



University of
Massachusetts
Amherst

A Comparison of Techniques for Handling Missing Data in Longitudinal Studies

| | |
|---------------|---|
| Item Type | Thesis (Open Access) |
| Authors | Bogdan, Alexander R |
| DOI | 10.7275/9053361 |
| Download date | 2026-03-14 14:05:18 |
| Link to Item | https://hdl.handle.net/20.500.14394/33420 |

A Comparison of Techniques for Handling Missing Data in Longitudinal Studies

A Thesis Presented

by

ALEXANDER R. BOGDAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

September 2016

Biostatistics

A Comparison of Techniques for Handling Missing Data in Longitudinal Studies

A Thesis Presented

by

ALEXANDER R. BOGDAN

Approved as to style and content by:

Kenneth P. Kleinman, Chair

Brian W. Whitcomb, Member

Nicholas G. Reich, Member

Susan E. Hankinson, Department Head
Department of Biostatistics & Epidemiology

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Kenneth P. Kleinman, for his continuous support throughout the conception and production of this project. I would like to thank Brian W. Whitcomb for his expertise on the epidemiological aspects of this project and for his efforts in obtaining the data necessary to make this project a reality. Thank you both for always being available, both as mentors on this project and as professors in the classroom. Special thanks to Nicholas G. Reich, not only for his thoughtful feedback and criticisms on this project, but also for his support and mentorship over the last two years. From my undergraduate to my graduate education, you have been an invaluable role model.

I would like to extend my sincerest gratitude to Enrique F. Schisterman and the entire research team behind the BioCycle Study. This project would not have been possible without their commitment to public health research and their willingness to promote academic accomplishment by sharing the data from their study.

My deepest appreciation goes to all of the friends, family and colleagues who have supported me throughout this project. From brainstorming and proofreading, to dinner parties and general merriment, you were all essential to making this thesis a reality.

A final word of thanks to Gregory J. Matthews, the man who rekindled my love of programming and sparked my interest in Biostatistics. A professor, a mentor, and above all else, a friend.

ABSTRACT

A COMPARISON OF TECHNIQUES FOR HANDLING MISSING DATA IN LONGITUDINAL STUDIES

SEPTEMBER 2016

ALEXANDER R. BOGDAN, B.S., UNIVERSITY OF MASSACHUSETTS AMHERST

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Kenneth P. Kleinman

Missing data are a common problem in virtually all epidemiological research, especially when conducting longitudinal studies. In these settings, clinicians may collect biological samples to analyze changes in biomarkers, which often do not conform to parametric distributions and may be censored due to limits of detection. Using complete data from the BioCycle Study (2005-2007), which followed 259 premenopausal women over two menstrual cycles, we compared four techniques for handling missing biomarker data with non-Normal distributions. We imposed increasing degrees of missing data on two non-Normally distributed biomarkers under conditions of missing completely at random, missing at random, and missing not at random. Generalized estimating equations were used to obtain estimates from complete case analysis, multiple imputation using joint modeling, multiple imputation using chained equations, and multiple imputation using chained equations and predictive mean matching on Day 2, Day 13 and Day 14 of a standardized 28-day menstrual cycle. Estimates were compared against those obtained from analysis of the completely observed biomarker data. All techniques performed comparably when applied to a Normally distributed biomarker. Multiple imputation using joint modeling and multiple imputation using chained equations produced similar estimates across all types and degrees of missingness for each biomarker. Multiple imputation using chained equations and predictive mean matching consistently deviated from both the complete data estimates and the

other missing data techniques when applied to a biomarker with a bimodal distribution. When addressing missing biomarker data in longitudinal studies, special attention should be given to the underlying distribution of the missing variable. As biomarkers become increasingly Normal, the amount of missing data tolerable while still obtaining accurate estimates may also increase when data are missing at random. Future studies are necessary to assess these techniques under more elaborate missingness mechanisms and to explore interactions between biomarkers for improved imputation models.

TABLE OF CONTENTS

| | Page |
|---|------|
| ACKNOWLEDGEMENTS | iii |
| ABSTRACT | iv |
| LIST OF TABLES | ix |
| LIST OF FIGURES | x |
| CHAPTER | |
| 1. INTRODUCTION | 1 |
| A. Missing Data: The Problem | 1 |
| B. Types of Missing Data | 2 |
| 1. Missing Completely at Random (MCAR) | 3 |
| 2. Missing at Random (MAR) | 3 |
| 3. Missing Not at Random (MNAR) | 4 |
| C. Patterns of Missingness | 4 |
| D. Degrees of Missingness | 5 |
| E. Plan of the Thesis | 6 |
| 2. TECHNIQUES FOR HANDLING MISSING DATA | 9 |
| A. General Techniques | 9 |
| B. Complete Case Analysis | 10 |
| C. Multiple Imputation | 10 |

| | |
|--|----|
| D. Multiple Imputation using Joint Modeling (MI-JM)..... | 13 |
| E. Multiple Imputation using Chained Equations (MICE) | 14 |
| F. Multiple Imputation using Chained Equations and Predictive Mean Matching (MICE-PMM) | 15 |
| 3. A MOTIVATING EXAMPLE: THE BIOCYCLE STUDY | 17 |
| A. Study Population & Design | 17 |
| B. Study Data | 19 |
| 1. Exposure Variables | 19 |
| 2. Outcome Variable | 20 |
| 3. Confounding Variables | 20 |
| 4. MISSINGNESS MECHANISM & STATISTICAL ANALYSIS | 22 |
| A. Missingness Mechanism | 22 |
| B. Application of Missing Data Techniques | 23 |
| C. Statistical Analysis | 24 |
| 5. RESULTS | 27 |
| A. Complete Data Estimates & Biomarker Distributions | 27 |
| B. Missing Completely at Random | 29 |
| C. Missing at Random | 32 |
| D. Missing Not at Random | 35 |
| 6. DISCUSSION | 39 |

BIBLIOGRAPHY44

LIST OF TABLES

| Table | Page |
|--|------|
| 1. Selected Participant Characteristics, BioCycle Study (2005-2007) | 18 |
| 2. Generalized Estimating Equation Coefficients for Complete Data, Biocycle Study (2005-2007) | 27 |

LIST OF FIGURES

| Figure | Page |
|--|------|
| 1. Histograms and kernel density plots: SIM, VEGF, HGF | 28 |
| 2. Histograms and kernel density plots by day: HGF..... | 29 |
| 3. Change in coefficient estimates as a function of missingness: HGF, MCAR..... | 30 |
| 4. Change in coefficient estimates as a function of missingness: VEGF, MCAR | 31 |
| 5. Change in coefficient estimates as a function of missingness: SIM, MCAR | 32 |
| 6. Change in coefficient estimates as a function of missingness: HGF, MAR | 33 |
| 7. Change in coefficient estimates as a function of missingness: VEGF, MAR | 34 |
| 8. Change in coefficient estimates as a function of missingness: SIM, MAR | 35 |
| 9. Change in coefficient estimates as a function of missingness: HGF, MNAR | 36 |
| 10. Change in coefficient estimates as a function of missingness: VEGF, MNAR | 37 |
| 11. Change in coefficient estimates as a function of missingness: SIM, MNAR | 38 |

CHAPTER 1

INTRODUCTION

A. Missing Data: The Problem

Missing data is a nearly universal aspect of data collection and analysis. Whether surveying patients about dietary preferences or collecting information on internet browsing patterns, missing data is an issue that is not limited to any subset of fields or disciplines – if data are involved, missingness must be accounted for. Even the most meticulously designed studies and protocols, infused with considerable time, money, and resources, can fall victim to this complication. Improperly addressing missingness in data analysis can lead to biased estimation and invalid inference, obscuring the focal point of any analytical endeavor.

In 1976, Donald B. Rubin published a seminal work that provided the framework used today by statisticians worldwide when tackling missing data. In it, Rubin (1976) outlined three different types of missing data and the unique assumptions that must be made with each; he also later devised a set of rules for combining the estimates obtained from analysis of imputed data sets. Dubbed “Rubin’s Rules”, these guidelines offer a means to obtain unbiased estimators and make valid inferences when using imputed data sets if certain conditions have been satisfied (Little & Rubin, 2002) . However, despite years of research and consideration devoted to this topic, ideas regarding the most robust techniques to use, and when to use them, remain varied.

Longitudinal studies are frequently affected by missing data, either by chance or by design. Such research involves repeatedly obtaining data from subjects over a given observation period and conducting analyses to utilize the temporal nature of the information collected. This type of study is preferred in the biomedical sciences when investigators are interested in evaluating changes in biomarkers over a period of time. However, it is easy for patients to miss

clinic visits and it is common for some biomarkers to register below the limits of detection when analyzed. This high potential for missing data in longitudinal studies is of serious concern and must be handled appropriately during data analysis.

A myriad of statistical methods exist that can preserve the correlated nature of missing repeated data in the context of a longitudinal study (Engels & Diehr, 2003; Ferro, 2014; Linero & Daniels, 2015; Luo, Lawson, He, Elm, & Tilley, 2016; Twisk & de Vente, 2002), making it challenging to offer sound justification for selecting a particular technique. Utilizing a longitudinal data set with complete data, we propose to compare the performance of four commonly used techniques for handling missingness. Additionally, we propose to explore the potential impact made by the distributions of variables with missing data on obtained estimates. Certain predictors (such as biomarkers and environmental exposures) often do not adhere to commonly used parametric distributions and frameworks, which may complicate analyses and affect results. We will conduct analyses on two biomarkers with differing distributions and contrast results to examine the possible role of this factor.

B. Types of Missing Data

As part of their framework devised to address the handling of missing data, Little & Rubin (2002) outlined three distinct mechanisms by which data could be considered missing. Each possesses its own key assumptions that must be considered before attempting any statistical modeling. For the present study, we will be applying these different types of missing data to predictor variables.

1. Missing Completely at Random (MCAR)

Data are said to be missing completely at random (MCAR) if the probability of an observation being missing is independent of any observed or unobserved data. Mathematically, this can be denoted as:

$$P(R|Y, X) = P(R|Y, X^{obs}, X^{miss}) = P(R|\phi)$$

where \mathbf{R} represents the missingness indicator (such that $R=1$ when X is missing and $R=0$ when X is observed), \mathbf{Y} is a vector for the response variable, \mathbf{X}^{obs} is a vector of all predictors with observed values, \mathbf{X}^{miss} is a vector of all predictors with missing values, and ϕ is a vector of model parameters. An example of data that are MCAR would be a subject deciding whether or not to complete a questionnaire based on the result of a fair, random coin flip. The assumption that missingness is unrelated to any known or unknown factor in a study is quite strong, and in practice it is often untenable. However, if this assumption can be reasonably upheld, statistical analyses can proceed using only the completely observed data, as the information available still represents a random sample of the target population.

2. Missing at Random (MAR)

In comparison to MCAR, data can be defined as missing at random (MAR) if the probability of an observation being missing is dependent only on the observed data. This can be represented mathematically as:

$$P(R|Y, X) = P(R|Y, X^{obs}, X^{miss}) = P(R|Y, X^{obs}, \phi)$$

An example of data that are MAR could be a study testing the efficiency of a novel fitness device. If body mass index (BMI) is recorded for all participants, and subjects with higher BMI are more

likely to miss clinic visits or dropout of the study, then the data can be considered MAR since missingness is a function of the data that has been observed in the given study.

While often more reasonable than the MCAR assumption, MAR may be implausible if data are ascertained on only a small number of covariates. Inclusion of information on a diverse array of predictors in imputation models may allay this issue and make MAR a more realistic assumption (Collins, Schafer, & Kam, 2001). If held, this assumption allows for unbiased estimation when appropriate techniques for handling the missing data have been applied.

