROCEDURES

This project focuses on the missing data analysis investigation of two retrospective cohort studies;[17,18] consequently, it is retrospective in nature. The original findings of the two retrospective studies used complete-case analysis which will be labeled as the control group. Results from the missing data analysis will be categorized as the experimental group. Missing data mechanism for MCAR will be tested using Little's test;[23] however, this is not a confirmatory test for non-MCAR mechanism. There are no statistical tests to distinguish between MCAR/MAR and NMAR. Therefore, for the purpose of this project, MAR will be assumed in order to fulfill the assumptions of multiple imputation.

3.3.1. Case study 1:

The purpose of Case study 1 was to evaluate the impact of adherence on lipid panel changes (25% reduction from baseline for LDL, TC, and non-HDL). Adherence to statin medication was based on the medication possession ratio (MPR), the main exposure variable in the multivariate regression model. Lipid panel changes were measured at 12 months after initiation of a statin

prescription and included low-density lipoproteins (LDL), non-high-density lipoprotein (non-HDL), and triglycerides (TG).

Adherence was the main exposure variable of interest and was categorized into adherent or non-adherent based on an arbitrary MPR threshold level of 0.80. Patients who were at or above the threshold were considered adherent; patients who were below the threshold were considered non-adherent. MPR is a surrogate marker for adherence and does not indicate that the patient ingested the medication; however, several studies have reported that it provides an accurate depiction of patient adherence using pharmacy claims data.[63,64] MPR was calculated as the number of days supplied of prescription medication divided by days of observation.[63,64] MPR was right-skewed; however it was assumed that this would not violate the assumptions of the multiple logistic regression.

The dependent variable was change in lipid panel levels which included LDL, non-HDL, and TG. Although these outcomes were continuous, the study categorized them into achieving a 25% reduction from baseline which is considered a clinically significant change.[65,66] This level of reduction has been associated with improved clinical outcomes.

3.3.2. Case study 2:

The purpose of Case study 2 was to evaluate the association between drug use with exenatide or liraglutide and HbA1c reduction in the Veteran population. The main exposure variable in the multivariate regression model was drug use (exenatide or liraglutide) and the outcome variable was change in HbA1c at 12 months after initiation.

HbA1c is a surrogate measure for long-term glycemic control in patients with diabetes.[67,68] The glycation process of hemoglobin reflects the amount of unregulated glucose in the blood that is exposed to erythrocytes where the hemoglobin is attached. Therefore, when glucose plasma concentration is high, more glucose binds to the erythrocyte and glycated hemoglobin increases.[69,70] Because, glycated hemoglobin is irreversibly attached to the erythrocyte, it takes approximately 120 days for elimination which is the average life span of a red blood cell.[71] The Diabetes Control and Complications Trial Research Group (DCCT) demonstrated the clinical association between HbA1c and microvascular complications.[72] As a result, HbA1c has become a standard laboratory assay for diabetic outcomes and long-term glycemic control.

3.4. INSTRUMENTS

Data source for the cases was derived from the Veterans Affairs (VA) Regional Data Warehouse and Corporate Data Warehouse.[73] Figures 3 and 4 diagrams the schematic data architecture of the Regional Data Warehouse and Corporate Data Warehouses, respectively. The VA databases have a comprehensive capture of pharmacy utilization, medical records, demographic characteristics, and health plan coverage elements allowing robust analysis of health outcomes.

Access to both databases required Institutional Review Board approval and must be in compliance with the Human Insurance Portability and Accountability Act.[74] Patients were de-identified and protected health information was never available to any of the investigators.

Data query was performed using Microsoft® Structured Query Language (SQL) Server Management Studio 2008 (Redmond, WA) and then transferred into a Comma-Separated Values (CSV) file which was imported into a statistical software for analysis.

Figure 3. Schematic of the VA Regional Data Warehouse.[75]

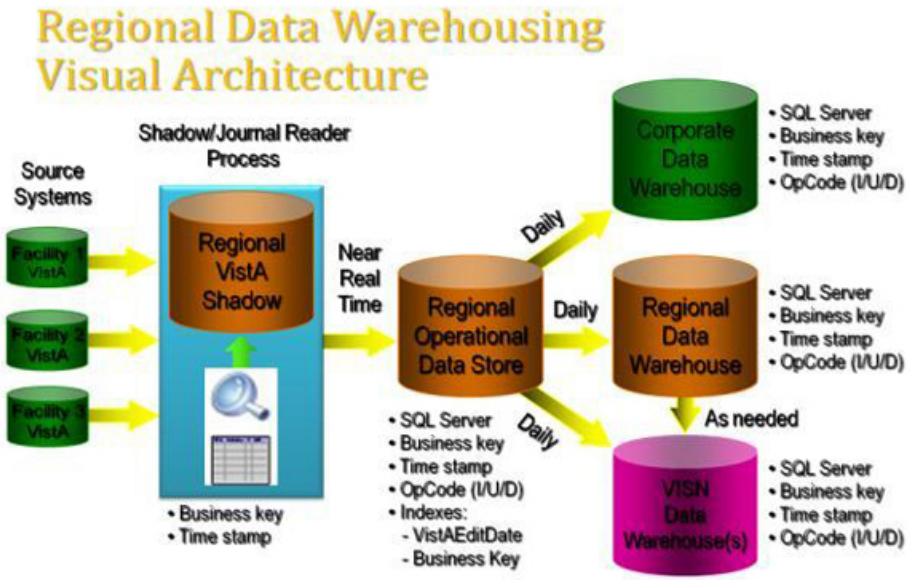
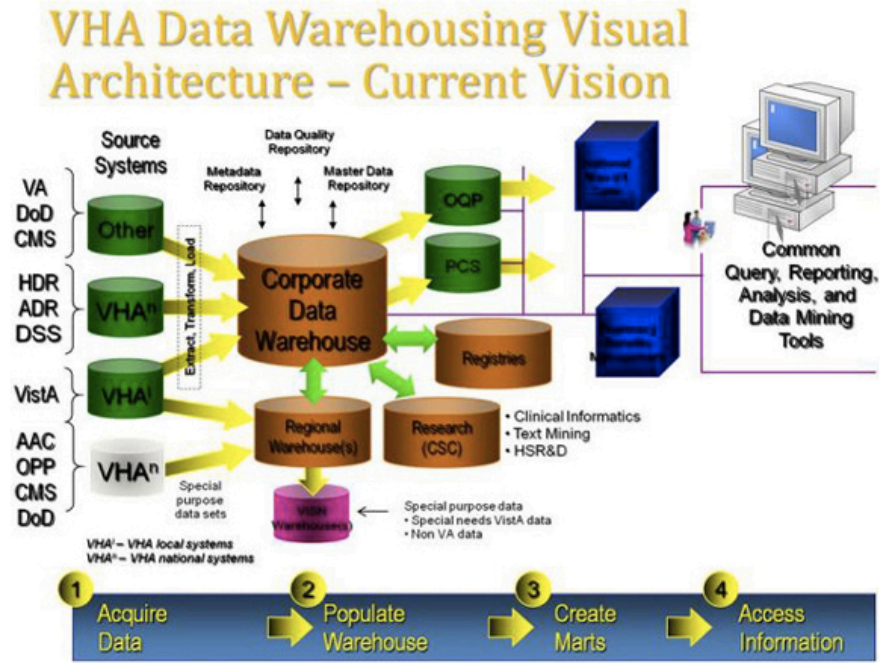


Figure 4. Schematic of the VA Corporate Data Warehouse.[75]



Case study 1 used data from the VA Regional Data Warehouse for VISN 22 who were new statin users. Care study 2 used data from the VA Corporate Data Warehouse to capture all patients who used exenatide and liraglutide for type II diabetes mellitus. Exenatide and liraglutide are not on the national formulary. Consequently, utilization is limited and a larger geographic area was required to capture an adequate sample size. Therefore, the national VA Corporate Data Warehouse was used to capture the small number of patients on exenatide and liraglutide.

3.5. PLANS FOR DATA ANALYSIS

The plan for missing data analysis will follow a systematic pattern that will be replicated for each study. The multivariate regression models of each study will be recreated and complete-case analysis will be performed. Parameter estimates and confidence intervals from the complete-case analysis will be compared to the multiple imputation methods.

The regression framework employed for Case study 1 was the logistic model expressed as:

$$P(D = 1|X_1, X_2, \dots, X_k) = P(X) = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{k}\beta_i X_i+\varepsilon)}},$$

where $i = 1, 2, 3, \dots, k$, $D = 1$ denotes the outcome of interest (e.g., achieving 25% in lipid reduction), X_i denote the k number of independent variables in the regression model, α and β denote unknown parameters, $P(X)$ denotes the probability of achieving the clinical goals (X=1) given that the following independent variables (X_k) are present, and ε denotes the error term. The logit form of the logistic model is expressed as:

$$Logit\ P(X) = \alpha + \sum_{i=1}^{k}\beta_i X_i + \varepsilon,$$

where $i = 1, 2, 3, \ldots, k$, X_k denotes the k number independent variables for in the regression model, and ε denotes the error term. The odds ratio (OR) is computed as the product of exponentials:

$$Odds\ ratio(OR) = \prod_{i=1}^{k} e^{\beta_i(X_{1i} - X_{0i})},$$

where X_1 and X_0 are two specifications of the collection of k independent variables $X_1, X_2, X_3, \ldots, X_k$.

The regression framework employed for Case study 2 was modeled as a multiple linear regression:

$$Y_i = \alpha + \sum_{i=1}^{k} \beta_i X_i + \varepsilon,$$

where Y_i denotes the outcome of interest (e.g., reduction in HbA1c), $i = 1, 2, 3, \ldots, k$, α denotes the Y-intercept or constant, β_i denotes the parameter coefficient for the independent variable (X_i), X_i denotes the k number of independent variables, and ε denotes the error term.

Multiple imputation methods will require an initial assumption that the missing data mechanism is MAR where the probability of the missing value depends of the observed value (Y_{obs}), not on the unobserved value (Y_{mis}). Multiple imputation will be carried out using Markov chain Monte Carlo (MCMC) simulation where a large number of samples are drawn from a posterior distribution yielding an estimate for the missing value.[6,20] Five imputed dataset ($m = 5$) will be generated and combined for the analysis of multivariate regression models using the methods developed by Little and Rubin.[6] Parameter estimates and confidence intervals will be compared to results of the complete-case analysis.

Sensitivity analysis will be performed comparing the parameter estimates and confidence intervals to those of the complete-case analysis and multiple imputation methods. Statistical significance will be set at a 5% (two-tailed) level. All data analysis will be performed using SAS version 9.3 (Cary, NC).

3.6. LIMITATIONS AND DELIMITATIONS

A noticeable limitation in this project is the lack of a "true" dataset. A "true" dataset is the complete dataset without any missing observations. This is impossible to acquire since missing data may be due to systematic issues, lack of response, miscoding, errors in imputation and a variety of other factors. Consequently, the mechanism of missing data needs to be investigated. Complete-case analysis assumes that the data is either MCAR or MAR. This assumption is robust to small amounts of missing data (<5%); however, large proportions of missing data are unlikely to be MCAR. This project assumes that the missing data mechanism was MAR. It is also possible that the missing data mechanism was NMAR.

An assumption of Case study 1 is that MPR accurately reflects patient adherence to their statin therapy. There are limitations with this assumption. MPR does not directly measure patient consumption of their statin therapy; instead, it provides an indirect estimate of adherence based on pharmacy refill data.[76] Other forms of adherence measurements are available which were not used in Case study 1. For example, Proportion of Days Covered (PDC) reflects the percentage of days the medication was available to the patient.[64] PDC is calculated as the total days the complete medication regimen was available divided by the total number of days evaluated capped at 1.0.[64] MPR can also be truncated or allowed to exceed a cap of 1.0. In Case study 1,

MPR was truncated at 1.0. It is unclear whether using the PDC would impact whether a patient was adherent or not. Hess, et al. reported that differences between MPR and PDC were negligible and provided similar answers in terms of categorizing patients as adherent or non-adherent.[64]

Another limitation of this project is the small sample size of the liraglutide group relative to the exenatide group in Case study 2.[18] The small sample is potentially sensitive to missing data which can result in inaccurate parameter estimates due to large uncertainties or variances. A larger sample size would mitigate this issue; however, there was not possible with the current design.