3. Missing Not at Random (MNAR)

Data are considered missing not at random (MNAR) if the probability of an observation being missing is dependent on unobserved data. This can be expressed as:

$$P(R|Y, X) = P(R|Y, X^{obs}, X^{miss}) = P(R|Y, X^{miss}, \phi)$$

Imagine if, after accounting for all observed data, younger patients were systematically more likely to miss early morning clinic visits during a drug trial than older patients. It may be that younger patients sleep longer than older patients. The resulting missingness would be classified as MNAR since the predictor causing the missing data was unobserved. It is impossible to rule out MNAR when dealing with missing data (Little & Rubin, 2002), especially if the number of covariates included in the analysis is limited. Estimates based on data that are MNAR may be biased even if suitable methods for handling missingness are used.

C. Patterns of Missingness

In addition to the three unique types of missing data, there are two distinct patterns of missingness that can occur: monotone and non-monotone. In longitudinal settings, a monotone

pattern of missingness occurs when, once unobserved, a subject is not observed again for the duration of the observation period. Said another way, whenever y_{ij} is missing, so too is y_{ik} for all $k \geq j$. In epidemiological studies, the most common sources of monotone missingness are dropout and death. Likelihood-based analysis approaches are often easier to apply when missingness is monotone (Kenward & Carpenter, 2007), and data can sometimes be reorganized to achieve a monotone pattern, however this will not be discussed.

More commonly, missing data assume an often arbitrary pattern of non-monotone missingness. This occurs when a subject is unobserved at a given time point, but is observed again at one or several later time points during the observation period; if a subject is missing at time point y_{ij} , some values for y_{ik} may be observed when $k \geq j$. This pattern arises frequently from missed clinic visits, forgetting to answer a question, and other potentially repeatable scenarios. Unlike monotone missingness, there is often no simple factorization applicable in non-monotone situations, making likelihood-based approaches computationally challenging. Markov Chain Monte Carlo (MCMC) can be used to impute sufficient information such that the remaining pattern of missing data is monotone and likelihood-based methods can then be applied.

D. Degree of Missingness

Once type and pattern of missingness have been accounted for, a prudent statistician must also consider how much of the data that they are handling is missing. The degree of missingness is generally quantified in one of two ways: the proportion of observations with missing values, or the proportion of missing values for a given predictor. The former metric offers a succinct measure of missingness when using complete case analysis, since any subject without complete data will be excluded. In most other cases and for the purposes of this paper, missingness will be denoted by the percentage of missing values for a select variable such that

$$\%x^{miss} = \frac{n_{x_i}^{miss}}{n_{x_i}^{total}} * 100$$

While there is technically no degree of missingness that would prohibit an analysis from being conducted, it is important to understand that as the proportion of missing data increases, the effect it may have on estimates increases simultaneously (Schafer & Olsen, 1998). When data are MCAR, complete case estimates should be unbiased without regard to the amount of missingness in the data set, however loss of power must still be considered. When data are MAR or MNAR, bias in complete case estimates can increase as a function of percent missingness. As will be discussed later, certain techniques for handling missing data draw insight on the unobserved data by examining the observed data; as the proportion of observed data decreases, estimates of the unobserved data become more prone to uncertainty and bias (Little & Rubin, 2002). It is ultimately the decision of the investigator to determine when the degree of missingness makes analysis using a select predictor unviable.

In contrast to predictors, debate exists surrounding how to properly handle missing response variables, often without regard to the percentage of missing data. Some researchers suggest exclusion of these individuals from analysis entirely, even in instances where values are imputed for unobserved outcomes, since their inclusion only contributes noise to obtained estimates (Little, 1992; Von Hippel, 2007). In the present study, all outcome data are observed.

E. Plan of the Thesis

I propose to compare the efficiency of four techniques for handling missing data arising from non-Normal predictors in the context of a longitudinal study. I will conduct an analysis of the relationship between select non-Normal biomarker exposures and anovulation over two menstrual cycle periods among a sample of adult, premenopausal women. Additionally, a

simulated “ideal” cytokine which follows a distinctly Normal distribution will be analyzed for comparative purposes. The data for this analysis are completely observed and will serve as the foundation for the main objective of this study.

Four techniques will be assessed in this comparative evaluation: complete case analysis, multiple imputation using chained equations and assuming normal distributions, multiple imputation using chained equations with predictive mean modeling, and multiple imputation using joint modeling. Each method will be applied to a non-monotone pattern of missing data under imposed types (missing completely at random [MCAR], missing at random [MAR] and missing not at random [MNAR]) and degrees of missingness (ranging from 5% - 50% incrementally by 5% intervals). The data from each technique will be analyzed using generalized linear models and generalized estimating equations to obtain estimates for performance comparison.

To evaluate each of the conditions specified above, 500 replications will be performed for each technique for each given set of conditions (e.g., 500 replications of multiple imputation using chained equations with 5% MAR data). Relative differences in obtained estimates between techniques will be used to evaluate performance. All obtained results will be compared against an analysis containing completely observed data from the BioCycle study (described in Chapter 3).

Chapter 2 will provide a detailed explanation of the techniques to be compared, offering a brief introduction to their respective mathematical foundations and highlighting key advantages and disadvantages of each unique approach. Chapter 3 will introduce the BioCycle study data to be used as the motivating example for this study. Chapter 4 will draw attention to the mechanisms used for imposing the various types and degrees of missingness, as well as the statistical procedures used for the subsequent analysis. Chapter 5 will present the results of these analyses and provide figures for visual comparison of the techniques assessed. Chapter 6 will offer a

discussion of the findings, rationale for the relative superiority/inferiority of certain methods over others, and an examination of the strengths and limitations of the present study.

CHAPTER 2

TECHNIQUES FOR HANDLING MISSING DATA

A. General Techniques

Despite the significant risks posed by missing data to accurate analyses, many researchers regard missing data as a nuisance rather than a problem meriting a thoughtful response. As such, there exist several common techniques for dealing with missed observations that have gained widespread traction owed to their simplicity and ease of implementation. These include mean imputation, last observation/value carried forward (LOCF/LVCF), and exclusion of variables with a high degree of missingness from analysis.

Mean imputation is a practice in which the mean value of the observed data for a given variable is imputed to all unobserved data for that variable. This arbitrary assignment of values has the potential to both induce bias in subsequent estimates and to understate the true variability of the data (Greenland & Finkle, 1995). LOCF/LVCF is often seen in longitudinal studies and involves imputation of the last observed value to all subsequent missing values for a selected predictor. As described by Cook, Zeng, & Yi (2004), this approach may also introduce bias and underestimate the true variability of the variable in question. Exclusion of predictors with a high degree of missingness is perhaps the most harmful of the techniques listed; this method can induce bias, artificially inflate standard errors, and unnecessarily discard informative variables.

Given the hazards associated with each of these techniques, none are recommended for addressing the problem of missing data. Thankfully, advances in computational hardware and statistical software have popularized more sophisticated methods that can yield asymptotically unbiased estimators and standard errors. These alternative techniques have become increasingly easy to implement and are often able to efficiently handle different types of data. Such techniques

will be the focus of this study, however first we will discuss what is perhaps the most common method of accounting for missingness: complete case analysis.

B. Complete Case Analysis

Complete case analysis is a simple and sometimes unbiased approach for handling missing data. This technique owes its popularity to its ease of implementation and its inexpensive nature in terms of time and computational intensity. Using complete case analysis, any observation containing missing data for any covariate is excluded from statistical analysis, meaning that estimates are based entirely on the observed data. When data are MCAR, this can yield unbiased estimates, however even if this assumption can be met, the exclusion of observations results in diminished power (Little & Rubin, 2002). When data are not MCAR, this method produces biased estimates in addition to decreasing power. With more sophisticated techniques for handling missing data readily available in modern software, it is unadvisable to perform complete case analysis; however, given the extensive prevalence of this approach in published literature (Karahalios et al., 2013), it will be included in our comparison of techniques.

C. Multiple Imputation

Multiple imputation is a well-known statistical approach for handling missing data with several common variations. The core idea behind this technique is to use the distribution of the observed data to obtain estimates of plausible values for the missing data. Developed by Rubin (2004), the process for utilizing this technique can be summarized in three distinct steps:

Step 1: Random draws from the posterior predictive distribution are used to accurately reflect uncertainty in the parameters (described below), and m completed data sets are constructed.

Step 2: Each imputed data set is analyzed individually to produce a series of parameter estimates.

Step 3: Following Rubin's rules, (described below) these estimates are combined to yield estimates and standard errors that are asymptotically unbiased and efficient.

For all techniques using multiple imputation herein, $m = 20$ imputed data sets and 20 burn-in iterations are used per imputation. Below, the first step of the multiple imputation process is broken down.

To begin, all predictors with missingness are identified. Proper imputation proceeds by replacing all missing values with random draws from the posterior predictive distribution of the missing predictor conditional on the observed data. Let us assume that $\mathbf{x} = (x_1, \dots, x_p)'$ is a vector containing the intercept and all predictors with complete data. For a continuous predictor with missing values, z , an imputation model is specified where observations with observed values of z are linearly regressed conditionally on \mathbf{x} such that

$$z|\mathbf{x}; \boldsymbol{\beta} \sim N(\boldsymbol{\beta}\mathbf{x}, \sigma^2)$$

From this model, let $\hat{\boldsymbol{\beta}}$ be a row vector of length p containing the intercept and estimated parameters from all subjects with observed z . Let $\boldsymbol{\Sigma}$ be the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, and let $\hat{\sigma}$ represent the estimated root mean squared error. Using the joint posterior distribution of σ , $\boldsymbol{\beta}$ as outlined by Rubin (2004), the imputation parameters σ^* and $\boldsymbol{\beta}^*$ can be obtained as follows:

$$\sigma^* = \hat{\sigma} \sqrt{\frac{(n_{obs,z} - p)}{q}}$$

where $n_{obs,z}$ is the number of subjects with observed z values, p is the number of covariates with complete data, and q is a random draw from a χ^2 distribution with $n_{obs,z} - p$ degrees of freedom, and

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}} \mathbf{u}_1 \boldsymbol{\Sigma}^{1/2}$$

where \mathbf{u}_l is a row vector of p random draws from the standard Normal distribution and $\Sigma^{1/2}$ represents the Cholesky decomposition of Σ . Using these approaches, σ^* approximates the sample standard deviation of the “complete” data while incorporating a degree of uncertainty based on the sample size and the number of predictors. Similarly, β^* approximates the estimated parameters of the complete data, accounting for differences in the sample standard deviation between the observed and “complete” data, and the covariances of the observed data with randomness added by \mathbf{u}_l . The imputation parameters can then be used to acquire imputed values for each subject with missing z from the posterior predictive distribution such that

$$z_i^* = \beta^* x_i + u_{2i} \sigma^*$$

where u_{2i} represents a random draw from the standard Normal distribution.

This process is repeated for each predictor with missing data, until all missing values have been imputed. This constitutes a single imputed dataset, and the first step is repeated for each predictor with missingness m times to produce the desired number of imputed data sets. It is worth noting that while we have outlined the imputation model for a continuous missing variable, parametric models for other data types including binary variables and both ordered and unordered categorical variables can be used as well. These distinctions allow a parametric probability distribution to accompany each imputation model so that appropriate assumptions may be upheld. However, problems may arise when handling data with unique features such as bounds or when imputed values are rounded (Horton, Lipsitz, & Parzen, 2003); for example, when imputing age, negative values and implausibly large values may be produced.

There are a number of statistics that can be combined using Rubin’s rules (Little & Rubin, 2002), however we will focus specifically on coefficient and variance estimates. The overall coefficient estimate $\hat{\theta}$ represents the average of each imputation-specific estimate such that

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^m \hat{\theta}_i$$

where m is the number of imputed data sets generated. Similarly, the overall variance estimate $var(\hat{\theta})$ reflects the within-imputation variance \mathbf{W} and the between-imputation variance \mathbf{B} such that

$$var(\hat{\theta}) = \mathbf{W} + \left(1 + \frac{1}{m}\right) \mathbf{B}$$

$$\mathbf{W} = \frac{1}{m} \sum_{i=1}^m \mathbf{W}_i \quad \mathbf{B} = \frac{1}{m-1} \sum_{i=1}^m (\hat{\theta}_i - \hat{\theta})^2$$

, where \mathbf{W}_i is the variance of $\hat{\theta}_i$. Using these rules, valid inferences can be made from the estimates obtained using multiple imputation.