In observational studies, unmeasured variables can be potential confounders despite controlling for all measurable variables. Propensity score matching may be considered in this situation, however it is highly sensitive to unmeasured confounders.[77] It is not an absolute answer in the absence of a randomized controlled trial. Future investigation using propensity score matching should be pursued and compared to the result of the missing data analyses in these two case studies.

The Veteran population may not be generalizable to the non-veteran community. For both case studies, a majority of the sample subjects were male and white. It is debatable whether these factors affect the internal validity of the multivariate models for the case studies. However, their effect on the methods used for missing data analysis is likely to be trivial and are therefore ignored.

CHAPTER 4 - RESULTS

4.1. INTRODUCTION

The results for each Case study included the crude estimate, complete-case analysis, and the multiple imputation results. In Case study 1, the crude odds ratio (OR) and the OR from the complete-case analysis were compared to the OR from the multiple imputation method. In Case study 2, the crude estimate and the estimate from the complete-case analysis were compared to the estimate from the multiple imputation method. The SAS codes for each case are provided in APPENDIX A.

4.2. FINDINGS

4.2.1. Case study 1:

The purpose of Case study 1 was to evaluate the impact of adherence on lipid panel changes (LDL, TC, and non-HDL). Lipid panel changes were considered significant if there was a reduction of 25% at one year from baseline. Measurements were taken at one year and at baseline for LDL, TC, and non-HDL.

There were a total of 7,739 patients that were identified based on the inclusion and exclusion criteria. However, 6,074 (79%) patients had complete values and were eligible for inclusion in the complete-case analysis. For patients with complete data, 2,827 (47%) were adherent (MPR $\geq$ 0.80) and 3,247 (54%) were non-adherent (MPR < 0.80) (Table 4). A large proportion of the patients were male (N=5,786, 95%), white (N=2,948, 49%), prescribed simvastatin (N=85%), and had hypertension (N=4,366, 72%). Patients in the adherent group were older (64 versus 62

years, P<0.0001), had higher starting medication count (7.9 versus 6.8, P<0.0001), and lower LDL (133.9 versus 141.0 mg/dL, P<0.0001), TC (209.3 versus 217.6 mg/dL, P<0.0001), and non-HDL (167.3 versus 174.9 mg/dL, P<0.0001). In addition, patients who were adherent had lower number of patients with diabetes (P=0.0051), hypertension (P<0.0001), and vascular disease (P=0.0009); but higher number of patients with congestive heart failure (P=0.0055).

Table 4. Baseline demographics between adherent and non-adherent subjects.

Variables	Adherent	Non-Adherent	P-value
	N=2827	N=3247	
Age (years), mean (SD)	64.07 (10.79)	62.28 (11.29)	<0.0001
Male, No. (%)	2701 (95.54)	3085 (95.01)	0.3303
Starting medication count, mean (SD)	7.94 (4.62)	6.79 (4.14)	<0.0001
LDL baseline, mean (SD)	133.89 (40.32)	141.04 (39.22)	<0.0001
Non-HDL baseline, mean (SD)	167.30 (46.43)	174.85 (45.32)	<0.0001
TG baseline, mean (SD)	169.72 (150.19)	172.47 (144.95)	0.8263
TC baseline, mean (SD)	209.25 (48.39)	217.57 (47.03)	<0.0001
Ethnicity			<0.0001
White	1474 (52.14)	1474 (45.40)	
Black	345 (12.20)	548 (16.88)	
Hispanic	266 (9.41)	393 (12.10)	
Asian	91 (3.22)	121 (3.73)	
American Indian	37 (1.31)	46 (1.42)	
Unknown	614 (21.72)	665 (20.48)	
Statin			0.057
Simvastatin	2378 (84.12)	2776 (85.49)	
Atorvastatin	9 (0.32)	23 (0.71)	
Rosuvastatin	187 (6.61)	170 (5.24)	
Lovastatin	187 (6.61)	212 (6.53)	
Pravastatin	34 (1.20)	30 (0.92)	
Fluvastatin	32 (1.13)	36 (1.11)	
Copayment	1847 (65.33)	2196 (67.63)	0.0583
Diabetes	1113 (39.37)	1165 (35.88)	0.0051
Hypertension	2142 (75.77)	2224 (68.49)	<0.0001
Vascular disease	969 (34.38)	984 (30.30)	0.0009
Congestive heart failure	157 (5.55)	131 (4.03)	0.0055
History of myocardial infarction	84 (2.97)	99 (3.05)	0.8599
Angina	70 (2.48)	66 (2.03)	0.2439

LDL, low-density lipoprotein
TG, triglyceride
TC, total cholesterol
HDL, high-density lipoprotein

There was a moderate amount of missing data, mainly with the lipid panel. Baseline LDL, HDL, and TC values had 16.1%, 16.6%, and 17.9% missing data (Table 5). Follow-up LDL, HDL, and TC values had 19.2%, 19.6%, and 19.4% missing data. Baseline BMI had 6.8% missing data. Since the percent of missing data was greater than 5%, it was reasonable to assume that there was a potential for bias. The missing data pattern was suggestive of MAR; thereby allowing for multiple imputation to be performed.

Table 5. Missing data pattern for Case study 1.

Variables	Number missing	Percent missing
HDL FU	1520	19.6%
TC FU	1502	19.4%
LDL FU	1484	19.2%
TC baseline	1383	17.9%
HDL baseline	1286	16.6%
LDL baseline	1248	16.1%
BMI baseline	522	6.8%
Adherence status	0	0.0%
Copayment status	0	0.0%
Age	0	0.0%
Gender	0	0.0%
Ethnicity	0	0.0%
Starting medication count	0	0.0%
Diabetes	0	0.0%
Hypertension	0	0.0%
Vascular disease	0	0.0%
CHF	0	0.0%
History of MI	0	0.0%
Angina	0	0.0%

LDL, low-density lipoprotein
TC, total cholesterol
HDL, high-density lipoprotein

Comparison between groups with missing data reported no apparent differences except for the number of compliant patients (Table 6). There was approximately an equal proportion of patients who were male for groups that had missing data for baseline LDL, HDL, and TC compared to

groups with complete data for baseline LDL, HDL, and TC. Patient who had complete data had a higher proportion categorized as adherent compared to those with missing data for baseline LDL, HDL, and TC.

Table 6. Comparison between groups with missing and non-missing data.

Missing variable of interest	Missing			Non-Missing		
	LDL at baseline	HDL at baseline	TC at baseline	LDL at baseline	HDL at baseline	TC at baseline
Number	1248	1286	1239	6491	6453	6500
Age (years), mean (SD)	65.49 (11.82)	65.59 (11.85)	65.82 (11.81)	63.13 (11.15)	63.09 (11.14)	63.07 (11.14)
Gender (males), number (%)	1194 (95.67%)	1230 (95.65%)	1188 (95.88%)	6187 (95.32%)	6151 (95.32%)	6193 (95.28%)
Adherent, number (%)	483 (38.70%)	501 (38.96%)	485 (39.14%)	3003 (46.26%)	2985 (46.26%)	3001 (46.17%)

LDL, low-density lipoprotein
HDL, high-density lipoprotein
TC, total cholesterol

In the univariate analysis, there were statistically significant differences in the proportion of patients who had 25% or greater reduction in LDL, TC, and non-HDL for the adherent compared to the non-adherent patients (Table 7). Adherent patients had higher number who achieved a 25% or greater reduction in LDL compared to non-adherent patients (1,475 versus 776, P<0.0001). Similarly, adherent patients had a higher number who achieved a 25% or greater reduction in TC (1,005 versus 485, P<0.0001) and non-HDL (1,320 versus 650, P<0.0001) compared to non-adherent patients.

Table 7. Univariate analysis with lipid outcomes, Case study 1.

Outcome	Adherent		Non-Adherent		P-value
	N	%	N	%	
LDL >=25%	1475	52.18	776	23.90	<0.0001
TC >=25%	1005	35.55	485	14.94	<0.0001
Non-HDL >=25%	1320	46.69	650	20.02	<0.0001

LDL, low-density lipoprotein
TC, total cholesterol
HDL, high-density lipoprotein

Logistic regression model for the multivariate analysis controlled for age, BMI, gender, baseline lipid values, comorbid conditions (diabetes, hypertension, congestive heart failure, history of myocardial infarction, angina, vascular disease), statin use, ethnicity, and starting medication count. The regression results for the crude, complete-case analysis, and multiple imputation method were similar and did not vary significantly (Table 8). There was a bias away from the null with both the complete-case analysis and multiple imputation results for the odd of achieving a 25% or greater reduction in LDL compared to the crude OR. The crude OR (3.47; 95% CI: 3.11, 3.88) was lower than the OR for the complete-case analysis (4.34; 95% CI: 3.84, 4.90) and multiple imputation method (4.10; 95% CI: 3.66, 4.58). Similar bias away from the null with both the complete-case analysis and multiple imputation results was reported for the odd of achieving a 25% or greater reduction in TC compared to the crude OR. The crude OR (3.14; 95% CI: 2.78, 3.55) was lower than the OR for the complete-case analysis (4.00; 95% CI: 3.49, 4.59) and multiple imputation method (4.07; 95% CI: 3.58, 4.63). Bias away from the null was reported for the odds of achieving a 25% or greater reduction in non-HDL compared to the crude OR. The crude OR (3.50; 95% CI: 3.13, 3.92) was lower than the OR for the complete-case analysis (4.54; 95% CI: 4.00, 5.16) and multiple imputation method (4.37; 95% CI: 3.88, 4.91).

Table 8. Odds of achieving a ≥25% reduction in lipid panel levels for adherent versus non-adherent patients on a statin in the VASDHS, Case study 1.

	Crude N=6,074	Complete-case analysis* N=6,074	Multiple imputation* N=7,739
Outcome	Odds ratio (95% CI)	Odds ratio (95% CI)	Odds ratio (95% CI)
25% or greater reduction in LDL	3.47 (3.11, 3.88)	4.34 (3.84, 4.90)	4.10 (3.66, 4.58)
25% or greater reduction in Total Cholesterol	3.14 (2.78, 3.55)	4.00 (3.49, 4.59)	4.07 (3.58, 4.63)
25% or greater reduction in non-HDL	3.50 (3.13, 3.92)	4.54 (4.00, 5.16)	4.37 (3.88, 4.91)

VASDHS, Veterans Affairs San Diego Healthcare System
95% CI, 95% Confidence Interval
LDL, low-density lipoprotein
HDL, high-density lipoprotein
*Adjusted for age, gender, BMI, baseline lipid values, comorbid conditions (diabetes, hypertension, congestive heart failure, history of myocardial infarction, angina, vascular disease), statin use, ethnicity, and starting medication count.

4.2.2. Case study 2:

The purpose of Case study 2 was to evaluate the association between two glucagon-like peptide (GLP)-1 agonist (exenatide and liraglutide) and HbA1c reduction in the Veteran population using national data. The main exposure variable of interest was the GLP-1 agonist prescribed and the main outcome variable was change inHbA1c reduction at 12 months from baseline.

A total of 1,094 patients met inclusion and exclusion criteria. Exenatide had a larger proportion of patients compared to liraglutide (1,054 and 40, respectively). However, there were only 585 (53.5%) patients with complete data. Patients in the exenatide group (N=572) had a larger number of complete cases relative to the liraglutide group (N=13) with complete data.

Among the complete cases, there were no statistically significant differences in baseline demographics (Table 9). Exenatide and liraglutide patients were approximately 60 to 61 years old and mostly white males. Although the differences were not statistically significant, there was a higher proportion of dyslipidemia (100% versus 88%, P=0.3838) and obesity (77% versus 67%, P=0.1884) in liraglutide patients compared to exenatide patients. Conversely, There was a higher proportion of hypertension in exenatide patients compared to liraglutide patients (89% versus 77%, P=0.5607).

Baseline HbA1c, BMI, and race had 86 (7.9%), 108 (9.9%), and 291 (26.6%) patients with missing data, respectively (Table 10). The dependent variable, HbA1c at 12 months, had 188 (17.2%) patients with missing data. Patients who were in the exenatide arm of the study had higher proportion with missing baseline BMI values compared to liraglutide patients (10.1%

versus 5.0%). However, patients who were in the liraglutide arm of the study had higher proportions with missing baseline HbA1c (25.0 % versus 7.2%), race (32.5% versus 26.4%), and HbA1c at 12 months values (41.7% versus 16.3%) compared to exenatide.

Table 9. Baseline demographics for exenatide and liraglutide groups, Case study 2.

Variable	Exenatide (N=572)	Liraglutide (N=13)	P-value
Age (years), mean (SD)	60.28 (7.83)	60.93 (8.50)	0.6843
BMI (kg/m^2), mean (SD)	37.82 (7.55)	35.23 (5.83)	0.1753
Baseline HbA1c, mean (SD)	8.39 (1.65)	8.62 (1.21)	0.4665
Charlson Comorbidity Index, mean (SD)	1.66 (0.95)	1.31 (0.63)	0.1531
Female, N (%)	29 (5)	0 (0)	1.0000
Ethnicity, N (%)			0.7762
White	529 (92)	13 (100)	
Asian	6 (1)	0 (0)	
Unknown	22 (4)	0 (0)	
Declined to answer	17 (3)	0 (0)	
Congestive heart failure, N (%)	40 (7)	1 (8)	0.6138
Depression, N (%)	143 (25)	2 (15)	0.7448
Dyslipidemia, N (%)	503 (88)	13 (100)	0.3838
Hypertension, N (%)	508 (89)	10 (77)	0.1884
History of myocardial infarction, N (%)	6 (1)	0 (0)	1.0000
Obesity, N (%)	384 (67)	10 (77)	0.5607

SD, standard deviation

Since the percent of missing data was greater than 5%, it was reasonable to assume that there was a potential for bias. The missing data pattern was suggestive of MAR; thereby allowing for multiple imputation to be performed.

Table 10. Number of missing data, Case study 2.

	Total		Exenatide		Liraglutide	
Variable	Number missing	Percent missing	Number missing	Percent missing	Number missing	Percent missing
HbAc1 at 1 year	188	17.2%	170	16.1%	18	45.0%
BMI at baseline	108	9.9%	106	10.1%	2	5.0%
HbA1c at baseline	86	7.9%	76	7.2%	10	25.0%
Race	291	26.6%	278	26.4%	13	32.5%
Age	0	0.0%	0	0.0%	0	0.0%
Gender	0	0.0%	0	0.0%	0	0.0%
Charlson Comorbidity Index	0	0.0%	0	0.0%	0	0.0%
Congestive heart failure	0	0.0%	0	0.0%	0	0.0%
Depression	0	0.0%	0	0.0%	0	0.0%
Hypertension	0	0.0%	0	0.0%	0	0.0%
Obesity	0	0.0%	0	0.0%	0	0.0%
Dyslipidemia	0	0.0%	0	0.0%	0	0.0%

BMI, Body mass index (kb/m^2)

Comparisons between groups with missing data and complete data reported that no clear differences were present (Table 11). Patients who had missing HbA1 baseline values were similar in age and gender compared with patients who had non-missing HbA1 baseline values. There was a larger proportion of patients receiving liraglutide who had missing baseline HbA1c values compared to the non-missing group (12% versus 3%). In addition, patients with missing baseline HbA1c had a lower proportion of patients receiving exenatide compared to patients with complete data (88% versus 97%). For groups with missing and non-missing data for race, average age, gender and drug exposure were similar.

Table 11. Comparison between groups with missing and non-missing data.

	Missing		Non-Missing	
Missing variable of interest	HbA1c at baseline	Race	HbA1c at baseline	Race
Number	86	291	1008	803
Age (years), mean (SD)	60.93 (10.41)	61.07 (9.25)	62.74 (8.40)	63.15 (8.26)
Gender (males), number (%)	80 (93.02)	269 (92.44)	949 (94.15)	760 (94.65)
Drug: Exenatide, number (%)	76 (88.37)	278 (95.53)	978 (97.02)	776 (96.64)
Drug: Liraglutide, number (%)	10 (11.63)	13 (4.47)	30 (2.98)	27 (3.36)

HbA1c, hemoglobin A1c.

In the univariate analyses, liraglutide had a higher but non-significant reduction in HbA1c compared to exenatide (1.17 versus 0.65, P=0.3432). Results of the regression models reported that there were no statistically significant association between GLP-1 agonists (exenatide or liraglutide) and change in HbA1c relative to each other while controlling for age, gender, race, baseline HbA1c, BMI, CCI, history of myocardial infarction, congestive heart failure, hypertension, obesity, and dyslipidemia (Table 12). There was a higher but non-significant HbA1c reduction with exenatide relative to liraglutide in the crude estimate (0.37; 95% CI: -0.29, 1.03). Similarly, findings were reported with exenatide relative to liraglutide in the complete-case analysis (0.36; 95% CI: -0.35, 1.08) and multiple imputation method (0.25; 95% CI: -0.33, 0.82). The confidence band of the estimate generated by the multiple imputation method was narrower than the complete-case analysis and crude methods. Both missing data analyses resulted in an estimate that was biased towards the null.

Table 12. Percent change in HbA1 at 2 years from baseline for exenatide relative to liraglutide, Case study 2.

	Crude N=585	Complete-case analysis* N=585	Multiple imputation* N=1,094
Comparison	Change in HbA1c (95% CI)	Change in HbA1c (95% CI)	Change in HbA1c (95% CI)
Exenatide versus liraglutide	0.37% (-0.29%, 1.03%)	0.36% (-0.35%, 1.08%)	0.25% (-0.33%, 0.82%)

*Adjusted for age, gender, race, baseline HbA1c, BMI, CCI, history of myocardial infarction, congestive heart failure, hypertension, obesity, and dyslipidemia.

4.3. SUMMARY

In Case study 1, there were statistically significant differences in the odds of achieving a 25% in LDL, TC, and non-HDL between adherent and non-adherent patients to their statin prescriptions. However, the moderate amount of missing data created some potential for bias. Crude OR was estimated in order to provide a reference with which to compare other missing data analyses.

Complete-case analysis was performed to generate odds ratios with 95% confidence intervals for LDL, TC, and non-HDL outcomes. However, there were no differences in conclusions from the crude ORs. Similarly, there were no differences between the ORs and 95% CIs generated by multiple imputation methods and complete-case analysis.

For Case study 1, there was no apparent impact by missing data on the study conclusions, whether using crude or complete-case analysis. Multiple imputation method allowed for the use of the entire sample (N=7,739), thereby, maximizing the sample size. However, the improvement in precision was unnecessary since the point estimate and confidence limits did not change the study conclusions.

In Case study 2, the impact of missing data had a greater potential for influence since there was a disproportionate number of patients in the exenatide and liraglutide groups. There were 1,054 patients in the exenatide group and 40 patients in the liraglutide group. When complete-case analysis was performed, these numbers were reduced to 574 and 13 for the exenatide and liraglutide groups, respectively. The small sample size in the liraglutide group made it susceptible to potentially large changes in the outcomes.

There were no statistically significant differences in HbA1c reduction between exenatide and liraglutide at 1 year. Although, patients in the exenatide group had a larger HbA1c reduction relative to patients prescribed liraglutide, this difference was not statistically significant in the crude analysis (P=0.2674). Similarly, the complete-case analysis reported that there was a higher but non-significant HbA1c reduction with exenatide relative to liraglutide (P=0.3187).

Multiple imputation provided a method to maximize the already small sample size of liraglutide. When compared to the crude and complete-case analysis, no significant differences in conclusions were observed. Although there were no differences in conclusions between the different missing data analysis used, multiple imputation was biased towards the null compared to the crude and complete-case analysis.

In both case studies, there were no significant impact by the missing data; and multiple imputation provided the same answers as the complete-case analysis. Although there were some improvements in precisions with multiple imputation in Case study 2, this was not observed in Case study 2. The benefits of multiple imputation includes increased sample size for both case studies, especially Case study 2; however, there were no changes in the conclusions based on the complete-case analyses.

CHAPTER 5: CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS

5.1. INTRODUCTION

This section summarizes the critical elements of the study thesis. The study conclusions are provided as well as potential implications of the results in the field of epidemiology and missing data analysis. Several recommendations for the use of multiple imputation for missing data analysis are also provided.

5.2. SUMMARY OF STUDY

Missing data in epidemiologic studies may impact the results of multivariate models. Depending on the pattern and amount of missing data, missing data can influence the conclusions of studies, especially those that use multivariate models. In most statistical software, multivariate models remove cases from analysis when missing data is present. This reduces the sample size and may lead to biased results. In situations where this is highly probable, an investigation of the impact of missing data will need to be performed.

Most studies do not evaluate the impact of missing data. A study published by Fielding, et al.[58] reported that 59% (36 out of 61 randomly selected studies) of studies with missing data published in the *New England Journal of Medicine*, *Journal of the American Medical Association*, *BMJ*, and *Lancet* from 2005 to 2006 did not account for the potential impact that missing data may have on the conclusions of individual studies. Similarly, Wood, et al.[8] reported that 89% (63 out of 71 randomly selected studies reviewed) published in the *New England Journal of Medicine*, *Journal of the American Medical Association*, *BMJ*, and *Lancet* between

July and December 2001 had missing data. Most of the studies (92%) applied complete-case analysis which is the default mode for statistical software when dealing with missing data.

The purpose of this study was to evaluate the impact of missing data on two case studies that used multivariate models. In Case study 1, Watanabe, et al.[17] used multivariate models to evaluate the association between adherence and achievement of 25% lipid panel reductions (e.g., LDL, TC, and non-HDL). In Case study 2, Bounthavong, et al.[18] used multiple linear regression to evaluate the association between GLP-1 agonists (exenatide or liraglutide) and HbA1c reduction. In both cases, complete-case analysis was performed in handling missing data; however, the impact of missing data was not evaluated.

5.2.1. Case Study 1:

In Case study 1, there was a moderate amount of missing data that ranged from 16% to 19% for individual variables. In the multivariate models, the initial sample size of 7,739 subjects was reduced to 6,074 (~22% missing) due to missing data. Crude analysis was performed to provide a reference for comparison with the complete-case analysis and multiple imputation method. The conclusions from the complete-case analysis were similar to the crude analysis. Further comparisons between the conclusions of the multiple imputation method and complete-case analysis were also similar controlling for age, BMI, gender, baseline lipid values, comorbid conditions (diabetes, hypertension, congestive heart failure, history of myocardial infarction, angina, vascular disease), statin use, ethnicity, and starting medication count. In all scenarios, there were significant associations between adherence and achieving 25% reduction in lipid panel levels.