D. Multiple Imputation using Joint Modeling (MI-JM)

Joint modeling, developed extensively by Schafer (1997), is one of two main approaches for constructing imputation models when multiply imputing data. To begin, observations are partitioned to create a group for each unique pattern of missingness. A joint model is then constructed which is shared by all observations in each group such that

$$P(Y, X, R | \theta)$$

where θ contains a set of model parameters from a prior distribution; common distributions include multivariate normal and log-linear. From this prior distribution, sub-models for each pattern of missing data are created. Mathematical integrations from all of the group sub-models can quickly become analytically and computationally intensive. To address this problem, modern adaptations of methods such as expectation-maximization (EM) algorithms (Dempster, Laird, & Rubin, 1977) and Bayesian methods based on Markov Chain Monte Carlo (McCulloch, 1997) are

commonly used to obtain estimates when using joint modeling. In the present study, EM algorithms were used to compute maximum likelihood estimates for the means and covariance matrices under the multivariate Normal distribution.

A notable disadvantage that arises from selecting a singular parametric multivariate density is the inability of joint modeling to handle different data types (Lee & Carlin, 2010). For example, if a multivariate Normal distribution is selected as the prior distribution, any variables which contain missing values and do not adhere to the Normal distribution will be mis-specified. It is quite common for variables with missing data to represent a host of different data types, and so upholding the assumptions of a distribution such as the multivariate Normal may be unfeasible and obtained estimates may be biased. Despite this shortcoming, Schafer (1997) has indicated that inferences may still be credible even when these assumptions are violated. Ultimately, joint modeling is a valid technique for imputation based firmly in statistical theory.

E. Multiple Imputation using Chained Equations (MICE)

Multiple imputation using chained equations (MICE) follows the same three step format of multiple imputation using joint modeling, however there is a key difference which occurs in the first step. Chained equations, also referred to as fully conditional specification or regression switching, generates values by utilizing a set of distinct, imputation models which are often univariate (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). These models are iteratively processed to distance the imputed values from the arbitrary random values initially drawn using what are called “burn-in” iterations or simply “cycles”. Each cycle uses the imputed values from the previous cycle to impute the missing variables in the current cycle, incrementally improving the imputed values obtained. After a specified number of cycles, the imputed values are retained and this constitutes one imputed data set. This repeated processing of unique

imputation models is in contrast to the single processing of the multivariate distribution that governs all imputation models in joint modeling. Let us assume that $\mathbf{x} = (x_1, \dots, x_k)$, where each x_i may or may not contain missing values. For each variable with missing values x_i^{miss} , a unique imputation model with z predictors is specified such that the conditional density

$$P(x_i^{miss} | z, x_i^{obs}, r)$$

can be ascertained. Regression of x_i^{miss} is completed using all individuals with observed values x_i^{obs} . In this way, chained equations model a joint conditional density for each predictor with missing values following the appropriate distribution of the predictor. As described by van Buuren (2007), this method also allows tailoring of imputation models to incorporate unique features of the data that would otherwise be discarded, such as interactions and bounded values.

MICE is advantageous when the variables to be imputed span different data types, as a unique imputation model is specified for each predictor. However, creating each imputation model can quickly become a time intensive process in the case of large data sets with many variables containing missing values. Furthermore, chained equations lack the same formal foundation in statistical theory as joint modeling, however published literature suggests that this approach produces unbiased estimates (Horton & Lipsitz, 2001; Raghunathan, Lepkowski, Van Hoewyk, & Peter Solenberger, 2001).

F. Multiple Imputation using Chained Equations and Predictive Mean Matching

(MICE-PMM)

Similar to MICE, multiple imputation using chained equations and predictive mean matching (MICE-PMM) is asymptotically efficient and yields unbiased estimates and standard errors. This method closely follows the outline of MICE and hosts the same advantages and

disadvantages, with one key difference in the first step of the multiple imputation process. As described below, once the imputation parameters have been determined, predictive mean matching uses this information to produce imputed values by an alternative approach.

Predictive mean matching (PMM) is an *ad hoc* method for imputation of a predictor with missing values z , such that imputed values are obtained only from the observed values of non-missing z (Little & Rubin, 2002). This can prove to be advantageous if z and \mathbf{x} share a non-linear relationship or, in the case of z being a continuous variable, if the assumption of Normality cannot reasonably be met. Conversely, if imputation of z requires extrapolation outside the range of observed values for the variable, this method may be disadvantageous.

To obtain PMM imputed values, the same process as described previously to obtain the imputation parameter $\boldsymbol{\beta}^*$ is used. However, instead of using this value to draw from a Normal distribution with mean $\boldsymbol{\beta}^* \mathbf{x}_i$, k subjects with the smallest values of

$$|\widehat{\boldsymbol{\beta}} \mathbf{x}_j - \boldsymbol{\beta}^* \mathbf{x}_i|, (j=1, \dots, n_{obs})$$

are distinguished. One subject from the k closest individuals, say k' , is randomly selected and the imputed value for z_i becomes $z_{k'}$. The number of closest subjects to be considered for sampling observed values is set by the researcher; in the present study, $k = 5$ for all techniques utilizing predictive mean matching.

CHAPTER 3

A MOTIVATING EXAMPLE: THE BIOCYCLE STUDY

A. Study Population & Design

The BioCycle Study was a prospective cohort study of 259 healthy premenopausal women conducted from 2005-2007 in Western New York state (Wactawski-Wende et al., 2009). The primary purpose of this study was to examine the relationship of endogenous reproductive hormones with oxidative stress levels and antioxidants longitudinally during the menstrual cycle. Enrolled participants completed questionnaires and submitted urine samples at baseline, and were followed over one (n=9) or two (n=250) menstrual cycles. Baseline surveys included information on a number of anthropometric, demographic, medical and lifestyle factors; eligibility criteria are discussed elsewhere. To coordinate eight cycle visits per menstrual cycle, fertility monitors continuously assessing endogenous reproductive hormone levels and an algorithm controlling for individual cycle length were utilized. The same method was repeated over two menstrual cycles for a total of 16 clinic visits per subject. At each clinic visit, blood and urine samples were collected from each subject for biomarker analysis; the eight visits per cycle corresponded approximately to days 2, 7, 12, 13, 14, 18, 21 & 27 of a standardized 28-day menstrual cycle.

All subjects were able to attend each clinic visit and provide biological samples and complete responses to questionnaires when requested, resulting in a data set with no missing values during data collection. Table 1 contains summary statistics for select characteristics of BioCycle study participants.

Table 1. Selected Participant Characteristics, BioCycle Study (2005-2007)

| N = 259 Characteristic | Ovulatory Status | | | | | |
|---------------------------|------------------|------|-----|-------------|------|-----|
| | Ovulatory | | | Anovulatory | | |
| | n | Mean | SD | n | Mean | SD |
| Age | 235 | 27.7 | 8.3 | 24 | 22.8 | 5.2 |
| BMI | 235 | 24.2 | 3.8 | 24 | 23.2 | 4.0 |
| Characteristic | n | % | n | % | | |
| Race | | | | | | |
| White | 140 | 59.6 | 14 | 58.4 | | |
| Black | 46 | 19.6 | 5 | 20.8 | | |
| Other | 49 | 20.8 | 5 | 20.8 | | |
| Ever used birth control | | | | | | |
| Yes | 129 | 56.1 | 6 | 33.3 | | |
| Parity | | | | | | |
| 0 | 165 | 71.4 | 22 | 100.0 | | |
| 1 | 14 | 6.1 | 0 | | | |
| ≥2 | 52 | 24.5 | 0 | | | |
| Physical activity | | | | | | |
| Low | 22 | 9.4 | 3 | 12.5 | | |
| Moderate | 83 | 35.3 | 9 | 37.5 | | |
| High | 130 | 55.3 | 12 | 50.0 | | |
| Current Smoker | | | | | | |
| Yes | 9 | 3.8 | 1 | 4.2 | | |
| Marital Status | | | | | | |
| Married | 65 | 27.7 | 1 | 4.2 | | |
| Educational Attainment | | | | | | |
| >High school | 205 | 87.2 | 21 | 87.5 | | |

SD: standard deviation, BMI: body mass index

B. Study Data

1. Exposure Variables

Exposure variables for this study include hepatocyte growth factor (HGF) and vascular endothelial growth factor (VEGF). HGF is a protein known to regulate cellular growth, motility, and morphogenesis in endothelial and epithelial cells (Funakoshi & Nakamura, 2003). VEGF is a cytokine that mediates vasculogenesis and angiogenesis, and is important for signal transduction (Hoeben et al., 2004). Each exposure variable was quantified using BioSource 30-plex human cytokine assays (Invitrogen; Carlsbad, CA). This standard panel included a large number of cytokines with suspected involvement in menstrual cycle function and implantation.

While the study was able to collect complete data for all subject responses and biological samples, data on biomarkers was imprecise due to values falling below the limit of detection during chemical analysis. Limit of detection thresholds are determined based on the sensitivity and specificity of the apparatus being used; when values fall below this threshold, they are often not reported and are treated as missing. To account for this, HGF and VEGF were selected as exposure variables based on their minimal degrees of censorship due to limit of detection issues (HGF: 0.2%, VEGF: 3.34%) and their bivariate relationships with the outcome variable (Wilcoxon signed rank-sum test; $P_{\text{HGF}}=0.4$, $P_{\text{VEGF}}<0.0001$) compared to other biomarkers assessed. Both cytokines exhibited heavy positive skew and were natural log-transformed to approximate Normality for analysis.

A simulated cytokine (henceforth abbreviated SIM) was created to replicate an “ideal” Normally distributed cytokine after natural log-transformation. The mean of SIM ($\mu_{\text{SIM}} = -2.67$) closely resembled VEGF ($\mu_{\text{VEGF}} = -2.62$), however SIM was more dispersed ($\sigma_{\text{SIM}} = 1.18$; $\sigma_{\text{VEGF}} = 0.46$) than VEGF, and had negligible skew and kurtosis. SIM exhibited no censored data due to

limit of detection issues, and a significant bivariate relationship with anovulation similar to VEGF ($P_{SIM} < 0.0001$).

2. Outcome Variable

Anovulatory status was used as the outcome variable for this study. Briefly, anovulation is a cycle-specific event that occurs if an egg is not released from the ovaries following the luteinizing hormone surge that occurs on Day 14 of a standard 28-day menstrual cycle. For this study, anovulatory status was determined based on the Bio-P5-LH algorithm developed by Lynch et al. (2014), which accounts for serum progesterone, serum luteinizing hormone, and approximate menstrual cycle phase based on fertility monitor information. Using this method, participants were either ovulatory or anovulatory for each menstrual cycle within the observation period, resulting in a binary outcome variable. Outcome status was determined on Day 14 of each cycle, and so data predicting anovulatory status were restricted to days 2, 13 & 14 of each menstrual cycle. Days 2, 13 & 14 represent menses, the luteinizing hormone surge, and expected ovulation, respectively, which are three periods of substantial changes in endogenous reproductive hormones. Day 7 and Day 12 represent the mid-follicular phase and the late follicular phase, respectively; these two periods are relatively stable in terms of hormonal changes and were thus excluded from the present analysis.

3. Confounding Variables

Age, body mass index (BMI), physical activity, smoking status, parity, ever use of birth control, race, marital status and educational attainment were all considered as potential confounding variables for the present study. All other biomarkers available in the data set were

also considered for inclusion. Bivariate relationships between these predictors and both the outcome and the exposures were evaluated for inclusion in analysis and imputation models.