5.2.2. Case Study 2:

In Case study 2, there was a moderate-large amount of missing data that ranged from 5% to 45% for individual variables. The study had 1,094 patients who met the inclusion and exclusion criteria; however, only 585 (53.5% of the original sample) patients had complete data for analysis. Multiple linear regression was performed to evaluated the association between GLP-1 agonists and reduction in HbA1c levels controlling for age, gender, race, baseline HbA1c, BMI, CCI, history of myocardial infarction, congestive heart failure, hypertension, obesity, and dyslipidemia. Complete-case analysis and multiple imputation method were performed and the results were compared to the crude analysis. In all three methods, there were no significant association between exenatide and reduction in HbA1c relative to liraglutide. Multiple imputation method provided the same conclusion as the complete-case analysis and crude analysis.

5.3. CONCLUSIONS

5.3.1. Case Study 1:

In Case study 1, adherent patients were reported to have a higher odd of achieving 25% or more reduction in lipid panel levels relative to non-adherent patients controlling for potential confounding factors. Kazerooni, et al.[36] reported similar findings in that patients who were adherence (MPR $\geq$ 0.80) had a significant reduction in LDL and non-HDL level relative to non-adherent patients in a veteran population.

The use of multiple imputation did not change the conclusion of the complete-case analysis and crude analysis. However, it increased the sample size of the study and improved precision on the confidence limits when compared to the complete-case analysis. It also provided supporting evidence that the conclusions were appropriate and unaffected by missing data.

5.3.2. Case Study 2:

In Case study 2, patients who were prescribed exenatide had no significant association with change in HbA1c compared to liraglutide while controlling for potential confounders. In contrast, Buse, et al.[78] reported that liraglutide once daily had a significantly greater reduction in HbA1c compared with exenatide twice daily in a 26-week, open-labeled, parallel group, multinational study [-1.12% (SE, 0.08) versus -0.79% (0.08), P<0.0001]. The small sample size of the liraglutide cohort may reflect confounding by indication which is a form of selection bias.[79] It is plausible that patients who were prescribed liraglutide may have some other indication than diabetes that is different from exenatide patients. The VHA national formulary does not include GLP-1 agonists. However, during the study period of interest, exenatide twice daily had a criteria for use document that formulary managers in the VHA used to determine whether or not the patient is eligible to receive the liraglutide based on specific criteria; liraglutide did not (http://www.pbm.va.gov). Therefore, it is speculated, that providers and pharmacists preferred to use exenatide as a consequence of formulary guidelines being available. This may indicate that patients on lirgalutide could have been approved based on a specific clinical indication (other than diabetes) that would prevent them from using exenatide (e.g., previous history of exenatide, contraindication, or patient/provider preferences). Future studies

will need to randomize exenatide and liraglutide to diabetic patients based on baseline HbA1c in order to reduce bias and potential confounding.

Regardless, multiple imputation provided similar conclusions as the complete-case analysis and crude analysis. The confidence limit from the multiple imputation method was narrower compared to the complete-case analysis and crude analysis; however, this improvement in precision did not change the conclusion that there was no significant association between exenatide and change in HbA1c compared to liraglutide. In addition, multiple imputation provided benefit by increasing the sample size of the small cohort of liraglutide patients.

5.3.3. Limitations

In both case studies, multiple imputation and complete-case analysis for the missing data methods performed. However, there are other methods available that were not investigated, thereby, limiting the generalizability of these results. Maximum likelihood estimator is another method that could be applied to handling missing data. Unlike multiple imputation where *m* number of datasets need to be generated through imputation methods, maximum likelihood is a simple method that does not require much computational power or time to perform.[29–31]

This study was performed using data from the veteran population; as a result, generalizability to a non-veteran population may be limited. It is unclear whether or not veterans would be more adherent compared to non-veterans; or if veterans would be less likely to have major reductions in the HbA1c after using a GLP-1 agonist. There have been no studies to demonstrate these changes. Moreover, the veteran demographic is predominantly white male which limits

generalizability to the non-white female population. Fitzgerald, et al. reported that there were no differences in attitude but not adherence between male and female in patients with Type 2 diabetes mellitus.[80] However, they reported that for type 1 diabetes mellitus, there were differences in patient attitudes.[80] Men were more likely to follow the provider's orders while women were more likely to take type 1 diabetes mellitus serious and agree that diabetes has a significant effect on quality of life.[80] Future studies will need to recruit equal number of male and female subjects to eliminate this potential confounder.

Another limitation of this study is the absence of an evaluation for NMAR. Heckman selection procedure is used when the missing data pattern is reflective of NMAR.[81] Since this study assumed MAR, multiple imputation was performed. Heckman selection procedure tends to provide conclusions that are different from both the complete-case analysis and multiple imputation. Missing data pattern that is NMAR is unlikely; therefore, Heckman selection procedure was not performed in this study.

5.4. IMPLICATIONS

Multiple imputation improved the sample size and increased precision in cases where multivariate regression models were employed. In both case studies, multiple imputation provided reassurance for the conclusions obtained by complete-case analysis.

Currently, most statistical software (e.g., SAS, SPSS, and STATA) by default perform complete-case analysis for multivariate models. Researchers who are unaware of this default setting may find that the sample size has been reduced and potential bias present. It will also be difficult to

convince researchers to perform additional analysis, especially if the benefits of missing data analysis provide minimal incremental benefits. In the two case studies of this thesis, using multiple imputation to handle missing data did not outperform complete-case analysis methods. However, multiple imputation performed an important role in providing additional support for the initial conclusions based on complete-case analysis. This has important implications when it comes to using the results from the complete-case analysis for making policy decisions.

Situations where the conclusions are divergent between complete-case analysis and multiple imputation method would benefit from the analyses. In these situations, researchers will need to address the impact that missing data have on their study conclusions. More importantly, they would need to discuss the potential compromise in internal validity due to missing data.

5.5. RECOMMENDATIONS

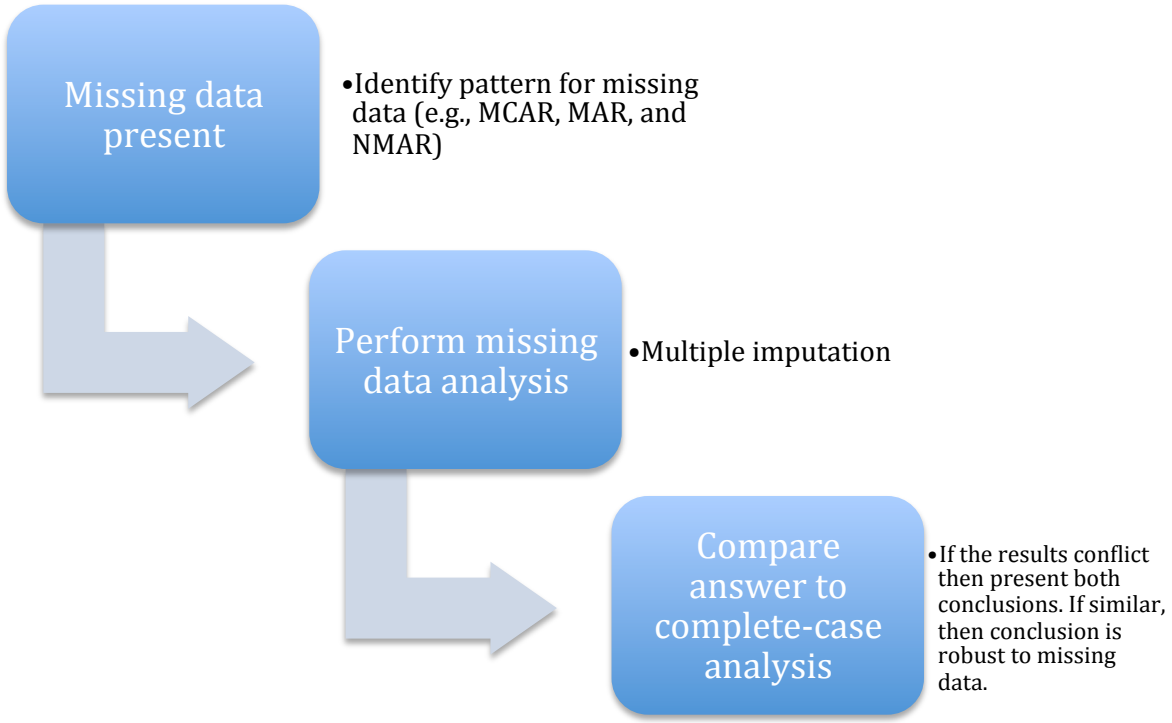
Based on the results of this study, multiple imputation provided similar conclusions that were based on complete-case anlysis. Researchers who are involved with multivariate models may consider using multiple imputation to address missing data in order to provide additional support for results obtained from complete-case analysis.

However, if differences are present, researchers will need to address the impact of missing data on their conclusions. According to the European Medicines Agency (EMA), there is no standard guideline that can be applied in all scenarios where missing data is present.[10] According to the EMA, each case will require the researcher to apply the appropriate methods in handling missing data on a case-by-case basis. An important element of missing data analysis is the performance

of sensitivity analysis to support the initial study conclusions. In our case studies, multiple imputation was performed in order to support the conclusions based on complete-case analysis.

Researchers facing problems with missing data should perform complete-case analysis followed by another missing data analysis method (e.g., multiple imputation, maximum likelihood estimation) in order to support the initial conclusions. If there are conflicts, then the researcher will need to highlight this as a potential problem of the study. Transparency is critical when it comes to making informed decisions about based on study conclusions hampered by missing data. Figure 5 summarizes the recommendations for performing missing data analysis.

Figure 5. Recommended guideline for validating study conclusions.



REFERENCES

1. Ferguson, L. External validity, generalizability, and knowledge utilization. *J. Nurs. Scholarsh. Off. Publ. Sigma Theta Tau Int. Honor Soc. Nurs. Sigma Theta Tau* **36,** 16–22 (2004).
2. Campbell, D. T. & Stanley, J. *Experimental and quasiexperimental designs for research.* (Houghton Mifflin, 1963).
3. Grimes, D. A. & Schulz, K. F. Bias and causal associations in observational research. *Lancet* **359,** 248–252 (2002).
4. Tripepi, G., Jager, K. J., Dekker, F. W., Wanner, C. & Zoccali, C. Bias in clinical research. *Kidney Int.* **73,** 148–153 (2008).
5. Kleinbaum, D. G., Sullivan, K. M. & Barker, N. D. *ActiveEpi Companion Textbook: A supplement for use with the ActiveEPI CD-ROM*. (2003).
6. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data*. (John Wiley & Sons, Inc., 2002).
7. Rubin, D. B. Inference and missing data. *Biometrika* **63,** 581–592 (1976).
8. Wood, A. M., White, I. R. & Thompson, S. G. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin. Trials Lond. Engl.* **1,** 368–376 (2004).
9. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338,** b2393 (2009).
10. European Medicines Evaluation Agency. Guideline on Missing Data in Confirmatory Clinical Trials. Committee for Medical Products for Human Use. (2009). at <http://www.ema.europa.eu/pdfs/human/ewp/177699endraft.pdf>
11. The Prevention and Treatment of Missing Data in Clinical Trials. at <http://www.nap.edu/openbook.php?record_id=12955>
12. Liu, M., Wei, L. & Zhang, J. Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study. *Pharm. Stat.* **5,** 7–18 (2006).
13. Li, T. *et al. Minimal Standards in the Prevention and Handling of Missing Data in Observational and Experimental Patient Centered Outcomes Research*. (2012). at <http://www.pcori.org/assets/Minimal-Standards-in-the-Prevention-and-Handling-of-Missing-Data-in-Observational-and-Experimental-Patient-Centered-Outcomes-Research1.pdf>
14. Vandenbroucke, J. P. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Med* **4,** e297 (2007).
15. Moher, D. *et al.* CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *J. Clin. Epidemiol.* **63,** e1–37 (2010).
16. Gallo, V. *et al.* STrengthening the Reporting of OBservational studies in Epidemiology--Molecular Epidemiology STROBE-ME: an extension of the STROBE statement. *J. Clin. Epidemiol.* **64,** 1350–1363 (2011).
17. Watanabe, J. H., Bounthavong, M. & Chen, T. Revisiting the medication possession ratio threshold for adherence in lipid management. *Curr. Med. Res. Opin.* **29,** 175–180 (2013).
18. Bounthavong, M., Tran, J. N., Watanabe, J. H. & Chen, T. C. PDB15 - Retrospective Cohort Study Evaluating Liraglutide And Exenatide In A Veteran Population. *Value Health* **16,** A158 (2013).

19. Haukoos, J. S. & Newgard, C. D. Advanced statistics: missing data in clinical research--part 1: an introduction and conceptual framework. *Acad. Emerg. Med.* **14,** 662–668 (2007).
20. Rubin, D. B. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **91,** 473–489 (1996).
21. Raghunathan, T. E. What do we do with missing data? Some options for analysis of incomplete data. *Annu. Rev. Public Health* **25,** 99–117 (2004).
22. McKnight, P. E., McKnight, K. M., Sidani, S. & Figueredo, A. J. *Missing Data: A Gentle Introduction*. (The Guildorf Press, 2007).
23. Little, R. J. A. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J. Am. Stat. Assoc.* **83,** 1198–1202 (1988).
24. Newgard, C. D. & Haukoos, J. S. Advanced statistics: missing data in clinical research--part 2: multiple imputation. *Acad. Emerg. Med.* **14,** 669–678 (2007).
25. Kneipp, S. M. & McIntosh, M. Handling missing data in nursing research with multiple imputation. *Nurs. Res.* **50,** 384–389 (2001).
26. Sinharay, S., Stern, H. S. & Russell, D. The use of multiple imputation for the analysis of missing data. *Psychol. Methods* **6,** 317–329 (2001).
27. Schafer, J. L. Multiple imputation: a primer. *Stat. Methods Med. Res.* **8,** 3–15 (1999).
28. Héraud-Bousquet, V., Larsen, C., Carpenter, J., Desenclos, J.-C. & Le Strat, Y. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Med. Res. Methodol.* **12,** 73 (2012).
29. Chen, M. H. & Ibrahim, J. G. Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* **57,** 43–52 (2001).
30. Horton, N. J. & Laird, N. M. Maximum likelihood analysis of generalized linear models with missing covariates. *Stat. Methods Med. Res.* **8,** 37–50 (1999).
31. Shih, W. J. Maximum likelihood estimation and likelihood ratio test for square tables with missing data. *Stat. Med.* **6,** 91–97 (1987).
32. Li, L., Shen, C., Li, X. & Robins, J. M. On weighting approaches for missing data. *Stat. Methods Med. Res.* **22,** 14–30 (2013).
33. Seaman, S. R. & White, I. R. Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **22,** 278–295 (2013).
34. Collins, L. M., Schafer, J. L. & Kam, C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* **6,** 330–351 (2001).
35. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. (John Wiley & Sons, Inc., 1987).
36. Kazerooni, R., Watanabe, J. H. & Bounthavong, M. Association Between Statin Adherence and Cholesterol Level Reduction from Baseline in a Veteran Population. *Pharmacotherapy* (2013). doi:10.1002/phar.1305
37. Messer, K. & Natarajan, L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Stat. Med.* **27,** 6332–6350 (2008).
38. Buhi, E. R., Goodson, P. & Neilands, T. B. Out of sight, not out of mind: strategies for handling missing data. *Am. J. Health Behav.* **32,** 83–92 (2008).
39. Choi, Y. J., Nam, C. M. & Kwak, M. J. Multiple imputation technique applied to appropriateness ratings in cataract surgery. *Yonsei Med. J.* **45,** 829–837 (2004).
40. Balise, R. R. *et al.* Imputation of missing ages in pedigree data. *Hum. Hered.* **63,** 168–174 (2007).

41. Moons, K. G. M., Donders, R. A. R. T., Stijnen, T. & Harrell, F. E., Jr. Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.* **59,** 1092–1101 (2006).
42. Rue, T., Thompson, H. J., Rivara, F. P., Mackenzie, E. J. & Jurkovich, G. J. Managing the common problem of missing data in trauma studies. *J. Nurs. Scholarsh. Off. Publ. Sigma Theta Tau Int. Honor Soc. Nurs. Sigma Theta Tau* **40,** 373–378 (2008).
43. Moore, L. *et al.* Multiple imputation of the Glasgow Coma Score. *J. Trauma* **59,** 698–704 (2005).
44. Croiseau, P., Génin, E. & Cordell, H. J. Dealing with missing data in family-based association studies: a multiple imputation approach. *Hum. Hered.* **63,** 229–238 (2007).
45. Souverein, O. W., Zwinderman, A. H. & Tanck, M. W. T. Multiple imputation of missing genotype data for unrelated individuals. *Ann. Hum. Genet.* **70,** 372–381 (2006).
46. Cordell, H. J. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genet. Epidemiol.* **30,** 259–275 (2006).
47. Siddique, J. & Belin, T. R. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Stat. Med.* **27,** 83–102 (2008).
48. Bono, C., Ried, L. D., Kimberlin, C. & Vogel, B. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. *Res. Soc. Adm. Pharm. RSAP* **3,** 1–27 (2007).
49. Moore, L., Hanley, J. A., Turgeon, A. F., Lavoie, A. & Emond, M. A multiple imputation model for imputing missing physiologic data in the national trauma data bank. *J. Am. Coll. Surg.* **209,** 572–579 (2009).
50. Mulla, Z. D., Seo, B., Kalamegham, R. & Nuwayhid, B. S. Multiple imputation for missing laboratory data: an example from infectious disease epidemiology. *Ann. Epidemiol.* **19,** 908–914 (2009).
51. Van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T. & Moons, K. G. M. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J. Clin. Epidemiol.* **59,** 1102–1109 (2006).
52. Yang, X., Belin, T. R. & Boscardin, W. J. Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61,** 498–506 (2005).
53. Greenland, S. & Finkle, W. D. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am. J. Epidemiol.* **142,** 1255–1264 (1995).
54. Newgard, C. D. The validity of using multiple imputation for missing out-of-hospital data in a state trauma registry. *Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med.* **13,** 314–324 (2006).
55. O'Loughlin, R. E. *et al.* The epidemiology of invasive group A streptococcal infection and potential vaccine implications: United States, 2000-2004. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **45,** 853–862 (2007).
56. Van Beek, E. J. *et al.* A normal perfusion lung scan in patients with clinically suspected pulmonary embolism. Frequency and clinical validity. *Chest* **108,** 170–173 (1995).
57. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).

58. Fielding, S., Maclennan, G., Cook, J. A. & Ramsay, C. R. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* **9,** 51 (2008).
59. US Department of Veterans Affairs. VISN 22: Desert Pacific Healthcare Network. at <http://www.va.gov/directory/guide/region.asp?ID=1022>
60. US Department of Veterans Affairs. Veterans Health Administration. About the VHA. at <http://www.va.gov/health/aboutVHA.asp>
61. *Department of Veterans Affairs. 2007 Veterans Health Care Handbook. Military Handbooks. 2007.* (Department of Veterans Affairs, 2007). at <http://www.militaryhandbooks.com>
62. Department of Veterans Affairs. Veteran Population. National Center for Veterans Analysis and Statistics. at <http://www.va.gov/vetdata/Veteran_Population.asp>
63. Andrade, S. E., Kahler, K. H., Frech, F. & Chan, K. A. Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiol. Drug Saf.* **15,** 565–574; discussion 575–577 (2006).
64. Hess, L. M., Raebel, M. A., Conner, D. A. & Malone, D. C. Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. *Ann. Pharmacother.* **40,** 1280–1288 (2006).
65. Grundy, S. M. *et al.* Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III Guidelines. *J. Am. Coll. Cardiol.* **44,** 720–732 (2004).
66. Gotto, A. M., Jr & Grundy, S. M. Lowering LDL cholesterol: questions from recent meta-analyses and subset analyses of clinical trial DataIssues from the Interdisciplinary Council on Reducing the Risk for Coronary Heart Disease, ninth Council meeting. *Circulation* **99,** E1–7 (1999).
67. Saudek, C. D. & Brick, J. C. The clinical use of hemoglobin A1c. *J. Diabetes Sci. Technol.* **3,** 629–634 (2009).
68. Kilpatrick, E. S. Haemoglobin A1c in the diagnosis and monitoring of diabetes mellitus. *J. Clin. Pathol.* **61,** 977–982 (2008).
69. Jovanovic, L. & Peterson, C. M. The clinical utility of glycosylated hemoglobin. *Am. J. Med.* **70,** 331–338 (1981).
70. McDonald, J. M. & Davis, J. E. Glycosylated hemoglobins and diabetes mellitus. *Hum. Pathol.* **10,** 279–291 (1979).
71. Bunn, H. F., Haney, D. N., Kamin, S., Gabbay, K. H. & Gallop, P. M. The biosynthesis of human hemoglobin A1c. Slow glycosylation of hemoglobin in vivo. *J. Clin. Invest.* **57,** 1652–1659 (1976).
72. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. *N. Engl. J. Med.* **329,** 977–986 (1993).
73. VA Informatics and Computing Infrastructure (VINCI). at <http://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm#.UfDEN20ueSo>
74. Summary of the HIPAA Privacy Rule. at <http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/>
75. *VA Information Resource Center. VIReC Resource Guide: VA Corporate Data Warehouse.* (Hines, IL: U.S. Dept. of Veterans Affairs, Health Services Research and Development Service, VA Information Resource Center, 2012).

76. Ho, P. M., Bryson, C. L. & Rumsfeld, J. S. Medication Adherence: Its Importance in Cardiovascular Outcomes. *Circulation* **119,** 3028–3035 (2009).
77. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55 (1983).
78. Buse, J. B. *et al.* Liraglutide once a day versus exenatide twice a day for type 2 diabetes: a 26-week randomised, parallel-group, multinational, open-label trial (LEAD-6). *Lancet* **374,** 39–47 (2009).
79. Salas, M., Hofman, A. & Stricker, B. H. Confounding by indication: an example of variation in the use of epidemiologic terminology. *Am. J. Epidemiol.* **149,** 981–983 (1999).
80. Fitzgerald, J. T., Anderson, R. M. & Davis, W. K. Gender differences in diabetes attitudes and adherence. *Diabetes Educ.* **21,** 523–529 (1995).
81. Sales, A. E., Plomondon, M. E., Magid, D. J., Spertus, J. A. & Rumsfeld, J. S. Assessing response bias from missing quality of life data: the Heckman method. *Health Qual. Life Outcomes* **2,** 49 (2004).