CHAPTER 4

MISSINGNESS MECHANISMS & STATISTICAL ANALYSIS

A. Missingness Mechanisms

Each type of missingness was imposed using the following equation:

$$I_{ijk} = \beta_1 * x_{age} + \beta_2 * x_{race} + \omega$$

where ω is a random draw from the standard Normal distribution ($\mu=0, \sigma=1$), x_{age} is the standardized age of the given subject, x_{race} is an indicator for the race of a given subject, and β_1 and β_2 are indicators for MAR and MNAR, respectively. Using this equation, I_{ijk} functions as a type-specific indicator of missingness for the i^{th} individual on the j^{th} day of the k^{th} cycle; each observation was assigned three values for each time point of I_{jk} , one for each type of missingness based on the values of β_1 and β_2 .

MCAR was imposed by assigning values of zero to both β_1 and β_2 . This allowed missingness to be determined solely on a random draw ω from the standard Normal distribution, independent from all observed and unobserved data. The randomness introduced by this draw also allowed I to be different at each time point for each observation, which in turn produced a non-monotone pattern of missing data for all three types of missing data. Missingness was imposed depending on the value of I_{ijk} (as described below).

MAR was imposed by assigning values of one and zero to β_1 and β_2 , respectively. x_{age} will appear in all imputation models, and so the probability of missingness will be associated with the observed data. MNAR was implemented by assigning values of one to both β_1 and β_2 . Unlike x_{age} , x_{race} was included in the imputation models, allowing the probability of missingness to be based on unobserved data. Omission of x_{race} from the imputation model thus imposed MNAR

conditions, even when imputation is performed; if x_{race} were included in the imputation model, the data would be MAR.

It is important to note that the probability of missingness for the MNAR data was partially a function of the observed data due to the inclusion of x_{age} in the equation and subsequent modelling steps. In practice, it is unlikely that missing values are due entirely to unobserved data (barring a very sparse data set), but rather a combination of factors both observed and unobserved. Furthermore, any statistician using an imputation-based approach will likely include several or many plausible predictors to fortify the assumption of MAR. Inclusion of both observed and unobserved variables therefore offered an imposition of MNAR that was not limited to simulated statistical inquiry and was more in line with real world situations.

Means and standard deviations for I_{ijk} were determined for each type of missing data; these were then used to calculate percentile cut-offs representing 5% - 50% of the total data in increments of 5%, taking into account the degree of censorship affecting the selected biomarker due to limit of detection. For a given percentile, any observations with an I value greater than the absolute value of the cut-off was assigned missing, resulting in a proportion of missing values equal to the desired percentage.

B. Application of Missing Data Techniques

Complete case analysis was conducted by exclusion of all observations with missing values. All imputation-based techniques (except multiple imputation using joint modeling) used linear regression and were conducted using the following imputation model for HGF:

$$X_{hgf}^{miss} = \beta_{anov}Y_{anov} + \beta_{age}X_{age} + \beta_{rantes}X_{rantes}$$

Chemokine ligand-5 (RANTES) was a cytokine that exhibited the highest correlation with HGF (Spearman $r = -0.21$; $P < 0.0001$) and the lowest degree of censorship (2.54%) due to limit of detection. For VEGF, the imputation model was constructed such that

$$X_{vegf}^{miss} = \beta_{anov}Y_{anov} + \beta_{age}X_{age} + \beta_{ifna}X_{ifna} + \beta_{il17}X_{il17} + \beta_{rantes}X_{rantes}$$

Interferon- α (IFNA), interleukin-17 (IL-17) and RANTES each exhibited Spearman correlation coefficients with VEGF greater than $r = 0.3$ where $P < 0.0001$. The imputation model for SIM was developed such that

$$X_{sim}^{miss} = \beta_{anov}Y_{anov} + \beta_{age}X_{age} + \beta_{cyto2}X_{cyto2} + \beta_{cyto3}X_{cyto3}$$

where simulated cytokine 2 (CYTO2) and simulated cytokine 3 (CYTO3) represented two synthetic biomarkers designed based on the relationship between VEGF, RANTES, and IL-17 (Spearman $r > 0.3$ for both; $P < 0.0001$). The SIM cytokine was intended to represent an “ideal” Normally distributed cytokine; because VEGF much more closely approximated Normality than HGF, bivariate relationships for VEGF were used to model the ancillary simulated cytokines.

MICE and MICE-PMM were also subjected to 20 burn-in iterations before each imputation. Multiple imputation using joint modeling (MI-JM) was conducted following the multivariate Normal distribution and the initial estimates for the expectation-maximization were obtained from complete cases. All multiple imputation techniques produced 20 completed data sets each, and multiple imputation using chained equations and predictive mean matching was conducted where $k = 5$ for PMM.

C. Statistical Analysis

As previously discussed, HGF and VEGF were selected due to their relative lack of censored values due to limit of detection (0.2% and 3.34% missing, respectively) and their strong

relationship with anovulatory status as compared to other biomarkers in the completely observed data set. For bivariate analyses, all remaining biomarkers in the data set with an acceptable degree of values below the limit of detection (<5%) were evaluated for associations with anovulatory status using Wilcoxon signed rank-sum tests. Continuous demographic variables, which were Normally distributed, were assessed used two sample t -tests. Chi-square tests were used for all categorical variables.

For HGF and VEGF, Spearman correlation coefficients were used to evaluate bivariate relationships with other biomarkers as well as continuous demographic variables. Associations between binary categorical variables were assessed using Wilcoxon signed rank-sum tests; Kruskal-Wallis tests were used for multi-level categorical variables. Bivariate relationships for SIM were investigated using the same techniques as for HGF and VEGF.

Generalized linear models were used to investigate the relationship between the selected biomarkers and anovulation, such that for the i^{th} subject during the j^{th} cycle,

$$\text{logit}(y_{ij}) = x_{ij}\beta$$

The binomial distribution and its canonical link (logit) were used for these models. Model parameters from generalized linear models were used as initial estimates for generalized estimating equations, which were utilized to account for repeated measures. Analyses were stratified by clinic visit day, and so repeated measures for each subject were used to account for measurements over two menstrual cycles. An exchangeable working correlation matrix was used and the maximum number of iterations per computation was set to 75. This model was analyzed for each type of missingness and each percentile cut-off.

Parameter estimates from generalized estimating equations were aggregated and used to obtain unbiased estimates following Rubin's Rules for combining multiply imputed datasets. A total of 500 replications were performed for all missing data techniques and multivariable

analyses. Obtained estimates were compared against estimates from the completely observed data to assess the relative performance of each technique for each type of missingness. All analyses and accompanying graphics were written in SAS 9.4 (Cary, NC).

CHAPTER 5

RESULTS

A. Complete Data Estimates & Biomarker Distributions

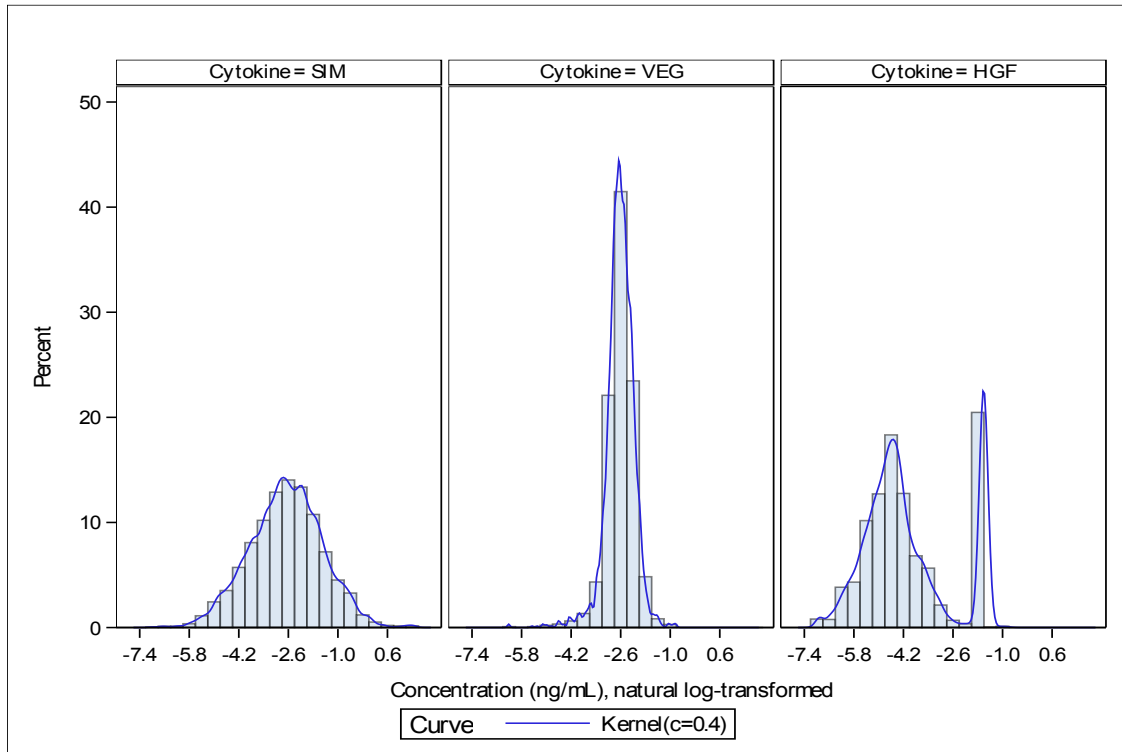
Over the two menstrual cycle observation period, a total of 34 anovulatory cycles were reported where values for HGF or VEGF were not censored due to limit of detection. Complete data estimates for HGF, VEGF, and SIM by clinic visit day are presented in Table 2; kernel density plots for each biomarker are presented in Figure 1. The distribution of HGF varied noticeably between study days; comparative histograms showing these differences are presented in Figure 2. With the exception of Day 2 when data were MCAR, for all types of missingness over all days of observation, estimates obtained for SIM were comparable for all techniques assessed (Figures 5, 8 & 11).

Table 2. Generalized Estimating Equation Coefficients for Complete Data, Biocycle Study (2005-2007)

| Clinic Visit Day | Cytokine | | |
|------------------------------------|---------------------|----------------------|---------------------|
| | HGF β (SE) | VEGF β (SE) | SIM β (SE) |
| Day 2 (Menses) | -0.013 (0.109) | 0.364 (0.376) | -1.479 (0.120) |
| Day 13 (Luteinizing hormone surge) | -0.030 (0.119) | 0.688 (0.443) | -0.844 (0.143) |
| Day 14 (Expected ovulation) | -0.030 (0.103) | 0.286 (0.431) | -0.829 (0.143) |