APPENDIX

Appendix 1. SAS codes for Case study 1.

```
*************************************************************************;
*STUDENT: MARK BOUNTHAVONG - THESIS;
*TITLE: MISSING DATA ANALYSIS WITH MULTIVARIATE MODELS CASE 1;
*THESIS CHAIR: DR. KEVIN SULLIVAN;
*THESIS COMMITTEE MEMBER: DR. JONATHAN H. WATANABE;
*DATE OF ANALYSIS: 17 SEPTEMBER 2013;
*************************************************************************;

*THERE ARE A COUPLE OF THINGS THAT NEED TO BE CLARIFIED FOR THIS;
*THE OUTCOMES ARE 25% REDUCTION IN LIPID PROFILE FOR LDL, HDL AND NON-HDL;
*NON-HDL WAS CALCULATED AS: NON-HDL = TC - HDL;
*************************************************************************;
*START;
*************************************************************************;
*************************************************************************;
*STEP 1: IMPORTING DATA;
*************************************************************************;

libname s "H:\Courses\Thesis\copay.xlsx";

proc contents data=s.'sheet1$'n;
run;

data copay;
set s.'sheet1$'n;
run;

proc contents data=copay;
run;


*************************************************************************;
*STEP 2: VERIFYING THE TOTAL SAMPLE SIZE;
*************************************************************************;
*TOTAL N = 7739;
PROC FREQ DATA=COPAY;
TABLES RK_ETHNICITY;
RUN;


*************************************************************************;
*STEP 3: BUILDING A DATASET THAT ONLY HAS THE COMPLETE CASES;
*************************************************************************;
DATA COPAY2;
SET COPAY;
      NON_HDL_BASE = TC_BASE - HDL_BASE;
      NON_HDL_FU = TC_FU - HDL_FU;
      CHANGE_LDL = LDL_FU-LDL_BASE;
      CHANGE_TC = TC_FU-TC_BASE;
      CHANGE_NON_HDL = NON_HDL_FU - NON_HDL_BASE;

      PERCENT_LDL = -(CHANGE_LDL)/LDL_BASE;
      PERCENT_TC = -(CHANGE_TC)/TC_BASE;
      PERCENT_NON_HDL = -(CHANGE_NON_HDL)/NON_HDL_BASE;
```

```
      IF PERCENT_LDL >=0.25 THEN LDL25 = 1;
      IF PERCENT_LDL <0.25 THEN LDL25 = 0;

      IF PERCENT_TC >=0.25 THEN TC25 = 1;
      IF PERCENT_TC <0.25 THEN TC25 = 0;

      IF PERCENT_NON_HDL >=0.25 THEN NONHDL25 = 1;
      IF PERCENT_NON_HDL <0.25 THEN NONHDL25 = 0;

RUN;

*************************************************************************;
*STEP 3-a: DESCRIPTIVE ANALYSIS BETWEEN PATIENTS WITH AND W/O MISSING DATA;
*************************************************************************;
*LDL AT BASELINE IS GROUPED MISSING AND NON-MISSING;
DATA COPAY9;
SET COPAY2;
      IF ldl_base =. THEN MISSING = 1; ELSE MISSING = 0;
/*    IF tc_base =. THEN MISSING = 1; ELSE MISSING = 0;
      IF hdl_base =. THEN MISSING = 1; ELSE MISSING = 0;
      IF LDL25 =. THEN MISSING = 1; ELSE MISSING = 0;;
      IF TC25 =. THEN MISSING = 1; ELSE MISSING = 0;
      IF NONHDL25 =. THEN MISSING = 1; ELSE MISSING = 0;
*/
RUN;

PROC SORT DATA=COPAY9;
BY MISSING;
RUN;

PROC MEANS DATA=COPAY9 N MEAN STD MEDIAN QRANGE;
CLASS MISSING;
VAR AGE LDL_BASE TC_BASE HDL_BASE BMIBASE;
RUN;

PROC FREQ DATA=COPAY9;
TABLES (GENDER RK_COMPLIANT)*MISSING / CHISQ;
RUN;


*HDL-LDL AT BASELINE IS GROUPED MISSING AND NON-MISSING;
DATA COPAY10;
SET COPAY2;
/*    IF ldl_base =. THEN MISSING = 1; ELSE MISSING = 0;
*/    IF tc_base =. THEN MISSING = 1; ELSE MISSING = 0;
      IF hdl_base =. THEN MISSING = 1; ELSE MISSING = 0;
/*    IF LDL25 =. THEN MISSING = 1; ELSE MISSING = 0;;
      IF TC25 =. THEN MISSING = 1; ELSE MISSING = 0;
      IF NONHDL25 =. THEN MISSING = 1; ELSE MISSING = 0;
*/
RUN;

PROC SORT DATA=COPAY10;
BY MISSING;
RUN;
```

```
PROC MEANS DATA=COPAY10 N MEAN STD MEDIAN QRANGE;
CLASS MISSING;
VAR AGE LDL_BASE TC_BASE HDL_BASE BMIBASE;
RUN;

PROC FREQ DATA=COPAY10;
TABLES (GENDER RK_COMPLIANT)*MISSING / CHISQ;
RUN;


*HDL-LDL AT BASELINE IS GROUPED MISSING AND NON-MISSING;
DATA COPAY11;
SET COPAY2;
/*     IF ldl_base =. THEN MISSING = 1; ELSE MISSING = 0;*/
       IF tc_base =. THEN MISSING = 1; ELSE MISSING = 0;
/*     IF hdl_base =. THEN MISSING = 1; ELSE MISSING = 0;
       IF LDL25 =. THEN MISSING = 1; ELSE MISSING = 0;;
       IF TC25 =. THEN MISSING = 1; ELSE MISSING = 0;
       IF NONHDL25 =. THEN MISSING = 1; ELSE MISSING = 0;
*/
RUN;

PROC SORT DATA=COPAY11;
BY MISSING;
RUN;

PROC MEANS DATA=COPAY11 N MEAN STD MEDIAN QRANGE;
CLASS MISSING;
VAR AGE LDL_BASE TC_BASE HDL_BASE BMIBASE;
RUN;

PROC FREQ DATA=COPAY11;
TABLES (GENDER RK_COMPLIANT)*MISSING / CHISQ;
RUN;

*************************************************************************;
*STEP 3-b: *NEW DATA SET FOR TOTAL N = 6074 FOR COMPLETE-CASE ANALYSIS;
*************************************************************************;
DATA COPAY3;
SET COPAY2;

     IF MPR ~= . ;
     IF RK_MALE ~= . ;
     IF angina ~= . ;
     IF chf ~= . ;
     IF copd ~= . ;
     IF dm ~= . ;
     IF depres ~= . ;
     IF hxmi ~= . ;
     IF vascdis ~= . ;
     IF htn ~= . ;
     IF bipol ~= . ;
     IF rk_copay ~= . ;
     IF age ~= . ;
     IF ldl_base ~= . ;
     IF tc_base ~= . ;
     IF hdl_base ~= . ;
```

```
      IF tg_base ~= . ;
      IF bmibase ~= . ;
        IF RK_COMPLIANT ~=.;
        IF LDL25 ~=.;
        IF TC25 ~=.;
        IF NONHDL25 ~=.;

RUN;

*************************************************************************
*STEP 3-c: VERIFY THAT THE CALCULATIONS WERE HANDLED APPROPRIATELY;
*************************************************************************
PROC SORT DATA=COPAY3;
BY RK_COMPLIANT;
RUN;
PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM PERCENT_LDL;
RUN;
PROC MEANS DATA=COPAY3;
CLASS RK_COMPLIANT;
VAR PERCENT_LDL;
RUN;
PROC CONTENTS DATA=COPAY3;
RUN;
PROC PRINT DATA=COPAY3 (OBS=10);
VAR PERCENT_LDL;
RUN;


*************************************************************************;
*STEP 4: BASELINE DEMOGRAPHICS;
*************************************************************************;
PROC MEANS DATA=COPAY3 MEAN N STD MEDIAN QRANGE;
CLASS RK_COMPLIANT;
VAR AGE STARTMEDCNT LDL_BASE NON_HDL_BASE TG_BASE TC_BASE PERCENT_LDL
PERCENT_TC PERCENT_NON_HDL;
RUN;

PROC FREQ DATA=COPAY3;
TABLES (RK_MALE RK_ETHNICITY STATIN RK_COPAY DM HTN VASCDIS CHF HXMI
ANGINA)*RK_COMPLIANT / CHISQ;
RUN;

*VISUAL INSPECTION OF DISTRIBUTIONS;
PROC SORT DATA=COPAY3;
BY RK_COMPLIANT;
RUN;

PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM AGE;
RUN;

PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM LDL_BASE;
```

```
RUN;

*SKEWED TO THE RIGHT;
PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM NON_HDL_BASE;
RUN;

*SKEWED TO THE RIGHT;
PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM TG_BASE;
RUN;

PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM TC_BASE;
RUN;

*SKEWED TO THE RIGHT;
PROC SGPLOT DATA=COPAY3;
BY RK_COMPLIANT;
HISTOGRAM STARTMEDCNT;
RUN;

*TTEST;
PROC TTEST DATA=COPAY3;
CLASS RK_COMPLIANT;
VAR AGE LDL_BASE TC_BASE NON_HDL_BASE;
RUN;

PROC NPAR1WAY DATA=COPAY3 WILCOXON;
CLASS RK_COPAY;
VAR STARTMEDCNT TG_BASE;
RUN;


*************************************************************************;
*STEP 5: UNIVARIATE ANALYSIS;
*************************************************************************;
*PART A: OUTCOME=PERCENT LDL CHANGE >= 25%;
PROC FREQ DATA=COPAY3;
TABLES RK_COMPLIANT*LDL25 / CHISQ;
RUN;
*PART B: OUTCOME=PERCENT TC CHANGE >= 25%;
PROC FREQ DATA=COPAY3;
TABLES RK_COMPLIANT*TC25 / CHISQ;
RUN;

*PART C: OUTCOME=PERCENT NON-HDL CHANGE >= 25%;
PROC FREQ DATA=COPAY3;
TABLES RK_COMPLIANT*NONHDL25 / CHISQ;
RUN;


*************************************************************************;
*STEP 6: PERFORMING THE LOGISTIC REGRESSION MODEL -- COMPLETE-CASE ANALYSIS;
```

```
*************************************************************************;
*PART A: LDL25 = 1;
PROC LOGISTIC DATA=COPAY3;
CLASS       rk_ethnicity (PARAM=REF REF='WHITE')
            rk_copay (param=ref ref='0')
            statin (PARAM=REF REF='SIMVASTATIN')
            RK_MALE (PARAM=REF REF='0')
            dm (param=ref ref='0')
            htn (param=ref ref='0')
            vascdis (param=ref ref='0')
            chf (param=ref ref='0')
            COPD (param=ref ref='0')
            HXMI (param=ref ref='0')
            ANGINA(param=ref ref='0');
MODEL LDL25 (EVENT = '1') = RK_COMPLIANT RK_MALE age startmedcnt
LDL_BASE NON_HDL_base TG_base TC_BASE rk_ethnicity statin dm htn bmibase
vascdis chf HXMI ANGINA;
RUN;

*PART B: TC25 = 1;
PROC LOGISTIC DATA=COPAY3;
CLASS       rk_ethnicity (PARAM=REF REF='WHITE')
            rk_copay (param=ref ref='0')
            statin (PARAM=REF REF='SIMVASTATIN')
            RK_MALE (PARAM=REF REF='0')
            dm (param=ref ref='0')
            htn (param=ref ref='0')
            vascdis (param=ref ref='0')
            chf (param=ref ref='0')
            COPD (param=ref ref='0')
            HXMI (param=ref ref='0')
            ANGINA(param=ref ref='0');
MODEL TC25 (EVENT = '1') = RK_COMPLIANT RK_MALE age startmedcnt
LDL_BASE NON_HDL_base TG_base TC_BASE rk_ethnicity statin dm htn bmibase
vascdis chf HXMI ANGINA;
RUN;

*PART C: NONHDL25 = 1;
PROC LOGISTIC DATA=COPAY3;
CLASS       rk_ethnicity (PARAM=REF REF='WHITE')
            rk_copay (param=ref ref='0')
            statin (PARAM=REF REF='SIMVASTATIN')
            RK_MALE (PARAM=REF REF='0')
            dm (param=ref ref='0')
            htn (param=ref ref='0')