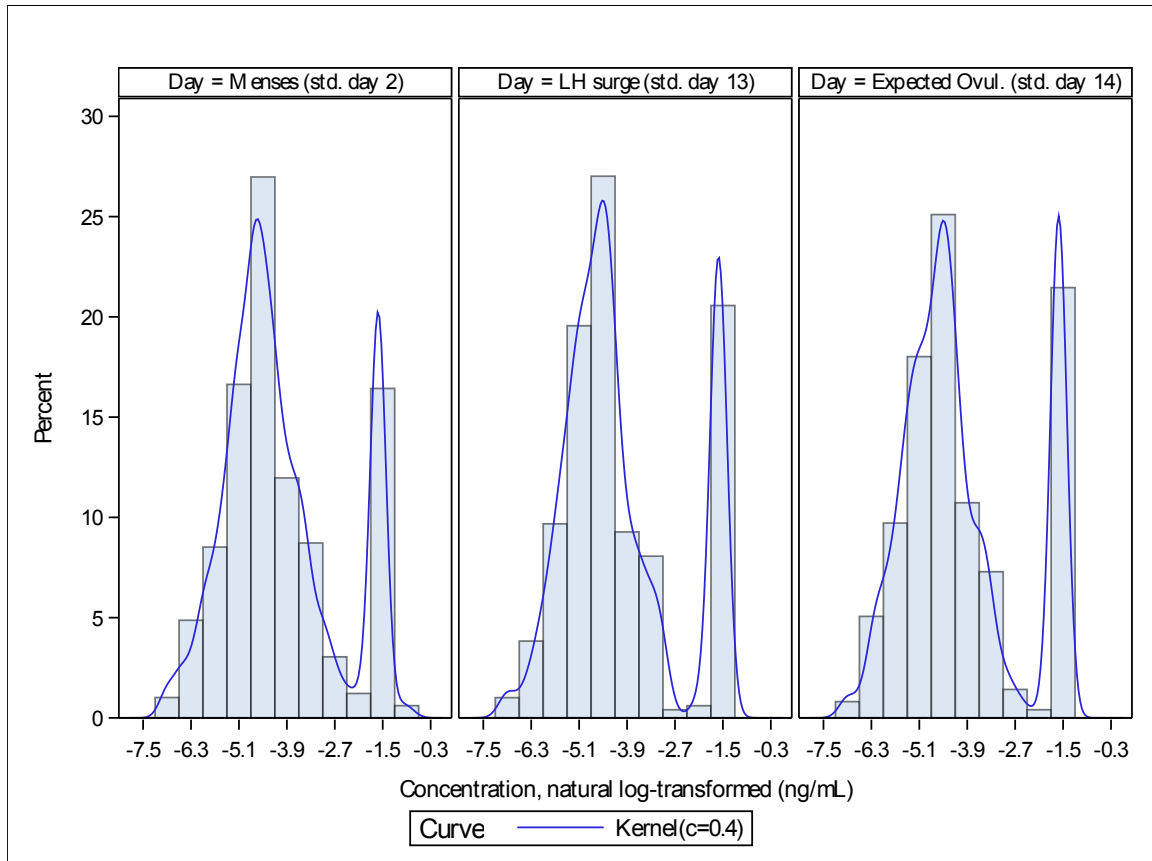
HGF: Hepatocyte growth factor, VEGF: Vascular endothelial growth factor, SIM: Simulated cytokine

Figure 1. Histograms and kernel density plots: SIM, VEGF, HGF



SIM: Simulated cytokine, VEGF: Vascular endothelial growth factor, HGF: Hepatocyte growth factor

Figure 2. Histograms and kernel density plots by day: HGF



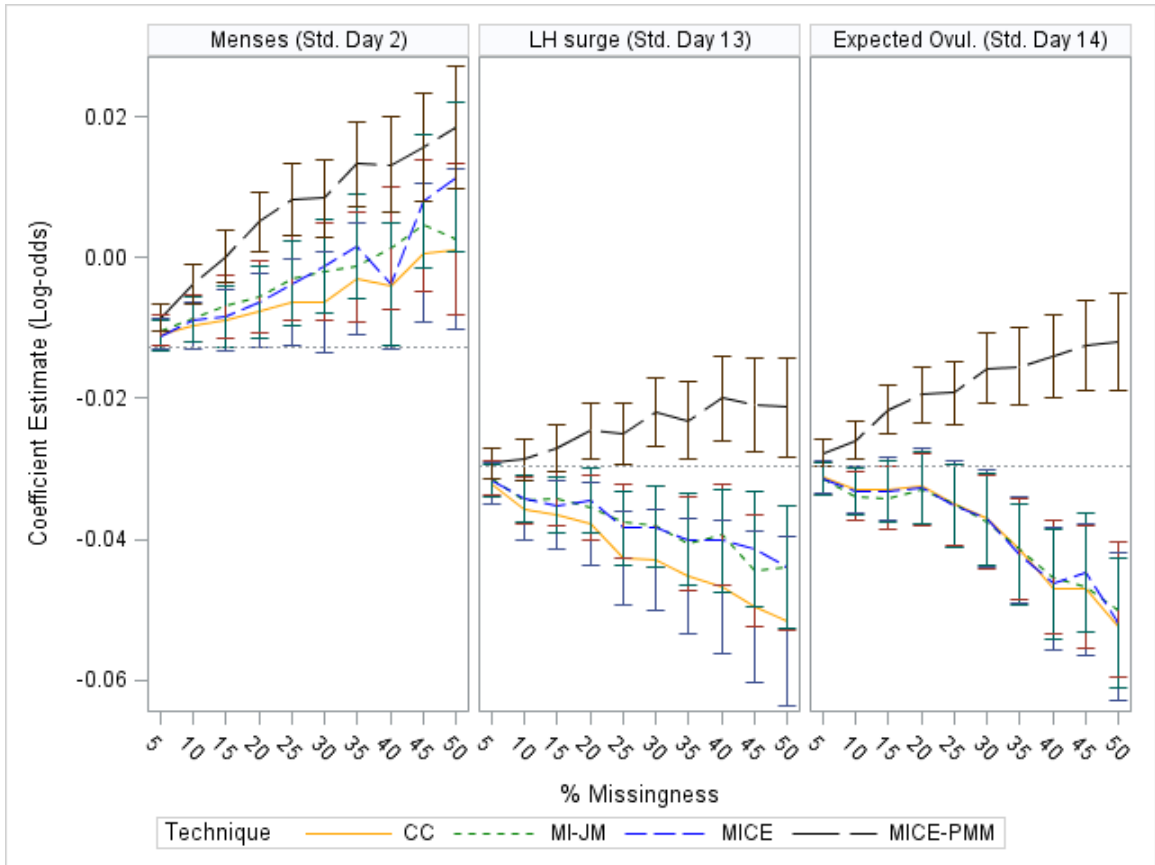
HGF: Hepatocyte growth factor, LH: Luteinizing hormone, Ovul: Ovulation

Days correspond to key events during a standardized 28-day menstrual cycle

B. Missing Completely at Random

Figure 3 shows the performance of the selected techniques under MCAR conditions for HGF. Estimates obtained using MI-JM closely resembled those of MICE on all days. Complete case analysis performed poorly on Day 13 relative to the other techniques, however produced similar estimates to MI-JM and MICE on Day 2. On all days, MICE-PMM estimates noticeably diverged from the both the complete estimates and the estimates produced by the other methods. All techniques diverged consistently from the complete data estimates after 40% missing data on Day 2.

Figure 3. Change in coefficient estimates as a function of missingness: HGF, MCAR

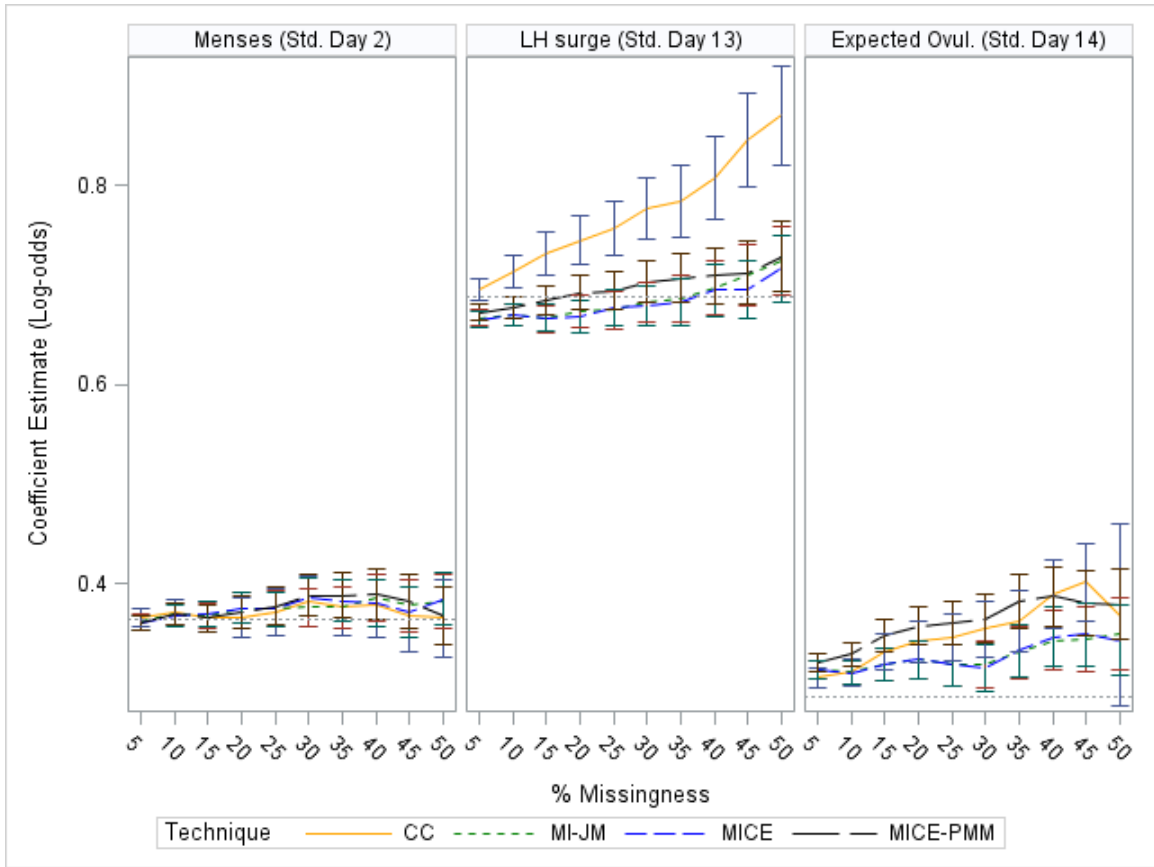


Complete data estimates (denoted by gray dotted line): Day 2 = -0.013, Day 13 = -0.030, Day 14 = -0.030; standard error bars denote $\pm 2(SE)$

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

The performance of each technique under MCAR conditions for VEGF are presented in Figure 4. All methods performed comparably on Day 2 across all degrees of missingness. MICE and MI-JM outperformed MICE-PMM and complete case analysis on Day 14. Estimates on Day 2 did not diverge substantially from the completely observed data values for all approaches; they appeared to diverge from the complete data from 15% - 40% missing data and converge thereafter. On Day 13, all techniques except complete case performed similarly with estimates beginning to diverge after 45% missingness.

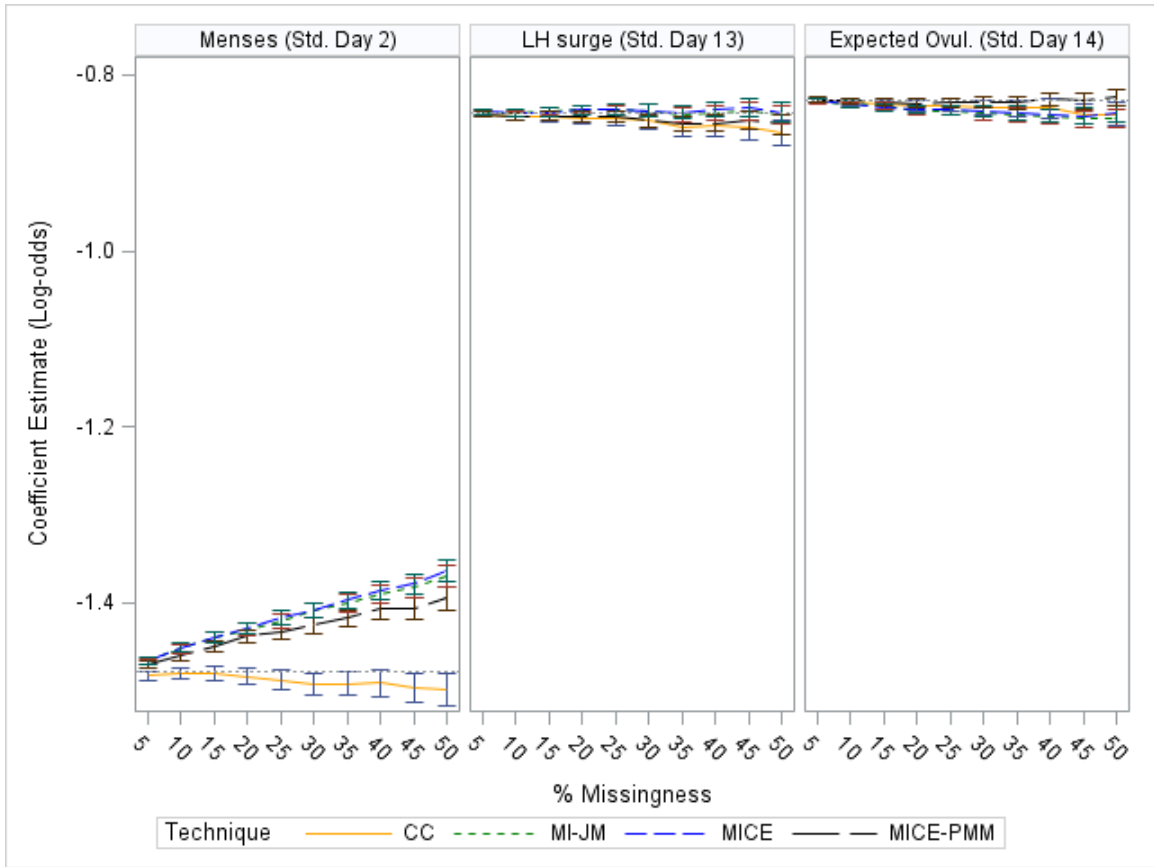
Figure 4. Change in coefficient estimates as a function of missingness: VEGF, MCAR