            vascdis (param=ref ref='0')
            chf (param=ref ref='0')
            COPD (param=ref ref='0')
            HXMI (param=ref ref='0')
            ANGINA(param=ref ref='0');
MODEL NONHDL25 (EVENT = '1') = RK_COMPLIANT RK_MALE age startmedcnt
LDL_BASE NON_HDL_base TG_base TC_BASE rk_ethnicity statin dm htn bmibase
vascdis chf HXMI ANGINA;
RUN;


*************************************************************************;
```

```
*STEP 7: ANALYZING MISSING DATA PATTERNS;
*************************************************************************;
*DETERMINING THE TYPE OF MISSING VALUE PRESENT (IDEALLY, IT SHOULD BE 1);
OPTIONS NOFMTERR NOCENTER NODATE NOLABEL;
PROC FREQ DATA=COPAY2 NLEVELS;
TABLES RK_COMPLIANT RK_MALE RK_ETHNICITY STATIN age startmedcnt LDL_BASE
NON_HDL_base
TG_base TC_BASE LDL_FU NON_HDL_FU TC_FU dm htn bmibase vascdis chf HXMI
ANGINA / NOPRINT MISSING;
RUN;

*CALCULATING THE PERCENT MISSING FROM N=7739;
PROC MEANS DATA=COPAY2 NMISS N;
VAR RK_COMPLIANT RK_MALE age startmedcnt LDL_BASE NON_HDL_base
TG_base TC_BASE LDL_FU NON_HDL_FU TC_FU dm htn bmibase vascdis chf HXMI
ANGINA;
OUTPUT OUT=MISSINGPATTERN (DROP=_TYPE_ _FREQ_) NMISS= / AUTONAME;
RUN;
PROC TRANSPOSE DATA=MISSINGPATTERN PREFIX=NMISS OUT=MISSPATT_1;
VAR _NUMERIC_;
RUN;
DATA MISSPATT2;
SET MISSPATT_1;
PMISS=NMISS1/7739*100;
RUN;
PROC PRINT DATA=MISSPATT2;
RUN;


*************************************************************************;
*STEP 8: PERFORMING MULTIPLE IMPUTATION - PROC MI;
*************************************************************************;
*ASSUMPTION: NOT ASSUMING MONOTONE DATA PATTERN MISSINGNESS;
*APPLYING THE FCS OPTION FOR ARBITARY MISSING PATTERN;
PROC MI DATA=COPAY2 SEED=42037921
NIMPUTE=5 OUT=MIOUT2;
CLASS       STATIN RK_ETHNICITY LDL25 RK_MALE DM HTN VASCDIS CHF COPD HXMI
ANGINA;
FCS NBITER=5 DISCRIM (STATIN RK_ETHNICITY LDL25 RK_MALE DM HTN VASCDIS
CHF COPD HXMI ANGINA/DETAILS);
VAR   MPR RK_MALE angina chf copd dm hxmi vascdis htn LDL25 age
LDL_BASE tc_base NON_hdl_base tg_base bmibase RK_COMPLIANT STATIN
RK_ETHNICITY
NONHDL25 TC25;
RUN;


*************************************************************************;
*STEP 9: PERFORMING MULTIPLE IMPUTATION - PROC LOGISTIC WITH 5 IMPUTATIONS;
*************************************************************************;

*************************************************************************;
*PART A: LDL25;
*************************************************************************;
*PEFORMING LOGISTIC REGRESSION WITH _IMPUTATION_;
PROC LOGISTIC DATA=MIOUT2 DESCENDING;
BY _IMPUTATION_;
```

```
CLASS       rk_ethnicity (PARAM=REF REF='WHITE')
            LDL25 (param=ref ref='0')
            statin (PARAM=REF REF='SIMVASTATIN')
            RK_MALE (PARAM=REF REF='0')
            dm (param=ref ref='0')
            htn (param=ref ref='0')
            vascdis (param=ref ref='0')
            chf (param=ref ref='0')
            COPD (param=ref ref='0')
            HXMI (param=ref ref='0')
            ANGINA(param=ref ref='0');
MODEL LDL25 (EVENT = '1') = RK_COMPLIANT age RK_MALE startmedcnt LDL_BASE
NON_HDL_BASE TC_BASE TG_base rk_ethnicity statin dm htn bmibase vascdis chf
HXMI
ANGINA;
ODS OUTPUT PARAMETERESTIMATES=LOGPARMS;
RUN;
ODS OUTPUT PARAMETERESTIMATES = LOGOUT;

*************************************************************************;
*PART B: TC25;
*************************************************************************;
*PEFORMING LOGISTIC REGRESSION WITH _IMPUTATION_;
PROC LOGISTIC DATA=MIOUT2 DESCENDING;
BY _IMPUTATION_;
CLASS       rk_ethnicity (PARAM=REF REF='WHITE')
            TC25 (param=ref ref='0')
            statin (PARAM=REF REF='SIMVASTATIN')
            RK_MALE (PARAM=REF REF='0')
            dm (param=ref ref='0')
            htn (param=ref ref='0')
            vascdis (param=ref ref='0')
            chf (param=ref ref='0')
            COPD (param=ref ref='0')
            HXMI (param=ref ref='0')
            ANGINA(param=ref ref='0');
MODEL TC25 (EVENT = '1') = RK_COMPLIANT age RK_MALE startmedcnt LDL_BASE
NON_HDL_BASE TC_BASE TG_base rk_ethnicity statin dm htn bmibase vascdis chf
HXMI
ANGINA;
ODS OUTPUT PARAMETERESTIMATES=LOGPARMS2;
RUN;
ODS OUTPUT PARAMETERESTIMATES = LOGOUT2;

*************************************************************************;
*PART C: NONHDL25;
*************************************************************************;
*PEFORMING LOGISTIC REGRESSION WITH _IMPUTATION_;
PROC LOGISTIC DATA=MIOUT2 DESCENDING;
BY _IMPUTATION_;
CLASS       rk_ethnicity (PARAM=REF REF='WHITE')
            NONHDL25 (param=ref ref='0')
            statin (PARAM=REF REF='SIMVASTATIN')
            RK_MALE (PARAM=REF REF='0')
            dm (param=ref ref='0')
            htn (param=ref ref='0')
            vascdis (param=ref ref='0')
```

```
            chf (param=ref ref='0')
            COPD (param=ref ref='0')
            HXMI (param=ref ref='0')
            ANGINA(param=ref ref='0');
MODEL NONHDL25 (EVENT = '1') = RK_COMPLIANT age RK_MALE startmedcnt LDL_BASE
NON_HDL_BASE TC_BASE TG_base rk_ethnicity statin dm htn bmibase vascdis chf
HXMI
ANGINA;
ODS OUTPUT PARAMETERESTIMATES=LOGPARMS3;
RUN;
ODS OUTPUT PARAMETERESTIMATES = LOGOUT3;


*************************************************************************;
*STEP 10: PERFORMING MULTIPLE IMPUTATION - PROC MIANALYZE;
*************************************************************************;

*************************************************************************;
*PART A: LDL25;
*************************************************************************;
*SUMMARIZING THE IMPUTATION INTO ONE RESULT USING PROC MIANALYZE;
PROC MIANALYZE PARMS=LOGPARMS;
MODELEFFECTS INTERCEPT RK_COMPLIANT age RK_MALE startmedcnt LDL_BASE
NON_HDL_base TC_BASE TG_base rk_ethnicity statin dm htn bmibase vascdis chf
HXMI ANGINA;
RUN;

*************************************************************************;
*PART B: TC25;
*************************************************************************;
*SUMMARIZING THE IMPUTATION INTO ONE RESULT USING PROC MIANALYZE;
PROC MIANALYZE PARMS=LOGPARMS2;
MODELEFFECTS INTERCEPT RK_COMPLIANT age RK_MALE startmedcnt LDL_BASE
NON_HDL_base TC_BASE TG_base rk_ethnicity statin dm htn bmibae vascdis chf
HXMI ANGINA;
RUN;

*************************************************************************;
*PART C: NONHDL25;
*************************************************************************;
*SUMMARIZING THE IMPUTATION INTO ONE RESULT USING PROC MIANALYZE;
PROC MIANALYZE PARMS=LOGPARMS3;
MODELEFFECTS INTERCEPT RK_COMPLIANT age RK_MALE startmedcnt LDL_BASE
NON_HDL_base TC_BASE TG_base rk_ethnicity statin dm htn bmibase vascdis chf
HXMI ANGINA;
RUN;

*************************************************************************;
*STEP 11: CALCULATING ODDS RATIOS AND 95% CI;
*************************************************************************;

*************************************************************************;
*PART A: LDL25;
*************************************************************************;
DATA LOGOUT_A;
SET LOGOUT;
IF PARM ^="INTERCEPT";
```

```
ODDS_RATIO = EXP(ESTIMATE);
ODDS_LCLMEAN = EXP(LCLMEAN);
ODDS_UCLMEAN = EXP(UCLMEAN);
RUN;
PROC PRINT DATA=LOGOUT_A NOOBS;
VAR PARM ODDS:;
RUN;

*************************************************************************;
*PART B: TC25;
*************************************************************************;
DATA LOGOUT_B;
SET LOGOUT2;
IF PARM ^="INTERCEPT";
ODDS_RATIO = EXP(ESTIMATE);
ODDS_LCLMEAN = EXP(LCLMEAN);
ODDS_UCLMEAN = EXP(UCLMEAN);
RUN;
PROC PRINT DATA=LOGOUT_B NOOBS;
VAR PARM ODDS:;
RUN;

*************************************************************************;
*PART C: NONHDL25;
*************************************************************************;
DATA LOGOUT_C;
SET LOGOUT3;
IF PARM ^="INTERCEPT";
ODDS_RATIO = EXP(ESTIMATE);
ODDS_LCLMEAN = EXP(LCLMEAN);
ODDS_UCLMEAN = EXP(UCLMEAN);
RUN;
PROC PRINT DATA=LOGOUT_C NOOBS;
VAR PARM ODDS:;
RUN;


*************************************************************************;
*STEP 12: PERFORMING THE CRUDE ODDS RATIO AND CHI-SQUARED TEST;
*************************************************************************;
*DIFFERENCES IN THE CRUDE OR;

*PART A: LDL25;
PROC LOGISTIC DATA=COPAY2;
CLASS       LDL25 (PARAM=REF REF='0')
            RK_COMPLIANT (PARAM=REF REF='0');
MODEL LDL25 (EVENT='1') = RK_COMPLIANT;
RUN;
*CHI-SQUARE TEST;
PROC FREQ DATA=COPAY2;
TABLES LDL25*RK_COMPLIANT / NOROW NOCOL NOPERCENT CHISQ;
RUN;

*PART B: TC25;
PROC LOGISTIC DATA=COPAY2;
CLASS       TC25 (PARAM=REF REF='0')
            RK_COMPLIANT (PARAM=REF REF='0');
```

```
MODEL TC25 (EVENT='1') = RK_COMPLIANT;
RUN;
*CHI-SQUARE TEST;
PROC FREQ DATA=COPAY2;
TABLES TC25*RK_COMPLIANT / NOROW NOCOL NOPERCENT CHISQ;
RUN;

*PART C: NONHDL25;
PROC LOGISTIC DATA=COPAY2;
CLASS       NONHDL25 (PARAM=REF REF='0')
            RK_COMPLIANT (PARAM=REF REF='0');
MODEL NONHDL25 (EVENT='1') = RK_COMPLIANT;
RUN;
*CHI-SQUARE TEST;
PROC FREQ DATA=COPAY2;
TABLES NONHDL25*RK_COMPLIANT / NOROW NOCOL NOPERCENT CHISQ;
RUN;
*************************************************************************;
*END;
*************************************************************************;
```