Complete data estimates (denoted by gray dotted line): Day 2 = 0.364, Day 13 = 0.688, Day 14 = 0.286; standard error bars denote $\pm 2(SE)$

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

Figure 5. Change in coefficient estimates as a function of missingness: SIM, MCAR



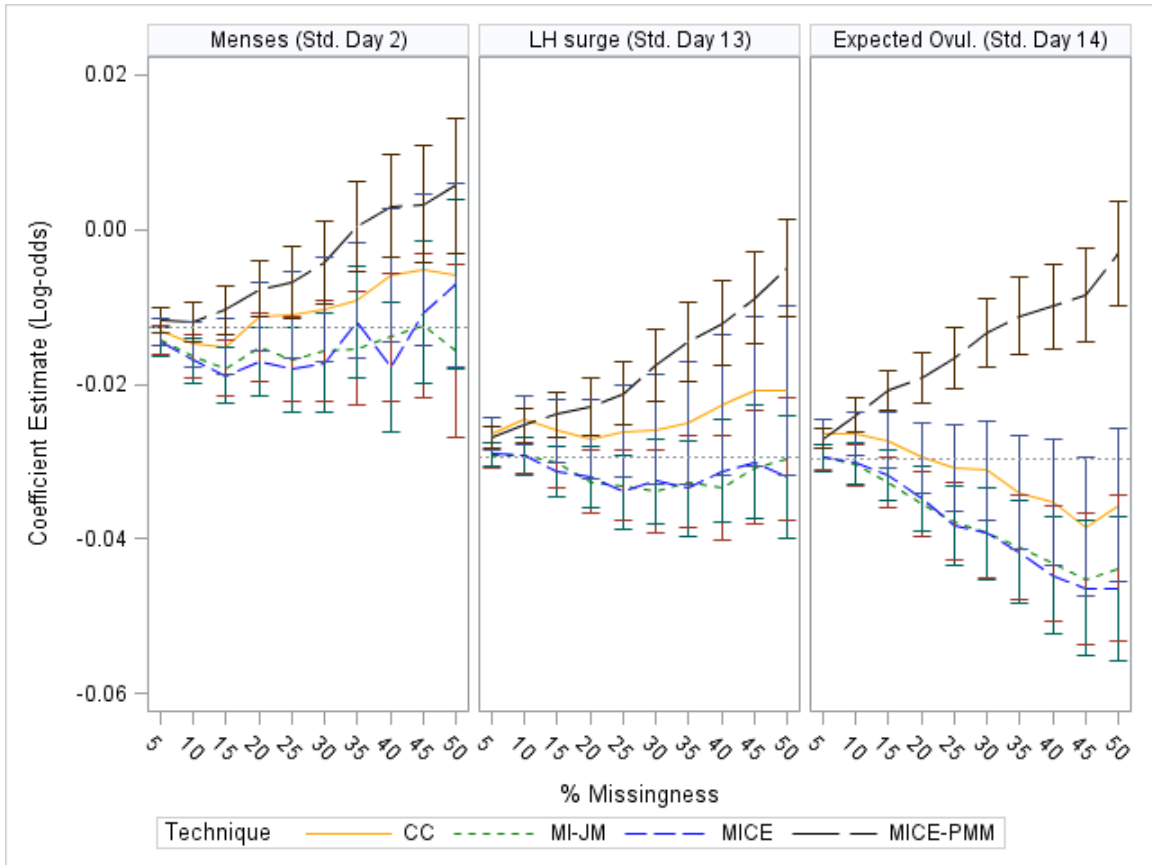
Complete data estimates (denoted by gray dotted line): Day 2 = -1.478, Day 13 = -0.844, Day 14 = -0.829; standard error bars denote $\pm 2(\text{SE})$

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

C. Missing at Random

All approaches performed poorly for HGF on Day 2 when data were missing at random (Figure 6). MICE-PMM deviated consistently from both the complete data estimate and the other technique estimates on Day 14. Similarly, MI-JM and MICE outperformed complete case and MICE-PMM on Day 13 under all degrees of missingness.

Figure 6. Change in coefficient estimates as a function of missingness: HGF, MAR

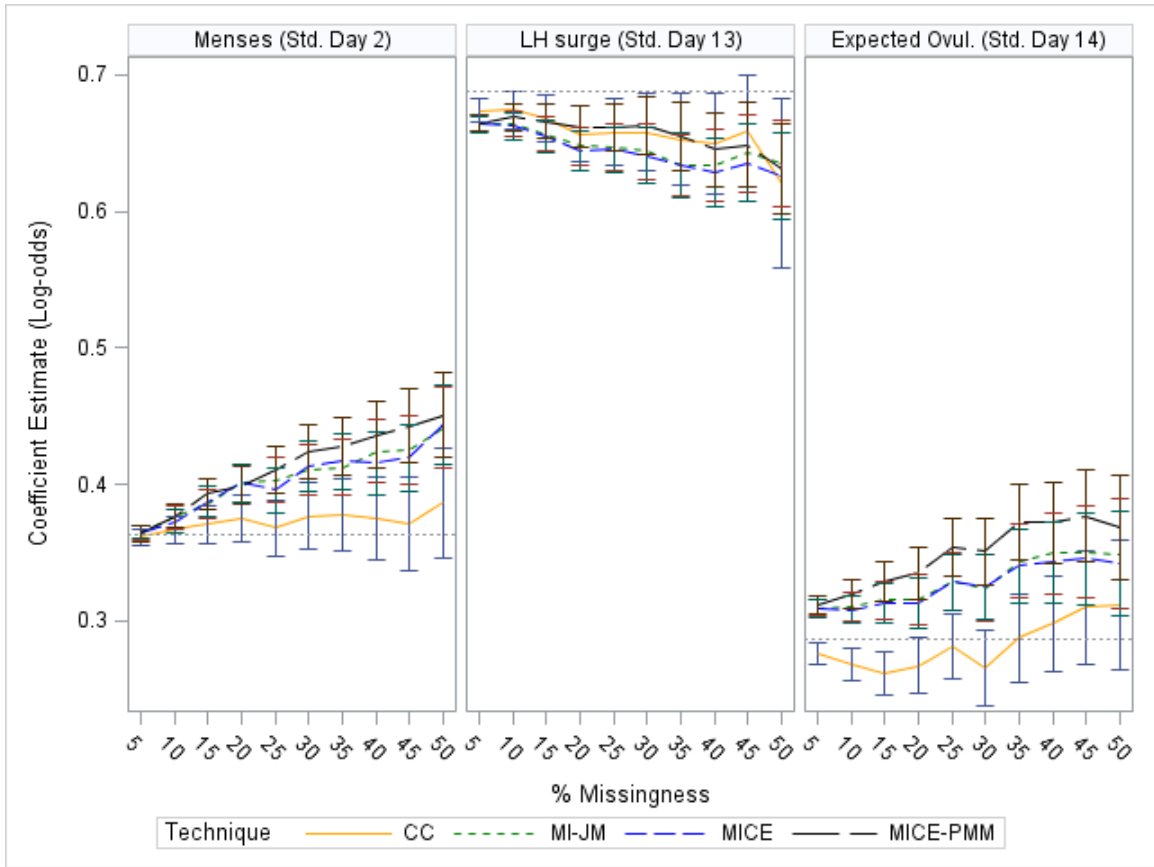


Complete data estimates (denoted by gray dotted line): Day 2 = -0.013, Day 13 = -0.030, Day 14 = -0.030; standard error bars denote +/- 2(SE)

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

Under MAR conditions, there was minimal distinction between techniques on Day 2 and Day 13 for VEGF (Figure 7). However, complete case appeared to converge on the complete data estimates after 30% missing data. On Day 14, all techniques exhibited similar trajectories, however complete case appeared to under-estimate the coefficient whereas MI-JM, MICE, and MICE-PMM appeared to over-estimate.

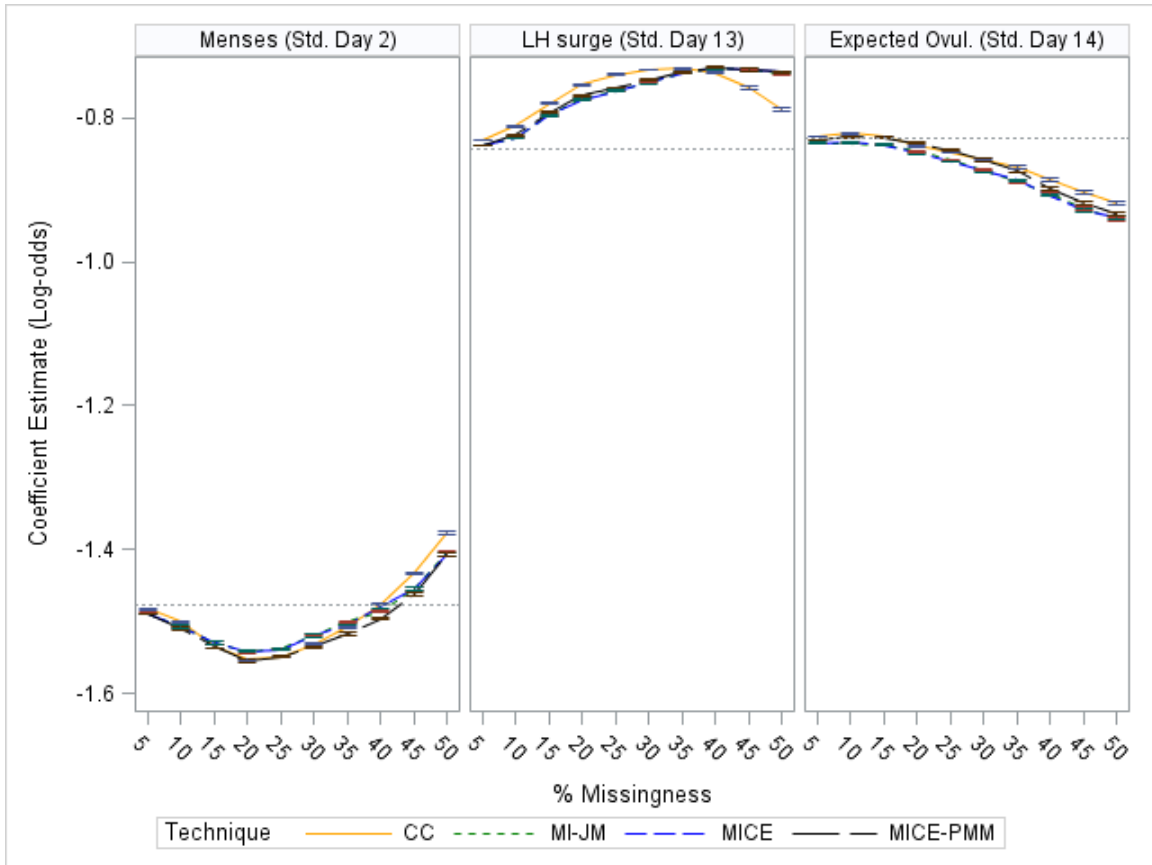
Figure 7. Change in coefficient estimates as a function of missingness: VEGF, MAR