Appendix 2. SAS codes for Case study 2.

```
*************************************************************;
*STUDENT: MARK BOUNTHAVONG - THESIS;
*TITLE: MISSING DATA ANALYSIS WITH MULTIVARIABLE MODELS CASE 2;
*THESIS CHAIR: DR. KEVIN SULLIVAN;
*THESIS COMMITTEE MEMBER: DR. JONATHAN H. WATANABE;
*DATE OF ANALYSIS: 19 SEPTEMBER 2013;
*************************************************************;

*************************************************************;
*STEP 1: IMPORT DATA;
*************************************************************;
*IMPORT Data_liraglutide_exenatide_07192013.CSV;

PROC CONTENTS DATA=DATA;
RUN;

*************************************************************;
*STEP 2: CREATING A DATASET FOR COMPLETE-CASE ANALYSIS;
*************************************************************;

DATA DATA2;
SET DATA;

IF INDEXDRUGNAME = 'EXENATIDE' THEN DRUGNAME = 1;
IF INDEXDRUGNAME = 'LIRAGLUTI' THEN DRUGNAME = 2;

IF GENDER = 'M' THEN FEMALE = 0;
IF GENDER = 'F' THEN FEMALE = 1;
IF RACE = 'BLACK, NON HISPANIC' THEN RACE2 = 2;
IF RACE = 'BLACK, NOT OF HISPANIC ORIGIN' THEN RACE2 = 2;
IF RACE = 'BLACK,NOT OF HISPANIC ORIGIN' THEN RACE2 = 2;
IF RACE = 'HISPANIC, WHITE' THEN RACE2 = 4;
IF RACE = 'HISPANIC,WHITE' THEN RACE2 = 4;
IF RACE = 'WHITE, NOT OF HISPANIC ORIGIN' THEN RACE2 = 1;
IF RACE = 'WHITE,NOT OF HISPANIC ORIGIN' THEN RACE2 = 1;
IF RACE = 'AMERICAN INDIAN OR ALASKA NATIVE' THEN RACE2 = 6;
IF RACE = 'ASIAN' THEN RACE2 = 3;
IF RACE = 'BLACK OR AFRICAN AMERICAN' THEN RACE2 = 2;
IF RACE = 'CAUCASIAN' THEN RACE2 = 1;
IF RACE = 'DECLINED TO ANSWER' THEN RACE2 = 8;
IF RACE = 'NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER' THEN RACE2 = 5;
IF RACE = 'NULL' THEN RACE2 = .;
IF RACE = 'UNKNOWN' THEN RACE2 = 7;
IF RACE = 'UNKNOWN BY PATIENT' THEN RACE2 = 7;
IF RACE = 'WHITE' THEN RACE2 = 1;
IF RACE = 'WHITE NOT OF HISP ORIG' THEN RACE2 = 1;

HBA1CDIFF = PST365HBA1C - PREHBA1C;

RUN;

PROC CONTENTS DATA=DATA2;
RUN;

PROC FORMAT;
```

```
*DISTRIBUTION IS NORMAL FOR EXENATIDE;
*DISTRIBUTION IS SKEWED FOR LIRAGLUTIDE;
PROC SGPLOT DATA=DATA4;
BY DRUGNAME;
HISTOGRAM PREHBA1C;
RUN;

*DISTRIBUTION IS NOT NORMAL FOR BOTH GROUPS;
PROC SGPLOT DATA=DATA4;
BY DRUGNAME;
HISTOGRAM PRECCI;
RUN;

*DISTRIBUTION IS NOT NORMAL FOR BOTH GROUPS;
PROC SGPLOT DATA=DATA4;
BY DRUGNAME;
HISTOGRAM PREBMI;
RUN;


      VALUE RACE_VALUE        1 = 'WHITE'
                                2 = 'BLACK'
                                3 = 'ASIAN'
                                4 = 'HISPANIC'
                                5 = 'PACIFIC ISLANDER'
                                6 = 'NATIVE AMERICAN'
                                7 = 'UNKNOWN'
                                8 = 'DECLINED TO ANSWER'
                                9 = .;
 RUN;
 DATA DATA3;
 SET DATA2;
 FORMAT RACE2      RACE_VALUE.;
 RUN;
************************************************************;
*STEP 2-a: GROUPING MISSING AND NON-MISSING DATA;
************************************************************;
*SET MISSING GROUPING VARIABLE BASED ON MISSING BASELINE HBA1C;
DATA DATA9;
SET DATA2;
      IF PREHBA1C =. THEN MISSING=1; ELSE MISSING=0;
RUN;

*GENERATING MEANS FOR AGE FOR MISSING/NON-MISSING GROUPS;
PROC MEANS DATA=DATA9 N MEAN STD;
CLASS MISSING;
VAR AGE;
RUN;

*GENERATING FREQ FOR GENDER AND DRUGNAME;
PROC FREQ DATA=DATA9;
TABLES (FEMALE DRUGNAME)*MISSING/CHISQ;
RUN;
```

```
 *SET MISSING GROUPING VARIABLE BASED ON MISSING RACE2;
 DATA DATA10;
 SET DATA2;
       IF RACE2 =. THEN MISSING=1; ELSE MISSING=0;
 RUN;

 *GENERATING MEANS FOR AGE FOR MISSING/NON-MISSING GROUPS;
 PROC MEANS DATA=DATA10 N MEAN STD;
 CLASS MISSING;
 VAR AGE;
 RUN;

 *GENERATING FREQ FOR GENDER AND DRUGNAME;
 PROC FREQ DATA=DATA10;
 TABLES (FEMALE DRUGNAME)*MISSING/CHISQ;
 RUN;

*FINAL COMPLETE-CASE ANALYSIS;
DATA DATA4;
SET DATA3;
IF PST365HBA1C ~=.;
IF PREHBA1C ~=.;
IF INDEXAGE ~=.;
IF PRECCI ~=.;
IF PREBMI ~=.;
IF RACE2 ~=.;
IF FEMALE ~=.;
RUN;

************************************************************;
*STEP 3: VISUAL INSPECTION OF CONTINUOUS DATA;
************************************************************;
ODS LISTING GPATH = "H:\LIRAGLUTIDE STUDY";
*DISTRIBUTION IS SKEWED TO THE RIGHT FOR EXENATIDE;
*DISTRIBUTION IS UNIFORMED FOR LIRAGLUTIDE;
PROC SGPLOT DATA=DATA4;
BY DRUGNAME;
HISTOGRAM PST365HBA1C;
RUN;
```

```
*DISTRIBUTION IS NORMAL FOR EXENATIDE;
*DISTRIBUTION IS NOT NORMAL FOR LIRAGLUTIDE;
PROC SGPLOT DATA=DATA4;
BY DRUGNAME;
HISTOGRAM INDEXAGE;
RUN;


************************************************************;
*STEP 4: DETERMINE THE NUMBER IN EACH GROUP;
************************************************************;
PROC FREQ DATA=DATA4;
TABLES DRUGNAME;
RUN;


************************************************************;
*STEP 5: BASELINE DEMOGRAPHICS FOR EACH GROUP;
************************************************************;

PROC MEANS DATA=DATA4 N MEAN STD;
CLASS DRUGNAME;
VAR INDEXAGE PREHBA1C PRECCI PREBMI;
RUN;

*UNIVARIATE ANALYSIS HBA1C;
PROC NPAR1WAY DATA=DATA4 WILCOXON;
CLASS INDEXDRUGNAME;
VAR PREHBA1C PREBMI PRECCI INDEXAGE;
RUN;

*DISCRETE ANALYSIS;
PROC FREQ DATA=DATA4;
TABLES (FEMALE RACE2 CHF_00 DEPRESS_00 DYSLIPID_00 HTN_00 MI_00
OBS_00)*DRUGNAME / CHISQ;
RUN;

************************************************************;
*STEP 6: ANALYZING MISSING DATA PATTERNS;
************************************************************;
*LOOKING AT MISSING DATA PATTERNS;
PROC FREQ DATA=DATA;
TABLES INDEXDRUGNAME;
RUN;

OPTIONS NOFMTERR NOCENTER NODATE NOLABEL;
PROC FREQ DATA=DATA2 NLEVELS;
TABLES _ALL_ / NOPRINT MISSING;
RUN;

*PROPORTI0N OF MISSING VALUES;
*LOOK AT THE TOTAL (BOTH GROUPS: EXENATIDE + LIRAGLUTIDE);
PROC MEANS DATA = DATA2 NMISS N;
VAR PST365HBA1C DRUGNAME RACE2 indexage prebmi precci PREHBA1C mi_00 chf_00
htn_00 obs_00 dyslipid_00 FEMALE;
OUTPUT OUT=T (DROP = _TYPE_ _FREQ__) NMISS = /AUTONAME;
```

```
RUN;
PROC TRANSPOSE DATA = T PREFIX = NMISS OUT=S1;
VAR _NUMERIC_;
RUN;
DATA S2;
SET S1;
PMISS1 = NMISS1/1094*100;
RUN;
PROC PRINT DATA=S2;
RUN;
*BROKEN BY GROUPS (EXENATIDE VERSUS LIRAGLUTIDE);
PROC MEANS DATA = DATA2 NMISS N;
BY DRUGNAME;
VAR PST365HBA1C DRUGNAME RACE2 indexage prebmi precci PREHBA1C mi_00 chf_00
htn_00 obs_00 dyslipid_00 FEMALE;
OUTPUT OUT=T (DROP = _TYPE_ _FREQ_) NMISS = /AUTONAME;
RUN;
PROC TRANSPOSE DATA = T PREFIX = NMISS OUT=S1;
VAR _NUMERIC_;
RUN;
DATA S2;
SET S1;
PMISS1 = NMISS1/1054*100;
PMISS2 = NMISS2/40*100;
RUN;
PROC PRINT DATA=S2;
RUN;
```

```
************************************************************;
*STEP 7: UNIVARIATE ANALYSIS;
************************************************************;
PROC MEANS DATA=DATA4 N MEAN STD;
CLASS DRUGNAME;
VAR HBA1CDIFF PREHBA1C PST365HBA1C;
RUN;
PROC SGPLOT DATA=DATA4;
BY DRUGNAME;
VBAR HBA1CDIFF;
RUN;
*USE NON-PARAMETERIC - WILCOXON;
PROC NPAR1WAY DATA=DATA4 WILCOXON;
CLASS DRUGNAME;
VAR HBA1CDIFF;
RUN;



************************************************************;
*STEP 8: PERFORMING THE CRUDE ANALYSIS;
************************************************************;
PROC GLM DATA=DATA2;
CLASS DRUGNAME;
MODEL HBA1CDIFF = DRUGNAME / SOLUTION CLPARM;
RUN;

************************************************************;
*STEP 9: PERFORMING THE COMPLETE-CASE ANALYSIS - MULTIVARIATE MODEL;
************************************************************;
PROC GLM DATA=DATA3;
CLASS       DRUGNAME
            FEMALE
            RACE2
            CHF_00
            HTN_00
            OBS_00
            DYSLIPID_00;
MODEL HBA1CDIFF = DRUGNAME RACE2 indexage prebmi precci PREHBA1C mi_00 chf_00
htn_00 obs_00 dyslipid_00 FEMALE / SOLUTION CLPARM;
RUN;

    ************************************************************;
    *STEP 10: PERFORMING MULTIPLE IMPUTATION;
    ************************************************************;
    *MULTIPLE IMPUTATION START;
    PROC MI DATA=DATA2 seed=42037921
    nimpute=5 out=miout2;
    mcmc timeplot(mean(pst365hba1c)) acfplot(mean(pst365hba1c));
    var pst365hba1c DRUGNAME RACE2 indexage prebmi precci PREHBA1C mi_00 chf_00
    htn_00 obs_00 dyslipid_00 FEMALE;
    run;

    PROC GLM DATA=MIOUT2;
    BY _IMPUTATION_;
    MODEL HBA1CDIFF = DRUGNAME RACE2 indexage prebmi precci PREHBA1C mi_00 chf_00
    htn_00 obs_00 dyslipid_00 FEMALE / INVERSE;
    ODS OUTPUT  PARAMETERESTIMATES=GLMPARMS;
    RUN;

    PROC MIANALYZE PARMS=GLMPARMS EDF=195;
    MODELEFFECTS INTERCEPT DRUGNAME RACE2 indexage prebmi precci PREHBA1C mi_00
    chf_00
    htn_00 obs_00 dyslipid_00 FEMALE;
    RUN;
```