Complete data estimates (denoted by gray dotted line): Day 2 = 0.364, Day 13 = 0.688, Day 14 = 0.286; standard error bars denote +/- 2(SE)

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PM: Multiple imputation using chained equations and predictive mean matching

Figure 8. Change in coefficient estimates as a function of missingness: SIM, MAR



Complete data estimates (denoted by gray dotted line): Day 2 = -1.478, Day 13 = -0.844, Day 14 = -0.829; standard error bars denote $\pm 2(SE)$

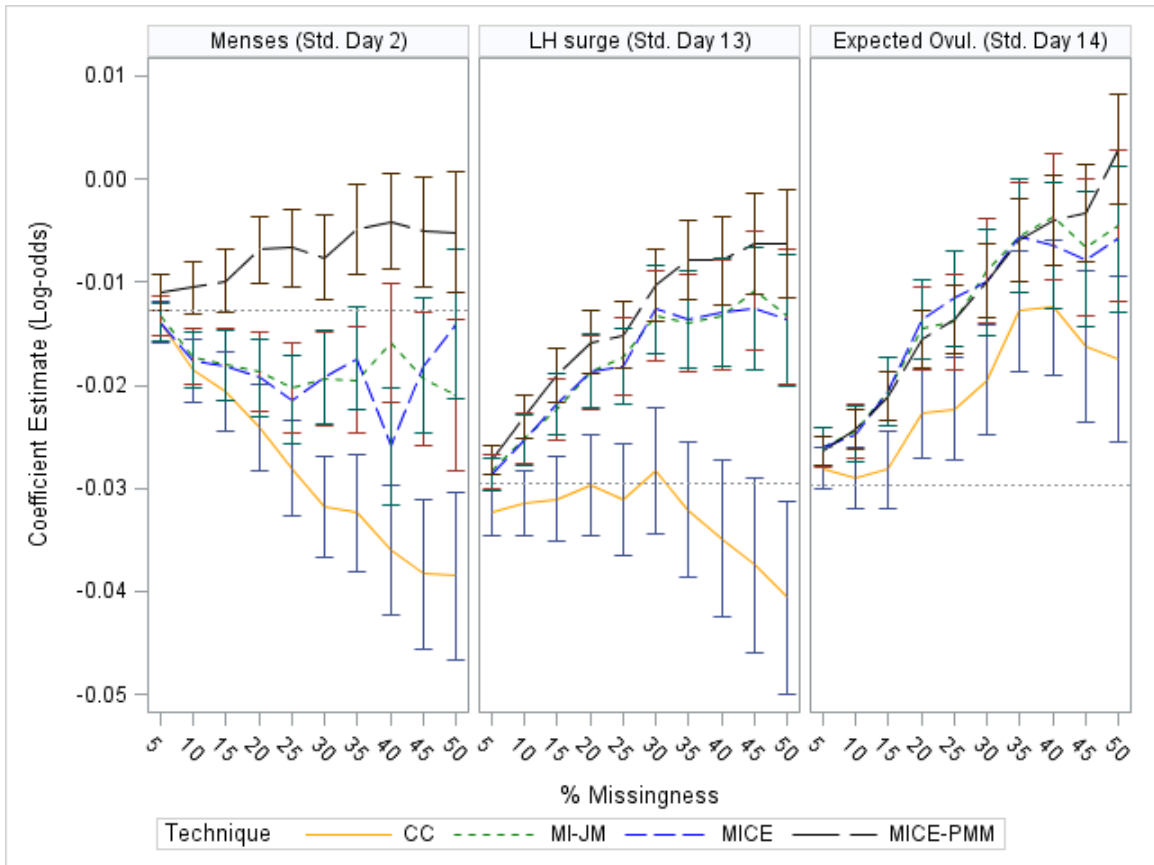
LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PM: Multiple imputation using chained equations and predictive mean matching

D. Missing Not at Random

There was no distinguishably superior technique when data are MNAR for HGF (Figure 9). Values for MICE-PM followed the complete data estimates on Day 2 more closely than all other methods over all degrees of missingness. Complete case analysis performed well on Day 13 until the missing data exceeds 35%, while MICE and MI-JM appeared to converge on the complete estimates as missingness increased from 35% to 50%. MI-JM, MICE AND MICE-

PMM only modestly under-estimated the coefficient until 10%, after which they began to consistently and increasingly under-estimate.

Figure 9. Change in coefficient estimates as a function of missingness: HGF, MNAR



Complete data estimates (denoted by gray dotted line): Day 2 = -0.013, Day 13 = -0.030, Day 14 = -0.030; standard error bars denote $\pm 2(SE)$

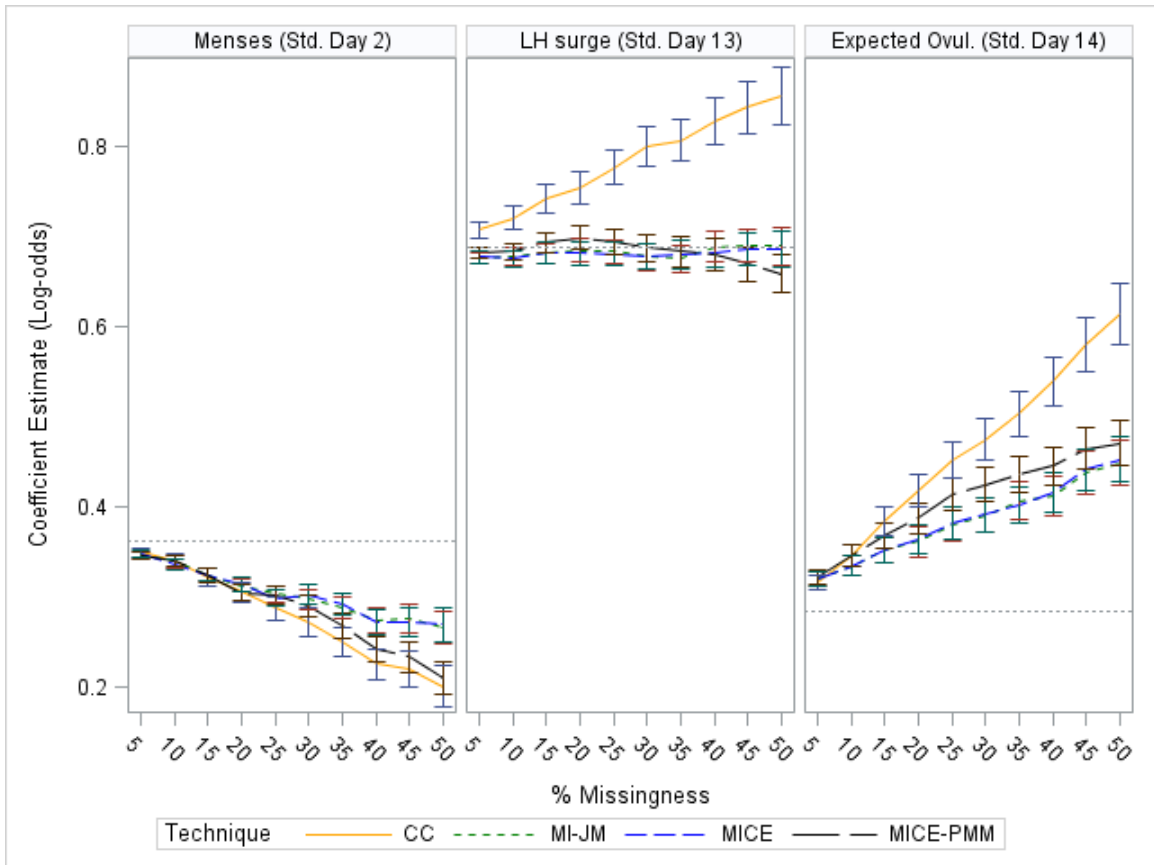
LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

Under MNAR conditions for VEGF, MI-JM, MICE, and MICE-PMM produced similar estimates across all days, however there was no consistency between the three days (Figure 10).

The coefficient was under-estimated on Day 2, closely resembled on Day 13 across all degrees of

missing data, and over-estimated on Day 14. Complete case analysis was outperformed in all instances, most noticeably on Day 13 and Day 14.

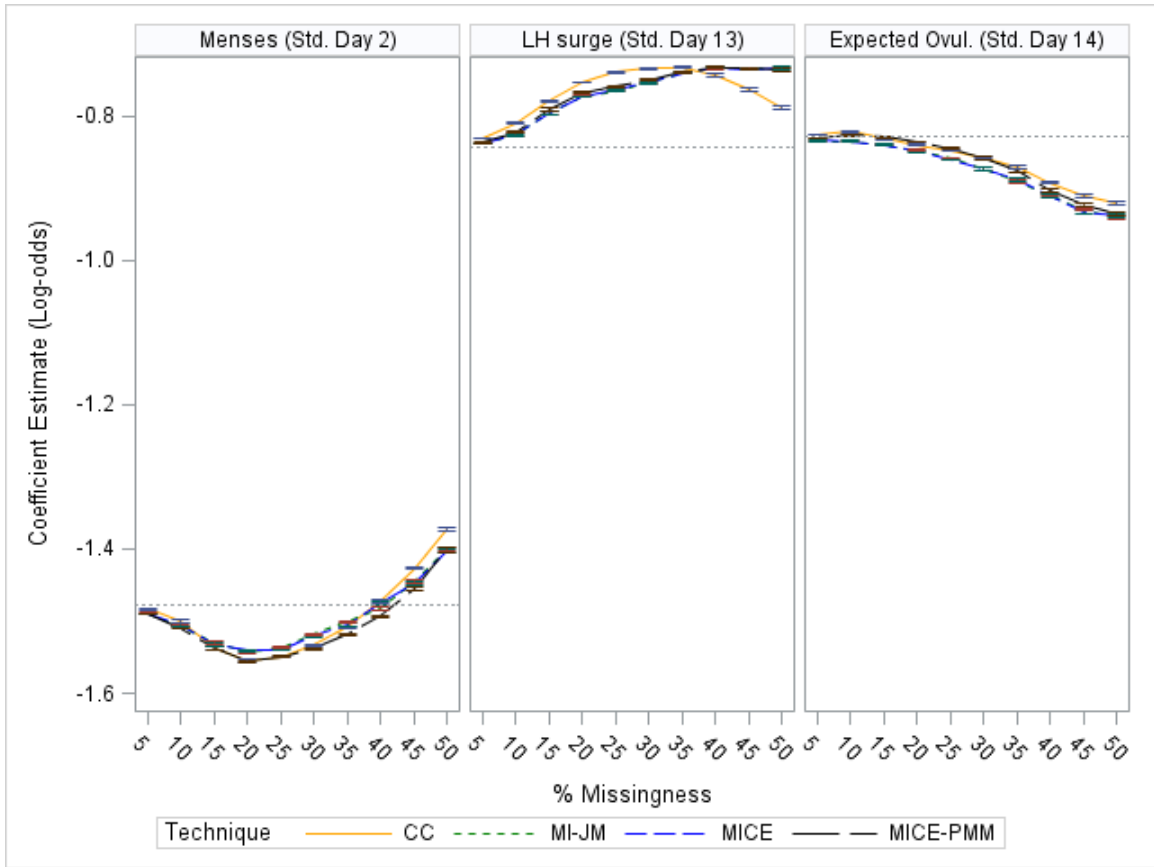
Figure 10. Change in coefficient estimates as a function of missingness: VEGF, MNAR



Complete data estimates (denoted by gray dotted line): Day 2 = 0.364, Day 13 = 0.688, Day 14 = 0.286; standard error bars denote +/- 2(SE)

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

Figure 11. Change in coefficient estimates as a function of missingness: SIM, MNAR



Complete data estimates (denoted by gray dotted line): Day 2 = -1.478, Day 13 = -0.844, Day 14 = -0.829; standard error bars denote $\pm 2(\text{SE})$

LH: Luteinizing hormone; Ovul: Ovulation; CC: Complete case; MI-JM: Multiple imputation using joint modeling; MICE: Multiple imputation using chained equations; MICE-PMM: Multiple imputation using chained equations and predictive mean matching

CHAPTER 6

DISCUSSION

The findings of this study suggest that violation of distributional assumptions can have a profound impact on the performance of specific methods for handling missing data. It is common to obtain biomarker data that are positively skewed, especially when the specimen being measured is present in trace quantities or subject to limit of detection censorship. Natural logarithm transformations were used to correct for this skewness in HGF and VEGF, however both resulting distributions retained some degree of skewness and substantial kurtosis. Application of power transformations such as the Box-Cox transformation (Box & Cox, 1964) may offer greater reductions in skewness, at the cost of interpretability, in future studies involving biomarker data. Kurtosis, however, would remain an issue even after these transformations.

HGF maintained a strong, asymmetrical bimodal distribution even after ln-transformation (Figures 1 & 2), which noticeably affected estimates from MICE-PMM when data were MCAR and MAR. This technique produces imputed values based on the observed values of the given predictor, which generally proves to be advantageous when the data have unique features such as bounds or when the underlying distribution deviates from Normality (Little & Rubin, 2002). However, our findings showed that MICE-PMM consistently biased estimates when the underlying distribution exhibited extreme bimodality. Specifically, as the degree of missingness increased, a substantial portion of the remaining observed values were represented by a single mode at -1.6 ng/mL (natural log-transformed), which constituted approximately 20% of the original distribution. This limited the variety of observed values from which MICE-PMM could produce imputed values, and inflated the influence of this mode, causing it to contribute disproportionately to the imputed distribution. This suggests that MICE-PMM, sometimes viewed

as a panacea for missing data, should be applied only after careful review of univariate distributions and appropriate transformations when dealing with biomarkers.

In contrast to HGF, all techniques performed comparably for VEGF with the exception of complete case analysis when data were MNAR. The distribution of VEGF after natural log-transformation retained a slight negative skew, however the biomarker was distinctly unimodal and approximated Normality. When data were MCAR and MAR, MICE-PMM yielded estimates which were substantially closer to those of the complete data and the other techniques. When viewed with the findings for HGF, these results highlight the important role that the distribution of a biomarker can play when using imputation-based techniques which rely on predictive mean matching.

MI-JM and MICE performed consistently and similarly for both HGF and VEGF. With respect to HGF, the unrestricted nature of these two techniques allowed the bimodal distribution to be maintained as the amount of missing data increased. This resulted in estimates which were closer to the complete data than MICE-PMM and complete case analysis. Estimates produced by MI-JM do not seem to have suffered despite HGF violating the assumptions of the multivariate Normal distribution. Furthermore, the ability of MICE to produce accurate estimates offers additional evidence to implicate predictive mean matching as a poor choice for highly bimodal data.

Results from SIM demonstrate that, when transformed to achieve an “ideal” Normal distribution, all techniques perform quite comparably. Under Normal conditions, it is reasonable to expect both MICE and MICE-PMM to obtain similar results; when upholding the assumptions of the multivariate Normal distribution, we would also expect similar estimates from MI-JM. The similarity between MAR and MNAR estimates was unusual, however as will be explained below, this is likely due to complications with the missingness mechanism. Taken together, these findings support the notion that when natural log-transformed biomarkers approach Normality,

multiple imputation in any of the forms evaluated offer similar and notably more accurate estimates.

While the imputation models specified appear to have yielded sufficiently accurate estimates when data were MCAR or MAR for VEGF, construction of appropriate models proved to be very challenging for these biomarker data. Generally, predictors included in an imputation model are selected based on *a priori* knowledge, literature review, or bivariate analyses. Kenward & Carpenter (2007) explain the benefits of having distinct analysis and imputation models, which allow additional covariates to be used for imputation which do not appear in the substantive model. However, even if known relationships between the selected cytokines and some available biomarkers had been suggested by literature, significant amounts of censorship due to limit of detection prohibited inclusion of most biological variables in our imputation models. Bivariate analyses mainly identified weak correlations with HGF or VEGF and other cytokines ($r \leq |0.3|$) as well as demographic characteristics, which likely would have contributed noise to the imputed values. Interactions were also not assessed; circulating biomarkers are often components of intricate and overlapping biological pathways, making interaction effects both very common and very complex. Without justification founded in research, any attempts to explore interactions between cytokines must be accompanied with appropriate statistical controls for multiple comparisons. Fully conditional specification offers perhaps the best opportunity to accurately capture all potential predictors and interactions that might improve biomarker imputation.

When developing complete imputation models for each biomarker, researchers may also improve their models by accounting for the temporal component of their longitudinal studies. In the present analysis, we provided comparative estimates for three clinically important days in a standardized 28-day menstrual cycle. On Day 13 and Day 14 specifically, there are a myriad of chemical shifts occurring among female endogenous reproductive hormones. To obtain the most

accurate imputed values, it may be useful to develop unique imputation models corresponding to critical time points in addition to specific biomarkers.

Our ability to accurately simulate MAR and MNAR conditions in this study was limited. To appropriately replicate these phenomena, relationships between each biomarker and predictors which could plausibly have produced such conditions were evaluated. While some of the continuous variables exhibited statistically significant correlations, none exceeded a Pearson or Spearman correlation coefficient of $r = |0.2|$ with either biomarker; only race showed significant difference between categories, for HGF but not VEGF. After considering all bivariate associations, age (HGF: Spearman $r = 0.05$, $P=0.01$; VEGF: Spearman $r = -0.16$, $P<0.0001$) and race (HGF: $P=0.002$; VEGF: $P=0.13$) were selected to contribute to the missingness mechanism. Utilization of covariates with stronger relationships to HGF and VEGF may have yielded a more vivid example of data which are MAR and MNAR.

Despite poor associations with the biomarkers of interest, the predictors available are representative of real data. Access to the BioCycle study data was an overall strength of this analysis, allowing findings to be framed in the context of a real-world longitudinal study. Replications of the analyses conducted improved the accuracy and stability of estimates obtained from each of the selected techniques. Missing biomarker data due to limit of detection issues restricted our ability to develop richer imputation models and explore possible interactions documented in published literature. Furthermore, we did not evaluate the performance of likelihood-based approaches or fully Bayesian methods of handling missing data.

Accounting for missing data is an essential step in any data analysis which cannot be ignored. In the context of a longitudinal study, we have demonstrated that when biomarkers are successfully transformed to closely approximate Normality, estimates obtained from several different multiple imputation-based techniques resemble the original data under MCAR and MAR conditions with a non-monotone pattern of missingness. Special attention must be given to

transformations applied to biomarkers before analysis and the resulting distribution of the predictor when choosing an imputation method; estimates can quickly become biased when using predictive mean matching if transformed variables deviate significantly from Normality. Adhering to these precautions will assist researchers handling missing longitudinal data in obtaining accurate estimates and making valid statistical inferences.

BIBLIOGRAPHY

- Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211–252.
- Buuren, S. V., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064. <http://doi.org/10.1080/10629360600810434>
- Buuren, S. van. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219–242. <http://doi.org/10.1177/0962280206074463>
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. <http://doi.org/10.1037/1082-989X.6.4.330>
- Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal Analysis of Incomplete Longitudinal Binary Data: A Cautionary Note on LOCF Imputation. *Biometrics*, 60(3), 820–828. <http://doi.org/10.1111/j.0006-341X.2004.00234.x>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.
- Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56(10), 968–976. [http://doi.org/10.1016/S0895-4356\(03\)00170-7](http://doi.org/10.1016/S0895-4356(03)00170-7)
- Ferro, M. A. (2014). Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Annals of Epidemiology*, 24(1), 75–77. <http://doi.org/10.1016/j.annepidem.2013.10.007>
- Funakoshi, H., & Nakamura, T. (2003). Hepatocyte growth factor: from diagnosis to clinical applications. *Clinica Chimica Acta*, 327(1–2), 1–23. [http://doi.org/10.1016/S0009-8981\(02\)00302-9](http://doi.org/10.1016/S0009-8981(02)00302-9)
- Greenland, S., & Finkle, W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, 142(12), 1255–1264.

- Hoeben, A., Landuyt, B., Highley, M. S., Wildiers, H., Oosterom, A. T. V., & Bruijn, E. A. D. (2004). Vascular Endothelial Growth Factor and Angiogenesis. *Pharmacological Reviews*, 56(4), 549–580. <http://doi.org/10.1124/pr.56.4.3>
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple Imputation in Practice. *The American Statistician*, 55(3), 244–254. <http://doi.org/10.1198/000313001317098266>
- Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A Potential for Bias When Rounding in Multiple Imputation. *The American Statistician*, 57(4), 229–232. <http://doi.org/10.1198/0003130032314>
- Karahalios, A., Baglietto, L., Lee, K. J., English, D. R., Carlin, J. B., & Simpson, J. A. (2013). The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study. *Emerging Themes in Epidemiology*, 10, 6. <http://doi.org/10.1186/1742-7622-10-6>
- Kenward, M. G., & Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16(3), 199–218. <http://doi.org/10.1177/0962280206075304>
- Lee, K. J., & Carlin, J. B. (2010). Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology*, 171(5), 624–632. <http://doi.org/10.1093/aje/kwp425>
- Linero, A. R., & Daniels, M. J. (2015). A Flexible Bayesian Approach to Monotone Missing Data in Longitudinal Studies with Nonignorable Missingness with Application to an Acute Schizophrenia Clinical Trial. *Journal of the American Statistical Association*, 110(509), 45–55. <http://doi.org/10.1080/01621459.2014.969424>
- Little, R. J. A. (1992). Regression With Missing X's: A Review. *Journal of the American Statistical Association*, 87(420), 1227–1237. <http://doi.org/10.2307/2290664>
- Little, R. J. A., Rubin, D. B., Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. In *Statistical Analysis with Missing Data* (2nd ed., pp. 1–23). Hoboken, NJ: John Wiley & Sons, Inc. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9781119013563.ch1/summary>
- Luo, S., Lawson, A. B., He, B., Elm, J. J., & Tilley, B. C. (2016). Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. *Statistical Methods in Medical Research*, 25(2), 821–837. <http://doi.org/10.1177/0962280212469358>

- Lynch, K. E., Mumford, S. L., Schliep, K. C., Whitcomb, B. W., Zarek, S. M., Pollack, A. Z., ... Schisterman, E. F. (2014). Assessment of anovulation in eumenorrhic women: comparison of ovulation detection algorithms. *Fertility and Sterility*, *102*(2), 511–518.e2. <http://doi.org/10.1016/j.fertnstert.2014.04.035>
- McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *92*(437), 162–170. <http://doi.org/10.1080/01621459.1997.10473613>
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, *27*(1), 85-95.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. <http://doi.org/10.2307/2335739>
- Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC Press.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, *33*(4), 545–571. http://doi.org/10.1207/s15327906mbr3304_5
- Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology*, *55*(4), 329–337. [http://doi.org/10.1016/S0895-4356\(01\)00476-0](http://doi.org/10.1016/S0895-4356(01)00476-0)
- Von Hippel, P. T. (2007). Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*, *37*(1), 83–117. <http://doi.org/10.1111/j.1467-9531.2007.00180.x>
- Wactawski-Wende, J., Schisterman, E. F., Hovey, K. M., Howards, P. P., Browne, R. W., Hediger, M., ... Trevisan, M. (2009). BioCycle study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatric and Perinatal Epidemiology*, *23*(2), 171–184. <http://doi.org/10.1111/j.1365-3016.2008.00985.x>